

Lectures on Financial Economics

© by Antonio Mele

University of Lugano

and

swiss:finance:institute

June 2012



Front cover explanations

Top: Illustration of the increased efficiency in maritime routing allowed by the Suez Canal (right panel) opened in 1869, and the Panama Canal (left panel) opened in 1913, two amongst the most enduring technological marvels with global economic and political implications.

Bottom: A 75 year 3% coupon bearing bond issued by the Panama Canal Company (“Compagnie Universelle du Canal Interocéanique de Panama”) in October 1884. The company defaulted in 1889 under the leadership of the Count Ferdinand de Lesseps, who during 1858 had also founded the Suez Canal Company (“Compagnie Universelle du Canal Maritime de Suez”).

Preface

These *Lectures on Financial Economics* are based on notes I wrote in support of advanced undergraduate and graduate lectures in financial economics, macroeconomic dynamics, financial econometrics and financial engineering.

Part I, “Foundations,” develops the fundamental tools of analysis used in Part II and Part III. These tools span such disparate topics as classical portfolio selection, dynamic consumption- and production- based asset pricing, in both discrete and continuous-time, the intricacies underlying incomplete markets and some other market imperfections and, finally, econometric tools comprising maximum likelihood, methods of moments, and the relatively more modern simulation-based inference methods.

Part II, “Applied asset pricing theory,” is about identifying the main empirical facts in finance and the challenges they pose to financial economists: from excess price volatility and countercyclical stock market volatility, to cross-sectional puzzles such as the value premium. This second part reviews the main models aiming to take these puzzles on board.

Part III, “Asset pricing and reality,” aims just to this: to use the main tools in Part I and the lessons drawn from Part II, so as to cope with the main challenges occurring in actual capital markets, arising from option pricing and trading, interest rate modeling and credit risk and their associated derivatives. In a sense, Part II is about the big puzzles we face in fundamental research, while Part III is about how to live within our current and certainly unsatisfactory paradigms, so as to cope with demand for intellectual expertise.

These notes are still underground. The economic motivation and intuition are not always developed as deeply as they deserve, some derivations are inelegant, and sometimes, the English is a bit informal. Moreover, I still have to include material on asset pricing with asymmetric information, monetary models of asset prices, theories about the nominal and the real term structure of interest rates, bubbles, asset prices implications of overlapping generations models, or financial frictions and their interconnections with business cycle developments. Finally, I need to include more extensive surveys for each topic I cover, especially in Part II. I plan to

revise these notes to fill these gaps. Meanwhile, any comments on this version are more than welcome.

Antonio Mele
June 2012

“Antonio Mele does not accept any liability for any losses related to the use of the models, data, and methods described or developed in these lectures.”

Contents

I	Foundations	12
1	The classic capital asset pricing model	13
1.1	Introduction	13
1.2	Portfolio selection	13
1.2.1	The wealth constraint	13
1.2.2	Portfolio choice	14
1.2.3	Without the safe asset	15
1.2.4	The market portfolio	17
1.3	The CAPM	19
1.4	The APT	22
1.4.1	A first derivation	22
1.4.2	The APT with idiosyncratic risk and a large number of assets	23
1.4.3	Empirical evidence	24
1.5	Appendix 1: Analytical details relating to portfolio choice	25
1.5.1	The primal program	25
1.5.2	The dual program	26
1.6	Appendix 2: The market portfolio	28
1.6.1	The tangent portfolio is the market portfolio	28
1.6.2	Tangency condition	28
1.7	Appendix 3: An alternative derivation of the SML	30
1.8	Appendix 4: Liquidity traps, portfolio selection and the demand for money	31
1.8.1	Dichotomy choices and aggregate money demand	31
1.8.2	Money demand in a theory of portfolio selection	32
	References	34

2	The CAPM in general equilibrium	35
2.1	Introduction	35
2.2	The static general equilibrium in a nutshell	35
2.2.1	Walras' Law	36
2.2.2	Competitive equilibrium	36
2.2.3	Optimality	37
2.3	Time and uncertainty	41
2.4	Financial assets	42
2.5	Absence of arbitrage	42
2.5.1	How to price a financial asset?	42
2.5.2	The Land of Cockaigne	44
2.6	Equivalent martingales and equilibrium	48
2.6.1	The rational expectations assumption	48
2.6.2	Stochastic discount factors	49
2.6.3	Optimality and equilibrium	50
2.7	Consumption-CAPM	54
2.7.1	The risk premium	54
2.7.2	The beta relation	55
2.7.3	CCAPM & CAPM	55
2.8	Infinite horizon	55
2.9	Further topics on incomplete markets	56
2.9.1	Nominal assets and real indeterminacy of the equilibrium	56
2.9.2	Nonneutrality of money	57
2.10	Appendix 1	58
2.11	Appendix 2: Proofs of selected results	59
2.12	Appendix 3: The multicommodity case	62
	References	64
3	Infinite horizon economies	65
3.1	Introduction	65
3.2	Consumption-based asset evaluation	65
3.2.1	Recursive plans: introduction	65
3.2.2	The marginalist argument	66
3.2.3	Intertemporal elasticity of substitution	67
3.2.4	Lucas' model	68
3.3	Production: foundational issues	71
3.3.1	Decentralized economy	72
3.3.2	Centralized economy	73
3.3.3	Dynamics	74
3.3.4	Stochastic economies	76
3.4	Production-based asset pricing	80
3.4.1	Firms	80
3.4.2	Consumers	84

3.4.3	Equilibrium	85
3.5	Money, production and asset prices in overlapping generations models	85
3.5.1	Introduction: endowment economies	85
3.5.2	Diamond's model	88
3.5.3	Money	88
3.5.4	Money in a model with real shocks	92
3.6	Optimality	93
3.6.1	Models with productive capital	93
3.6.2	Models with money	94
3.7	Appendix 1: Finite difference equations, with economic applications	95
3.8	Appendix 2: Neoclassic growth in continuous-time	99
3.8.1	Convergence from discrete-time	99
3.8.2	The model	100
3.9	Appendix 3: Notes on optimization of continuous time systems	102
	References	104
4	Continuous time models	105
4.1	Introduction	105
4.2	On lambdas and betas	106
4.2.1	Prices	106
4.2.2	Expected returns	107
4.2.3	Risk-adjusted discount rates	108
4.3	An introduction to methods or, the origins: Black & Scholes	109
4.3.1	Time	109
4.3.2	Asset prices as Feynman-Kac representations	110
4.3.3	Girsanov theorem	113
4.4	An introduction to no-arbitrage and equilibrium	115
4.4.1	Self-financed strategies	115
4.4.2	No-arbitrage in Lucas tree	116
4.4.3	Equilibrium with CRRA	117
4.4.4	Bubbles	119
4.4.5	Reflecting barriers and absence of arbitrage	120
4.5	Martingales and arbitrage	121
4.5.1	The information framework	121
4.5.2	Viability	122
4.5.3	Market completeness	124
4.6	Equilibrium with a representative agent	126
4.6.1	Consumption and portfolio choices: martingale approaches	126
4.6.2	The older, Merton's approach: dynamic programming	128
4.6.3	Equilibrium	129
4.6.4	Continuous time Consumption-CAPM	130
4.7	Market imperfections and portfolio choice	131
4.8	Jumps	132

4.8.1	Poisson jumps	132
4.8.2	Interpretation	133
4.8.3	Properties and related distributions	134
4.8.4	Asset pricing implications	135
4.8.5	An option pricing formula	136
4.9	Continuous time Markov chains	136
4.10	Appendix 1: Self-financed strategies	137
4.11	Appendix 2: An introduction to stochastic calculus for finance	138
4.11.1	Stochastic integrals	138
4.11.2	Stochastic differential equations	148
4.12	Appendix 3: Proof of selected results	153
4.12.1	Proof of Theorem 4.2	153
4.12.2	Proof of Eq. (4.53).	153
4.12.3	Walras's consistency tests	154
4.13	Appendix 4: The Green's function	155
4.13.1	Setup	155
4.13.2	The PDE connection	156
4.14	Appendix 5: Portfolio constraints	157
4.15	Appendix 6: Models with final consumption only	159
4.16	Appendix 7: Topics on jumps	161
4.16.1	The Radon-Nikodym derivative	161
4.16.2	Arbitrage restrictions	162
4.16.3	State price density: introduction	162
4.16.4	State price density: general case	163
	References	165
5	Taking models to data	166
5.1	Introduction	166
5.2	Data generating processes	166
5.2.1	Basics	166
5.2.2	Restrictions on the DGP	167
5.2.3	Parameter estimators	168
5.2.4	Basic properties of density functions	168
5.2.5	The Cramer-Rao lower bound	169
5.3	Maximum likelihood estimation	169
5.3.1	Basics	169
5.3.2	Factorizations	169
5.3.3	Asymptotic properties	170
5.4	M-estimators	172
5.5	Pseudo, or quasi, maximum likelihood	173
5.6	GMM	174
5.7	Simulation-based estimators	177
5.7.1	Three simulation-based estimators	178

5.7.2	Asymptotic normality	180
5.7.3	A fourth simulation-based estimator: Simulated maximum likelihood . . .	183
5.7.4	Advances	184
5.7.5	In practice? Latent factors and identification	184
5.8	Asset pricing, prediction functions, and statistical inference	185
5.9	Appendix 1: Proof of selected results	189
5.10	Appendix 2: Collected notions and results	190
5.11	Appendix 3: Theory for maximum likelihood estimation	193
5.12	Appendix 4: Dependent processes	194
5.12.1	Weak dependence	194
5.12.2	The central limit theorem for martingale differences	194
5.12.3	Applications to maximum likelihood	194
5.13	Appendix 5: Proof of Theorem 5.4	196
	References	197
II	Applied asset pricing theory	200
6	Neo-classical kernels and puzzles	201
6.1	Introduction	201
6.2	The equity premium puzzle	202
6.2.1	A single-factor model	202
6.2.2	Extensions	205
6.2.3	The puzzles	205
6.3	Hansen-Jagannathan cup	207
6.4	Multifactor extensions	209
6.4.1	Exponential affine pricing kernels	209
6.4.2	Lognormal returns	211
6.5	Pricing kernels and Sharpe ratios	213
6.5.1	Market portfolios and pricing kernels	213
6.5.2	Pricing kernel bounds	214
6.6	Conditioning bounds	216
6.7	The cross section of stock returns and volatilities	217
6.7.1	Returns	217
6.7.2	Volatilities	218
6.8	Appendix	219
	References	222
7	Aggregate fluctuations in equity markets	224
7.1	Introduction	224
7.2	The empirical evidence: bird's eye view	225
7.3	Volatility: a business cycle perspective	231
7.3.1	Volatility cycles	231

7.3.2	Understanding the empirical evidence	233
7.3.3	What to do with stock market volatility?	238
7.3.4	What did we learn?	244
7.4	Rational market fluctuations	245
7.4.1	The dynamics of asset returns	245
7.4.2	Volatility, options and convexity	246
7.5	Time-varying discount rates or uncertain growth?	251
7.5.1	Markov pricing kernels	252
7.5.2	External habit formation	253
7.5.3	Large price swings as a learning induced phenomenon	257
7.5.4	Linearity-generating processes	263
7.6	Appendix 1: Calibration of the tree in Section 7.3	267
7.7	Appendix 2: Asset prices in a multifactor model	269
7.8	Appendix 3: Arrow-Debreu PDEs	270
7.9	Appendix 4: The maximum principle	271
7.10	Appendix 5: Stochastic dominance	273
7.10.1	Classics	273
7.10.2	Dynamic	274
7.11	Appendix 6: Proof of Theorem 7.1	276
7.12	Appendix 7: Dynamics of habit in Campbell and Cochrane (1999)	277
7.13	Appendix 8: An algorithm to simulate discrete-time pricing models	279
7.14	Appendix 9: Heuristic details of learning in continuous time	280
7.15	Appendix 10: Linear regime-switching economies	281
7.16	Appendix 11: Bond price convexity revisited	282
	References	283
8	Tackling the puzzles	287
8.1	Introduction	287
8.2	Non-expected utility	289
8.2.1	Recursive formulation	289
8.2.2	Testable restrictions	290
8.2.3	Risk premiums and interest rates	290
8.2.4	Campbell-Shiller approximation	292
8.2.5	Risks for the long-run	292
8.3	Heterogeneous agents and “catching up with the Joneses”	293
8.4	Idiosyncratic risk	295
8.5	Incomplete markets and heterogenous agents	298
8.6	Economies with production	301
8.7	Leverage and volatility	302
8.7.1	Model	303
8.8	Multiple trees and the cross-section of asset returns	306
8.9	The term-structure of interest rates	306
8.10	Prices, quantities and the separation hypothesis	308

8.11	Appendix 1: Non-expected utility	309
8.11.1	Detailed derivation of optimality conditions and selected relations	309
8.11.2	Details concerning models of long-run risks	312
8.11.3	Continuous time	312
8.12	Appendix 2: Economies with heterogenous agents	313
	References	317
9	Information and other market frictions	320
9.1	Introduction	320
9.2	Prelude: imperfect information in macroeconomics	321
9.3	Grossman-Stiglitz paradox	323
9.4	Noisy rational expectations equilibrium	323
9.4.1	Differential information	323
9.4.2	Asymmetric information	323
9.4.3	Information acquisition	323
9.5	Strategic trading	323
9.6	Dealers markets	324
9.7	Noise traders	324
9.8	Demand-based derivative prices	324
9.8.1	Options	324
9.8.2	Preferred habitat and the yield curve	324
9.9	Over-the-counter markets	324
	References	325
III	Asset pricing and reality	326
10	Options and volatility	327
10.1	Introduction	327
10.2	Forwards	327
10.2.1	Pricing	327
10.2.2	Forwards as a means to borrow money	327
10.2.3	A pricing formula	328
10.2.4	Forwards and volatility	328
10.3	Optionality and no-arb bounds	328
10.3.1	Model-free properties	329
10.3.2	A case study: accumulators, decumulators	332
10.4	Evaluation and hedging	334
10.4.1	Spanning and cloning	334
10.4.2	Black & Scholes	335
10.4.3	Surprising cancellations and “preference-free” formulae	337
10.4.4	Future options and Black’s formula	337
10.4.5	Hedging	337

10.4.6	Endogenous volatility	338
10.4.7	Marking to market	340
10.4.8	Properties of options in diffusive models	340
10.5	Stochastic volatility	343
10.5.1	Statistical models of changing volatility	343
10.5.2	Implied volatility, smiles and skews	344
10.5.3	Option pricing with stochastic volatility	349
10.6	Trading volatility with options	356
10.6.1	Payoffs	356
10.6.2	P&Ls of Δ -hedged strategies	360
10.7	Local volatility	362
10.7.1	Issues	362
10.7.2	The perfect fit	363
10.7.3	Relations with implied volatility	364
10.8	The price of volatility	366
10.8.1	Evaluation	367
10.8.2	Forward volatility trading	370
10.8.3	Marking to market	370
10.8.4	Stochastic interest rates	371
10.8.5	Hedging	371
10.9	Skewness contracts	372
10.10	American options	374
10.10.1	Real options theory	374
10.10.2	Perpetual puts	375
10.10.3	Perpetual calls	376
10.11	A few exotics	378
10.12	Market imperfections	378
10.13	Appendix 1: The original arguments underlying the Black & Scholes formula	379
10.14	Appendix 2: Black (1976)	380
10.15	Appendix 3: Stochastic volatility	381
10.15.1	Hull & White equation	381
10.15.2	Extensions	381
10.15.3	Smile analytics	382
10.16	Appendix 4: Local volatility	384
10.17	Appendix 5: Volatility contracts	386
10.18	Appendix 6: Skewness contracts	389
	References	390
11	The engineering of fixed income securities	393
11.1	Introduction	393
11.1.1	Relative pricing in fixed income markets	394
11.1.2	Many evaluation paradigms	394
11.1.3	Plan of the chapter	394

11.2	Markets and interest rate conventions	395
11.2.1	Markets for interest rates	395
11.2.2	Mathematical definitions of interest rates	397
11.2.3	Yields to maturity on coupon bearing bonds	399
11.3	Curve fitting	399
11.3.1	Extracting zeros from bond prices	399
11.3.2	Bootstrapping	400
11.3.3	Splines	401
11.3.4	Arbitrage	402
11.4	Duration and convexity hedging and trading	405
11.4.1	Duration	406
11.4.2	Convexity	407
11.4.3	Asset-liability management	407
11.5	Foundational issues in interest rate modeling	415
11.5.1	Tree representation of the short-term rate	416
11.5.2	Tree pricing	420
11.5.3	Introduction to calibration	422
11.5.4	Calibrating probabilities throught derivative data	436
11.6	The Ho and Lee model	444
11.6.1	The tree	445
11.6.2	The price movements and the martingale restriction	445
11.6.3	The recombining condition	446
11.6.4	Calibration of the model	449
11.6.5	An example	449
11.6.6	Continuous-time approximations, with an application to barbell trading	454
11.7	Beyond Ho and Lee: Calibration	458
11.7.1	Arrow-Debreu securities	459
11.7.2	The algorithm in two examples	461
11.8	Callables, puttable and convertibles with trees	470
11.8.1	Callable bonds	471
11.8.2	Convertible bonds	476
11.9	Appendix 1: Proof of Eq. (11.18)	480
11.10	Appendix 2: The Ho and Lee price representation	482
	References	484
12	Interest rates	485
12.1	Introduction	485
12.2	Prices and interest rates	486
12.2.1	Bond prices	486
12.2.2	Forward martingale probabilities	489
12.2.3	Stochastic duration	491
12.3	Stylized facts	492
12.3.1	The expectation hypothesis, and bond returns predictability	492

12.3.2	The yield curve and the business cycle	494
12.3.3	Additional stylized facts about the US yield curve	496
12.3.4	Common factors affecting the yield curve	496
12.4	Models of the short-term rate	499
12.4.1	Models versus representations	499
12.4.2	The bond pricing equation	500
12.4.3	Some famous short-term rate models	503
12.4.4	Multifactor models	509
12.4.5	Affine and quadratic term-structure models	514
12.4.6	Short-term rates as jump-diffusion processes	515
12.4.7	Some stylized facts and estimation strategies	517
12.5	No-arbitrage models: early formulations	522
12.5.1	Fitting the yield-curve, perfectly	522
12.5.2	Ho & Lee	524
12.5.3	Hull & White	525
12.6	The Heath-Jarrow-Morton framework	526
12.6.1	Framework	526
12.6.2	The model	527
12.6.3	The dynamics of the short-term rate	528
12.6.4	Embedding	528
12.7	Stochastic string shocks models	529
12.7.1	Addressing stochastic singularity	530
12.7.2	No-arbitrage restrictions	531
12.8	Interest rate derivatives	532
12.8.1	Introduction	532
12.8.2	A put-call parity for fixed income markets	532
12.8.3	European options on bonds	533
12.8.4	Callable and puttable bonds	537
12.8.5	Related fixed income products	540
12.8.6	Market models	546
12.9	Appendix 1: The FTAP for bond prices	552
12.10	Appendix 2: Certainty equivalent interpretation of forward prices	554
12.11	Appendix 3: Additional results on T -forward martingale probabilities	555
12.12	Appendix 4: Principal components analysis	556
12.13	Appendix 5: A few analytics for the Hull and White model	557
12.14	Appendix 6: Expectation theory and embedding in selected models	558
12.15	Appendix 7: Additional results on string models	560
12.16	Appendix 8: Changes of numéraire and Jamshidian's (1989) formula	561
	References	562
13	Risky debt and credit derivatives	566
13.1	Introduction	566
13.2	The classics: Modigliani-Miller irrelevance results	566

13.3	Conceptual approaches to valuation of defaultable securities	568
13.3.1	Firm's value, or structural, approaches	568
13.3.2	An application of the structural approach: the pricing of convertible bonds	582
13.3.3	Reduced form approaches: rare events, or intensity, models	585
13.3.4	Ratings	589
13.4	Credit derivatives, and structured products based thereon	593
13.4.1	A brief history of credit risk and financial innovation	593
13.4.2	Options and spreads	596
13.4.3	Credit Default Swaps	597
13.4.4	Collateralized Debt Obligations (CDOs)	614
13.5	Procyclicality, credit crunches and quantitative easing	626
13.5.1	Regulatory framework	627
13.5.2	The 2007 subprime crisis	630
13.5.3	Top tier capital ratio targets and endogenous volatility	634
13.5.4	Credit crunches and quantitative easing	640
13.6	A few hints on the risk-management practice	643
13.6.1	Value at Risk (VaR)	643
13.6.2	Backtesting	646
13.6.3	Stress testing	647
13.6.4	Credit risk and VaR	648
13.7	Appendix 1: Present values contingent on future bankruptcy	650
13.8	Appendix 2: Proof of selected results	651
13.9	Appendix 3: Transition probability matrices and pricing	652
13.10	Appendix 4: Bond spreads in markets with stochastic default intensity	654
13.11	Appendix 5: Conditional probabilities of survival	655
13.12	Appendix 6: Details regarding CDS index swaps and swaptions	656
13.13	Appendix 7: Modeling correlation with copulae functions	658
13.14	Appendix 8: Details on CDO pricing with imperfect correlation	660
	References	661

Part I

Foundations

1

The classic capital asset pricing model

1.1 Introduction

An investor is concerned with the choice of assets to include in a portfolio. Which weights does each asset need to bear for the investor to maximize some utility criterion? What are the asset pricing implications of market-wide optimal portfolio choices? How do these choices relate to the basic requirement that there are no arbitrage opportunities left available in the markets? This chapter deals with these issues within the context of a static market, one where the notion of time does not affect choices and prices. The next section deals with portfolio selection problems when our investor maximizes a mean-variance criterion, as in the seminal approach of Markovitz (1952). Optimal portfolio choices like these naturally lead to a notion of market-wide market portfolio, and asset pricing implications, summarized by the CAPM (capital asset pricing model), and developed in Section 1.3. The CAPM predicts that each asset expected return links to the market portfolio. It is, of course, a quite coarse description of asset markets. Section 1.4 develops the APT (arbitrage pricing theory) model, which provides refinements of the CAPM, predicting that each asset return does relate to a number of factors, under the assumption of absence of arbitrage.

1.2 Portfolio selection

We begin with the derivation of wealth constraint. Second, we illustrate the main results of the model, with and without a safe asset. Third, we introduce the notion of market portfolio.

1.2.1 *The wealth constraint*

The space choice comprises m risky assets, and some safe asset. Let $S = [S_1, \dots, S_m]$ be the risky assets price vector, and let S_0 be the price of the riskless asset. We wish to evaluate the value of a portfolio that contains all these assets. Let $\theta = [\theta_1, \dots, \theta_m]$, where θ_i is the number of the i -th risky asset, and let θ_0 be the number of the riskless assets, in this portfolio. The

initial wealth is, $w = S_0\theta_0 + S \cdot \theta$. Terminal wealth is $w^+ = x_0\theta_0 + x \cdot \theta$, where x_0 is the payoff promised by the riskless asset, and $x = [x_1, \dots, x_m]$ is the vector of the payoffs pertaining to the risky assets, i.e. x_i is the payoff of the i -th asset.

The following pieces of notation considerably simplify the presentation. Let $R \equiv \frac{x_0}{S_0}$, and $\tilde{R}_i \equiv \frac{x_i}{S_i}$. In words, R is the gross interest rate obtained by investing in a safe asset, and \tilde{R}_i is the gross return obtained by investing in the i -th risky asset. Accordingly, we define $r \equiv R - 1$ as the safe interest rate; $\tilde{b} = [\tilde{b}_1, \dots, \tilde{b}_m]$, where $\tilde{b}_i \equiv \tilde{R}_i - 1$ is the rate of return on the i -th asset; and $b \equiv E(\tilde{b})$, the vector of the expected returns on the risky assets. Finally, we let $\pi = [\pi_1, \dots, \pi_m]$, where $\pi_i \equiv \theta_i S_i$ is the wealth invested in the i -th asset. We have,

$$w^+ = x_0\theta_0 + \sum_{i=1}^m x_i\theta_i \equiv R\pi_0 + \sum_{i=1}^m \tilde{R}_i\pi_i \quad \text{and} \quad w = \pi_0 + \sum_{i=1}^m \pi_i. \quad (1.1)$$

Combining the two expressions for w^+ and w , we obtain, after a few computations,

$$w^+ = \pi^\top (\tilde{R} - \mathbf{1}_m R) + R w = \pi^\top (b - \mathbf{1}_m r) + R w + \pi^\top (\tilde{b} - b).$$

We use the decomposition, $\tilde{b} - b = a \cdot \tilde{u}$, where a is a $m \times d$ “volatility” matrix, with $m \leq d$, and \tilde{u} is a random vector with expectation zero and variance-covariance matrix equal to the identity matrix. With this decomposition, we can rewrite the budget constraint in Eq. (1.1) as follows:

$$w^+ = \pi^\top (b - \mathbf{1}_m r) + R w + \pi^\top a \tilde{u}. \quad (1.2)$$

We now use Eq. (1.2) to compute the expected return and the variance of the portfolio value. We have,

$$E[w^+(\pi)] = \pi^\top (b - \mathbf{1}_m r) + R w \quad \text{and} \quad \text{var}[w^+(\pi)] = \pi^\top \Sigma \pi \quad (1.3)$$

where $\Sigma \equiv a a^\top$. Let $\sigma_i^2 \equiv \Sigma_{ii}$. We assume that Σ has full-rank, and that,

$$\sigma_i^2 > \sigma_j^2 \Rightarrow b_i > b_j \quad \text{all } i, j,$$

which implies that $r < \min_j(b_j)$.

1.2.2 Portfolio choice

We assume that the investor maximizes the expected return on his portfolio, given a certain level of the variance of the portfolio’s value, which we set equal to $w^2 \cdot v_p^2$. We use Eq. (1.3) to set up the following program

$$\hat{\pi}(v_p) = \arg \max_{\pi \in \mathbb{R}^m} E[w^+(\pi)] \quad \text{s.t.} \quad \text{var}[w^+(\pi)] = w^2 \cdot v_p^2. \quad [1.P1]$$

The first order conditions for [1.P1] are,

$$\hat{\pi}(v_p) = (2\nu)^{-1} \Sigma^{-1} (b - \mathbf{1}_m r) \quad \text{and} \quad \hat{\pi}^\top \Sigma \hat{\pi} = w^2 \cdot v_p^2,$$

where ν is a Lagrange multiplier for the variance constraint. By plugging the first condition into the second, we obtain, $(2\nu)^{-1} = \mp \frac{w \cdot v_p}{\sqrt{\text{Sh}}}$, where

$$\text{Sh} \equiv (b - \mathbf{1}_m r)^\top \Sigma^{-1} (b - \mathbf{1}_m r), \quad (1.4)$$

is the *Sharpe market performance*. To ensure efficiency, we take the positive solution. Substituting the positive solution for $(2\nu)^{-1}$ into the first order condition, we obtain that the portfolio that solves [1.P1] is

$$\frac{\hat{\pi}(v_p)}{w} \equiv \frac{\Sigma^{-1}(b - \mathbf{1}_m r)}{\sqrt{\text{Sh}}} \cdot v_p. \quad (1.5)$$

We are now ready to calculate the value of [1.P1], $E[w^+(\hat{\pi}(v_p))]$ and, hence, the expected portfolio return, defined as,

$$\mu_p(v_p) \equiv \frac{E[w^+(\hat{\pi}(v_p))] - w}{w} = r + \sqrt{\text{Sh}} \cdot v_p, \quad (1.6)$$

where the last equality follows by simple computations. Eq. (1.6) describes what is known as the *Capital Market Line* (CML).

1.2.3 Without the safe asset

Next, let us suppose the investor's space choice does not include the riskless asset. In this case, his current wealth is $w = \sum_{i=1}^m \pi_i$, and his terminal wealth is $w^+ = \sum_{i=1}^m \tilde{R}_i \pi_i$. By the definition of $\tilde{b}_i \equiv \tilde{R}_i - 1$, and by a few simple computations,

$$w^+ = \sum_{i=1}^m \tilde{b}_i \pi_i + \sum_{i=1}^m \pi_i = \pi^\top b + w + \pi^\top a \tilde{u}, \quad (1.7)$$

where a and \tilde{u} are as defined as in Eq. (1.2). We can use Eq. (1.7) to compute the expected return and the variance of the portfolio value, which are:

$$E[w^+(\pi)] = \pi^\top b + w, \text{ where } w = \pi^\top \mathbf{1}_m \text{ and } \text{var}[w^+(\pi)] = \pi^\top \Sigma \pi. \quad (1.8)$$

The program our investor solves, now, is:

$$\hat{\pi}(v_p) = \arg \max_{\pi \in \mathbb{R}} E[w^+(\pi)] \quad \text{s.t. } \text{var}[w^+(\pi)] = w^2 \cdot v_p^2 \text{ and } w = \pi^\top \mathbf{1}_m. \quad [1.P2]$$

In the appendix, we show that provided $\alpha\gamma - \beta^2 > 0$ (a second order condition), the solution to [1.P2] is,

$$\frac{\hat{\pi}(v_p)}{w} = \frac{\gamma\mu_p(v_p) - \beta}{\alpha\gamma - \beta^2} \Sigma^{-1} b + \frac{\alpha - \beta\mu_p(v_p)}{\alpha\gamma - \beta^2} \Sigma^{-1} \mathbf{1}_m, \quad (1.9)$$

where $\alpha \equiv b^\top \Sigma^{-1} b$, $\beta \equiv \mathbf{1}_m^\top \Sigma^{-1} b$ and $\gamma \equiv \mathbf{1}_m^\top \Sigma^{-1} \mathbf{1}_m$, and $\mu_p(v_p)$ is the expected portfolio return, defined as in Eq. (1.6). In the appendix, we also show that,

$$v_p^2 = \frac{1}{\gamma} \left[1 + \frac{1}{\alpha\gamma - \beta^2} (\gamma\mu_p(v_p) - \beta)^2 \right]. \quad (1.10)$$

Therefore, the *global minimum variance portfolio* achieves a variance equal to $v_p^2 = \gamma^{-1}$ and an expected return equal to $\mu_p = \beta/\gamma$.

Note that for each v_p , there are two values of $\mu_p(v_p)$ that solve Eq. (1.10). The optimal choice for our investor is that with the highest μ_p . We define the *efficient portfolio frontier* as the set of values (v_p, μ_p) that solve Eq. (1.10) with the highest μ_p . It has the following expression,

$$\mu_p(v_p) = \frac{\beta}{\gamma} + \frac{1}{\gamma} \sqrt{(\gamma v_p^2 - 1)(\alpha\gamma - \beta^2)}. \quad (1.11)$$

Clearly, the efficient portfolio frontier is an increasing and concave function of v_p . It can be interpreted as a sort of “production function,” one that produces “expected returns” through inputs of “levels of risk” (see, e.g., Figure 1.1). The choice of which portfolio has effectively to be selected depends on the investor’s preference toward risk.

EXAMPLE 1.1. Let the number of risky assets $m = 2$. In this case, we do not need to optimize anything, as the budget constraint, $\frac{\pi_1}{w} + \frac{\pi_2}{w} = 1$, pins down an unique relation between the expected portfolio return and the variance of the portfolio’s value. We simply have, $\mu_p = \frac{E[w^+(\pi)] - w}{w} = \frac{\pi_1}{w}b_1 + \frac{\pi_2}{w}b_2$, or,

$$\begin{cases} \mu_p = b_1 + (b_2 - b_1)\frac{\pi_2}{w} \\ v_p^2 = \left(1 - \frac{\pi_2}{w}\right)^2 \sigma_1^2 + 2\left(1 - \frac{\pi_2}{w}\right)\frac{\pi_2}{w}\sigma_{12} + \left(\frac{\pi_2}{w}\right)^2 \sigma_2^2 \end{cases}$$

whence:

$$v_p = \frac{1}{b_2 - b_1} \sqrt{(b_2 - \mu_p)^2 \sigma_1^2 + 2(b_2 - \mu_p)(\mu_p - b_1)\rho\sigma_1\sigma_2 + (\mu_p - b_1)^2 \sigma_2^2}$$

When $\rho = 1$,

$$\mu_p = b_1 + \frac{(b_1 - b_2)(\sigma_1 - v_p)}{\sigma_2 - \sigma_1}.$$

In Appendix 4, we use an even simpler version of this model to explain how Tobin (1958) reformulated Keynesian theories predicting that money demand is inversely related to the nominal interest rate.

In the general case, diversification pays, provided that asset returns are not perfectly positively correlated. As Figure 1.1 reveals, we may even achieve a portfolio less risky than the less risky asset. Moreover, risk can be zeroed when $\rho = -1$, which corresponds to $\frac{\pi_1}{w} = \frac{\sigma_2}{\sigma_2 - \sigma_1}$ and $\frac{\pi_2}{w} = -\frac{\sigma_1}{\sigma_2 - \sigma_1}$ or, alternatively, to $\frac{\pi_1}{w} = -\frac{\sigma_2}{\sigma_2 - \sigma_1}$ and $\frac{\pi_2}{w} = \frac{\sigma_1}{\sigma_2 - \sigma_1}$.

Let us return to the general case. The portfolio in Eq. (1.9) can be decomposed into two components, as follows:

$$\frac{\hat{\pi}(v_p)}{w} = \ell(v_p) \frac{\pi_d}{w} + [1 - \ell(v_p)] \frac{\pi_g}{w}, \quad \ell(v_p) \equiv \frac{\beta(\mu_p(v_p)\gamma - \beta)}{\alpha\gamma - \beta^2},$$

where

$$\frac{\pi_d}{w} \equiv \frac{\Sigma^{-1}b}{\beta}, \quad \frac{\pi_g}{w} \equiv \frac{\Sigma^{-1}\mathbf{1}_m}{\gamma}.$$

Hence, we see that $\frac{\pi_g}{w}$ is the *global minimum variance portfolio*, for we know from Eq. (1.10) that the minimum variance occurs at $(v_p, \mu_p) = \left(\sqrt{\frac{1}{\gamma}}, \frac{\beta}{\gamma}\right)$, in which case $\ell(v_p) = 0$.¹ More generally, we can span any portfolio on the frontier by just choosing a convex combination of $\frac{\pi_d}{w}$ and $\frac{\pi_g}{w}$, with weight equal to $\ell(v_p)$. It’s a *mutual fund separation theorem*.

¹It is easy to show that the covariance of the global minimum variance portfolio with any other portfolio equals γ^{-1} .

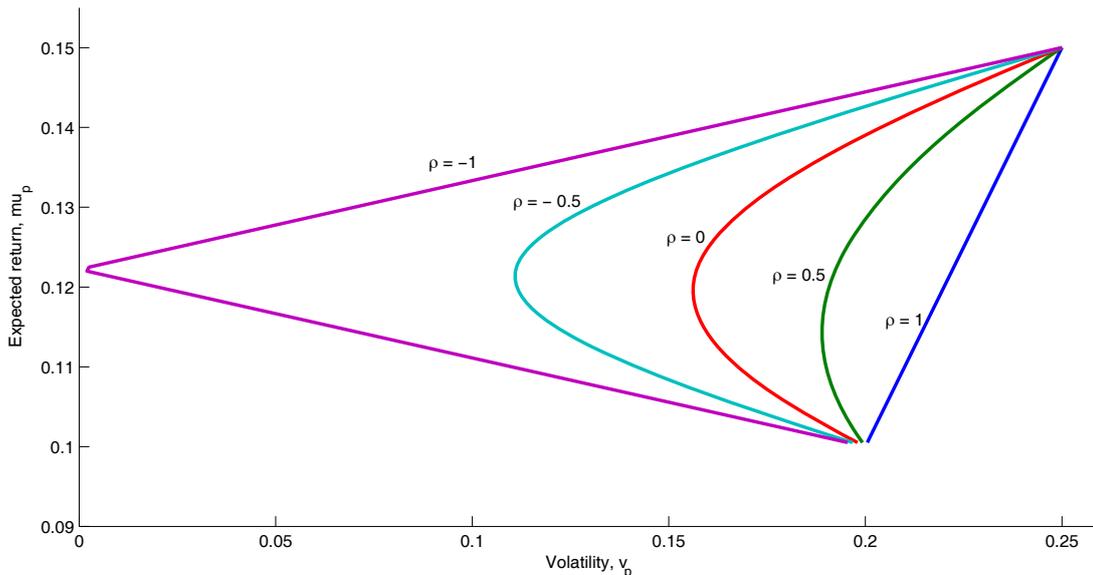


FIGURE 1.1. From top to bottom: portfolio frontiers corresponding to $\rho = -1, -0.5, 0, 0.5, 1$. Parameters are set to $b_1 = 0.10$, $b_2 = 0.15$, $\sigma_1 = 0.20$, $\sigma_2 = 0.25$. For each portfolio frontier, the efficient portfolio frontier includes those portfolios which yield the lowest volatility for a given expected return.

1.2.4 The market portfolio

The *market portfolio* is the portfolio at which the CML in Eq. (1.6) and the efficient portfolio frontier in Eq. (1.11) intersect. In fact, the market portfolio is the point at which the CML is *tangent* at the efficient portfolio frontier. For this reason, the market portfolio is also referred to as the “tangent” portfolio. In Figure 1.2, the market portfolio corresponds to the point M (the portfolio with volatility equal to v_M and expected return equal to μ_M), which is the point at which the CML is tangent to the efficient portfolio frontier, AMC .²

As Figure 1.2 illustrates, the CML dominates the efficient portfolio frontier AMC . This is because the CML is the value of the investor’s problem, [1.P1], obtained using all the risky assets *and* the riskless asset, and the efficient portfolio frontier is the value of the investor’s problem, [1.P2], obtained using *only* all the risky assets.³ For the same reason, the CML and the efficient portfolio frontier can only be tangent with each other. For suppose not. Then, there would exist a point on the efficient portfolio frontier that dominates some portfolio on the CML, a contradiction. Likewise, the CML must have a portfolio in common with the efficient portfolio frontier - the portfolio that does not include the safe asset. Below, we shall use this insight to characterize, analytically, the market portfolio.

Why is the market portfolio called in this way? Figure 1.2 reveals that any portfolio on the CML can be obtained as a combination of the safe asset and the market portfolio M (a portfolio

²The existence of the market portfolio requires a restriction on r , derived in Eq. (1.12) below.

³Figure 1.2 also depicts the dotted line MZ , which is the value of the investor’s problem when he invests a proportion higher than 100% in the market portfolio, leveraged at an interest rate for borrowing higher than the interest rate for lending. In this case, the CML coincides with rM , up to the point M . From M onwards, the CML coincides with the highest between MZ and MA .

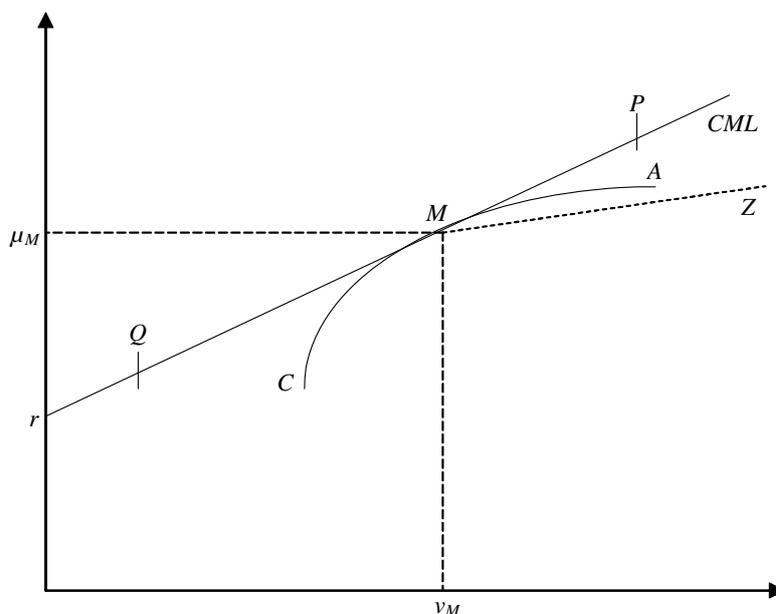


FIGURE 1.2.

containing only the risky assets). An investor with high risk-aversion would like to choose a point such as Q , say. An investor with low risk-aversion would like to choose a point such as P , say. But no matter how risk averse an individual is, the optimal solution for him is to choose a combination of the safe asset and the market portfolio M . Thus, the market portfolio plays an instrumental role. It obviously does not depend on the risk attitudes of any investor - it is a mere convex combination of all the existing assets in the economy. Instead, the optimal course of action for any investor is to use those proportions of this portfolio that make his overall exposure to risk consistent with his risk appetite. It's a *two fund separation theorem*.

The equilibrium implications of this separation theorem are as follows. As we have explained, any portfolio can be attained by lending or borrowing funds in zero net supply, and in the portfolio M . In equilibrium, then, every investor must hold some proportions of M . But since in aggregate, there is no net borrowing or lending, one has that in aggregate, all investors must have portfolio holdings that sum up to the market portfolio, which is therefore the value-weighted portfolio of all the existing assets in the economy. This argument is formally developed in the appendix.

We turn to characterize the market portfolio. We need to assume that the interest rate is sufficiently low to allow the CML to be tangent at the efficient portfolio frontier. The technical condition that ensures this is that the return on the safe asset be less than the expected return on the global minimum variance portfolio, viz

$$r < \frac{\beta}{\gamma}. \quad (1.12)$$

Let π_M be the market portfolio. To identify π_M , we note that it belongs to *AMC* if $\pi_M^\top \mathbf{1}_m = w$, where π_M also belongs to the CML and, therefore, by Eq. (1.5), is such that:

$$\frac{\pi_M}{w} = \frac{\Sigma^{-1}(b - \mathbf{1}_m r)}{\sqrt{\text{Sh}}} \cdot v_M. \quad (1.13)$$

Therefore, we must be looking for the value v_M that solves

$$w = \mathbf{1}_m^\top \pi_M = w \cdot \mathbf{1}_m^\top \frac{\Sigma^{-1}(b - \mathbf{1}_m r)}{\sqrt{\text{Sh}}} \cdot v_M,$$

i.e.

$$v_M = \frac{\sqrt{\text{Sh}}}{\beta - \gamma r}. \quad (1.14)$$

Then, we plug this value of v_M into the expression for π_M in Eq. (1.13) and obtain,⁴

$$\frac{\pi_M}{w} = \frac{1}{\beta - \gamma r} \Sigma^{-1}(b - \mathbf{1}_m r). \quad (1.15)$$

Once again, the market portfolio belongs to the efficient portfolio frontier. Indeed, on the one hand, the market portfolio can not be above the efficient portfolio frontier, as this would contradict the efficiency of the *AMC* curve, which is obtained by investing in the risky assets only; on the other hand, the market portfolio can not be below the efficient portfolio frontier, for by construction, it belongs to the CML which, as shown before, dominates the efficient portfolio frontier. In the appendix, we confirm, analytically, that the market portfolio does indeed enjoy the tangency condition.

1.3 The CAPM

The Capital Asset Pricing Model (CAPM) provides an asset evaluation formula. In this section, we derive the CAPM through arguments that have the same flavor as the original derivation of Sharpe (1964). The first step is the creation of a portfolio including a proportion α of wealth invested in any asset i and the remaining proportion $1 - \alpha$ invested in the market portfolio. Mathematically, we are considering an α -parametrized portfolio, with expected return and volatility given by:

$$\begin{cases} \tilde{\mu}_p \equiv \alpha b_i + (1 - \alpha)\mu_M \\ \tilde{v}_p \equiv \sqrt{(1 - \alpha)^2 \sigma_M^2 + 2(1 - \alpha)\alpha \sigma_{iM} + \alpha^2 \sigma_i^2} \end{cases} \quad (1.16)$$

where we have defined $\sigma_M \equiv v_M$. Clearly, the market portfolio, M , belongs to the α -parametrized portfolio. By the Example 1.1, the curve in (1.16) has the same shape as the curve $A'Mi$ in Figure 1.3. The curve $A'Mi$ lies below the efficient portfolio frontier *AMC*. This is because the efficient portfolio frontier is obtained by optimizing a mean-variance criterion over all the existing assets and, hence, dominates any portfolio that only comprises the two assets i and M . Suppose, for example, that the $A'Mi$ curve intersects the *AMC* curve; then, a feasible combination of assets (including some proportion α of the i -th asset and the remaining proportion

⁴While the market portfolio depends on r , this portfolio does not obviously include any share in the safe asset.

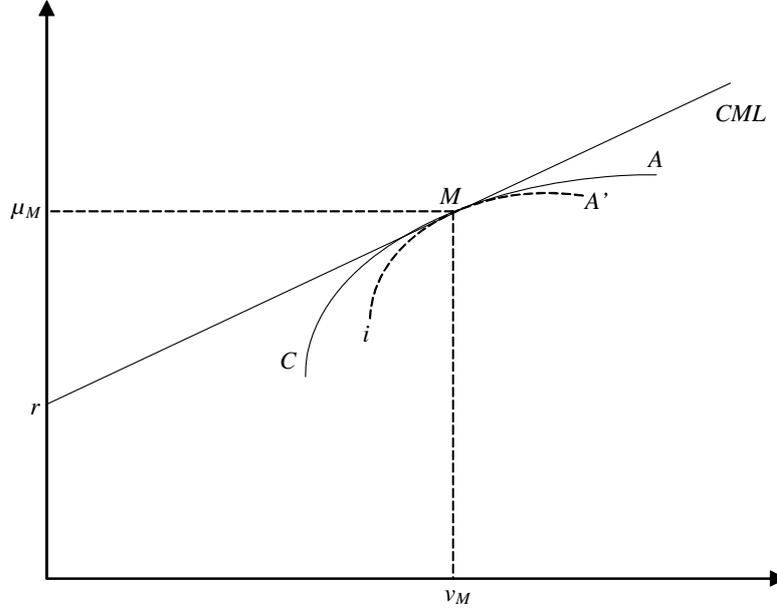


FIGURE 1.3.

$1 - \alpha$ of the market portfolio) would dominate AMC , a contradiction, given that AMC is the most efficient feasible combination of all the assets. On the other hand, the $A'Mi$ curve has a point in common with the AMC , which is M , in correspondence of $\alpha = 0$. Therefore, the curve $A'Mi$ is tangent to the efficient portfolio frontier AMC at M , which in turn, as we already know, is tangent to the CML at M .

Let us equate, then, the two slopes of the $A'Mi$ curve and the efficient portfolio frontier AMC at M . We shall show that this condition provides a restriction on the expected return b_i on any asset i . Because (1.16) is, mathematically, an α -parametrized curve, we may compute its slope at M through the computation of $d\tilde{\mu}_p/d\alpha$ and $d\tilde{v}_p/d\alpha$, at $\alpha = 0$. We have,

$$\frac{d\tilde{\mu}_p}{d\alpha} = b_i - \mu_M, \quad \left. \frac{d\tilde{v}_p}{d\alpha} \right|_{\alpha=0} = -\frac{-(1-\alpha)\sigma_M^2 + (1-2\alpha)\sigma_{iM} + \alpha\sigma_i^2|_{\alpha=0}}{\tilde{v}_p|_{\alpha=0}} = \frac{1}{\sigma_M} (\sigma_{iM} - \sigma_M^2).$$

Therefore,

$$\left. \frac{d\tilde{\mu}_p(\alpha)}{d\tilde{v}_p(\alpha)} \right|_{\alpha=0} = \frac{b_i - \mu_M}{\frac{1}{\sigma_M} (\sigma_{iM} - \sigma_M^2)}. \quad (1.17)$$

On the other hand, the slope of the CML is $(\mu_M - r)/\sigma_M$ which, equated to the slope in Eq. (1.17), yields,

$$b_i - r = \beta_i (\mu_M - r), \quad \beta_i \equiv \frac{\sigma_{iM}}{\sigma_M^2}, \quad i = 1, \dots, m. \quad (1.18)$$

Eq. (1.18) is the celebrated *Security Market Line* (SML). The appendix provides an alternative derivation of the SML. Assets with $\beta_i > 1$ are called “aggressive” assets. Assets with $\beta_i < 1$ are called “conservative” assets.

Note, the SML can be interpreted as a projection of the excess return on asset i (i.e. $\tilde{b}_i - r$) on the excess returns on the market portfolio (i.e. $\tilde{b}_M - r$). In other words,

$$\tilde{b}_i - r = \beta_i (\tilde{b}_M - r) + \varepsilon_i, \quad i = 1, \dots, m. \quad (1.19)$$

The previous relation leads to the following decomposition of the volatility (or risk) related to the i -th asset return:

$$\sigma_i^2 = \beta_i^2 v_M^2 + \text{var}(\varepsilon_i), \quad i = 1, \dots, m.$$

The quantity $\beta_i^2 v_M^2$ is usually referred to as *systematic* risk. The quantity $\text{var}(\varepsilon_i) \geq 0$, instead, is what we term *idiosyncratic* risk. In the next section, we shall show that idiosyncratic risk can be eliminated through a “well-diversified” portfolio - roughly, a portfolio that contains a large number of assets. Naturally, economic theory does not tell us anything substantial about how important idiosyncratic risk is for any particular asset.

The CAPM can be usefully interpreted within a classical hedging framework. Suppose we hold an asset that delivers a return equal to \tilde{z} - perhaps, a nontradable asset. We wish to hedge against movements of this asset by purchasing a portfolio containing a percentage of α in the market portfolio, and a percentage of $1 - \alpha$ units in a safe asset. The hedging criterion we wish to use is the variance of the overall exposure of the position, which we minimize by $\min_{\alpha} \text{var}[\tilde{z} - ((1 - \alpha)r + \alpha\tilde{b}_M)]$. It is straight forward to show that the solution to this basic problem is, $\hat{\alpha} \equiv \beta_{\tilde{z}} \equiv \text{cov}(\tilde{z}, \tilde{b}_M)/v_m^2$. That is, the proportion to hold is simply the beta of the asset to hedge with the market portfolio.

The CAPM is a model for the required return for any asset and so, it is a very first tool we can use to evaluate risky projects. Let

$$V = \text{value of a project} = \frac{E(C^+)}{1 + r_C},$$

where C^+ is future cash flow and r_C is the risk-adjusted discount rate for this project. We have:

$$\begin{aligned} \frac{E(C^+)}{V} &= 1 + r_C \\ &= 1 + r + \beta_C (\mu_M - r) \\ &= 1 + r + \frac{\text{cov}\left(\frac{C^+}{V} - 1, \tilde{x}_M\right)}{v_M^2} (\mu_M - r) \\ &= 1 + r + \frac{1}{V} \frac{\text{cov}(C^+, \tilde{x}_M)}{v_M^2} (\mu_M - r) \\ &= 1 + r + \frac{1}{V} \text{cov}(C^+, \tilde{x}_M) \frac{\lambda}{v_M}, \end{aligned}$$

where $\lambda \equiv \frac{\mu_M - r}{v_M}$, the unit market risk-premium.

Rearranging terms in the previous equation leaves:

$$V = \frac{E(C^+) - \frac{\lambda}{v_M} \text{cov}(C^+, \tilde{x}_M)}{1 + r}. \quad (1.20)$$

The certainty equivalent \bar{C} is defined as:

$$\bar{C} : V = \frac{E(C^+)}{1 + r_C} = \frac{\bar{C}}{1 + r},$$

or,

$$\bar{C} = (1 + r) V,$$

and using Eq. (1.20),

$$\bar{C} = E(C^+) - \frac{\lambda}{v_M} \text{cov}(C^+, \tilde{x}_M).$$

1.4 The APT

1.4.1 A first derivation

Suppose that the m asset returns we observe are generated by the following *linear factor model*,

$$\tilde{b}_{m \times 1} = a_{m \times 1} + B_{m \times k} \cdot f_{k \times 1} \equiv a + \text{cov}(\tilde{b}, f)[\text{var}(f)]^{-1} \cdot f \quad (1.21)$$

where a and B are a vector and a matrix of constants, and f is a k -dimensional vector of factors supposed to affect the asset returns, with $k \leq m$. Let us normalize $[\text{var}(f)]^{-1} = I_{k \times k}$, so that $B = \text{cov}(\tilde{b}, f)$. With this normalization, we have,

$$\tilde{b} = a + \begin{bmatrix} \text{cov}(\tilde{b}_1, f) \\ \vdots \\ \text{cov}(\tilde{b}_m, f) \end{bmatrix} \cdot f = a + \begin{bmatrix} \sum_{j=1}^k \text{cov}(\tilde{b}_1, f_j) f_j \\ \vdots \\ \sum_{j=1}^k \text{cov}(\tilde{b}_m, f_j) f_j \end{bmatrix}.$$

Next, let us consider a portfolio π including the m risky assets. The return of this portfolio is,

$$\pi^\top \tilde{b} = \pi^\top a + \pi^\top B f,$$

where as usual, $\pi^\top \mathbf{1}_m = 1$. An arbitrage opportunity arises if there exists some portfolio π such that the return on the portfolio is certain, and different from the safe interest rate r , i.e. if $\exists \pi : \pi^\top B = 0$ and $\pi^\top a \neq r$. Mathematically, this is ruled out whenever $\exists \lambda \in \mathbb{R}^k : a = B\lambda + \mathbf{1}_m r$. Substituting this relation into Eq. (1.21) leaves,

$$\tilde{b} = \mathbf{1}_m r + B\lambda + Bf = \mathbf{1}_m r + \text{cov}(\tilde{b}, f)\lambda + \text{cov}(\tilde{b}, f)f.$$

Taking the expectation,

$$b_i = r + (B\lambda)_i = r + \sum_{j=1}^k \underbrace{\text{cov}(\tilde{b}_i, f_j)}_{\equiv \beta_{i,j}} \lambda_j, \quad i = 1, \dots, m. \quad (1.22)$$

The APT collapses to the CAPM, once we assume that the only factor affecting the returns is the market portfolio. To show this, we must normalize the market portfolio return so that its variance equals one, consistently with Eq. (1.22). So let \tilde{r}_M be the normalized market return, defined as $\tilde{r}_M \equiv v_M^{-1} \tilde{b}_M$, so that $\text{var}(\tilde{r}_M) = 1$. We have,

$$\tilde{b}_i = a + \beta_i \tilde{r}_M, \quad i = 1, \dots, m,$$

where $\beta_i = \text{cov}(\tilde{b}_i, \tilde{r}_M) = v_M^{-1} \text{cov}(\tilde{b}_i, \tilde{b}_M)$. Then, we have,

$$b_i = r + \beta_i \lambda, \quad i = 1, \dots, m. \quad (1.23)$$

In particular, $\beta_M = \text{cov}(\tilde{b}_M, \tilde{r}_M) = v_M^{-1} \text{var}(\tilde{b}_M) = v_M$, and so, by Eq. (1.23),

$$\lambda = \frac{b_M - r}{v_M},$$

which is known as the *Sharpe ratio* for the market portfolio, or the market price of risk.

By replacing $\beta_i = v_M^{-1} \text{cov}(\tilde{b}_i, \tilde{b}_M)$ and the expression for λ above into Eq. (1.23), we obtain,

$$b_i = r + \frac{\text{cov}(\tilde{b}_i, \tilde{b}_M)}{v_M^2} (b_M - r), \quad i = 1, \dots, m.$$

This is simply the SML in Eq. (1.18).

1.4.2 The APT with idiosyncratic risk and a large number of assets

[Ross (1976), and Connor (1984), Huberman (1983).]

How can idiosyncratic risk be eliminated? Consider, for example, Eq. (1.19). Intuitively, we may form portfolios with a large number of assets, so as to make idiosyncratic risk negligible, by the law of large numbers. But would the beta-relation still hold, in this case? More in general, would the APT relation in Eq. (1.22) be still valid? The answer is in the affirmative, although it deserves some qualifications.

Consider the APT equation (1.21), and “add” a vector of idiosyncratic returns, ε , which are independent of f , and have mean zero and variance σ_ε^2 :

$$\tilde{b} = a + B \cdot f + \varepsilon.$$

We wish to show that in the absence of arbitrage, to be defined below, it must be that the number of assets such that Eq. (1.22) does *not* hold, $N(m)$ say, is bounded as m gets large, i.e.:

$$|a_i - ((B\lambda)_i + r)| > 0, \quad i = 1, \dots, N(m), \quad (1.24)$$

where

$$\lim_{m \rightarrow \infty} N(m) < \infty. \quad (1.25)$$

In other words, we wish to show that in a “large” market, Eq. (1.22) does indeed hold for most of the assets, an approach close to that in Huang and Litzenberger (1988, p. 106-108).

By the same arguments leading to Eq. (1.1), the wealth generated by a portfolio of the assets satisfying (1.24), $w_{N(m)}^+$ say, is,

$$w_{N(m)}^+ = \pi_{N(m)}^\top (a_{N(m)} - \mathbf{1}_{N(m)} r) + R w_{N(m)} + \pi_{N(m)}^\top (B_{N(m)} f + \varepsilon_{N(m)}),$$

where a_N , B_N and ε_N are (i) the vector of the expected returns, (ii) the return volatility (or factor exposures) matrix and (iii) the vector of idiosyncratic return components affecting these assets, and, finally, π_N and w_N are the portfolio and the initial wealth invested in these assets.

In this context, we may define an arbitrage as the portfolio $\pi_{N(m)}$ that in the limit, as the number of *all* the existing assets m gets large, is riskless and yet delivers an expected return strictly larger than the safe interest rate, viz

$$\lim_{m \rightarrow \infty} \frac{E[w_{N(m)}^+]}{w_{N(m)}} > R, \quad \text{and} \quad \lim_{m \rightarrow \infty} \text{var}[w_{N(m)}^+] \rightarrow 0. \quad (1.26)$$

We want to show that this situation does not arise, under the condition in (1.25), thereby establishing that the linear APT relation in Eq. (1.22) is valid for most of the assets, in a large market.

So suppose the linear relation, $a_N - \mathbf{1}_N r = B_N \lambda$, doesn't hold. Then, there exists a portfolio $\underline{\pi}$ such that,

$$\underline{\pi}^\top B_N = 0 \quad \text{and} \quad \underline{\pi}^\top (a_N - \mathbf{1}_N r) \neq 0. \quad (1.27)$$

Consider the portfolio:

$$\hat{\pi}_N = \frac{1}{N} \cdot \text{sign}(\underline{\pi}^\top (a_N - \mathbf{1}_N r)) \cdot \underline{\pi},$$

where $\underline{\pi}$ is as in (1.27). With this portfolio we have, clearly, that $E[w_N^+] = \hat{\pi}_N^\top (a_N - \mathbf{1}_N r) + R w_N > R w_N$, for each N , and even for N large. That is, $\lim_{m \rightarrow \infty} E[w_{N(m)}^+] / w_{N(m)} > R$, which is the first condition in (1.26). As regards the second condition in (1.26), we have that

$$\text{var}[w_N^+] = \hat{\pi}_N^\top (B_N B_N^\top + \sigma_\varepsilon^2 I_{N \times N}) \hat{\pi}_N = \sigma_\varepsilon^2 \hat{\pi}_N^\top \hat{\pi}_N,$$

where the second equality follows by the first relation in (1.27). Clearly, $\lim_{m \rightarrow \infty} \text{var}[w_{N(m)}^+] \rightarrow 0$ as $N(m) \rightarrow \infty$. Hence, in the absence of arbitrage, the condition in (1.25) must hold.

1.4.3 Empirical evidence

How to estimate Eq. (1.19)? Consider a slightly more general version of Eq. (1.19), where the safe interest rate is time-varying:

$$\tilde{b}_{i,t} - r_t = \beta_i (\tilde{b}_{M,t} - r_t) + \varepsilon_{i,t}, \quad i = 1, \dots, m,$$

where $\varepsilon_{i,t}$ denote “time-series residuals.” Fama and MacBeth (1973) consider the following procedure. In a first step, one obtains estimates of the exposures to the market, $\hat{\beta}_i$ say, for all stocks, using, for example, monthly returns, and approximating the market portfolio with some broad stock market index.⁵ In a second step, one runs cross-sectional regressions, one for each month,

$$\tilde{b}_{i,t} - r_t = \alpha_{it} + \lambda_t \hat{\beta}_i + \eta_{i,t}, \quad t = 1, \dots, T,$$

where T is the sample size and $\eta_{i,t}$ denote “cross-sectional residuals.” The time-series of cross-sectional estimates of the intercept $\alpha_{i,t}$ and the price of risk λ_t , $\hat{\alpha}_{i,t}$ and $\hat{\lambda}_t$ say, are, then, used to make statistical inference. For example, time-series averages and standard errors of $\hat{\alpha}_{i,t}$ and $\hat{\lambda}_t$ lead to point estimates and standard errors for $\alpha_{i,t}$ and λ_t . If the CAPM holds, estimates of α_i should not be significantly different from zero.

Chen, Roll and Ross (1986) use the Fama-MacBeth two-step procedure to estimate a multi-factor APT model, such as that in Section 1.4. They identify “macroeconomic forces” driving asset returns with the innovations in variables such as the term spread, expected and unexpected inflation, industrial production growth, or the corporate spread. They find that these sources of variation in the cross-section of asset returns are significantly priced.

⁵In tests of the CAPM, one uses proxies of the market portfolio, such as, say, the S&P 500. However, the market portfolio is unobservable. Roll (1977) points out that as a result, the CAPM is inherently untestable, as any test of the CAPM is a joint test of the model itself and of the closeness of the proxy to the market portfolio.

1.5 Appendix 1: Analytical details relating to portfolio choice

We derive Eq. (1.9), which is the solution to the portfolio choice, when the space choice does not include a safe asset. We derive the solution by solving two programs: (i) the primal program [1.P2] in the main text, amounting to maximizing the expected portfolio return, with a given variance of the portfolio value; and (ii) a dual program, to be introduced below, where we minimize the variance of the portfolio's value, with given portfolio expected return.

1.5.1 The primal program

Given Eq. (1.8), the Lagrangian function associated to [1.P2] is,

$$L = \pi^\top b + w - \nu_1(\pi^\top \Sigma \pi - w^2 \cdot v_p^2) - \nu_2(\pi^\top \mathbf{1}_m - w),$$

where ν_1 and ν_2 are two Lagrange multipliers. The first order conditions are,

$$\hat{\pi} = \frac{1}{2\nu_1} \Sigma^{-1} (b - \nu_2 \mathbf{1}_m), \quad \hat{\pi}^\top \Sigma \hat{\pi} = w^2 \cdot v_p^2, \quad \hat{\pi}^\top \mathbf{1}_m = w. \quad (1A.1)$$

Using the first and the third conditions, we obtain,

$$w = \mathbf{1}_m^\top \hat{\pi} = \frac{1}{2\nu_1} (\underbrace{\mathbf{1}_m^\top \Sigma^{-1} b}_{\equiv \beta} - \nu_2 \underbrace{\mathbf{1}_m^\top \Sigma^{-1} \mathbf{1}_m}_{\equiv \gamma}) \equiv \frac{1}{2\nu_1} (\beta - \nu_2 \gamma).$$

We can solve for ν_2 , obtaining,

$$\nu_2 = \frac{\beta - 2w\nu_1}{\gamma}.$$

By replacing the solution for ν_2 into the first condition in (1A.1) leaves,

$$\hat{\pi} = \frac{w}{\gamma} \Sigma^{-1} \mathbf{1}_m + \frac{1}{2\nu_1} \Sigma^{-1} \left(b - \frac{\beta}{\gamma} \mathbf{1}_m \right). \quad (1A.2)$$

Next, we derive the value of the program [1.P2]. We have,

$$E[w^+(\hat{\pi})] - w = \hat{\pi}^\top b = \frac{w}{\gamma} \underbrace{\mathbf{1}_m^\top \Sigma^{-1} b}_{\equiv \beta} + \frac{1}{2\nu_1} (\underbrace{b^\top \Sigma^{-1} b}_{\equiv \alpha} - \frac{\beta}{\gamma} \underbrace{\mathbf{1}_m^\top \Sigma^{-1} b}_{\equiv \beta}) = \frac{w}{\gamma} \beta + \frac{1}{2\nu_1} \left(\alpha - \frac{\beta^2}{\gamma} \right). \quad (1A.3)$$

It is easy to check that

$$\begin{aligned} \text{var}[w^+(\hat{\pi})] &= w^2 \cdot v_p^2 \\ &= \hat{\pi}^\top \Sigma \hat{\pi} \\ &= \left[\frac{w}{\gamma} \mathbf{1}_m^\top \Sigma^{-1} + \frac{1}{2\nu_1} \left(b^\top - \frac{\beta}{\gamma} \mathbf{1}_m^\top \right) \Sigma^{-1} \right] \left[\frac{w}{\gamma} \mathbf{1}_m + \frac{1}{2\nu_1} \left(b - \frac{\beta}{\gamma} \mathbf{1}_m \right) \right] \\ &= \frac{w^2}{\gamma} + \left(\frac{1}{2\nu_1} \right)^2 \left(\alpha - \frac{\beta^2}{\gamma} \right). \end{aligned} \quad (1A.4)$$

Let us gather Eqs. (1A.3) and (1A.4),

$$\begin{cases} \mu_p(v_p) \equiv \frac{E[w^+(\hat{\pi})] - w}{w} = \frac{\beta}{\gamma} + \frac{1}{2\nu_1 w} \left(\alpha - \frac{\beta^2}{\gamma} \right) \\ v_p^2 = \frac{1}{\gamma} + \left(\frac{1}{2\nu_1 w} \right)^2 \left(\alpha - \frac{\beta^2}{\gamma} \right) \end{cases} \quad (1A.5)$$

where we have emphasized the dependence of μ_p on v_p , which arises through the presence of the Lagrange multiplier ν_1 .

Let us rewrite the first equation in (1A.5) as follows,

$$\frac{1}{2\nu_1 w} = (\alpha\gamma - \beta^2)^{-1} (\gamma\mu_p(v_p) - \beta). \quad (1A.6)$$

We can use this expression for ν_1 to express $\hat{\pi}$ in Eq. (13.62) in terms of the portfolio expected return, $\mu_p(v_p)$. We have,

$$\frac{\hat{\pi}}{w} = \frac{\Sigma^{-1}\mathbf{1}_m}{\gamma} + (\alpha\gamma - \beta^2)^{-1} (\gamma\mu_p(v_p) - \beta) \left(\Sigma^{-1}b - \frac{\Sigma^{-1}\beta}{\gamma}\mathbf{1}_m \right).$$

By rearranging terms in the previous equation, we obtain Eq. (1.9) in the main text.

Finally, we substitute Eq. (1A.6) into the second equation in (1A.5), and obtain:

$$v_p^2 = \frac{1}{\gamma} \left[1 + (\alpha\gamma - \beta^2)^{-1} (\gamma\mu_p(v_p) - \beta)^2 \right],$$

which is Eq. (1.10) in the main text. Note, also, that the second condition in (1A.5) reveals that,

$$\left(\frac{1}{2\nu_1 w} \right)^2 = \frac{\gamma v_p^2 - 1}{\alpha\gamma - \beta^2}.$$

Given that $\alpha\gamma - \beta^2 > 0$, the previous equation confirms the properties of the *global minimum variance portfolio* stated in the main text.

1.5.2 The dual program

We now solve the dual program, defined as follows,

$$\hat{\pi} = \arg \min_{\pi \in \mathbb{R}^m} \text{var} \left[\frac{w^+(\pi)}{w} \right] \quad \text{s.t.} \quad E[w^+(\pi)] = E_p \quad \text{and} \quad w = \pi^\top \mathbf{1}_m, \quad [1A.P2-dual]$$

for some constant E_p . The first order conditions are

$$\frac{\hat{\pi}}{w} = \frac{\nu_1 w}{2} \Sigma^{-1} b + \frac{\nu_2 w}{2} \Sigma^{-1} \mathbf{1}_m \quad ; \quad \hat{\pi}^\top b = E_p - w \quad ; \quad w = \hat{\pi}^\top \mathbf{1}_m; \quad (1A.7)$$

where ν_1 and ν_2 are two Lagrange multipliers. By replacing the first condition in (8A.14) into the second one,

$$E_p - w = \hat{\pi}^\top b = w^2 \left(\underbrace{\frac{\nu_1}{2} b^\top \Sigma^{-1} b}_{\equiv \alpha} + \underbrace{\frac{\nu_2}{2} \mathbf{1}_m^\top \Sigma^{-1} b}_{\equiv \beta} \right) \equiv w^2 \left(\frac{\nu_1}{2} \alpha + \frac{\nu_2}{2} \beta \right). \quad (1A.8)$$

By replacing the first condition in (8A.14) into the third one,

$$w = \hat{\pi}^\top \mathbf{1}_m = w^2 \left(\underbrace{\frac{\nu_1}{2} b^\top \Sigma^{-1} \mathbf{1}_m}_{\equiv \beta} + \underbrace{\frac{\nu_2}{2} \mathbf{1}_m^\top \Sigma^{-1} \mathbf{1}_m}_{\equiv \gamma} \right) \equiv w^2 \left(\frac{\nu_1}{2} \beta + \frac{\nu_2}{2} \gamma \right). \quad (1A.9)$$

Next, let $\mu_p \equiv \frac{E_p - w}{w}$. By Eqs. (1A.8) and (1A.9), the solutions for ν_1 and ν_2 are,

$$\frac{\nu_1 w}{2} = \frac{\mu_p \gamma - \beta}{\alpha\gamma - \beta^2} \quad ; \quad \frac{\nu_2 w}{2} = \frac{\alpha - \beta \mu_p}{\alpha\gamma - \beta^2}$$

Therefore, the solution for the portfolio in Eq. (8A.14) is,

$$\frac{\hat{\pi}}{w} = \frac{\gamma\mu_p - \beta}{\alpha\gamma - \beta^2} \Sigma^{-1} b + \frac{\alpha - \beta\mu_p}{\alpha\gamma - \beta^2} \Sigma^{-1} \mathbf{1}_m.$$

Finally, the value of the program is,

$$\text{var} \left[\frac{w^+(\hat{\pi})}{w} \right] = \frac{1}{w^2} \hat{\pi}^\top \Sigma \hat{\pi} = \frac{1}{w} \hat{\pi}^\top \frac{\mu_p \gamma - \beta}{\alpha\gamma - \beta^2} b + \frac{1}{w} \hat{\pi}^\top \frac{\alpha - \mu_p \beta}{\alpha\gamma - \beta^2} \mathbf{1}_m = \frac{\gamma\mu_p^2 - 2\beta\mu_p + \alpha}{\alpha\gamma - \beta^2} = \frac{(\gamma\mu_p - \beta)^2}{(\alpha\gamma - \beta^2)\gamma} + \frac{1}{\gamma},$$

which is exactly Eq. (1.10) in the main text.

1.6 Appendix 2: The market portfolio

1.6.1 The tangent portfolio is the market portfolio

Let us define the market capitalization for any asset i as the value of all the assets i that are outstanding in the market, viz

$$\text{Cap}_i \equiv \bar{\theta}_i S_i, \quad i = 1, \dots, m,$$

where $\bar{\theta}_i$ is the number of assets i outstanding in the market. The market capitalization of all the assets is simply

$$\text{Cap}_M \equiv \sum_{i=1}^m \text{Cap}_i.$$

The market portfolio, then, is the portfolio with relative weights given by,

$$\bar{\pi}_{M,i} \equiv \frac{\text{Cap}_i}{\text{Cap}_M}, \quad i = 1, \dots, m.$$

Next, suppose there are N investors and that each investor j has wealth w_j , which he invests in two funds, a safe asset and the tangent portfolio. Let w_j^f be the wealth investor j invests in the safe asset and $w_j - w_j^f$ the remaining wealth the investor invests in the tangent portfolio. The tangent portfolio is defined as $\bar{\pi}_T \equiv \left(\frac{\pi_T}{w_j}\right)$, for some π_T solution to [1.P2], and is obviously independent of w_j (see Eq. (1.15) in the main text). The equilibrium in the stock market requires that

$$\text{Cap}_M \cdot \bar{\pi}_M = \sum_{j=1}^N (w_j - w_j^f) \bar{\pi}_T = \sum_{j=1}^N w_j \cdot \bar{\pi}_T = \text{Cap}_M \cdot \bar{\pi}_T.$$

where the second equality follows because the safe asset is in zero net supply and, hence, $\sum_{j=1}^N w_j^f = 0$; and the third equality holds because all the wealth in the economy is invested in stocks, in equilibrium.

1.6.2 Tangency condition

We check that the CML and the efficient portfolio frontier have the same slope in correspondence of the market portfolio. Let us impose the following tangency condition of the CML to the efficient portfolio frontier in Figure 1.2, *AMC*, at the point M :

$$\sqrt{\text{Sh}} = \frac{\alpha\gamma - \beta^2}{\gamma\mu_M - \beta} v_M. \quad (1A.10)$$

The left hand side of this equation is the slope of the CML, obtained through Eq. (1.6). The right hand side is the slope of the efficient portfolio frontier, obtained by differentiating $\mu_p(v)$ in the expression for the portfolio frontier in Eq. (1.11), and setting $v = v_M$ in

$$\frac{d\mu_p(v)}{dv} = \sqrt{(\gamma v^2 - 1)^{-1} (\alpha\gamma - \beta^2)} v = \frac{\alpha\gamma - \beta^2}{\gamma\mu_p(v) - \beta} v,$$

and where the second equality follows, again, by Eq. (1.11). By Eqs. (1A.10) and (1.14), we need to show that,

$$\frac{\gamma\mu_M - \beta}{\alpha\gamma - \beta^2} = \frac{1}{\beta - \gamma r}.$$

By plugging $\mu_M = r + \sqrt{\text{Sh}} \cdot v_M$ into the previous equality and rearranging terms,

$$v_M = \frac{\sqrt{\text{Sh}}}{\beta - \gamma r},$$

where we have made use of the equality $\text{Sh} = \alpha - 2\beta r + \gamma r^2$, obtained by elaborating on the definition of the Sharpe market performance Sh given in Eq. (1.4). This is indeed the variance of the market portfolio given in Eq. (1.14).

1.7 Appendix 3: An alternative derivation of the SML

The vector of covariances of the m asset returns with the market portfolio are:

$$\text{cov}(\tilde{x}, \tilde{x}_M) = \text{cov}\left(\tilde{x}, \tilde{x} \cdot \frac{\pi_M}{w}\right) = \Sigma \frac{\pi_M}{w} = \frac{1}{\beta - \gamma r} (b - \mathbf{1}_m r), \quad (1A.11)$$

where we have used the expression for the market portfolio given in Eq. (1.15). Next, premultiply the previous equation by $\frac{\pi_M^\top}{w}$ to obtain:

$$v_M^2 = \frac{\pi_M^\top}{w} \Sigma \frac{\pi_M}{w} = \frac{\pi_M^\top}{w} \frac{1}{\beta - \gamma r} (b - \mathbf{1}_m r) = \frac{1}{(\beta - \gamma r)^2} \text{Sh}, \quad (1A.12)$$

or $v_M = \frac{\sqrt{\text{Sh}}}{\beta - \gamma r}$, which confirms Eq. (1.14).

Let us rewrite Eq. (1A.11) component by component. That is, for $i = 1, \dots, m$,

$$\sigma_{iM} \equiv \text{cov}(\tilde{x}_i, \tilde{x}_M) = \frac{1}{\beta - \gamma r} (b_i - r) = \frac{v_M}{\sqrt{\text{Sh}}} (b_i - r) = \frac{v_M^2}{\mu_M - r} (b_i - r),$$

where the last two equalities follow by Eq. (1A.12) and by the relation, $\sqrt{\text{Sh}} = \frac{\mu_M - r}{v_M}$. By rearranging terms, we obtain Eq. (1.18).

1.8 Appendix 4: Liquidity traps, portfolio selection and the demand for money

Tobin (1958) used portfolio theory to model money demand, with the purpose of clarifying a few issues pertaining to the monetary theory in Keynes (1936). In Tobin’s interpretation, Keynesian theory entails that each agent can only make dichotomic choices, leading him to hold either money or bonds. A dichotomy could be easily avoided by adapting Markovitz portfolio selection theory to deal with these issues. It is useful to remind how dichotomic choices arise in the Keynesian world, according to Tobin. The first section of this appendix accomplishes this task. It also develops a simple example where in spite of a dichotomy at a microeconomic level, money demand is still negatively sloped with respect to the nominal rate, although flat in correspondence of small values of this nominal rate— a “liquidity trap.” The second section of this appendix summarizes Tobin’s contribution to money demand, developed by hinging upon Markovitz portfolio selection theory.

1.8.1 Dichotomy choices and aggregate money demand

Consider a perpetuity with coupons fixed and equal to 1, priced as $b_0 = \sum_{h=1}^{\infty} 1 \cdot (1 + i_0)^{-h} = i_0^{-1}$, where i_0 is the current nominal rate, assumed to be flat over all maturities. A crucial assumption is that each agent j “expects” that within a certain reference period, i_0 will converge to a “normal” rate, say $i_e(j) \equiv b_e^{-1}(j)$, with probability one, for all maturities. From the perspective of this agent, the capital gain from holding the perpetuity over this period is,

$$g_j(i_0) \equiv \frac{b_e(j) - b_0 + 1}{b_0} = \frac{i_0}{i_e(j)} - 1 + i_0.$$

It is easy to see that there exists a value of i_0 for each j , such that $g_j(i_0) = 0$, given by,

$$\hat{i}_0(j) \equiv \frac{i_e(j)}{1 + i_e(j)}.$$

It is a critical rate in that agent j would only invest in bonds if $i_0 > \hat{i}_0(j)$, would only demand cash if $i_0 < \hat{i}_0(j)$ and, finally, would be indifferent about the two choices if $i_0 = \hat{i}_0(j)$. Next, define $\xi \equiv \min_j \hat{i}_0(j) \equiv \hat{i}_0(j^*)$. For small ξ ,

$$\xi = \hat{i}_0(j^*) = \frac{i_e(j^*)}{1 + i_e(j^*)} \simeq i_e(j^*).$$

Given this approximation, when $i_0 = \xi$, every agent believes rates can only rise within the reference period, such that no one is willing to purchase any bond, as this purchase would lead to a sure loss. This situation is known as a *liquidity trap*: when $i_0 = \xi$, changes in money supply, be they positive or negative, do not affect interest rates. Indeed, at $i_0 = \xi$, the only investor holding bonds is simply the marginal investors j^* , who is indifferent about whether to hold money or bonds. If the central bank increases money supply by purchasing the bonds, this marginal investor would be perfectly ready to accept this new money and tender the bonds, as he is obviously indifferent between investing in bonds or hoarding money. Likewise, if the central bank decreases money supply through a bonds sale, the marginal investor would buy these bonds.

Yet an important point of Keynesian theory is that money demand is negatively sloped, at a macroeconomic level. We now develop an analytical example where this property holds true. Assume there are a continuum of agents on $[0, 1]$, ordered such that the distribution of $i_e(j)$ is uniform:

$$i_e(j) = \underline{\xi} + (\bar{\xi} - \underline{\xi})j, \quad j \in [0, 1],$$

for some two constants $\bar{\xi}$ and $\underline{\xi}$. Then,

$$\hat{i}_0(j) = \frac{\underline{\xi} + (\bar{\xi} - \underline{\xi})j}{1 + \underline{\xi} + (\bar{\xi} - \underline{\xi})j}.$$

The j -th agent choice relating to money demand, $m^d(j)$ say, is dichotomic, in that:

$$m^d(j) = \begin{cases} 1, & \text{if } \hat{i}_0(j) > i_0 \\ 0, & \text{otherwise} \end{cases}$$

Yet aggregate money demand is, after denoting $f(i_0) \equiv \frac{(1+\underline{\xi})i_0 - \underline{\xi}}{(\bar{\xi} - \underline{\xi})(1-i_0)}$,

$$M^d \equiv \int_0^1 m^d(j) dj = \int_0^1 \mathbb{I}_{\hat{i}_0(j) > i_0} dj = \int_{\{j: \hat{i}_0(j) > i_0\}} dj = \int_{f(i_0)}^1 dj = 1 - f(i_0) = \frac{\bar{\xi} - (1 + \bar{\xi})i_0}{(\bar{\xi} - \underline{\xi})(1 - i_0)},$$

where \mathbb{I} is the indicator function. Note that M^d is always positive, provided $i_0 \in \left(\frac{\underline{\xi}}{1+\underline{\xi}}, \frac{\bar{\xi}}{1+\bar{\xi}}\right)$. The interest rate relating to the liquidity trap is $\frac{\underline{\xi}}{1+\underline{\xi}}$. The next section explains how Tobin (1958) coped with degeneracy of interest rate expectations.

1.8.2 Money demand in a theory of portfolio selection

We consider a special case of Example 1.1 in the main text, arising when money is taken to be an asset, a safe asset such that its return and volatility are $b_1 \equiv 0$ and $\sigma_1 \equiv 0$. Bonds are, instead, risky, in that they offer a superior return but, also, a positive volatility. Therefore, the expected return and volatility of a portfolio of money and bonds are $\mu_p = b_2\pi_2$ and $v_p = \pi_2\sigma_2$, with straight forward notation. Therefore, we have that:

$$\mu_p = \frac{b_2}{\sigma_2} \cdot v_p, \quad \pi_2 = \frac{1}{\sigma_2} \cdot v_p. \quad (1A.13)$$

The top panels of next two pictures plots the first of Eqs. (1A.13), along with indifference curves. The optimum is achieved at point P , the point of tangency between the first of Eqs. (1A.13) and the indifference curve UU . Money demand, $1 - \pi_2$, is determined by the second of Eqs. (1A.13), and is shown in the bottom panels. As the expected return on the bond, b_2 , increases, the new optimum shifts to P' , the tangency point between the new relation in Eqs. (1A.13) and the indifference curve $U'U'$. Money demand decreases as a result.

This framework of analysis can be used to study the effects of a decreased interest rate volatility. Suppose the central bank has the power to lower both interest rates *and* interest rate volatility in such a way to keep the ratio b_2/σ_2 unchanged. As the second picture above illustrates, the optimum is still P , although then the second of Eqs. (1A.13) becomes steeper than before the policy action, by shifting from the line I to the line II . Money demand decreases as a result. We can say more. As is clear, money demand decreases as interest rate volatility, σ_2 , decreases. Therefore, the central bank might keep money supply constant and achieve lower interest rates by simply targeting low interest rate volatility.

[Show this in a sort of general equilibrium, and free up w , as we are implicitly assuming $w = 1$, for now.]

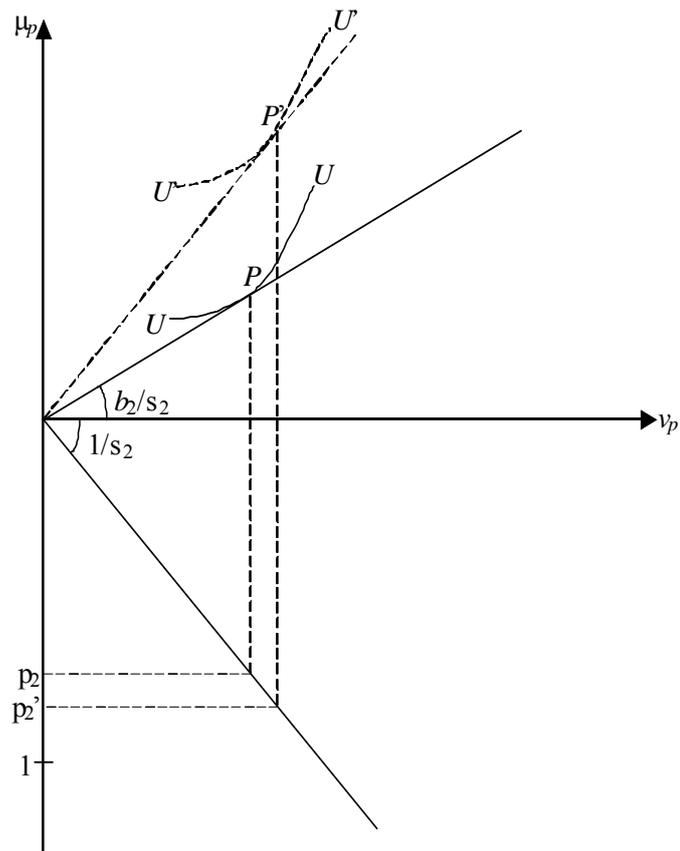


FIGURE 1.4.

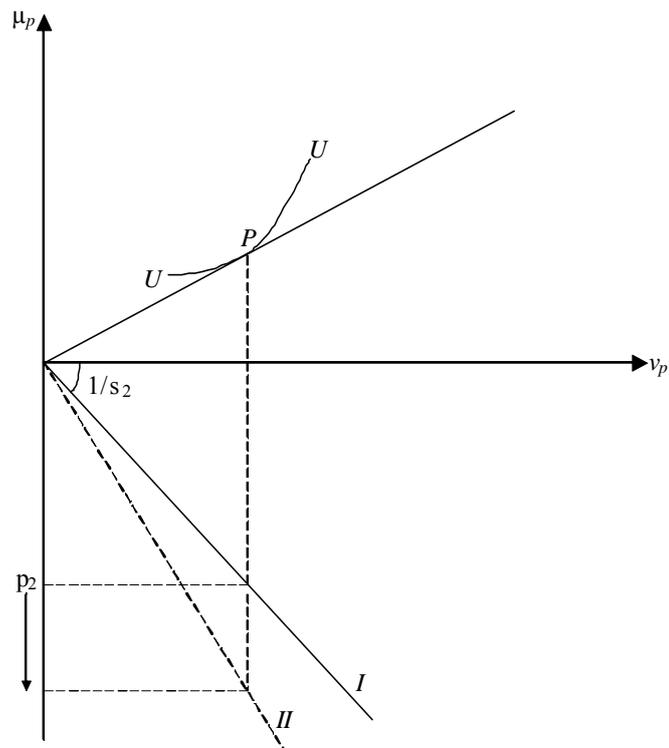


FIGURE 1.5.

References

- Chen, N-F., R. Roll and S.A. Ross (1986): "Economic Forces and the Stock Market." *Journal of Business* 59, 383-403.
- Connor, G. (1984): "A Unified Beta Pricing Theory." *Journal of Economic Theory* 34, 13-31.
- Fama, E.F. and J.D. MacBeth (1973): "Risk, Return, and Equilibrium: Empirical Tests." *Journal of Political Economy* 38, 607-636.
- Huang, C-f. and R.H. Litzenberger (1988): *Foundations for Financial Economics*. New York: North-Holland.
- Huberman, G. (1983): "A Simplified Approach to Arbitrage Pricing Theory." *Journal of Economic Theory* 28, 1983-1991.
- Keynes, J. M. (1936): *The General Theory of Employment, Interest and Money*. London: Palgrave Macmillan.
- Markovitz, H. (1952): "Portfolio Selection." *Journal of Finance* 7, 77-91.
- Roll, R. (1977): "A Critique of the Asset Pricing Theory's Tests Part I: On Past and Potential Testability of the Theory." *Journal of Financial Economics* 4, 129-176.
- Ross, S. (1976): "Arbitrage Theory of Capital Asset Pricing." *Journal of Economic Theory* 13, 341-360.
- Sharpe, W. F. (1964): "Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk." *Journal of Finance* 19, 425-442.
- Tobin, J. (1958): "Liquidity Preference as Behavior Towards Risk." *Review of Economic Studies* 25, 65-86.

2

The CAPM in general equilibrium

2.1 Introduction

This chapter develops the general equilibrium foundations to the CAPM, within a framework that abstracts from the production sphere of the economy. For this reason, we usually refer the resulting model to as the “Consumption-CAPM.” First, we review the static model of general equilibrium, without uncertainty. Then, we illustrate the economic rationale behind the existence of financial assets in an uncertain world. Finally, we derive the Consumption-CAPM.

2.2 The static general equilibrium in a nutshell

We consider an economy with n agents and m commodities. Let w_{ij} denote the amount of the i -th commodity the j -th agent is endowed with, and let $w^j = [w_{1j}, \dots, w_{mj}]$. Let the price vector be $p = [p_1, \dots, p_m]$, where p_i is the price of the i -th commodity. Let $w_i = \sum_{j=1}^n w_{ij}$ be the total endowment of the i -th commodity in the economy, and $W = [w_1, \dots, w_m]$ the corresponding endowments bundle in the economy.

The j -th agent has utility function $u_j(c_{1j}, \dots, c_{mj})$, where $(c_{ij})_{i=1}^m$ denotes his consumption bundle. We assume the following standard conditions for the utility functions u_j :

ASSUMPTION 2.1 (Preferences). *The utility functions u_j satisfy the following properties: (i) Monotonicity; (ii) Continuity; and (iii) Quasi-concavity: $u_j(x) \geq u_j(y)$, and $\forall \alpha \in (0, 1)$, $u_j(\alpha x + (1 - \alpha)y) > u_j(y)$ or, $\frac{\partial u_j}{\partial c_{ij}}(c_{1j}, \dots, c_{mj}) \geq 0$ and $\frac{\partial^2 u_j}{\partial c_{ij}^2}(c_{1j}, \dots, c_{mj}) \leq 0$.*

Let $B_j(p_1, \dots, p_m) = \{(c_{1j}, \dots, c_{mj}) : \sum_{i=1}^m p_i c_{ij} \leq \sum_{i=1}^m p_i w_{ij} \equiv R_j\}$, a bounded, closed and convex set, hence a convex set. Each agent maximizes his utility function subject to the budget constraint:

$$\max_{\{c_{ij}\}} u_j(c_{1j}, \dots, c_{mj}) \quad \text{subject to } (c_{1j}, \dots, c_{mj}) \in B_j(p_1, \dots, p_m). \quad [\text{P1}]$$

This problem has certainly a solution, for B_j is compact set and by Assumption 2.1, u_j is continuous, and a continuous function attains its maximum on a compact set. Moreover, the Appendix shows that this maximum is unique.

The first order conditions to [P1] are, for each agent j ,

$$\begin{cases} \frac{\partial u_j}{\partial c_{1j}} = \frac{\partial u_j}{\partial c_{2j}} = \dots = \frac{\partial u_j}{\partial c_{mj}} \\ \frac{p_1}{m} & p_2 & \dots & p_m \\ \sum_{i=1}^m p_i c_{ij} = \sum_{i=1}^m p_i w_{ij} \end{cases} \quad (2.1)$$

These conditions form a system of m equations with m unknowns. Let us denote the solution to this system with $[\hat{c}_{1j}(p, w^j), \dots, \hat{c}_{mj}(p, w^j)]$. The total demand for the i -th commodity is,

$$\hat{c}_i(p, w) = \sum_{j=1}^n \hat{c}_{ij}(p, w^j), \quad i = 1, \dots, m.$$

We emphasize the economy we consider in this chapter is one that completely abstracts from production. Here, prices are the key determinants of how resources are allocated in the end. The perspective is, of course, radically different from that taken by the Classical school (Ricardo, Marx and Sraffa), for which prices and resources allocation cannot be disentangled from the production side of the economy. In the next chapter and more advanced parts of the lectures, we consider the asset pricing implications of production, following the Neoclassical perspective.

2.2.1 Walras' Law

Let us plug the demand functions of the j -th agent into the constraint of [P1], to obtain,

$$\forall p, \quad 0 = \sum_{i=1}^m p_i (\hat{c}_{ij}(p, w^j) - w_{ij}). \quad (2.2)$$

Next, define the total excess demand for the i -th commodity as $e_i(p, w) \equiv \hat{c}_i(p, w) - w_i$. By aggregating the budget constraint across all the agents,

$$\forall p, \quad 0 = \sum_{j=1}^n \sum_{i=1}^m p_i (\hat{c}_{ij}(p, w^j) - w_{ij}) = \sum_{i=1}^m p_i e_i(p, w).$$

The previous equality is the celebrated *Walras' law*.

Next, multiply p by $\lambda \in \mathbb{R}_{++}$. Since the constraint to [P1] does not change, the excess demand functions are the same, for each value of λ . In other words, *the excess demand functions are homogeneous of degree zero in the prices*, or $e_i(\lambda p, w) = e_i(p, w)$, $i = 1, \dots, m$. This property of the excess demand functions is also referred to as *absence of monetary illusion*.

2.2.2 Competitive equilibrium

A *competitive equilibrium* is a vector \bar{p} in \mathbb{R}_+^m such that $e_i(\bar{p}, w) \leq 0$ for all $i = 1, \dots, m$, with at least one component of \bar{p} being strictly positive. Furthermore, if there exists a $j : e_j(\bar{p}, w) < 0$, then $\bar{p}_j = 0$.

2.2.2.1 Back to Walras' law

Walras' law holds by the mere aggregation of the agents' constraints. But the agents' constraints are accounting identities. In particular, Walras' law holds for any price vector and, a fortiori, it holds for the equilibrium price vector,

$$0 = \sum_{i=1}^m \bar{p}_i e_i(\bar{p}, w) = \sum_{i=1}^{m-1} \bar{p}_i e_i(\bar{p}, w) + \bar{p}_m e_m(\bar{p}, w). \quad (2.3)$$

Now suppose that the first $m-1$ markets are in equilibrium, or $e_i(\bar{p}, w) \leq 0$, for $i = 1, \dots, m-1$. By the definition of an equilibrium, we have that $\text{sign}(e_i(\bar{p}, w)) \bar{p}_i = 0$. Therefore, by Eq. (2.3), we conclude that if $m-1$ markets are in equilibrium, then, the remaining market is also in equilibrium.

2.2.2.2 The notion of numéraire

The excess demand functions are homogeneous of degree zero. Walras' law implies that if $m-1$ markets are in equilibrium, then, the m -th remaining market is also in equilibrium. We wish to link these two results. A first remark is that by Walras' law, the equations that define a competitive equilibrium are not independent. Once $m-1$ of these equations are satisfied, the m -th remaining equation is also satisfied. In other words, there are $m-1$ independent relations and m unknowns in the equations that define a competitive equilibrium. So, there exists an infinity of solutions.

Suppose, then, that we choose the m -th price to be a sort of exogeneous datum. The result is that we obtain a system of $m-1$ equations with $m-1$ unknowns. Provided it exists, such a solution is a function f of the m -th price, $\bar{p}_i = f_i(\bar{p}_m)$, $i = 1, \dots, m-1$. Then, we may refer to the m -th commodity as the *numéraire*. In other words, general equilibrium can only determine a structure of *relative* prices. The scale of these relative prices depends on the price level of the numéraire. It is easily checked that if the functions f_i are homogeneous of degree one, multiplying p_m by a strictly positive number λ does not change the relative price structure. Indeed, by the equilibrium condition, for all $i = 1, \dots, m$,

$$\begin{aligned} 0 &\geq e_i(\bar{p}_1, \bar{p}_2, \dots, \lambda \bar{p}_m, w) = e_i(f_1(\lambda \bar{p}_m), f_2(\lambda \bar{p}_m), \dots, \lambda \bar{p}_m, w) \\ &= e_i(\lambda \bar{p}_1, \lambda \bar{p}_2, \dots, \lambda \bar{p}_m, w) = e_i(\bar{p}_1, \bar{p}_2, \dots, \bar{p}_m, w), \end{aligned}$$

where the second equality is due to the homogeneity property of the functions f_i , and the last equality holds because the excess demand functions e_i are homogeneous of degree zero. In particular, by defining relative prices as $\hat{p}_j = p_j/p_m$, one has that $p_j = \hat{p}_j \cdot p_m$ is a function that is homogeneous of degree one. In other words, if $\lambda \equiv \bar{p}_m^{-1}$, then,

$$0 \geq e_i(\bar{p}_1, \dots, \bar{p}_m, w) = e_i(\lambda \bar{p}_1, \dots, \lambda \bar{p}_m, w) \equiv e_i\left(\frac{\bar{p}_1}{\bar{p}_m}, \dots, 1, w\right).$$

2.2.3 Optimality

Let $c^j = (c_{1j}, \dots, c_{mj})$ be the allocation to agent j , $j = 1, \dots, n$. The following definition is the well-known concept of a desirable resource allocation within a society, according to Pareto.

DEFINITION 2.2 (Pareto optimum). *An allocation $\bar{c} = (\bar{c}^1, \dots, \bar{c}^n)$ is a Pareto optimum if it is feasible, $\sum_{j=1}^n (\bar{c}^j - w^j) \leq 0$, and if there are no other feasible allocations $c = (c^1, \dots, c^n)$ such that $u_j(c^j) \geq u_j(\bar{c}^j)$, $j = 1, \dots, n$, with one strict inequality for at least one agent.*

We have the following fundamental result:

THEOREM 2.3 (First welfare theorem). *Every competitive equilibrium is a Pareto optimum.*

PROOF. Let us suppose on the contrary that \bar{c} is an equilibrium but not a Pareto optimum. Then, there exists a $c : u_{j^*}(c^{j^*}) > u_{j^*}(\bar{c}^{j^*})$, for some j^* . Because \bar{c}^{j^*} is optimal for agent j^* , $c^{j^*} \notin B_j(\bar{p})$, or $\bar{p}c^{j^*} > \bar{p}w_{j^*}$ and, by aggregating: $\bar{p}\sum_{j=1}^n c^j > \bar{p}\sum_{j=1}^n w^j$, which is unfeasible. It follows that c can not be an equilibrium. \parallel

Next, we show that any Pareto optimal allocation can be “decentralized.” That is, corresponding to a given Pareto optimum \bar{c} , there exist ways of redistributing endowments around, and a price vector $\bar{p} : \bar{p}\bar{c} = \bar{p}w$, which is an equilibrium for the initial set of resources.

THEOREM 2.4 (Second welfare theorem). *Every Pareto optimum can be decentralized.*

PROOF. In the appendix.

The previous theorem can be interpreted as one that supports an *equilibrium with transfer payments*. For any given Pareto optimum \bar{c}^j , a social planner can always give $\bar{p}w^j$ to each agent (with $\bar{p}\bar{c}^j = \bar{p}w^j$, where w^j is chosen by the planner), and agents choose \bar{c}^j . Figure 2.1 illustrates such a decentralization procedure within the Edgeworth’s box. Suppose that the objective is to achieve \bar{c} . Given an initial allocation w chosen by the planner, each agent is given $\bar{p}w^j$. Under *laissez faire*, \bar{c} will obtain. In other words, agents are given a constraint of the form $pc^j = \bar{p}w^j$. If w^j and \bar{p} are chosen so as to induce each agent to choose \bar{c}^j , then \bar{p} is a supporting equilibrium price. In this case, the marginal rates of substitutions are identical, as established by the following celebrated result:

THEOREM 2.5 (Characterization of Pareto optima: I). *A feasible allocation $\bar{c} = (\bar{c}^1, \dots, \bar{c}^n)$ is a Pareto optimum if and only if there exists a $\tilde{\phi} \in \mathbb{R}_{++}^{m-1}$ such that*

$$\tilde{\nabla}u_j = \tilde{\phi}, \quad j = 1, \dots, n, \quad \text{where } \tilde{\nabla}u_j \equiv \begin{pmatrix} \frac{\partial u_j}{\partial c_{2j}} \\ \frac{\partial u_j}{\partial c_{1j}} \end{pmatrix}. \quad (2.4)$$

PROOF. A Pareto optimum satisfies:

$$\bar{c} \in \arg \max_{c \in \mathbb{R}_+^{m \cdot n}} u_1(c^1) \quad \text{subject to} \quad \begin{cases} u_j(c^j) \geq \bar{u}_j, \quad j = 2, \dots, n & (\lambda_j, \quad j = 2, \dots, n) \\ \sum_{j=1}^n (c^j - w^j) \leq 0 & (\phi_i, \quad i = 1, \dots, m) \end{cases}$$

The Lagrangian function associated with this program is

$$L = u_1(c^1) + \sum_{j=2}^n \lambda_j (u_j(c^j) - \bar{u}_j) - \sum_{i=1}^m \phi_i \sum_{j=1}^n (c_{ij} - w_{ij}),$$

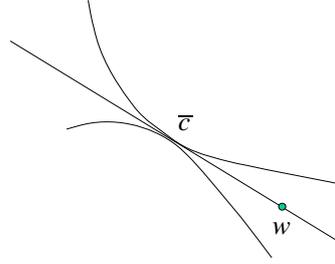


FIGURE 2.1. Decentralizing a Pareto optimum

and the first order conditions are

$$\begin{cases} \frac{\partial u_1}{\partial c_{11}} = \phi_1 \\ \dots \\ \frac{\partial u_1}{\partial c_{m1}} = \phi_m \end{cases}$$

and, for $j = 2, \dots, n$,

$$\begin{cases} \lambda_j \frac{\partial u_j}{\partial c_{1j}} = \phi_1 \\ \dots \\ \lambda_j \frac{\partial u_j}{\partial c_{mj}} = \phi_m \end{cases}$$

In each of the previous two systems, we divide each equation by the the first, obtaining exactly Eq. (2.4), with $\tilde{\phi} = \left(\frac{\phi_2}{\phi_1}, \dots, \frac{\phi_m}{\phi_1} \right)$. The converse is straight forward. ||

There is a simple and appealing interpretation of the Kuhn-Tucker multipliers ϕ on the constraints of Theorem 2.5. Note that by Eq. (2.1), in the competitive equilibrium,

$$\tilde{\nabla} u_j = \tilde{p} \equiv \left(\frac{p_2}{p_1}, \dots, \frac{p_m}{p_1} \right).$$

But because a competitive equilibrium is also a Pareto optimum, then, by Theorem 2.5,

$$\tilde{\nabla} u_j = \tilde{\phi} \equiv \left(\frac{\phi_2}{\phi_1}, \dots, \frac{\phi_m}{\phi_1} \right).$$

Hence, $\tilde{\phi}$ represents the vector of relative, shadow prices arising within the centralized allocation process.

We provide a further characterization of Pareto optimal allocations.

THEOREM 2.6 (Characterization of Pareto optima: II). *A feasible allocation $\bar{c} = (\bar{c}^1, \dots, \bar{c}^n)$ is a Pareto optimum if and only if there exists $\ell > 0$ such that \bar{c} is solution to the following program:*

$$u(w, \ell) = \max_{c^1, \dots, c^n} \sum_{j=1}^n \ell_j u_j(c^j) \quad \text{subject to} \quad \sum_{j=1}^n c^j \leq w \quad (\psi_j, j = 1, \dots, m) \quad [\text{P2}]$$

PROOF. The if part is simple and at the same time instructive. Let us solve the program in [P2]. The Lagrangian is,

$$L = \sum_{j=1}^n \ell_j u_j(c^j) - \sum_{i=1}^m \psi_i \sum_{j=1}^n (c_{ij} - w_{ij}),$$

and the first order conditions are, for $j = 1, \dots, n$,

$$\ell_j \nabla u_j = \psi \equiv (\psi_1, \dots, \psi_m)^\top, \quad \nabla u_j \equiv \left(\frac{\partial u_j}{\partial c_{1j}}, \dots, \frac{\partial u_j}{\partial c_{mj}} \right)^\top. \quad (2.5)$$

That is, $\tilde{\nabla} u_j$ equals the same vector of constants for all the agents, just as in Theorem 2.5. The converse to this theorem follows by an application of the usual separating theorem, as in Duffie (2001, Chapter 1). ||

Note, if $\ell_1 = 1$ and $\ell_j = \lambda_j$ for $j = 2, \dots, n$, then, $\psi_i = \phi_i$ ($i = 1, \dots, m$) and so the first order conditions in Theorem 2.5 and 2.6 would lead to the same allocation. More generally, we have:

THEOREM 2.7 (Centralization of competitive equilibrium through Pareto weightings). *The outcome of any competitive equilibrium can be obtained, through a central planner who maximizes the program in [P2], with system of social weights equal to $\ell_j = 1/\kappa_j$, where κ_j is the marginal utility of income for agent j .*

So agents with high marginal utility of income for a given price vector, will receive little social weight in the centralized planner allocation procedure. This result is particularly useful when it comes to study financial markets in economies with heterogeneous agents. Theorem 2.7 is also a point of reference, where to move from, when it comes to study asset prices in a world of incomplete markets. Chapter 8 contains several examples of these applications.

PROOF OF THEOREM 2.7. In the competitive equilibrium,

$$\nabla u_j = \kappa_j p, \quad p \equiv (p_1, \dots, p_m), \quad (2.6)$$

where κ_j are the Lagrange multipliers for the agents budget constraint, so that κ_j is the agent j marginal utility of income:

$$\kappa_j = \frac{\partial}{\partial m_j} u_j(\hat{c}_{1j}(p, w_{1j}, \dots, w_{mj}), \dots, \hat{c}_{mj}(p, w_{1j}, \dots, w_{mj})), \quad m_j \equiv \sum_{i=1}^m p_i w_{ij}.$$

By comparing the competitive equilibrium solution in Eq. (2.6) with the Pareto optimality property of the equilibrium in Eq. (2.5), we deduce that, a competitive equilibrium (\bar{c}, p) can

be implemented, by a social planner acting as in Theorem 2.6, when $\ell_j = 1/\kappa_j$. Then, it also follows that, necessarily, $\psi = p$, by the resources constraint, $\sum_{j=1}^n c^j \leq w$, which has to hold both in the competitive economy and the centralized one. Indeed, we have:

$$w_i = \sum_{j=1}^n f_{ij}(\kappa_j \psi) = \sum_{j=1}^n \hat{f}_{ij}(\kappa_j p), \quad i = 1, \dots, m,$$

where f_{ij} and \hat{f}_{ij} are the inverse functions for consumption, as implied by Eq. (2.5) and Eq. (2.6), respectively. The previous equality holds when $\psi = p$, in which case $f_{ij} = \hat{f}_{ij}$. This is indeed the only solution for ψ in the previous equation, given that f_{ij} is monotonically decreasing. ||

2.3 Time and uncertainty

“A commodity is characterized by its physical properties, the date and the place at which it will be available.”

Gerard Debreu (1959, Chapter 2)

General equilibrium theory can be used to study a variety of fields, by making an appropriate use of the previous definition - from the theory of international commerce to finance. To deal with uncertainty, Debreu (1959, Chapter 7) extended the previous definition, by emphasizing that a commodity should be described through a list of physical properties, with the structure of dates and places replaced by some event structure. The following example illustrates the difference between two contracts underlying delivery of corn arising under conditions of certainty (case A) and uncertainty (case B):

- A *The first agent will deliver 5000 tons of corn of a specified type to the second agent, who will accept the delivery at date t and in place ℓ .*
- B *The first agent will deliver 5000 tons of corn of a specified type to the second agent, who will accept the delivery in place ℓ and in the event s_t at time t . If s_t does not occur at time t , no delivery will take place.*

In both cases, the contract is paid at the time it is actually agreed.

The model of the previous section can be used to deal with contracts containing statements such as that in case B above. For example, consider a two-period economy. Suppose that in the second period, s_n mutually exhaustive and exclusive states of nature may occur. Then, we may recover the model of the previous section, once we replace m (the number of commodities described by physical properties, dates and places) with m^* , where $m^* = s_n \cdot m$. With m^* replacing m , the competitive equilibrium in this economy is defined as the competitive equilibrium in the economy of the previous section.

The important assumption underlying the previous simplifying trick is that markets exist, where commodities for all states of nature are traded. Such “contingent” markets are *complete* in that a market is open for *every* commodity in all states of nature. Therefore, the agents may implement any feasible action plan and, therefore, the resource allocation is Pareto-optimal. The presumed existence of $s_n \cdot m$ contingent markets is, however, very strong. We now show how the presence of financial assets helps us mitigate this assumption.

2.4 Financial assets

What role might be played by financial assets in an uncertainty world? Arrow (1953) developed the following interpretation. Rather than signing commodity-based contracts that are contingent on the realization of events, the agents might wish to sign contracts generating payoffs that are contingent on the realization of events. The payoffs delivered by the assets in the various states of the world could then be collected and used to satisfy the needs related to the consumption plans.

The simplest financial asset is the so-called Arrow-Debreu asset, i.e. an asset that pays some amount of numéraire in the state of nature s if the state s will prevail in the future, and nil otherwise. More generally, a financial asset is a function $x : \mathcal{S} \mapsto \mathbb{R}$, where \mathcal{S} is the set of all future events. Then, let m be the number of financial assets. To link financial assets to commodities, we note that if the state of nature s occurs, then, any agent could use the payoff $x_i(s)$ promised by the i -th assets \mathcal{A}_i to finance net transactions on the commodity markets, viz

$$p(s) \cdot e(s) = \sum_{i=1}^m \theta_i x_i(s), \quad \forall s \in \mathcal{S}, \quad (2.7)$$

where $p(s)$ and $e(s)$ denote some vectors of prices and excess demands related to the commodities, contingent on the realization of state s , and θ_i is the number of assets i held by the agent. In other words, the role of financial assets, here, is to transfer value from a state of nature to another to finance state-contingent consumption.

Unfortunately, Eq. (2.7) does not hold, in general. A condition is that the number of assets, m , be sufficiently high to let each agent cope with the number of future events in \mathcal{S} , s_n . Market completeness merely reduces to a size problem - the assets have to be sufficiently diverse to span all possible events in the future. Indeed, we shall show that if there are not payoffs that are perfectly correlated, then, markets are complete if and only if $m = s_n$. Note, also, that this reduces the dimension of our original problem, for we are then considering a competitive equilibrium in $s_n + m$ markets, instead of a competitive equilibrium in $s_n \cdot m$ markets.

2.5 Absence of arbitrage

2.5.1 How to price a financial asset?

Consider an economy in which uncertainty is resolved through the realization of the event: “Tomorrow it will rain.” A decision maker, an hypothetical Mr Law, must implement the following contingent plan: if tomorrow will be sunny, he will need $c_s > 0$ units of money, to buy sun-glasses; if tomorrow it will rain, Mr Law will need $c_r > 0$ units of money, to buy an umbrella. Mr Law has access to a financial market on which m assets are traded. He builds up a portfolio θ aimed to reproduce the structure of payments that he will need tomorrow:

$$\begin{cases} \sum_{i=1}^m \theta_i S_i (1 + x_i(r)) = c_r \\ \sum_{i=1}^m \theta_i S_i (1 + x_i(s)) = c_s \end{cases} \quad (2.8)$$

where S_i is the price of the i -th asset, θ_i is the number of assets to put in the portfolio, and $x_i(r)$ and $x_i(s)$ are the net returns of asset i in the two states of nature, which of course are

known by Mr Law. For now, we do need to assume anything as regards the resources needed to buy the assets, but we shall come back to this issue below (see Remark 2.6). Finally, and remarkably, we are not making any assumption regarding Mr Law's preferences.

Eqs. (2.8) form a system of two equations with m unknowns $(\theta_1, \dots, \theta_m)$. If $m < 2$, no perfect hedging strategy is possible - that is, the system (2.8) can not be solved to obtain the desired pair $(c_i)_{i=r,s}$. In this case, markets are incomplete. More generally, we may consider an economy with s_n states of nature, in which markets are complete if and only if Mr Law has access to s_n assets. More precisely, let us define the following "payoff matrix," defined as

$$X = \begin{bmatrix} S_1(1 + x_1(s_1)) & & S_m(1 + x_m(s_1)) \\ & \ddots & \\ S_1(1 + x_1(s_n)) & & S_m(1 + x_m(s_n)) \end{bmatrix},$$

where $x_i(s_j)$ is the payoff promised by the i -th asset in the state s_j . Then, to implement any state contingent consumption plan $c \in \mathbb{R}^{s_n}$, Mr Law has to be able to solve the following system,

$$c = X \cdot \theta,$$

where $\theta \in \mathbb{R}^m$, the portfolio. A unique solution to the previous system exists if $\text{rank}(X) = s_n = m$, and is given by $\hat{\theta} = X^{-1}c$. Consider, for example, the previous case, in which $s_n = 2$. Let us assume that $m = 2$, for any additional assets would be redundant here. Then, we have,

$$\begin{cases} \hat{\theta}_1 = \frac{(1 + x_2(r))c_s - (1 + x_2(s))c_r}{S_1[(1 + x_1(s))(1 + x_2(r)) - (1 + x_1(r))(1 + x_2(s))]} \\ \hat{\theta}_2 = \frac{(1 + x_1(s))c_r - (1 + x_1(r))c_s}{S_2[(1 + x_1(s))(1 + x_2(r)) - (1 + x_1(r))(1 + x_2(s))]} \end{cases}$$

Finally, assume that the second asset is safe, or that it yields the same return in the two states of nature: $x_2(r) = x_2(s) \equiv r$. Let $x_s = x_1(s)$ and $x_r = x_1(r)$. Then, the pair $(\hat{\theta}_1, \hat{\theta}_2)$ can be rewritten as,

$$\hat{\theta}_1 = \frac{c_s - c_r}{S_1(x_s - x_r)}, \quad \hat{\theta}_2 = \frac{(1 + x_s)c_r - (1 + x_r)c_s}{S_2(1 + r)(x_s - x_r)}.$$

As is clear, the issues we are dealing with relate to the *replication* of random variables. Here, the random variable is a state contingent consumption plan $(c_i)_{i=r,s}$, where c_r and c_s are known, which we want to replicate for hedging purposes. (Mr Law will need to buy either a pair of sun-glasses or an umbrella, tomorrow.)

In the previous two-state example, two assets with independent payoffs are able to generate any two-state variable. The next step, now, is to understand what happens when we assume that there exists a third asset, \mathcal{A} say, that delivers the same random variable $(c_i)_{i=r,s}$ we can obtain by using the previous pair $(\hat{\theta}_1, \hat{\theta}_2)$.

We claim that if the current price of the third asset \mathcal{A} is H , then, it must be that,

$$H = V \equiv \hat{\theta}_1 S_1 + \hat{\theta}_2 S_2, \tag{2.9}$$

for the financial market to be free of *arbitrage opportunities*, to be defined informally below. Indeed, if $V < H$, we can buy $\hat{\theta}$ and sell at the same time the third asset \mathcal{A} . The result is a sure profit, or an arbitrage opportunity, equal to $H - V$, for $\hat{\theta}$ generates c_r if tomorrow it will rain and c_s if tomorrow it will not rain. In both cases, the portfolio $\hat{\theta}$ generates the payments that

are necessary to honour the contract commitments related to the selling of \mathcal{A} . By a symmetric argument, the inequality $V > H$ would also generate an arbitrage opportunity. Hence, Eq. (2.9) must hold true.

It remains to compute the right hand side of Eq. (2.9), which in turn leads to an evaluation formula for the asset \mathcal{A} . We have:

$$H = \frac{1}{1+r} [P^*c_s + (1-P^*)c_r], \quad P^* = \frac{x_r - r}{x_s - x_r}. \quad (2.10)$$

Importantly, then, H can be understood as the discounted (by $1+r$) expectation of payoffs promised by \mathcal{A} , taken under some “artificial” probability P^* .

REMARK 2.8. In this introductory example, the asset \mathcal{A} can be priced without making reference to any agents’ preferences. The key observation to obtain this result is that the payoffs promised by \mathcal{A} can be obtained through the portfolio $\hat{\theta}$. This fact does not obviously mean that any agent should use this portfolio. For example, it may be the case that Mr Law is so poor that his budget constraint would not even allow him to implement the portfolio $\hat{\theta}$. The point underlying the previous example is that the portfolio $\hat{\theta}$ could be used to construct an arbitrage opportunity, arising when Eq. (2.9) does not hold. In this case, any penniless agent could implement the arbitrage described above.

The next step is to extend the results in Eq. (2.10) to a dynamic setting. Suppose that an additional day is available for trading, with the same uncertainty structure: the day after tomorrow, the asset \mathcal{A} will pay off c_{ss} if it will be sunny (provided the previous day was sunny), and c_{rs} if it will be sunny (provided the previous day was raining). By using the same arguments leading to Eq. (2.10), we obtain that:

$$H = \frac{1}{(1+r)^2} [P^{*2}c_{ss} + P^*(1-P^*)c_{sr} + (1-P^*)P^*c_{rs} + P^{*2}c_{rr}].$$

Finally, by extending the same reasoning to T trading days,

$$H = \frac{1}{(1+r)^T} E^*(c_T), \quad (2.11)$$

where E^* denotes the expectation taken under the probability P^* .

The key assumption we used to derive Eq. (2.11) is that markets are complete at each trading day. True, at the beginning of the trading period Mr Law faced 2^T mutually exclusive possible states of nature that would occur at the T -th date, which would seem to imply that we would need 2^T assets to replicate the asset \mathcal{A} . However, we have just seen that to price \mathcal{A} , we only need 2 assets *and* T trading days. To emphasize this fact, we say that the structure of assets and transaction dates makes the markets *dynamically complete* in the previous example. The presence of dynamically complete markets allows one to implement dynamic trading strategies aimed at replicating the value of the asset \mathcal{A} , period by period. Naturally, the asset \mathcal{A} could be priced without any assumption about the preferences of any agent, due to the assumption of dynamically complete markets.

2.5.2 The Land of Cockaigne

We provide a precise definition of the notion of absence of arbitrage opportunities, as well as a connection between this notion and the notion and properties of the competitive equilibrium described in Section 2.2. For simplicity, we consider a multistate economy with only

one commodity. The extension to the multicommodities case is dealt with very briefly in the appendix.

Let $v_i(\omega_s)$ be the payoff of asset i in the state ω_s , $i = 1, \dots, m$ and $s = 1, \dots, d$. Consider the payoff matrix:

$$V \equiv \begin{bmatrix} v_1(\omega_1) & & v_m(\omega_1) \\ & \ddots & \\ v_1(\omega_d) & & v_m(\omega_d) \end{bmatrix}.$$

Let $v_{si} \equiv v_i(\omega_s)$, $v_{s\cdot} \equiv [v_{s1}, \dots, v_{sm}]$, $v_{\cdot i} \equiv [v_{1i}, \dots, v_{di}]^\top$. We assume that $\text{rank}(V) = m \leq d$.

The budget constraint of each agent has the form:

$$\begin{cases} c_0 - w_0 = -S\theta = -\sum_{i=1}^m S_i\theta_i \\ c_s - w_s = v_s\theta = \sum_{i=1}^m v_{si}\theta_i, \quad s = 1, \dots, d \end{cases}$$

Let $x^1 = [x_1, \dots, x_d]^\top$. The second constraint can be written as:

$$c^1 - w^1 = V\theta.$$

We define an arbitrage opportunity as a portfolio that has a negative value at the first period, and a positive value in at least one state of world in the second period, or a positive value in all states of the world in the second period and a nonpositive value in the first period.

Notation: $\forall x \in \mathbb{R}^m$, $x > 0$ means that at least one component of x is strictly positive while the other components of x are nonnegative. $x \gg 0$ means that all components of x are strictly positive. **[Insert here further notes]**

DEFINITION 2.9. *An arbitrage opportunity is a strategy θ that yields¹ either $V\theta \geq 0$ with an initial investment $S\theta < 0$, or a strategy θ that produces² $V\theta > 0$ with an initial investment $S\theta \leq 0$.*

As we shall show below (Theorem 2.11), an arbitrage opportunity can not exist in a competitive equilibrium, for the agents' program would not be well defined in this case. Introduce, then, the $(d+1) \times m$ matrix,

$$W = \begin{bmatrix} -S \\ V \end{bmatrix},$$

the vector subspace of \mathbb{R}^{d+1} ,

$$\langle W \rangle = \{z \in \mathbb{R}^{d+1} : z = W\theta, \theta \in \mathbb{R}^m\},$$

and, finally, the null space of $\langle W \rangle$,

$$\langle W \rangle^\perp = \{x \in \mathbb{R}^{d+1} : xW = \mathbf{0}_m\}.$$

¹ $V\theta \geq 0$ means that $[V\theta]_j \geq 0$, $j = 1, \dots, d$, i.e. it allows for $[V\theta]_j = 0$, $j = 1, \dots, d$.

² $V\theta > 0$ means $[V\theta]_j \geq 0$, $j = 1, \dots, d$, with at least one j for which $[V\theta]_j > 0$.

The economic interpretation of the vector subspace $\langle W \rangle$ is that of the excess demand space for all the states of nature, generated by the “wealth transfers” generated by the investments in the assets. Naturally, $\langle W \rangle^\perp$ and $\langle W \rangle$ are orthogonal, as $\langle W \rangle^\perp = \{x \in \mathbb{R}^{d+1} : xz = \mathbf{0}_m, z \in \langle W \rangle\}$.

Mathematically, the assumption that there are no arbitrage opportunities is equivalent to the following condition,

$$\langle W \rangle \cap \mathbb{R}_+^{d+1} = \{0\}. \quad (2.12)$$

The interpretation of (2.12) is in fact very simple. In the absence of arbitrage opportunities, there should be no portfolios generating “wealth transfers” that are nonnegative and strictly positive in at least one state, i.e. $\nexists \theta : W\theta > 0$. Hence, $\langle W \rangle$ and the positive orthant \mathbb{R}_+^{d+1} can not intersect.

The following result provides a general characterization of how the no-arbitrage condition in (2.12) restricts the price of all the assets in the economy.

THEOREM 2.10. *There are no arbitrage opportunities if and only if there exists a $\phi \in \mathbb{R}_{++}^d : S = \phi^\top V$. If $m = d$, ϕ is unique, and if $m < d$, $\dim(\phi \in \mathbb{R}_{++}^d : S = \phi^\top V) = d - m$.*

PROOF. In the appendix.

The previous theorem provides the foundations for many developments in financial economics. To provide its intuition, let us pre-multiply the second constraint by ϕ^\top , obtaining,

$$\phi^\top (c^1 - w^1) = \phi^\top V\theta = S\theta = -(c_0 - w_0),$$

where the second equality follows by Theorem 2.10, and the third equality is due to the first period budget constraint. Critically, then, Theorem 2.10 shows that in the absence of arbitrage opportunities, each agent has access to the following budget constraint,

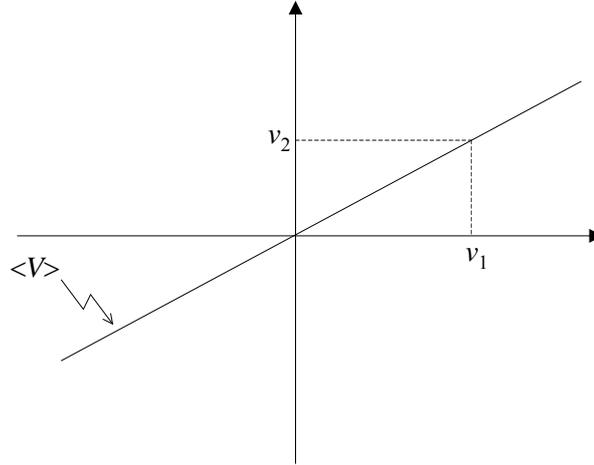
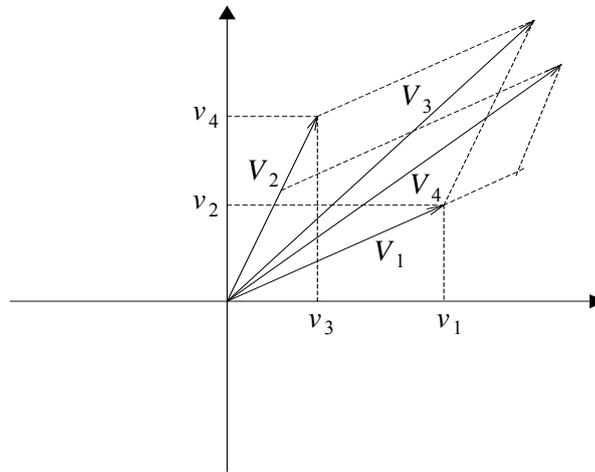
$$0 = c_0 - w_0 + \phi^\top (c^1 - w^1) = c_0 - w_0 + \sum_{s=1}^d \phi_s (c_s - w_s), \text{ with } (c^1 - w^1) \in \langle V \rangle. \quad (2.13)$$

The budget constraints in (2.13) reveal that ϕ can be interpreted as the vector of prices to the commodity in the future d states of nature, and that the numéraire in this economy is the first-period consumption. We usually refer ϕ to as the *state price* vector, or *Arrow-Debreu state price* vector. However, it would be misleading to say that the budget constraint in (2.13) is that we are used to see in the static Arrow-Debreu type model of Section 2.2. In fact, the Arrow-Debreu economy of Section 2.2 obtains when $m = d$, in which case $\langle V \rangle = \mathbb{R}^d$ in (2.13). This case, which according to Theorem 2.10 arises when markets are complete, also implies the remarkable property that there exists a unique ϕ that is compatible with the asset prices we observe.

The situation is radically different if $m < d$. In other terms, $\langle V \rangle$ is the subspace of excess demands agents have access to in the second period and can be “smaller” than \mathbb{R}^d if markets are incomplete. Indeed, $\langle V \rangle$ is the subspace generated by the payoffs obtained by the portfolio choices made in the first period,

$$\langle V \rangle = \{e \in \mathbb{R}^d : e = V\theta, \theta \in \mathbb{R}^m\}.$$

Consider, for example, the case $d = 2$ and $m = 1$. In this case, $\langle V \rangle = \{e \in \mathbb{R}^2 : e = V\theta, \theta \in \mathbb{R}\}$, with $V = V_1$, where $V_1 = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}$ say, and $\dim \langle V \rangle = 1$, as illustrated by Figure 2.2.

FIGURE 2.2. Incomplete markets, $d = 2$, $m = 1$.FIGURE 2.3. Complete markets, $\langle V \rangle = \mathbb{R}^2$.

Next, suppose we open a new market for a second financial asset with payoffs given by: $V_2 = \begin{pmatrix} v_3 \\ v_4 \end{pmatrix}$. Then, $m = 2$, $V = \begin{pmatrix} v_1 & v_3 \\ v_2 & v_4 \end{pmatrix}$, and $\langle V \rangle = \left\{ e \in \mathbb{R}^2 : e = \begin{pmatrix} \theta_1 v_1 + \theta_2 v_3 \\ \theta_1 v_2 + \theta_2 v_4 \end{pmatrix}, \theta \in \mathbb{R}^2 \right\}$, i.e. $\langle V \rangle = \mathbb{R}^2$. As a result, we can now generate any excess demand in \mathbb{R}^2 , just as in the Arrow-Debreu economy of Section 2.2. To generate any excess demand, we multiply the payoff vector V_1 by θ_1 and the payoff vector V_2 by θ_2 . For example, suppose we wish to generate the payoff the payoff vector V_4 in Figure 2.3. Then, we choose some $\theta_1 > 1$ and $\theta_2 < 1$. (The exact values of θ_1 and θ_2 are obtained by solving a linear system.) In Figure 2.3, the payoff vector V_3 is obtained with $\theta_1 = \theta_2 = 1$.

To summarize, if markets are complete, then, $\langle V \rangle = \mathbb{R}^d$. If markets are incomplete, $\langle V \rangle$ is only a subspace of \mathbb{R}^d , which makes the agents' choice space smaller than in the complete markets case.

We now present a fundamental result, about the “viability of the model.” Define the second period consumption $c_j^1 \equiv [c_{1j}, \dots, c_{dj}]^\top$, where c_{sj} is the second-period consumption in state s , and let,

$$(\hat{c}_{0j}, \hat{c}_j^1) \in \arg \max_{c_{0j}, c_j^1} [u_j(c_{0j}) + \beta_j E(\nu_j(c_j^1))], \quad \text{subject to} \quad \begin{cases} c_{0j} - w_{0j} &= -S\theta_j \\ c_j^1 - w_j^1 &= V\theta_j \end{cases} \quad [\text{P3}]$$

where u_j and ν_j are utility functions, both satisfying Assumption 2.1. Naturally, we could use more general formulations of utilities than that in [P3], and in fact we shall in more advanced parts of this book. For the sake of this introductory chapter, we only consider additive utility.

We have:

THEOREM 2.11. *The program [P3] has a solution if and only if there are no arbitrage opportunities.*

PROOF. Let us suppose on the contrary that the program [P3] has a solution $\hat{c}_{0j}, \hat{c}_j^1, \hat{\theta}_j$, but that there exists a $\theta : W\theta > 0$. The program constraint is, with straight forward notation, $\hat{c}_j = w_j + W\hat{\theta}_j$. Then, we may define a portfolio $\theta_j = \hat{\theta}_j + \theta$, such that $c_j = w_j + W(\hat{\theta}_j + \theta) = \hat{c}_j + W\theta > \hat{c}_j$, which contradicts the optimality of \hat{c}_j . For the converse, note that the absence of arbitrage opportunities implies that $\exists \phi \in \mathbb{R}_{++}^d : S = \phi^\top V$, which leads to the budget constraint in (2.13), for a given ϕ . This budget constraint is clearly a closed subset of the compact budget constraint B_j in [P1] (in fact, it is B_j restricted to $\langle V \rangle$). Therefore, it is a compact set and, hence, the program [P3] has a solution, as a continuous function attains its maximum on a compact set. ||

2.6 Equivalent martingales and equilibrium

We provide the definition of an equilibrium with financial markets, when the financial assets are in zero net supply.

DEFINITION 2.12. *An equilibrium is given by allocations and prices $\{(\hat{c}_{0j})_{j=1}^n, ((\hat{c}_{sj})_{j=1}^n)_{s=1}^d, (\hat{S}_i)_{i=1}^m \in R_+^n \times R_+^{nd} \times R_+^d\}$, where the allocations are solutions of the program [P3] and satisfy:*

$$0 = \sum_{j=1}^n (\hat{c}_{0j} - w_{0j}), \quad 0 = \sum_{j=1}^n (\hat{c}_{sj} - w_{sj}) \quad (s = 1, \dots, d), \quad 0 = \sum_{j=1}^n \theta_{ij} \quad (i = 1, \dots, d).$$

We now express demand functions in terms of the stochastic discount factor, and then look for an equilibrium by looking for the stochastic discount factor that clears the commodity markets. By Walras' law, this also implies the equilibrium on the financial market. Indeed, by aggregating the agent's constraints in the second period,

$$\sum_{j=1}^n (c_j^1 - w_j^1) = V \sum_{j=1}^n \theta_{ij}(m).$$

For simplicity, we also assume that $u_j'(x) > 0$, $u_j''(x) < 0 \forall x > 0$ and $\lim_{x \rightarrow 0} u_j'(x) = \infty$, $\lim_{x \rightarrow \infty} u_j'(x) = 0$ and that ν_j satisfies the same properties.

2.6.1 The rational expectations assumption

Lucas, Radner, Green. Every agent *correctly* anticipates the equilibrium price in each state of nature.

[Consider for example the models with asymmetric information that we will see later in these lectures. At some point we will have to compute, $E(\tilde{v}|p(\tilde{y}) = p)$. That is, the equilibrium is a

pricing function which takes some values $p(\tilde{y})$ depending on the state of nature. In this kind of models, $\lambda\theta_I(p(\tilde{y}), \tilde{y}) + (1 - \lambda)\theta_U(p(\tilde{y}), \tilde{y}) + \tilde{y} = 0$, and we look for a solution $p(\tilde{y})$ satisfying this equation.]

2.6.2 Stochastic discount factors

Theorem 2.10 states that in the absence of arbitrage opportunities,

$$S_i = \phi^\top v_{\cdot,i} = \sum_{s=1}^d \phi_s v_{s,i}, \quad i = 1, \dots, m. \quad (2.14)$$

Let us assume that the first asset is a safe asset, i.e. $v_{s,1} = 1 \forall s$. Then, we have

$$S_1 \equiv \frac{1}{1+r} = \sum_{s=1}^d \phi_s. \quad (2.15)$$

Eq. (2.15) confirms the economic interpretation of the state prices in (2.13). Recall, the states of nature are exhaustive and mutually exclusive. Therefore, ϕ_s can be interpreted as the price to be paid today for obtaining, for sure, one unit of numéraire, tomorrow, in state s . This is indeed the economic interpretation of the budget constraint in (2.13). Eq. (2.15) confirms this as it says that the prices of all these rights sum up to the price of a pure discount bond, i.e. an asset that yields one unit of numéraire, tomorrow, for sure.

Eq. (2.15) can be elaborated to provide us with a second interpretation of the state prices in Theorem 2.10. Define,

$$P_s^* \equiv (1+r)\phi_s,$$

which satisfies, by construction,

$$\sum_{s=1}^d P_s^* = 1.$$

Therefore, we can interpret $P^* \equiv (P_s^*)_{s=1}^d$ as a probability distribution. Moreover, by replacing P^* in Eq. (11.18) leaves,

$$S_i = \frac{1}{1+r} \sum_{s=1}^d P_s^* v_{s,i} = \frac{1}{1+r} E^{P^*}(v_{\cdot,i}), \quad i = 1, \dots, m. \quad (2.16)$$

Eq. (2.16) confirms Eq. (2.10), obtained in the introductory example of Section 2.5. It says that the price of any asset is the expectation of its future payoffs, taken under the probability P^* , discounted at the risk-free interest rate r . For this reason, we usually refer to the probability P^* as the *risk-neutral* probability. Eq. (2.16) can be extended to a dynamic context, as we shall see in later chapters. Intuitively, consider an asset that distributes dividends in every period, let $S(t)$ be its price at time t , and $D(t)$ the dividend paid off at time t . Then, the “payoff” it promises for the next period is $S(t+1) + D(t+1)$. By Eq. (2.16), $S(t) = (1+r)^{-1} E^{P^*}(S(t+1) + D(t+1))$ or, by rearranging terms,

$$E^{P^*} \left(\frac{S(t+1) + D(t+1) - S(t)}{S(t)} \right) = r. \quad (2.17)$$

That is, the expected return on the asset under P^* equals the safe interest rate, r . In a dynamic context, the risk-neutral probability P^* is also referred to as the *risk-neutral martingale measure*, or *equivalent martingale measure*, for the following reason. Define a money market account as an asset with value evolving over time as $M(t) \equiv (1+r)^t$. Then, Eq. (2.17) can be rewritten as $S(t)/M(t) = E^{P^*} [(S(t+1) + D(t+1))/M(t+1)]$. This shows that if $D(t+1) = 0$ for some t , then, the discounted process $S(t)/M(t)$ is a martingale under P^* .

Next, let us replace P^* into the budget constraint in (2.13), to obtain, for $(c^1 - w^1) \in \langle V \rangle$,

$$0 = c_0 - w_0 + \sum_{s=1}^d \phi_s (c_s - w_s) = c_0 - w_0 + \frac{1}{1+r} \sum_{s=1}^d P_s^* (c_s - w_s) = c_0 - w_0 + \frac{1}{1+r} E^{P^*} (c^1 - w^1). \quad (2.18)$$

For reasons developed below, it is also useful to derive an alternative representation of the budget constraint, in terms of the objective probability P (say). Let us introduce, first, the ratio ζ , defined as,

$$\zeta_s = \frac{P_s^*}{P_s}, \quad s = 1, \dots, d.$$

The ratio ζ_s indicates how far P^* and P are. We assume ζ_s is strictly positive, which means that P^* and P are *equivalent* measures, i.e. they assign the same weight to the null sets. Finally, let us introduce the *stochastic discount factor*, $m = (m_s)_{s=1}^d$, defined as,

$$m_s \equiv (1+r)^{-1} \zeta_s.$$

We have,

$$\frac{1}{1+r} E^{P^*} (c^1 - w^1) = \sum_{s=1}^d \frac{1}{1+r} P_s^* (c_s - w_s) = \sum_{s=1}^d \underbrace{\frac{1}{1+r} \zeta_s}_{=m_s} (c_s - w_s) P_s = E [m \cdot (c^1 - w^1)].$$

Hence, we can rewrite Eq. (2.18) as,

$$0 = c_0 - w_0 + E [m \cdot (c^1 - w^1)], \quad (c^1 - w^1) \in \langle V \rangle.$$

Similarly, by replacing the stochastic discount factor m into Eq. (2.16) we obtain,

$$S_i = \frac{1}{1+r} E^{P^*} (v_{\cdot,i}) = E (m \cdot v_{\cdot,i}), \quad i = 1, \dots, m. \quad (2.19)$$

Naturally, despite all such different ways to express budget constraints and asset prices, the key of the model is still ϕ ,

$$m_s = (1+r)^{-1} \zeta_s = (1+r)^{-1} \frac{P_s^*}{P_s} = \frac{\phi_s}{P_s},$$

which can be recovered, once we solve for the equilibrium stochastic discount factor m , as we shall illustrate in the next section.

2.6.3 Optimality and equilibrium

We have argued that in the absence of arbitrage opportunities, the program of any agent j is

$$\max_{(c_0, c^1)} [u_j(c_{0j}) + \beta_j \cdot E(\nu_j(c_j^1))] \quad \text{subject to } 0 = c_{0j} - w_{0j} + E [m \cdot (c_j^1 - w_j^1)], \quad (c^1 - w^1) \in \langle V \rangle. \quad [\text{P4}]$$

2.6.3.1 Complete markets and risk sharing

In the complete markets case, $\langle V \rangle = \mathbb{R}^d$, so that the first order conditions to the program [P4] are,

$$u'_j(\hat{c}_{0j}) = \lambda_j, \quad \beta_j \nu'_j(\hat{c}_{sj}) = \lambda_j m_s, \quad s = 1, \dots, d,$$

where λ_j is a Lagrange multiplier. So, really, the properties of this model are the same as those of the static model in Section 2.2. Formally, the complete markets economy in this section is the same as the static economy in Section 2.2, once we set $m = d$, where m is the dimension of the commodity space, in Section 2.2, and $p_s = \phi_s$, where p_s is the price of the s -th commodity in Section 2.2, with $p_1 = 1$ (the numéraire), and ϕ_s is the Arrow-Debreu state price in the unified budget constraint of Eq. (2.18).

These simple observations have profound implications: an economy subject to uncertainty can be understood through a static model, in the presence of complete markets! Under the conditions stated in Section 2.2, even complicated models with heterogeneous agents, with potentially interesting asset pricing implications, and still, apparently, so hopelessly difficult to analyze, can actually be “centralized,” through a dedicated design of Pareto’s weights, as formalized in Theorem 2.7. We can actually do much more. First, this centralization property is easily extended to a dynamic context, as we shall see in more advanced parts of these lectures (see Chapter 8), provided markets satisfy the property of being dynamically complete, a property explained in the next two chapters. Second, the assumption agents can exchange Arrow-Debreu securities for all future states of the world, is clearly unrealistic: markets are pretty likely to be incomplete, one possible reason why financial innovation is so pervasive, in practice. Yet the theory about centralization can be extended to an incomplete markets setting, through a system of “stochastic Pareto weights,” as we discuss in detail in Chapter 8. For now, let us proceed with the next simple and fundamental steps.

To illustrate the equilibrium implications of the first order conditions in a simple case, consider an economy with a single agent. In this economy, the first order conditions immediately lead to the following stochastic discount factor,

$$m_s = \beta \frac{\nu'(w_s)}{u'(w_0)}.$$

The economic interpretation of this stochastic discount factor is the following. In the autarchic state,

$$-\left. \frac{dc_0}{dc_s} \right|_{c_0=w_0, c_s=w_s} = \beta \frac{\nu'(w_s)}{u'(w_0)} P_s = m_s P_s = \phi_s$$

is the present consumption the agent is willing to give up to at $t = 0$, in order to obtain additional consumption at time $t = 1$, in state s . In other words, ϕ_s is the price, in terms of the present consumption numéraire, of one additional unit of consumption at time $t = 1$ and state s . So it is a state price, such that, the agent is happy to consume his own endowment, without any incentives to trade in the financial markets. The risk-neutral probability is,

$$P_s^* = \zeta_s P_s = (1+r) m_s P_s = (1+r) \beta \frac{\nu'(w_s)}{u'(w_0)} P_s.$$

By the first order conditions, and the pure discount bond evaluation formula, it is easily checked that $1 = \sum_{s=1}^d P_s^*$. Moreover,

$$\frac{P_s^*}{P_s} = m_s (1+r) = m_s \left[\beta E \left(\frac{\nu'(w_s)}{u'(w_0)} \right) \right]^{-1} = m_s \beta^{-1} \frac{u'(w_0)}{E[\nu'(w_s)]} = \frac{\nu'(w_s)}{E[\nu'(w_s)]},$$

where the second equality follows by the pure discount bond evaluation formula: $\frac{1}{1+r} = E(m)$.

In the multi-agent case, the situation is similar as soon as markets are complete. Indeed, consider the first order conditions of each agent,

$$\beta_j \frac{\nu'_j(\hat{c}_{sj})}{u'_j(\hat{c}_{0j})} = m_s, \quad s = 1, \dots, d, \quad j = 1, \dots, n.$$

The previous relation reveals that as soon as markets are complete, agents *must* have the same marginal rate of substitution, in equilibrium. This is because by Theorem 2.10, the state price vector ϕ is unique if and only if markets are complete, which then implies uniqueness of $m_s = \frac{\phi_s}{P_s}$ and, hence, the fact that each marginal rate of substitution $\beta_j \frac{\nu'_j(\hat{c}_{sj})}{u'_j(\hat{c}_{0j})}$ is independent of j . In this case, the equilibrium allocation is clearly a Pareto optimum, by the discussion at the beginning of this section, and Theorem 2.5.

The result that agents have the same marginal rate of substitution for each state of the world is known as *risk sharing*. It means that, given an initial endowment distribution among the agents, the market mechanism, through to a system of complete securities markets, is such that consumption risk is shifted around the economy, so that it is borne by the agents most willing to take it. For example, suppose that two agents 1 and 2 have the same discount rate, and utility functions $u_j = \nu_j$, with CRRA given by η_1 and η_2 , where $\eta_1 < \eta_2$. Then, $\text{Gr}_{s1} = (\text{Gr}_{s2})^{\eta_2/\eta_1}$, where Gr_{si} is consumption growth for the i -th agent in state s . In good times, when $\text{Gr}_{s2} > 1$, the more risk-averse agent experiences, ex-post, a lower consumption growth rate, $\text{Gr}_{s2} < \text{Gr}_{s1}$. In bad times, however, when $\text{Gr}_{s2} < 1$, the more risk-averse agent experiences, ex-post, a higher consumption growth rate, $\text{Gr}_{s2} > \text{Gr}_{s1}$. In other words, capital markets, when complete, operate in such a way to have the more risk-averse agent face a less volatile consumption growth.

2.6.3.2 Incomplete markets

If markets are incomplete, marginal rates of substitution cannot be equal, among agents, except perhaps on a set of endowments distribution with measure zero. The best outcome in this case, is a set of equilibria called *constrained Pareto optima*, i.e. constrained by ... the states of nature. As it turns out, there might not even exist constrained Pareto optima in multiperiod economies with incomplete markets—except perhaps those arising on a set of endowments distributions with zero measure.

When market are incomplete, the state price vector ϕ is not unique. That is, suppose that ϕ^\top is an equilibrium state price. Then, all the elements of

$$\Phi = \{\phi' \in \mathbb{R}_{++}^d : (\phi' - \phi)^\top V = 0\} \quad (2.20)$$

are also equilibrium state prices - there exists an infinity of equilibrium state prices that are consistent with absence of arbitrage opportunities. In other words, there exists an infinity of equilibrium state prices guaranteeing the same observable assets price vector S , for $\phi'^\top V = \phi^\top V = S$.

How do we proceed in this case? Introduce the following budget constraint:

$$\mathcal{C} = \{c \in \mathbb{R}_{++}^d : 0 = c_0 - w_0 + \phi^\top (c^1 - w^1), \quad (c^1 - w^1) \in \langle V \rangle, \quad \forall \phi \in \mathbb{R}_{++}^d : S = \phi^\top V\}. \quad (2.21)$$

This budget constraint, and the previous reasoning about the set Φ in (2.20) shows that in the context of incomplete markets, there exists many constraints to take care of, and the previous “martingale methods,” do not apply.

Yet let $\text{Val}(P_I)$ be the value of the following program in the incomplete markets at hand:

$$\max_{c \in \mathcal{C}} [u_j(c_{0j}) + \beta_j E(\nu_j(c_j^1))]. \quad [\text{P}_I]$$

Consider, next, the following constraint:

$$\mathcal{C}_\phi = \left\{ c \in \mathbb{R}_{++}^d : 0 = c_0 - w_0 + \phi^\top (c^1 - w^1), \quad (c^1 - w^1) \in \mathbb{R}^d, \right. \\ \left. \text{for some given } \phi \in \mathbb{R}_{++}^d : S = \phi^\top V \right\},$$

and let $\text{Val}(P_\phi)$ be the value of the program in some abstract complete markets case:

$$\max_{c \in \mathcal{C}_\phi} [u_j(c_{0j}) + \beta_j E(\nu_j(c_j^1))]. \quad [\text{P}_\phi]$$

Clearly, we have, $\text{Val}(P_I) \leq \text{Val}(P_\phi)$ for all ϕ , for the constraint in the incomplete markets case, \mathcal{C} , is more stringent than that in any complete market setting, \mathcal{C}_ϕ : the solution to the program in the incomplete markets case $[\text{P}_I]$, must satisfy the budget constraints in \mathcal{C} , formed using *all* of the possible Arrow-Debreu state prices (including the Arrow-Debreu state price ϕ given in \mathcal{C}_ϕ), as the constraint of Eq. (2.21) shows. Moreover, $(c^1 - w^1) \in \langle V \rangle$. These remarks suggest to define the following “min-max” Arrow-Debreu state price:

$$\phi^* = \arg \min_{\phi \in \Phi} \text{Val}(P_\phi).$$

The natural question is to know whether

$$\text{Val}(P_I) = \text{Val}(P_{\phi^*}). \quad (2.22)$$

This is indeed the case, given some regularity conditions. For the characterization of ϕ^* , suppose there exists $\hat{\phi} : \text{Val}(P_I) = \text{Val}(P_{\hat{\phi}})$. Then, $\hat{\phi} = \phi^*$. Indeed, suppose the contrary, i.e. there exists $\phi' : \text{Val}(P_{\phi'}) < \text{Val}(P_{\hat{\phi}})$. Then, we would have,

$$\text{Val}(P_I) \leq \text{Val}(P_{\phi'}) < \text{Val}(P_{\hat{\phi}}) = \text{Val}(P_I),$$

a contradiction. Note, again, this is a characterization result about ϕ^* , not an existence proof. But as mentioned earlier, Eq. (2.22) holds true, as shown in a dynamic setting by He and Pearson (1991). Chapter 4 provides general guidance about an even more general approach to solving problems of this kind, arising in a broader context of market imperfections, including incomplete markets as a special case.

2.6.3.3 Computation of the equilibrium

The first order conditions satisfied by any agent’s program are:

$$\hat{c}_{0j} = I_j(\lambda_j), \quad \hat{c}_{sj} = H_j(\beta_j^{-1} \lambda_j m_s), \quad (2.23)$$

where I_j and H_j denote the inverse functions of u'_j and ν'_j . By the assumptions we made on u_j and ν'_j , I_j and H_j inherit the same properties of u'_j and ν'_j . By replacing these functions into the constraint,

$$0 = \hat{c}_{0j} - w_{0j} + E[m \cdot (\hat{c}_j^1 - w_j^1)] = I_j(\lambda_j) - w_{0j} + \sum_{s=1}^d P_s [m_s \cdot (H_j(\beta_j^{-1} \lambda_j m_s) - w_{sj})].$$

Define the function,

$$z_j(\lambda_j) \equiv I_j(\lambda_j) + E[mH_j(\beta_j^{-1}\lambda_j m)] = w_{0j} + E(m \cdot w_j^1).$$

We see that $\lim_{x \rightarrow 0} z(x) = \infty$, $\lim_{x \rightarrow \infty} z(x) = 0$ and $z'(x) < 0$. Therefore, there exists a unique solution for λ_j :

$$\lambda_j \equiv \Lambda_j[w_{0j} + E(m \cdot w_j^1)],$$

where $\Lambda(\cdot)$ denotes the inverse function of z . By replacing back into Eqs. (2.23), we obtain:

$$\hat{c}_{0j} = I_j(\Lambda_j(w_{0j} + E(m \cdot w_j^1))), \quad \hat{c}_{sj} = H_j(\beta_j^{-1} m_s \Lambda_j(w_{0j} + E(m \cdot w_j^1))).$$

It remains to compute the general equilibrium. The kernel m must be determined. This means that we have d unknowns (m_s , $s = 1, \dots, d$). We have $d + 1$ equilibrium conditions (holding in the $d + 1$ markets). By Walras' law, only d of these are independent. Consider the equilibrium conditions in the d markets at the second period:

$$g_s(m_s; (m_{s'})_{s' \neq s}) \equiv \sum_{j=1}^n H_j(\beta_j^{-1} m_s \Lambda_j(w_{0j} + E(m \cdot w_j^1))) = \sum_{j=1}^n w_{sj} \equiv w_s, \quad s = 1, \dots, d.$$

These conditions determine the kernel $(m_s)_{s=1}^d$ which leads to compute prices and equilibrium allocations. Finally, once the optimal \hat{c}_s are computed, for $s = 0, 1, \dots, d$, the portfolio $\hat{\theta}$ generated them can be inferred through $\hat{\theta} = V^{-1}(\hat{c}^1 - w^1)$.

2.7 Consumption-CAPM

Consider the pricing equation (2.19). It states that for every asset with gross return $\tilde{R} \equiv S^{-1} \cdot \text{payoff}$,

$$1 = E(m \cdot \tilde{R}), \quad (2.24)$$

where m is some pricing kernel.

In the previous section, we learnt that in a complete markets economy, equilibrium leads to the following identification of the pricing kernel,

$$m_s = \beta \frac{v'(w_s)}{u'(w_0)}.$$

For a riskless asset, $1 = E(m \cdot R)$. By combining this equality with Eq. (2.24), leaves $E[m \cdot (\tilde{R} - R)] = 0$. By rearranging terms,

$$E(\tilde{R}) = R - \frac{\text{cov}(v'(w^+), \tilde{R})}{E[v'(w^+)]}. \quad (2.25)$$

2.7.1 The risk premium

Eq. (2.25) can be rewritten as,

$$E(\tilde{R}) - R = -\frac{\text{cov}(m, \tilde{R})}{E(m)} = -R \cdot \text{cov}(m, \tilde{R}). \quad (2.26)$$

The risk-premium to invest in the asset is high for securities which pay high returns when consumption is high (i.e. when we don't need high returns) and low returns when consumption is low (i.e. when we need high returns).

All in all, if the price $p = E(m \cdot \text{payoff}) = E(m) E(\text{payoff}) + \text{cov}(m, \text{payoff}) = R^{-1} E(\text{payoff}) - \text{Premium}$, where $\text{Premium} = -\text{cov}(m, \text{payoff})$, a discounting effect.

2.7.2 The beta relation

Suppose there is a \tilde{R}_m such that

$$\tilde{R}_m = -\gamma^{-1} \cdot \nu'(w_s), \quad \text{all } s.$$

In this case,

$$E(\tilde{R}) = R + \frac{\gamma \cdot \text{cov}(\tilde{R}_m, \tilde{R})}{E[\nu'(w^+)]} \quad \text{and} \quad E(\tilde{R}_m) = R + \frac{\gamma \cdot \text{var}(\tilde{R}_m)}{E[\nu'(w^+)]}.$$

These relations can be combined to yield,

$$E(\tilde{R}) - R = \beta \cdot [E(\tilde{R}_m) - R], \quad \beta \equiv \frac{\text{cov}(\tilde{R}_m, \tilde{R})}{\text{var}(\tilde{R}_m)}.$$

2.7.3 CCAPM & CAPM

Let \tilde{R}^p be the portfolio return which is the most highly correlated with the pricing kernel m . We have,

$$E(\tilde{R}^p) - R = -R \cdot \text{cov}(m, \tilde{R}^p). \quad (2.27)$$

Using Eqs. (2.26) and (2.27),

$$\frac{E(\tilde{R}) - R}{E(\tilde{R}^p) - R} = \frac{\text{cov}(m, \tilde{R})}{\text{cov}(m, \tilde{R}^p)},$$

and by rearranging terms,

$$E(\tilde{R}) - R = \frac{\beta_{\tilde{R},m}}{\beta_{\tilde{R}^p,m}} [E(\tilde{R}^p) - R] \quad [\mathbf{CCAPM}].$$

If \tilde{R}^p is perfectly correlated with m , i.e. if there exists $\gamma : \tilde{R}^p = -\gamma m$, then

$$\beta_{\tilde{R},m} = -\gamma \frac{\text{cov}(\tilde{R}^p, \tilde{R})}{\text{var}(\tilde{R}^p)} \quad \text{and} \quad \beta_{\tilde{R}^p,m} = -\gamma$$

and then

$$E(\tilde{R}) - R = \beta_{\tilde{R},\tilde{R}^p} [E(\tilde{R}^p) - R] \quad [\mathbf{CAPM}].$$

This is not the only way the CAPM obtains. As we shall explain in Chapter 6, the CAPM also obtains through the so-called “maximum correlation portfolio,” which is the portfolio that is the most highly correlated with the pricing kernel m .

2.8 Infinite horizon

We consider d states of the nature and $m = d$ Arrow securities. We write a unified budget constraint, as in the valuation equilibria approach of Debreu (1954).

We have,

$$\begin{cases} p_0(c_0 - w_0) = -S^{(0)}\theta^{(0)} = -\sum_{i=1}^m S_i^{(0)}\theta_i^{(0)} \\ p_s^1(c_s^1 - w_s^1) = \theta_s^{(0)}, \quad s = 1, \dots, d \end{cases}$$

or,

$$p_0(c_0 - w_0) + \sum_{i=1}^m S_i^{(0)} [p_i^1 (c_i^1 - w_i^1)] = 0.$$

The previous relation holds in a two-period economy. In a multiperiod economy, in the second period (as in the following periods) agents save indefinitely for the future. In the appendix, we show that,

$$0 = E \left[\sum_{t=0}^{\infty} m_{0,t} \cdot p^t (c^t - w^t) \right], \quad (2.28)$$

where $m_{0,t}$ are the state prices. From the perspective of time 0, at time t there exist d^t states of nature and, thus, d^t possible prices.

2.9 Further topics on incomplete markets

2.9.1 Nominal assets and real indeterminacy of the equilibrium

The equilibrium is a set of prices $(\hat{p}, \hat{S}) \in \mathbb{R}_{++}^{m \cdot (d+1)} \times \mathbb{R}_{++}^a$ such that:

$$0 = \sum_{j=1}^n e_{0j}(\hat{p}, \hat{S}), \quad 0 = \sum_{j=1}^n e_{1j}(\hat{p}, \hat{S}), \quad 0 = \sum_{j=1}^n \theta_j(\hat{p}, \hat{S}),$$

where the previous functions are the results of optimal plans of the agents. This system has $m \cdot (d+1) + a$ equations and $m \cdot (d+1) + a$ unknowns, where $a \leq d$. Let us aggregate the constraints of the agents,

$$p_0 \sum_{j=1}^n e_{0j} = -S \sum_{j=1}^n \theta_j, \quad p_1 \square \sum_{j=1}^n e_{1j} = B \sum_{j=1}^n \theta_j.$$

Suppose the financial markets clearing condition is satisfied, i.e. $\sum_{j=1}^n \theta_j = 0$. Then,

$$\begin{cases} 0 = p_0 \sum_{j=1}^n e_{0j} \equiv p_0 e_0 = \sum_{\ell=1}^m p_0^{(\ell)} e_0^{(\ell)} \\ \mathbf{0}_d = p_1 \square \sum_{j=1}^n e_{1j} \equiv p_1 \square e_1 = \left[\sum_{\ell=1}^m p_1^{(\ell)}(\omega_1) e_1^{(\ell)}(\omega_1), \dots, \sum_{\ell=1}^m p_1^{(\ell)}(\omega_d) e_1^{(\ell)}(\omega_d) \right]^\top \end{cases}$$

Therefore, there is one redundant equation for each state of nature, or $d+1$ redundant equations, in total. As a result, the equilibrium has less independent equations ($m \cdot (d+1) - 1$) than unknowns ($m \cdot (d+1) + d$), i.e., an indeterminacy degree equal to $d+1$. This result does not rely on whether markets are complete or not. In a sense, it is even not an indeterminacy result when markets are complete, as we may always assume agents would organize the exchanges at the beginning. In this case, only the suitably normalized Arrow-Debreu state prices would matter for agents.

The previous indeterminacy can be reduced to $d-1$, as we may use two additional homogeneity relations. To pin down these relations, let us consider the budget constraint of each agent j ,

$$p_0 e_{0j} = -S \theta_j, \quad p_1 \square e_{1j} = B \theta_j.$$

The first-period constraint is still the same if we multiply the spot price vector p_0 and the financial price vector S by a positive constant, λ (say). In other words, if $(\hat{p}_0, \hat{p}_1, \hat{S})$ is an equilibrium, then, $(\lambda\hat{p}_0, \lambda\hat{p}_1, \lambda\hat{S})$ is also an equilibrium, which delivers a first homogeneity relation. To derive the second homogeneity relation, we multiply the spot prices of the second period by a positive constant, λ and increase at the same time the first period agents' purchasing power, by dividing each asset price by the same constant, as follows:

$$p_0 e_{0j} = -\frac{S}{\lambda} \lambda \theta_j, \quad \lambda p_1 \square e_{1j} = B \lambda \theta_j.$$

Therefore, if $(\hat{p}_0, \hat{p}_1, \hat{S})$ is an equilibrium, then, $(\hat{p}_0, \lambda\hat{p}_1, \frac{\hat{S}}{\lambda})$ is also an equilibrium.

2.9.2 Nonneutrality of money

The previous indeterminacy arises because financial contracts are *nominal*, i.e. the asset payoffs are expressed in terms of some *unité de compte* that, among other things, we did not make precise. Such an indeterminacy vanishes if we were to consider *real* contracts, i.e. contracts with payoffs expressed in terms of the goods. To show this, note that in the presence of real contracts, the agents' constraints are

$$\begin{cases} p_0 e_{0j} = -S \theta_j \\ p_1(\omega_s) e_{1j}(\omega_s) = p_1(\omega_s) A_s \theta_j, \quad s = 1, \dots, d \end{cases}$$

where $A_s = [A_s^1, \dots, A_s^a]$ is the $m \times a$ matrix of the real payoffs. The previous constraint now reveals how to “recover” $d + 1$ homogeneity relations. For each strictly positive vector $\lambda = [\lambda_0, \lambda_1, \dots, \lambda_d]$, we have that if $[\hat{p}_0, S, p_1(\omega_1), \dots, p_1(\omega_s), \dots, p_1(\omega_d)]$ is an equilibrium, then, $[\lambda_0 \hat{p}_0, \lambda_0 S, p_1(\omega_1), \dots, p_1(\omega_s), \dots, p_1(\omega_d)]$ is also an equilibrium, and so is $[\hat{p}_0, S, p_1(\omega_1), \dots, \lambda_s p_1(\omega_s), \dots, p_1(\omega_d)]$, for $\lambda_s, s = 1, \dots, d$.

As is clear, the distinction between nominal and real assets has a precise meaning, when one considers a multi-commodity economy. Even in this case, however, such a distinction is not very interesting without a suitable introduction of a *unité de compte*. These considerations led Magill and Quinzii (1992) to solve the indeterminacy while still remaining in a framework with nominal assets. They simply propose to introduce money as a mean of exchange. The indeterminacy can then be resolved by “fixing” the prices via the $d + 1$ equations defining the money market equilibrium in all states of nature:

$$M_s = p_s \cdot \sum_{j=1}^n w_{sj}, \quad s = 0, 1, \dots, d.$$

Magill and Quinzii showed that the monetary policy $(M_s)_{s=0}^d$ is generically nonneutral.

2.10 Appendix 1

In this appendix we prove that the program [P1] has a unique maximum. Indeed, suppose on the contrary that we have two maxima:

$$\bar{c} = (\bar{c}_{1j}, \dots, \bar{c}_{mj}) \quad \text{and} \quad \bar{\bar{c}} = (\bar{\bar{c}}_{1j}, \dots, \bar{\bar{c}}_{mj}).$$

These two maxima would satisfy $u_j(\bar{c}) = u_j(\bar{\bar{c}})$, with $\sum_{i=1}^m p_i \bar{c}_{ij} = \sum_{i=1}^m p_i \bar{\bar{c}}_{ij} = R_j$. To check that this claim is correct, suppose on the contrary that $\sum_{i=1}^m p_i \bar{\bar{c}}_{ij} < R_j$. Then, the consumption bundle,

$$\bar{\bar{\bar{c}}} = (\bar{\bar{c}}_{1j} + \varepsilon, \dots, \bar{\bar{c}}_{mj}), \quad \varepsilon > 0,$$

would be preferred to $\bar{\bar{c}}$, by Assumption 2.1, and, at the same time, it would hold that, for sufficiently small ε ,

$$\sum_{i=1}^m p_i \bar{\bar{\bar{c}}}_{ij} = \varepsilon p_1 + \sum_{i=1}^m p_i \bar{\bar{c}}_{ij} < R_j.$$

[Indeed, we have, $A \equiv \sum_{i=1}^m p_i \bar{\bar{c}}_{ij}$. $A < R_j \Rightarrow \exists \varepsilon > 0 : A + \varepsilon p_1 < R_j$. E.g., $\varepsilon p_1 = R_j - A - \eta$, $\eta > 0$. The condition is then: $\exists \eta > 0 : R_j - A > \eta$.] Hence, $\bar{\bar{\bar{c}}}$ would be a solution to [P1], thereby contradicting the optimality of $\bar{\bar{c}}$. Therefore, the existence of two optima would imply a full use of resources. Next, consider a point y lying between \bar{c} and $\bar{\bar{c}}$, viz $y = \alpha \bar{c} + (1 - \alpha) \bar{\bar{c}}$, $\alpha \in (0, 1)$. By Assumption 2.1,

$$u_j(y) = u_j(\alpha \bar{c} + (1 - \alpha) \bar{\bar{c}}) > u_j(\bar{c}) = u_j(\bar{\bar{c}}).$$

Moreover,

$$\sum_{i=1}^m p_i y_i = \sum_{i=1}^m p_i (\alpha \bar{c}_{ij} + (1 - \alpha) \bar{\bar{c}}_{ij}) = \alpha \sum_{i=1}^m p_i \bar{c}_{ij} + \sum_{i=1}^m p_i \bar{\bar{c}}_{ij} - \alpha \sum_{i=1}^m p_i \bar{\bar{c}}_{ij} = \alpha R_j + R_j - \alpha R_j = R_j.$$

Hence, $y \in B_j(p)$ and is also strictly preferred to \bar{c} and $\bar{\bar{c}}$, which means that \bar{c} and $\bar{\bar{c}}$ are not optima, as initially conjectured. This establishes uniqueness of the solution to [P1].

2.11 Appendix 2: Proofs of selected results

We first provide a useful result, a well-known theorem on separation of two convex sets. We use this theorem to deal with the proof of the second welfare theorem (Theorem 2.4) and the existence of state prices tying up all asset prices together (Theorem 2.10). A final proof we provide in this appendix is that of Eq. (2.28).

MINKOWSKI'S SEPARATION THEOREM. *Let A and B be two non-empty convex subsets of \mathbb{R}^d . If A is closed, B is compact and $A \cap B = \emptyset$, then there exists a $\phi \in \mathbb{R}^d$ and two real numbers d_1, d_2 such that:*

$$a^\top \phi \leq d_1 < d_2 \leq b^\top \phi, \quad \forall a \in A, \forall b \in B.$$

We are now ready to prove Theorems 2.4 and 2.10.

PROOF OF THEOREM 2.4. Let \bar{c} be a Pareto optimum and $\tilde{B}_j = \{c^j : u_j(c^j) > u_j(\bar{c}^j)\}$. Let us consider the two sets $\tilde{B} = \bigcup_{j=1}^n \tilde{B}_j$ and $A = \left\{ (c^j)_{j=1}^n : c^j \geq 0 \forall j, \sum_{j=1}^n c^j = w \right\}$. A is the set of all possible combinations of feasible allocations. By the definition of a Pareto optimum, there are no elements in A that are simultaneously in \tilde{B} , or $A \cap \tilde{B} = \emptyset$. In particular, this is true for all compact subsets B of \tilde{B} , or $A \cap B = \emptyset$. Because A is closed, then, by the Minkowski's separating theorem, there exists a $p \in \mathbb{R}^m$ and two distinct numbers d_1, d_2 such that

$$p^\top a \leq d_1 < d_2 \leq p^\top b, \quad \forall a \in A, \forall b \in B.$$

This means that for all allocations $(c^j)_{j=1}^n$ preferred to \bar{c} , we have:

$$p^\top \sum_{j=1}^n w^j < p^\top \sum_{j=1}^n c^j,$$

or, by replacing $\sum_{j=1}^n w^j$ with $\sum_{j=1}^n \bar{c}^j$,

$$p^\top \sum_{j=1}^n \bar{c}^j < p^\top \sum_{j=1}^n c^j. \tag{2A.1}$$

Next we show that $p > 0$. Let $\bar{c}_i = \sum_{j=1}^n \bar{c}_{ij}$, $i = 1, \dots, m$, and partition $\bar{c} = (\bar{c}_1, \dots, \bar{c}_m)$. Let us apply the inequality in (2A.1) to $\bar{c} \in A$ and, for $\mu > 0$, to $c = (\bar{c}_1 + \mu, \dots, \bar{c}_m) \in B$. We have $p_1 \mu > 0$, or $p_1 > 0$. By reiterating the argument, $p_i > 0$ for all i . Finally, we choose $c^j = \bar{c}^j + \mathbf{1}_m \frac{\epsilon}{n}$, $j = 2, \dots, n$, $\epsilon > 0$ in (2A.1), $p^\top \bar{c}^1 < p^\top c^1 + p^\top \mathbf{1}_m \epsilon$ or,

$$p^\top \bar{c}^1 < p^\top c^1,$$

for ϵ sufficiently small. This means that $u_1(c^1) > u_1(\bar{c}^1) \Rightarrow p^\top c^1 > p^\top \bar{c}^1$. This means that $\bar{c}^1 = \arg \max_{c^1} u_1(c^1)$ s.t. $p^\top c^1 = p^\top \bar{c}^1$. By symmetry, $\bar{c}^j = \arg \max_{c^j} u_j(c^j)$ s.t. $p^\top c^j = p^\top \bar{c}^j$ for all j . \parallel

PROOF OF THEOREM 2.10. The condition in (2.12) holds for any compact subset of \mathbb{R}_+^{d+1} , and therefore it holds when it is restricted to the unit simplex in \mathbb{R}_+^{d+1} ,

$$\langle W \rangle \cap \mathcal{S}^d = \{0\}.$$

By the Minkowski's separation theorem, $\exists \tilde{\phi} \in \mathbb{R}^{d+1} : w^\top \tilde{\phi} \leq d_1 < d_2 \leq \sigma^\top \tilde{\phi}$, $w \in \langle W \rangle$, $\sigma \in \mathcal{S}^d$. By walking along the simplex boundaries, one finds that $d_1 < \tilde{\phi}_s$, $s = 1, \dots, d$. On the other hand,

$0 \in \langle W \rangle$, which reveals that $d_1 \geq 0$, and $\tilde{\phi} \in \mathbb{R}_{++}^{d+1}$. Next we show that $w^\top \tilde{\phi} = 0$. Assume the contrary, i.e. $\exists w_* \in \langle W \rangle$ that satisfies at the same time $w_*^\top \tilde{\phi} \neq 0$. In this case, there would be a real number ϵ with $\text{sign}(\epsilon) = \text{sign}(w_*^\top \tilde{\phi})$ such that $\epsilon w_* \in \langle W \rangle$ and $\epsilon w_*^\top \tilde{\phi} > d_2$, a contradiction. Therefore, we have $0 = \tilde{\phi}^\top W \theta = (\tilde{\phi}^\top (-S V)^\top) \theta = (-\tilde{\phi}_0 S + \tilde{\phi}_{(d)}^\top V) \theta$, $\forall \theta \in \mathbb{R}^m$, where $\tilde{\phi}_{(d)}$ contains the last d components of $\tilde{\phi}$. Whence $S = \phi^\top V$, where $\phi^\top = \left(\frac{\tilde{\phi}_1}{\tilde{\phi}_0}, \dots, \frac{\tilde{\phi}_d}{\tilde{\phi}_0} \right)$.

The proof of the converse is immediate (hint: multiply by θ): **shown in further notes.**

The proof of the second part is the following one. We have that “each point of \mathbb{R}^{d+1} is equal to each point of $\langle W \rangle$ plus each point of $\langle W \rangle^\perp$,” or $\dim \langle W \rangle + \dim \langle W \rangle^\perp = d + 1$. Since $\dim \langle W \rangle = \text{rank}(W)$, $\dim \langle W \rangle^\perp = d + 1 - \dim \langle W \rangle$, and since $S = \phi^\top V$ in the absence of arbitrage opportunities, $\dim \langle W \rangle = \dim \langle V \rangle = m$, whence:

$$\dim \langle W \rangle^\perp = d - m + 1.$$

In other terms, before we showed that $\exists \tilde{\phi} : \tilde{\phi}^\top W = 0$, or $\tilde{\phi}^\top \in \langle W \rangle^\perp$. Whence $\dim \langle W \rangle^\perp \geq 1$ in the absence of arbitrage opportunities. The previous relation provides more information. Specifically, $\dim \langle W \rangle^\perp = 1$ if and only if $d = m$. In this case, $\dim \{ \tilde{\phi} \in \mathbb{R}_+^{d+1} : \tilde{\phi}^\top W = 0 \} = 1$, which means that the relation $-\tilde{\phi}_0 S + \tilde{\phi}_d^\top V = 0$ also holds true for $\tilde{\phi}^* = \tilde{\phi} \cdot \lambda$, for every positive scalar λ , but there are no other possible candidates. Therefore, $\phi^\top = \left(\frac{\tilde{\phi}_1}{\tilde{\phi}_0}, \dots, \frac{\tilde{\phi}_d}{\tilde{\phi}_0} \right)$ is such that $\phi = \phi(\lambda)$, and then it is unique.

By a similar reasoning, $\dim \{ \tilde{\phi} \in \mathbb{R}_+^{d+1} : \tilde{\phi}^\top W = 0 \} = d - m + 1 \Rightarrow \dim \{ \phi \in \mathbb{R}_{++}^d : S = \phi^\top V \} = d - m$.
 \parallel

PROOF OF EQ. (2.28). Let $S_{s',s}^{(2)(\ell)}$ be the price at $t = 2$ in state s' if the state in $t = 1$ was s , for the Arrow security promising 1 unit of numéraire in state ℓ at $t = 3$. Let $S_{s',s}^{(2)} = [S_{s',s}^{(2)(1)}, \dots, S_{s',s}^{(2)(m)}]$. Let $\theta_i^{(1)(s)}$ be the quantity purchased at $t = 1$ in state i of Arrow securities promising 1 unit of numéraire if s at $t = 2$. Let $p_{s,i}^2$ be the price of the good at $t = 2$ in state s if the previous state at $t = 1$ was i . Let $S^{(0)(i)}$ and $S_s^{(1)(i)}$ correspond to $S_{s',s}^{(2)(\ell)}$; $S^{(0)}$ and $S_s^{(1)}$ correspond to $S_{s',s}^{(2)}$.

The budget constraint is

$$\begin{cases} p_0 (c_0 - w_0) = -S^{(0)} \theta^{(0)} = -\sum_{i=1}^m S^{(0)(i)} \theta^{(0)(i)} \\ p_s^1 (c_s^1 - w_s^1) = \theta^{(0)(s)} - S_s^{(1)} \theta_s^{(1)} = \theta^{(0)(s)} - \sum_{i=1}^m S_s^{(1)(i)} \theta_s^{(1)(i)}, \quad s = 1, \dots, d. \end{cases}$$

where $S_s^{(1)(i)}$ is the price to be paid at time 1 and in state s , for an Arrow security giving 1 unit of numéraire if the state at time 2 is i .

By replacing the second equation of (3.9) in the first one:

$$\begin{aligned} p_0 (c_0 - w_0) &= -\sum_{i=1}^m S^{(0)(i)} \left[p_i^1 (c_i^1 - w_i^1) + S_i^{(1)} \theta_i^{(1)} \right] \\ \iff \\ 0 &= p_0 (c_0 - w_0) + \sum_{i=1}^m S^{(0)(i)} p_i^1 (c_i^1 - w_i^1) + \sum_{i=1}^m S^{(0)(i)} S_i^{(1)} \theta_i^{(1)} \\ &= p_0 (c_0 - w_0) + \sum_{i=1}^m S^{(0)(i)} p_i^1 (c_i^1 - w_i^1) + \sum_{i=1}^m S^{(0)(i)} \sum_{j=1}^m S_i^{(1)(j)} \theta_i^{(1)(j)} \\ &= p_0 (c_0 - w_0) + \sum_{i=1}^m S^{(0)(i)} p_i^1 (c_i^1 - w_i^1) + \sum_{i=1}^m \sum_{j=1}^m S^{(0)(i)} S_i^{(1)(j)} \theta_i^{(1)(j)} \end{aligned}$$

At time 2,

$$p_{s,i}^2 (c_{s,i}^2 - w_{s,i}^2) = \theta_i^{(1)(s)} - S_{s,i}^{(2)} \theta_{s,i}^{(2)} = \theta_i^{(1)(s)} - \sum_{\ell=1}^m S_{s,i}^{(2)(\ell)} \theta_{s,i}^{(2)(\ell)}, \quad s = 1, \dots, d.$$

Here $S_{s,i}^{(2)}$ is the price vector, to be paid at time 2 in state s if the previous state was i , for the Arrow securities expiring at time 3. The other symbols have a similar interpretation.

By plugging (???) into (???),

$$\begin{aligned} 0 &= p_0 (c_0 - w_0) + \sum_{i=1}^m S^{(0)(i)} p_i^1 (c_i^1 - w_i^1) + \sum_{i=1}^m \sum_{j=1}^m S^{(0)(i)} S_i^{(1)(j)} \left[p_{j,i}^2 (c_{j,i}^2 - w_{j,i}^2) + S_{j,i}^{(2)} \theta_{j,i}^{(2)} \right] \\ &= p_0 (c_0 - w_0) + \sum_{i=1}^m S^{(0)(i)} p_i^1 (c_i^1 - w_i^1) + \sum_{i=1}^m \sum_{j=1}^m S^{(0)(i)} S_i^{(1)(j)} p_{j,i}^2 (c_{j,i}^2 - w_{j,i}^2) \\ &\quad + \sum_{i=1}^m \sum_{j=1}^m \sum_{\ell=1}^m S^{(0)(i)} S_i^{(1)(j)} S_{j,i}^{(2)(\ell)} \theta_{j,i}^{(2)(\ell)}. \end{aligned}$$

In the absence of arbitrage opportunities, $\exists \phi_{t+1,s'} \in \mathbb{R}_{++}^d$ - the state prices vector for $t+1$ if the state in t is s' - such that:

$$S_{s',s}^{(t)(\ell)} = \phi'_{t+1,s'} \cdot e_\ell, \quad \ell = 1, \dots, m,$$

where $e_\ell \in \mathbb{R}_+^d$ and has all zeros except in the ℓ -th component which is 1. Next, we restate the previous relation in terms of the kernel $m_{t+1,s'} = (m_{t+1,s'}^{(\ell)})_{\ell=1}^d$ and the probability distribution $P_{t+1,s'} = (P_{t+1,s'}^{(\ell)})_{\ell=1}^d$ of the events in $t+1$ when the state in t is s' :

$$S_{s',s}^{(t)(\ell)} = m_{t+1,s'}^{(\ell)} \cdot P_{t+1,s'}^{(\ell)}, \quad \ell = 1, \dots, m.$$

By replacing in (???), and imposing the transversality condition:

$$\sum_{\ell_1=1}^m \sum_{\ell_2=1}^m \sum_{\ell_3=1}^m \sum_{\ell_4=1}^m \dots \sum_{\ell_t=1}^m \dots S^{(0)(\ell_1)} S_{\ell_1}^{(1)(\ell_2)} S_{\ell_2, \ell_1}^{(2)(\ell_3)} S_{\ell_3, \ell_2}^{(3)(\ell_4)} \dots S_{\ell_{t-1}, \ell_{t-2}}^{(t-1)(\ell_t)} \dots \xrightarrow{t \rightarrow \infty} 0,$$

we get eq. (2.28). \parallel

2.12 Appendix 3: The multicommodity case

The multicommodity case is interesting, but at the same time is extremely delicate to deal with when markets are incomplete. While standard regularity conditions ensure the existence of an equilibrium in the static and complete markets case, only “generic” existence results are available for the incomplete markets cases. Hart (1974) built up well-chosen examples in which there exist sets of endowments distributions for which no equilibrium can exist. However, Duffie and Shafer (1985) showed that such sets have zero measure, which justifies the terminology of “generic” existence.

Here we only provide a derivation of the constraints. m_t commodities are traded in period t ($t = 0, 1$). The states of nature in the second period are d , and the number of traded assets is a . The first period budget constraint is:

$$p_0 e_{0j} = -S\theta_j, \quad e_{0j} \equiv c_{0j} - w_{0j}$$

where $p_0 = (p_0^{(1)}, \dots, p_0^{(m_1)})$ is the first period price vector, $e_{0j} = (e_{0j}^{(1)}, \dots, e_{0j}^{(m_1)})'$ is the first period excess demands vector, $S = (S_1, \dots, S_a)$ is the financial asset price vector, and $\theta_j = (\theta_{1j}, \dots, \theta_{aj})'$ is the vector of assets quantities that agent j buys at the first period.

The second period budget constraint is,

$$E_1 p_1' = B \cdot \theta_j$$

where

$$E_1 = \begin{bmatrix} e_1(\omega_1) & 0 & \cdots & 0 \\ 1 \times m_2 & 1 \times m_2 & & 1 \times m_2 \\ 0 & e_1(\omega_2) & \cdots & 0 \\ 1 \times m_2 & 1 \times m_2 & & 1 \times m_2 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & e_1(\omega_d) \\ 1 \times m_2 & 1 \times m_2 & & 1 \times m_2 \end{bmatrix}$$

is the matrix of excess demands, $p_1 = (p_1(\omega_1), \dots, p_1(\omega_d))$ is the matrix of spot prices, and

$$B = \begin{bmatrix} v_1(\omega_1) & v_a(\omega_1) \\ \vdots & \vdots \\ v_1(\omega_d) & v_a(\omega_d) \end{bmatrix}$$

is the payoffs matrix. We can rewrite the second period constraint as $p_1 \square e_{1j} = B \cdot \theta_j$, where e_{1j} is defined similarly as e_{0j} , and $p_1 \square e_{1j} \equiv (p_1(\omega_1)e_{1j}(\omega_1), \dots, p_1(\omega_d)e_{1j}(\omega_d))'$. The budget constraints are then,

$$p_0 e_{0j} = -S\theta_j, \quad p_1 \square e_{1j} = B\theta_j.$$

Now suppose that markets are complete, i.e., $a = d$ and B can be inverted. The second constraint is then: $\theta_j = B^{-1}p_1 \square e_{1j}$. Consider without loss of generality Arrow securities, or $B = I$. We have $\theta_j = p_1 \square e_{1j}$, and by replacing into the first constraint,

$$\begin{aligned} 0 &= p_0 e_{0j} + S\theta_j \\ &= p_0 e_{0j} + S p_1 \square e_{1j} \\ &= p_0 e_{0j} + S \cdot (p_1(\omega_1)e_{1j}(\omega_1), \dots, p_1(\omega_d)e_{1j}(\omega_d))' \\ &= p_0 e_{0j} + \sum_{i=1}^d S_i \cdot p_1(\omega_i)e_{1j}(\omega_i) \\ &= \sum_{h=1}^{m_1} p_0^{(h)} e_{0j}^{(h)} + \sum_{i=1}^d S_i \cdot \sum_{\ell=1}^{m_2} p_1^{(\ell)}(\omega_i) e_{1j}^{\ell}(\omega_i) \\ &= \sum_{h=1}^{m_1} p_0^{(h)} e_{0j}^{(h)} + \sum_{i=1}^d \sum_{\ell=1}^{m_2} \hat{p}_1^{(\ell)}(\omega_i) e_{1j}^{\ell}(\omega_i) \end{aligned}$$

where $\tilde{p}_1^{(\ell)}(\omega_i) \equiv S_i \cdot p_1^{(\ell)}(\omega_i)$. The price to be paid today for the obtention of a good ℓ in state i is equal to the price of an Arrow asset written for state i multiplied by the spot price $\tilde{p}_1^{(\ell)}(\omega_i)$ of this good in this state; here the Arrow-Debreu state price is $\tilde{p}_1^{(\ell)}(\omega_i)$. The general equilibrium can be analyzed by making reference to such state prices. From now on, we simplify and set $m_1 = m_2 \equiv m$. Then we are left with determining $m(d+1)$ equilibrium prices, i.e. $p_0 = (p_0^{(1)}, \dots, p_0^{(m)})$, $\tilde{p}_1(\omega_1) = (\tilde{p}_1^{(1)}(\omega_1), \dots, \tilde{p}_1^{(m)}(\omega_1))$, \dots , $\tilde{p}_1(\omega_d) = (\tilde{p}_1^{(1)}(\omega_d), \dots, \tilde{p}_1^{(m)}(\omega_d))$. By exactly the same arguments of the previous chapter, there exists one degree of indeterminacy. Therefore, there are only $m(d+1) - 1$ relations that can determine the $m(d+1)$ prices. (Price normalization can be done by letting one of the first period commodities be the numéraire.) On the other hand, in the initial economy we have to determine $m(d+1) + d$ prices $(\hat{p}, \hat{S}) \in \mathbb{R}_{++}^{m \cdot (d+1)} \times \mathbb{R}_{++}^d$ which are the solution to the system:

$$\sum_{j=1}^n e_{0j}(\hat{p}, \hat{S}) = 0, \quad \sum_{j=1}^n e_{1j}(\hat{p}, \hat{S}) = 0, \quad \sum_{j=1}^n \theta_j(\hat{p}, \hat{S}) = 0,$$

where the previous functions are obtained as solutions to the agents' programs. When we solve for Arrow-Debreu prices, in a second step we have to determine $m(d+1) + d$ prices starting from the knowledge of $m(d+1) - 1$ relations defining the Arrow-Debreu prices, which implies a price indeterminacy of the initial economy equal to $d+1$. In fact, it is possible to show that the degree of indeterminacy is only $d-1$.

References

- Arrow, K. J. (1953): “Le rôle des valeurs boursières pour la répartition la meilleure des risques.” *Econométrie* 41-48. CNRS, Paris. Translated and reprinted in 1964: “The Role of Securities in the Optimal Allocation of Risk-Bearing.” *Review of Economic Studies* 31, 91-96.
- Debreu, G. (1954): “Valuation Equilibrium and Pareto Optimum.” *Proceedings of the National Academy of Sciences* 40, 588-592.
- Debreu, G. (1959): *Theory of Value: An Axiomatic Analysis of Economic Equilibrium*. New Haven: Yale University Press.
- Duffie, D. (2001): *Dynamic Asset Pricing Theory*. Princeton: Princeton University Press.
- Duffie, D. and W. Shafer (1985): “Equilibrium in Incomplete Markets: I. A Basic Model of Generic Existence.” *Journal of Mathematical Economics* 13 285-300.
- Hart, O. (1974): “On the Existence of Equilibrium in a Securities Model.” *Journal of Economic Theory* 9, 293-311.
- He, H. and N. Pearson (1991): “Consumption and Portfolio Policies with Incomplete Markets and Short-Sales Constraints: The Infinite Dimensional Case.” *Journal of Economic Theory* 54, 259-304.

3

Infinite horizon economies

3.1 Introduction

We study asset prices in multiperiod economies, where agents either live forever, and have access to a set of complete markets, or belong to overlapping generations. We consider models without and with production, without and with money, and develop the fundamental tools we need in subsequent chapters, to analyze financial frictions, bubbles and sunspots in capital markets.

3.2 Consumption-based asset evaluation

3.2.1 Recursive plans: introduction

We consider a simple, benchmark case, arising in the absence of any risks for a decision maker. Consider an agent endowed with initial wealth equal to w_0 , who solves the following problem:

$$\begin{aligned} V(w_0) &\equiv \max_{(c_t)_{t=0}^{\infty}} \sum_{t=0}^{\infty} \beta^t u(c_t) \\ \text{s.t. } w_{t+1} &= (w_t - c_t)R_{t+1}, \quad (R_t)_{t=0}^{\infty} \text{ given} \end{aligned} \tag{3.P1}$$

The previous problem can be reformulated in a recursive format:

$$V(w_t) = \max_{c_t} [u(c_t) + \beta V(w_{t+1})] \quad \text{s.t. } w_{t+1} = (w_t - c_t)R_{t+1}. \tag{3.1}$$

By replacing the wealth constraint into the maximand, it is easily checked that the first-order condition for c leads to, $u'(c_t) = \beta V'(w_{t+1})R_{t+1}$. Therefore, the consumption policy is a function of both wealth and the interest rate, which for sake of simplicity we denote as $c(w_t)$. The value function and the first-order condition, then, can be written as:

$$V(w_t) = u(c(w_t)) + \beta V((w_t - c(w_t))R_{t+1}), \quad u'(c(w_t)) = \beta V'((w_t - c(w_t))R_{t+1})R_{t+1}.$$

By differentiating the value function, and using the first-order condition,

$$V'(w_t) = u'(c(w_t))c'(w_t) + \beta V'((w_t - c(w_t))R_{t+1})(1 - c'(w_t))R_{t+1} = u'(c(w_t)).$$

Therefore, $V'(w_{t+1}) = u'(c(w_{t+1}))$ too, and by substituting back into the first-order condition,

$$\beta \frac{u'(c(w_{t+1}))}{u'(c(w_t))} = \frac{1}{R_{t+1}}. \quad (3.2)$$

The economic intuition underlying Eq. (3.2) is the same as that we saw in the two-period economy analyzed in Chapter 2. Eq. (3.2) says that the present consumption I give up, at t , to obtain additional consumption at $t + 1$, has to equal a pure discount bond issued at t and expiring the next period, along an optimal consumption path. Therefore, the bond price represents the relative price of consumption tomorrow, relative to consumption today.

We can arrive to this conclusion through an alternative approach, based on Lagrange multipliers. This approach is useful when dealing with more intricate issues relating to production economies or economies with financial frictions, as we shall see in this and further chapters. So consider the constraint in program [3.P1]. Savings at time t are $\text{sav}_t \equiv w_t - c_t$. Using this definition, the constraint in [3.P1] is: $c_{t+1} + \text{sav}_{t+1} = R_{t+1}\text{sav}_t$, with $\text{sav}_{-1} = w_0$, given. Let λ_t be a sequence of Lagrange multipliers associated to these constraints. Consider the program,

$$\mathcal{L}(\text{sav}_{-1}) \equiv \max_{(c_t, \text{sav}_t)_{t=0}^{\infty}} \sum_{t=0}^{\infty} [\beta^t u(c_t) - \lambda_t (c_t + \text{sav}_t - R_t \text{sav}_{t-1})],$$

where λ_t is a sequence of Lagrange multipliers. The first-order condition for consumption c_t is, $\beta^t u'(c_t) = \lambda_t$, and the first-order condition for savings sav_t leads to: $\lambda_t = \lambda_{t+1} R_{t+1}$. Putting all together yields precisely Eq. (3.2). Note that the same program can be cast, and solved, in a recursive format,

$$\mathcal{L}(\text{sav}_{t-1}) = \max_{c_t, \text{sav}_t, \lambda_t} [u(c_t) - \lambda_t (c_t + \text{sav}_t - R_t \text{sav}_{t-1}) + \beta \mathcal{L}(\text{sav}_t)].$$

The first-order condition for consumption and savings are $u'(c_t) = \lambda_t$ and $\lambda_t = \beta \mathcal{L}'(\text{sav}_t)$, respectively. By replacing the first-order condition for λ_t , i.e. the budget constraint, and differentiating $\mathcal{L}(\text{sav}_{t-1})$, leaves $\mathcal{L}'(\text{sav}_{t-1}) = \beta \mathcal{L}'(\text{sav}_t) R_t$. These conditions lead to Eq. (3.2).

As a simple example, consider the case of a logarithmic utility function, $u(c) = \ln c$. Let us guess that the value function is $V(w_t) \equiv V(w_t; R_t) = a_t + b \ln w_t$. The first-order condition then yields $c(w) = b^{-1}w$. By Eq. (3.2), then, $w_{t+1} = \beta w_t R_{t+1}$. Comparing the right hand side of this equation with the right hand side of the constraint in the program [3.P1], leaves $c(w_t) = (1 - \beta) w_t$; in other terms, $b = (1 - \beta)^{-1}$.¹

Next, we introduce uncertainty.

3.2.2 The marginalist argument

Consider the following thought experiment. At time t , I give up to a small quantity of consumption equal to Δc_t . The reduction in the (current) utility is, then, equal to $\beta^t u'(c_t) \Delta c_t$. But by investing Δc_t in a safe asset, I can have access to $\Delta c_{t+1} = R_{t+1} \Delta c_t$ additional units of consumption at time $t + 1$. These additional consumption units lead to an expected utility gain equal to $\beta^{t+1} E_t (u'(c_{t+1}) \Delta c_{t+1})$, where E_t denote the expectation conditional upon the information up to time t . If c_t and c_{t+1} are part of an optimal consumption plan, I should be left

¹To pin down the coefficient series a_t , use the definition of the value function, $V(w_t; R_t) \equiv u(c(w_t)) + \beta V(w_{t+1}; R_{t+1})$. By plugging $V(w, R_t) = a_t + b \log w$ and $c(w) = (1 - \beta)^{-1} w$ into this definition leaves, $a_t = \ln(1 - \beta) + \beta a_{t+1} + \frac{\beta}{1 - \beta} \ln(\beta R_{t+1})$. If R is constant, a_t is also constant, and equal to $(\ln(1 - \beta) + \frac{\beta}{1 - \beta} \ln(\beta R)) / (1 - \beta)$.

with no incentives to implement these intertemporal consumption transfers. Therefore, along an optimal consumption plan, any reductions and gains in the welfare of the type considered above need to be identical:

$$u'(c_t) = \beta E_t(u'(c_{t+1})R_{t+1}).$$

This relation generalizes Eq. (3.2). Next, suppose that at time t , Δc_t can be invested in a risky asset whose price is S_t . I can buy $\Delta c_t/S_t$ units of this asset. Come time $t+1$, I could sell the asset for S_{t+1} , pocket its dividend D_{t+1} , if any, and finance additional units of consumption equal to $\Delta c_{t+1} = (\Delta c_t/S_t) \cdot (S_{t+1} + D_{t+1})$. The reduction in the current utility is $\beta^t u'(c_t) \Delta c_t$, and the boost in the expected utility at time $t+1$ is $\beta^{t+1} E_t(u'(c_{t+1}) \Delta c_{t+1})$. Again, if I am following an optimal consumption policy, the incentives for this kind of intertemporal transfers should not exist. Therefore, the celebrated Lucas asset pricing equation holds:

$$u'(c_t) = \beta E_t \left[u'(c_{t+1}) \frac{S_{t+1} + D_{t+1}}{S_t} \right]. \quad (3.3)$$

Section 3.2.3 derives Eq. (3.3) through dynamic programming methods, which are essential, once we wish to work through more complex models such as those including financial frictions. The next section, instead, elaborates on Eq. (3.2).

3.2.3 Intertemporal elasticity of substitution

The elasticity of substitution between two consumption goods, c_A and c_B , is defined as $\text{EIS}(c_A, c_B) = -\frac{\partial(\frac{c_B}{c_A})}{\partial(\frac{p_B}{p_A})} \frac{(\frac{p_B}{p_A})}{(\frac{c_B}{c_A})}$, where p_A and p_B are the prices of the two goods. We may define $\text{EIS}(c_t, c_s)$, the elasticity of intertemporal substitution of consumption c_t and c_s at any two points in time t and $s \geq t$, by identifying $c_t = c_A$ and $c_s = c_B$, and replacing $\beta \frac{u'(c_B)}{u'(c_A)} = \frac{p_B}{p_A}$ in the previous expression for $\text{EIS}(c_A, c_B)$, leaving:

$$\text{EIS}(c_t, c_s) = -\frac{d(c_s/c_t)}{d(u'(c_s)/u'(c_t))} \frac{u'(c_s)/u'(c_t)}{c_s/c_t} = -\frac{d(c_s/c_t)/(c_s/c_t)}{db/b},$$

where the zero coupon price, $b = \beta u'(c_s)/u'(c_t) \equiv R^{-1}$, R denotes the gross interest rate from t to s , and the second equality holds in the deterministic case.

The elasticity, $\text{EIS}(c_t, c_s)$, tracks, approximately, the percentage increase in the desired consumption tomorrow relative to today, after a percentage decrease of the price of consumption tomorrow relative to today. Intuitively, high values of $\text{EIS}(c_t, c_s)$ describe a situation where the agent is quite sensitive about consuming at t and s : even a small increase in the interest rate R from t to s and, hence, a small percentage drop in b , can induce him to a substantial relative increase of consumption in the future.

In fact, as $s \rightarrow t$, $\text{EIS}(c_t, c_s)$ collapses to the inverse of the elasticity of marginal utility with respect to consumption or, simply, the relative risk-aversion,

$$\begin{aligned} \frac{1}{\text{EIS}(c_t)} &\equiv \lim_{s \rightarrow t} \frac{1}{\text{EIS}(c_t, c_s)} = -\lim_{s \rightarrow t} \frac{c_s/c_t}{u'(c_s)/u'(c_t)} \frac{d(u'(c_s)/u'(c_t))}{d(c_s/c_t)} \\ &= -\lim_{s \rightarrow t} \frac{d\left(1 + c_t \frac{u''(c_t)}{u'(c_t)} \left(\frac{c_s}{c_t} - 1\right)\right)}{d\left(\frac{c_s}{c_t}\right)} \\ &= -c_t \frac{u''(c_t)}{u'(c_t)}, \end{aligned}$$

where the second equality follows by a first-order Taylor's expansion of the marginal utility of consumption at time s , $u'(c_s) = u'(c_t) + u''(c_t)(c_s - c_t) + O((c_s - c_t)^2)$. The expression, $EIS(c_t)$, is called “instantaneous elasticity of intertemporal substitution.”

For example, in the CRRA case, and in the deterministic case, we have that along an optimal consumption path, $\frac{c_{t+1}}{c_t} = (\beta R)^{1/\eta}$, where η is the CRRA: as R increases, it becomes more attractive to save and postpone consumption. In equilibrium, $\ln R = -\ln \beta + \eta g$, where g denotes the growth rate of the economy, say. When g is high, more consumption will be available in the future, which creates disincentives to save: in this case, the agent is happy to consume relatively more in the future when the price of consumption in the future, relative to today, $b = R^{-1}$, is low, that is when R is high.

An agent with a low EIS has a quite inelastic demand for bonds. Intuitively, when the price of consumption in the future relative to today, b , drops, desired consumption tomorrow relative to today increases. But for an agent with a low EIS, the desired relative increase in future consumption is quite limited, and so is his demand for bonds—the instruments that allow him to allocate intertemporal consumption.

3.2.4 Lucas' model

3.2.4.1 The optimality condition

We consider markets for m “trees,” and assume that the only source of risk stems from the dividends related to these trees: $D = (D_1, \dots, D_m)$. We assume D is a Markov process and denote its conditional distribution function with $P(D_{t+1}|D_t)$. A representative agent solves the following program:

$$\begin{aligned} V(\theta_t, D_t) &= \max_{(c_{t+i}, \theta_{t+i})_{i=0}^{\infty}} E_t \left[\sum_{i=0}^{\infty} \beta^i u(c_{t+i}) \middle| \mathcal{F}_t \right] \\ \text{s.t. } c_t + S_t \theta_{t+1} &= (S_t + D_t) \theta_t \end{aligned} \quad [3.P2]$$

where \mathcal{F}_t denotes the information set as of time t , $\theta_{t+1} \in \mathbb{R}^m$ is \mathcal{F}_t -measurable, that is, θ_{t+1} needs to be chosen at time t . We can solve the program [3.P2], using the same recursive approach in Section 3.2.1, once due account is made of uncertainty. The Bellman's equation is:

$$V(\theta_t, D_t) = \max_{c_t, \theta_{t+1}} E[u(c_t) + \beta V(\theta_{t+1}, D_{t+1}) | \mathcal{F}_t] \quad \text{s.t. } c_t + S_t \theta_{t+1} = (S_t + D_t) \theta_t. \quad (3.4)$$

Similarly as we did for Eq. (3.1), let us replace the budget constraint into the maximand. The following first-order condition holds for θ_i :

$$0 = E[-u'((S_t + D_t) \theta_t - S_t \theta_{t+1}) S_{i,t} + \beta V_{1i}(\theta_{t+1}, D_{t+1}) | \mathcal{F}_t], \quad (3.5)$$

where the subscript in the value function on the right hand side denotes a partial derivative: $V_{1i}(\theta, D) = \partial V(\theta, D) / \partial \theta_i$. The optimal policy, θ_{t+1} is a function of the current state, (θ_t, D_t) , say $\theta_{t+1} = \mathcal{T}(\theta_t, D_t)$. By differentiating the value function with respect to θ_i , and using the previous first-order condition, leaves:

$$\begin{aligned} V_{1i}(\theta_t, D_t) &= E_t \left[u'(c_t) \left(S_{i,t} + D_{i,t} - \sum_{j=1}^m S_{j,t} \mathcal{T}_{1j}^i(\theta_t, D_t) \right) + \beta \sum_{j=1}^m V_{1i}(\theta_{t+1}, D_{t+1}) \mathcal{T}_{1j}^i(\theta_t, D_t) \right] \\ &= u'(c_t) (S_{i,t} + D_{i,t}), \end{aligned}$$

where for brevity, we use E_t to denote the expectation operator conditional upon \mathcal{F}_t , and we have defined $\mathcal{T}_{1j}^i(\theta_t, D_t) = \partial \mathcal{T}_i(\theta, D) / \partial \theta_j$ and \mathcal{T}_i is the i -th component of the vector \mathcal{T} . Substituting this result into Eq. (3.5) yields precisely the Lucas equation (3.3), holding for each asset i :

$$u'(c_t) = \beta E_t \left[u'(c_{t+1}) \frac{S_{i,t+1} + D_{i,t+1}}{S_{i,t}} \right]. \quad (3.6)$$

It is easy to show to extend these conditions to the case where a representative agent can also invest into a locally riskless asset, that is, an asset that expires over the next period. The budget constraint in Eq. (3.4) is, in this case: $c_t + S_t \theta_{t+1} + S_{0,t} \theta_{0,t+1} = (S_t + D_t) \theta_t + \theta_{0,t}$, where $\theta_{0,t}$ denotes the amount of the locally riskless asset, $S_{0,t} \equiv e^{-r_t}$, and r_t is the riskless interest rate, and the Lucas equation for the r_t would be: $e^{-r_t} = \beta E_t [u'(c_{t+1})]$.

3.2.4.2 Rational expectations equilibrium

The asset market clears when for each t , $\theta_t = \mathbf{1}_m$ and $\theta_{0,t} = 0$. By the budget constraint, then, the market for goods also clears, $c_t = \sum_{i=1}^m D_{it} \equiv \bar{D}_t$. A *rational expectation equilibrium* is a sequence of asset prices $(S_t)_{t=0}^\infty$ such that the optimality condition in Eq. (3.6) holds, the markets clear, $c_t = \bar{D}_t$, and each asset price is a function of the state, $S_{i,t} = S_i(D_t)$ say. All in all,

$$u'(\bar{D}_t) S_i(D_t) = \beta \int u'(\bar{D}_{t+1}) (S_i(\bar{D}_{t+1}) + D_{i,t+1}) dP(D_{t+1} | D_t). \quad (3.7)$$

This is a functional equation in $S_i(\cdot)$. Let us focus, first, on the IID case: $P(D_{t+1} | D_t) = P(D_{t+1})$.

IID shocks

Eq. (3.7) simplifies to:

$$u'(\bar{D}_t) S_i(D_t) = \beta \int u'(\bar{D}_{t+1}) (S_i(\bar{D}_{t+1}) + D_{i,t+1}) dP(D_{t+1}).$$

Note that the right hand side of this equation is independent of D . Therefore, $u'(\bar{D}_t) S_i(D_t)$ equals some constant κ_i (say), which we can easily find by substituting it back into the previous equation, leaving:

$$\kappa_i = \frac{\beta}{1 - \beta} \int u'(\bar{D}_{t+1}) D_{i,t+1} dP(D_{t+1}).$$

Therefore, the solution for $S_i(D)$ is:

$$S_i(D_t) = \frac{\kappa_i}{u'(\bar{D}_t)}.$$

Note, the elasticity of the price to dividend equals $-\frac{u''(\bar{D})}{u'(\bar{D})} D_i$, which collapses to relative risk-aversion, once we assume only one tree exists. For example, if relative risk-aversion is constant and equal to η ,

$$S(D_t) = \kappa \cdot D_t^\eta, \quad \kappa \equiv \frac{\beta}{1 - \beta} \int D^{1-\eta} dP(D).$$

Figure 3.1 depicts the behavior of the asset price function $S(D)$, under the assumption that κ is not increasing in η . The asset price collapses to the constant, $\beta(1 - \beta)^{-1} E(D)$, in the special case where the representative agent is risk-neutral, $\eta = 0$.

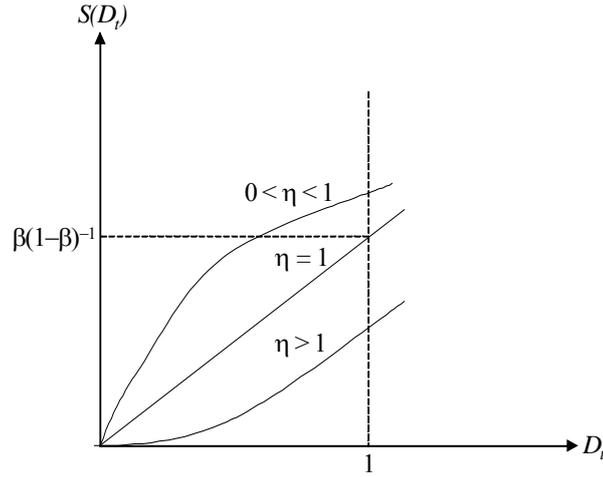


FIGURE 3.1. The asset pricing function $S(D_t)$ in the IID case and constant relative risk-aversion, equal to η .

Dependent shocks

Define $g_i(D) \equiv u'(\bar{D})S_i(D)$ and $h_i(D) \equiv \beta \int u'(\bar{D}_{t+1})D_{i,t+1}dP(D_{t+1}|D)$. In terms of these new functions, Eq. (3.7) is:

$$g_i(D) = h_i(D) + \beta \int g_i(D_{t+1}) dP(D_{t+1}|D).$$

It is a functional equation in g_i , which we can show it admits a unique solution, under the conditions contained in the celebrated Blackwell's theorem below:

THEOREM 3.1. *Let $\mathcal{B}(X)$ the Banach space of continuous bounded real functions on $X \subseteq \mathbb{R}^n$ endowed with the norm $\|f\| = \sup_X |f|$, $f \in \mathcal{B}(X)$. Introduce an operator $T : \mathcal{B}(X) \mapsto \mathcal{B}(X)$ with the following properties:*

- (i) *T is monotone: $\forall x \in X$ and $f_1, f_2 \in \mathcal{B}(X)$, $f_1(x) \leq f_2(x) \iff T[f_1](x) \leq T[f_2](x)$;*
- (ii) *$\forall x \in X$ and $c \geq 0$, $\exists \beta \in (0, 1) : T[f + c](x) \leq T[f](x) + \beta c$.*

Then, T is a β -contraction and, $\forall f_0 \in \mathcal{B}(X)$, it has a unique fixed point $\lim_{\tau \rightarrow \infty} T^\tau[f_0] = f = T[f]$.

So let us introduce the following operator:

$$T[g_i](D) = h_i(D) + \beta \int g_i(D') dP(D'|D).$$

The existence of g_i and, hence, S_i , relies on the existence of a fixed point of $T : g_i = T[g_i]$. It is easily checked that conditions (i) and (ii) in Theorem 3.1 hold here. To establish that $T : \mathcal{B}(D) \mapsto \mathcal{B}(D)$ as well, it is sufficient to show that $h_i \in \mathcal{B}(D)$. A sufficient condition given by Lucas (1978) is that u is bounded, and bounded away by a constant \bar{u} .² Note, a log-utility agent would not satisfy this condition, yet, this case can be easily solved in the case of a single tree, as shown next.

²In this case, concavity of u implies that for each D , $0 = u(0) \leq u(D) + u'(D)(-D) \leq \bar{u} - Du'(D)$, which implies that for each D , $Du'(D) \leq \bar{u}$ and, hence, $h_i(D) \leq \beta\bar{u}$. Then, it is possible to show that the solution is in $\mathcal{B}(D)$, which implies that $T : \mathcal{B}(D) \mapsto \mathcal{B}(D)$.

Suppose, then, that $u(c) = \ln c$, and that there is one single asset, such that Eq. (3.7) collapses to

$$\frac{S(D_t)}{D_t} = \beta \int \left(\frac{S(D_{t+1})}{D_{t+1}} + 1 \right) dP(D_{t+1}|D_t).$$

The solution to this equation is a constant price-dividend ratio,

$$\frac{S(D_t)}{D_t} = \frac{\beta}{1 - \beta}.$$

Note that at this level of generality, it cannot be said more about the price-dividend ratio, in the general CRRA case, even in the single asset case. Indeed, by Eq. (3.7),

$$\frac{S(D_t)}{D_t} = \beta \int \left(\frac{D_{t+1}}{D_t} \right)^{1-\eta} \left(\frac{S(D_{t+1})}{D_{t+1}} + 1 \right) dP(D_{t+1}|D_t).$$

It is easily seen that the solution to this functional equation is:

$$\frac{S(D_t)}{D_t} = \frac{\beta \int \left(\frac{D_{t+1}}{D_t} \right)^{1-\eta} dP(D_{t+1}|D_t)}{1 - \beta \int \left(\frac{D_{t+1}}{D_t} \right)^{1-\eta} dP(D_{t+1}|D_t)},$$

such that the price-dividend ratio is constant whenever the distribution of the consumption endowment growth rate is independent of D_t . In Chapter 6 of Part II of these lectures, we develop this case in more detail, assuming a log-normal distribution for D_{t+1}/D_t .

3.2.4.3 Arrow-Debreu state prices

We have the following consumption-based asset pricing equation, $S_{i,t} = E_t [m_{t+1}(S_{i,t+1} + D_{i,t+1})]$ where $m_{t+1} \equiv \beta \frac{u'(D_{t+1})}{u'(D_t)}$ is the *stochastic discounting factor*, using the terminology of Chapter 2. Under regularity conditions, the Radon-Nikodym derivative of the risk-neutral probability, P^* , with respect to P , is: $\frac{dP^*}{dP} \Big|_{D_{t+1}|D_t} \equiv \frac{u'(D_{t+1})}{E[u'(D_{t+1}|D_t])}$, such that the Arrow-Debreu state-price density is: $d\tilde{P}^*(D_{t+1}|D_t) = R_t^{-1} dP^*(D_{t+1}|D_t)$: the price to pay, in state D_t , to obtain one unit of the good the next period in state D_{t+1} .

3.2.4.4 Multiple trees

The Lucas model is extraordinary complex as in general, the price of any asset depends on the dividends paid by all the remaining assets, as Eq. (3.7) makes clear. The model can generate “contagion,” in that a shock in the fundamentals affecting some assets affects all the other asset evaluation, even when the dividends are not correlated. It is an interesting property, due to the simple circumstance that there is a representative agent who is pricing the same assets—markets are not segmented and a shock to the stochastic discounting factor, m_{t+1} , affects all the asset prices, $S_{i,t}$. We mention efforts made by the literature, discussed in deeper detail in Chapter 8 (Section 8.10): Menzly, Santos and Veronesi (2004), Cochrane, Longstaff and Santa-Clara (2008), Pavlova and Rigobon (2008), Martin (2011).

3.3 Production: foundational issues

In the economy of the previous section, the asset “reward,” is an exogenous datum. In this chapter, we lay down the foundations for the analysis of production-based economies, where

firms maximize their value and set dividends endogenously. In these economies, production and capital accumulation are endogenous. In this section, we review the foundational issues that arise in economies with productive capital. In the next section, we develop the asset pricing implications of these economies, in absence of frictions. In Part II, we extend the framework in this and the next section, and examine the asset price implications deriving from financial frictions.

3.3.1 Decentralized economy

A continuum of identical firms in $(0, 1)$ have access to capital and labor markets, and the following technology: $(K, N) \mapsto Y(K, N)$, where $Y_i(K, N) > 0$, $y_{ii}(K, N) < 0$, $\lim_{K \rightarrow 0^+} Y_1(K, N) = \lim_{N \rightarrow 0^+} Y_2(K, N) = \infty$, $\lim_{K \rightarrow \infty} Y_1(K, N) = \lim_{N \rightarrow \infty} Y_2(K, N) = 0$, and subscripts denote partial derivatives. We assume Y is homogeneous of degree one, i.e. $Y(\lambda K, \lambda N) = \lambda Y(K, N)$ for all $\lambda > 0$. Per capita production is $y(k) \equiv Y(K/N, 1)$, where $k \equiv K/N$ is per-capita capital, Population growth can be non-zero, i.e. N satisfies $N_t/N_{t-1} = (1+n)$. Firms purchase capital and labor at prices $R = Y_1(K, N)$ and $w = Y_2(K, N) = w$. We have,

$$R = y'(k), \quad w = y(k) - ky'(k).$$

The N_t consumers live forever. We assume each consumer offers inelastically one unit of labor, and that, for now, that $N_0 = 1$ and $n = 0$. The resource constraint for the consumer is:

$$c_t + s_t = R_t s_{t-1} + w_t N_t, \quad N_t \equiv 1, \quad t = 1, 2, \dots \quad (3.8)$$

At each time $t-1$, the consumer saves s_{t-1} units of capital, which he lends to the firm. At time t , the consumer receives the gross return on savings from the firm, $R_t s_{t-1}$, where $R_t = y'(k_t)$, plus the wage receipts $w_t N_t$. Then, he uses these resources to consume c_t and lend s_t to the firm. At time zero,

$$c_0 + s_0 = V_0 \equiv Y_1(K_0, N_0)K_0 + w_0 N_0, \quad N_0 \equiv 1.$$

Following the approach developed in Chapter 2, we can write down a single budget constraint, obtained iterating Eq. (3.8):

$$0 = c_0 + \sum_{t=1}^T \frac{c_t - w_t N_t}{\prod_{i=1}^t R_i} + \frac{s_T}{\prod_{i=1}^T R_i} - V_0,$$

and imposing the transversality condition:

$$\lim_{T \rightarrow \infty} s_T \prod_{i=1}^T R_i^{-1} = 0, \quad (3.9)$$

so as to have:

$$\max_{\{c_t\}_{t=0}^{\infty}} \sum_{t=1}^{\infty} \beta^t u(c_t), \quad \text{s.t.} \quad V_0 = c_0 + \sum_{t=1}^{\infty} \frac{c_t - w_t N_t}{\prod_{i=1}^t R_i}. \quad [3.P3]$$

The economic interpretation of the transversality condition (4.30) is the following. The first-order conditions of the program [3.P3] are:

$$\beta^t u'(c_t) = l \frac{1}{\prod_{i=1}^t R_i}, \quad (3.10)$$

where l is a Lagrange multiplier. In equilibrium, current savings equal next period capital, or $k_{t+1} = s_t$. Therefore, Eq. (4.30) is:

$$\lim_{T \rightarrow \infty} \beta^T u'(c_T) k_{T+1} = 0. \quad (3.11)$$

That is, the economic value of capital is capital weighted by discounted marginal utility, which needs to be zero, eventually.

The first-order condition (3.10) leads to the usual optimality condition in Eq. (3.2), where this time, $R_{t+1} = y'(k_{t+1})$. In this economy, an equilibrium is a sequence $((\hat{c}, \hat{k})_t)_{t=0}^{\infty}$ satisfying

$$\begin{cases} k_{t+1} = y(k_t) - c_t \\ \beta \frac{u'(c_{t+1})}{u'(c_t)} = \frac{1}{y'(k_{t+1})} \end{cases} \quad (3.12)$$

and the transversality condition in Eq. (3.11). The first equation in this system is simply this: capital available for producing the next period, k_{t+1} , is equal to savings, $s_t \equiv y(k_t) - c_t$.

3.3.2 Centralized economy

The market solution in (3.12) can be implemented by a social planner, who solves the following program:

$$\begin{aligned} V(k_0) &\equiv \max_{(c_t, k_t)_{t=0}^{\infty}} \sum_{i=0}^{\infty} \beta^i u(c_i) \\ \text{s.t. } &k_{t+1} = y(k_t) - c_t, \quad k_0 \text{ given} \end{aligned} \quad [3.P4]$$

under the further transversality condition in Eq. (3.11).

The program in [3.P4] is easily solved. By replacing the constraint into the utility function, and taking derivatives with respect to k_t , leads directly to the second equation in (3.12). Alternatively, let us introduce the Lagrangian,

$$\mathcal{L}(k_0) = \max_{(c_t, k_{t+1})_{t=0}^{\infty}} \sum_{t=0}^{\infty} [\beta^t u(c_t) - \lambda_t (k_{t+1} - y(k_t) + c_t)].$$

The first-order condition with respect to consumption is $\lambda_t = \beta^t u'(c_t)$, and the condition for capital is $\lambda_{t-1} = \lambda_t y'(k_t)$. Putting these conditions together, leads to the second equation in (3.12). The same argument can be made, following a recursive approach. We have:

$$\mathcal{L}(k_t) = \max_{c_t, k_{t+1}, \lambda_t} [u(c_t) - \lambda_t (k_{t+1} - y(k_t) + c_t) + \beta \mathcal{L}(k_{t+1})].$$

The first-order condition for consumption is $\lambda_t = u'(c_t)$, and that for capital is $\lambda_t = \beta \mathcal{L}'(k_{t+1})$. By replacing the first-order condition for λ_t (i.e., the constraint in program [3.P4]), and differentiating with respect to k_t , yields $\mathcal{L}'(k_t) = \beta \mathcal{L}'(k_{t+1}) y'(k_t)$. These three conditions lead, again, to the second equation in (3.12).

Finally, consider the Bellman's equation:

$$V(k_t) = \max_{c_t} [u(c_t) + \beta V(k_{t+1})], \quad \text{s.t. } k_{t+1} = y(k_t) - c_t.$$

The first-order condition leads to, $u'(c_t) = \beta V'(y(k_t) - c_t)$. Let us denote the policy with $c_t = c(k_t)$. In terms of the policy c function, the value function and the first-order conditions are:

$$V(k_t) = u(c(k_t)) + \beta V(y(k_t) - c(k_t)), \quad u'(c(k_t)) = \beta V'(y(k_t) - c(k_t)).$$

By differentiating the value function:

$$V'(k_t) = u'(c(k_t))c'(k_t) + \beta V'(y(k_t) - c(k_t))(y'(k_t) - c'(k_t)) = u'(c(k_t))y'(k_t).$$

By replacing back into the first-order condition, we obtain the second equation in (3.12).

3.3.3 Dynamics

We study the dynamics of the system in (3.12) in a small neighborhood of the stationary state, defined as the pair (c, k) , solution to:

$$c = y(k) - k, \quad \beta = \frac{1}{y'(k)}.$$

A first-order expansion of each equation in (3.12) around its stationary state, yields the following linear system:

$$\begin{pmatrix} \hat{k}_{t+1} \\ \hat{c}_{t+1} \end{pmatrix} = A \begin{pmatrix} \hat{k}_t \\ \hat{c}_t \end{pmatrix}, \quad A \equiv \begin{pmatrix} y'(k) & -1 \\ -\frac{u'(c)}{u''(c)}y''(k) & 1 + \beta\frac{u'(c)}{u''(c)}y''(k) \end{pmatrix}. \quad (3.13)$$

The solution to this system is obtained with the tools reviewed in Appendix 1 of this chapter. It is:

$$\hat{k}_t = v_{11}\kappa_1\lambda_1^t + v_{12}\kappa_2\lambda_2^t, \quad \hat{c}_t = v_{21}\kappa_1\lambda_1^t + v_{22}\kappa_2\lambda_2^t, \quad (3.14)$$

where: κ_i are constants that depend on the initial state, λ_i are the eigenvalues of A , and $\begin{pmatrix} v_{11} \\ v_{21} \end{pmatrix}$, $\begin{pmatrix} v_{12} \\ v_{22} \end{pmatrix}$ are the eigenvectors associated with λ_i . In Appendix 1, we show that $\lambda_1 \in (0, 1)$ and $\lambda_2 > 1$. The proof we provide in the appendix is important, as it illustrates precisely how the neoclassical model reviewed in this section, needs to be modified to induce indeterminacy in the dynamics of capital and consumption. A critical step in that proof relies on the assumption of diminishing returns, i.e. $y''(k) > 0$.

Let us return to the equations in (3.14). First, we need to rule out an explosive behavior of \hat{k}_t and \hat{c}_t , for otherwise we would contradict (i) that (c, k) is a stationary point, and (ii) the optimality of the trajectories. Since $\lambda_2 > 1$, the only possibility is to “lock” the initial state (\hat{k}_0, \hat{c}_0) in such a way that $\kappa_2 = 0$, which yields the following set of initial conditions: $\hat{k}_0 = v_{11}\kappa_1$ and $\hat{c}_0 = v_{21}\kappa_1$, or $\frac{\hat{c}_0}{\hat{k}_0} = \frac{v_{21}}{v_{11}}$.³ Therefore, the set of initial points that ensure a non-explosive path must lie on the line $c_0 = c + \frac{v_{21}}{v_{11}}(k_0 - k)$. Since k is a predetermined variable, there exists one, and only one, value of c_0 , which ensures a non-explosive path of the system around its steady state, as Figure 3.2 illustrates. In this figure, k_* is defined as the solution of $1 = y'(k_*) \Leftrightarrow k_* = (y')^{-1}[1]$, and $k = (y')^{-1}[\beta^{-1}]$.

The usual word of caution is in order. A linear approximation might turn out to be misleading. We develop one example where the dynamics of the system could be quite different from those analyze here, when we start away from the stationary state. Let $y(k) = k^\gamma$, $u(c) = \ln c$. It is easy to show that the exact solution is:

$$c_t = (1 - \beta\gamma)k_t^\gamma, \quad k_{t+1} = \beta\gamma k_t^\gamma.$$

Figure 3.3 depicts the nonlinear manifold associated with this system, and its linear approximation. For example, let $\beta = 0.99$ and $\gamma = 0.3$. Then, the (linear) saddlepath is, approximately,

$$c_t = c + 0.7101(k_t - k), \quad \text{where: } c = (1 - \gamma\beta)k^\gamma, \quad k = \gamma\beta^{1/(1-\gamma)},$$

where $\hat{k}_t = \lambda_1\hat{k}_{t-1}$, and $\lambda_1 = 0.3$.

³In fact, Appendix 1 shows that the converse is also true, i.e. $\frac{\hat{c}_0}{\hat{k}_0} = \frac{v_{21}}{v_{11}} \Rightarrow \kappa_2 = 0$.

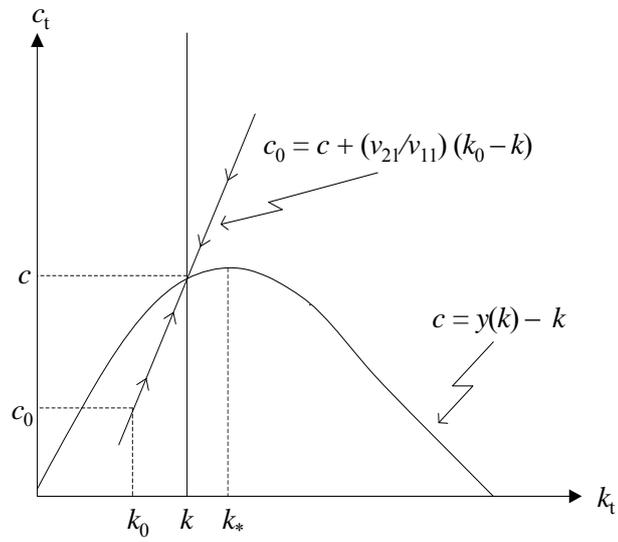


FIGURE 3.2.

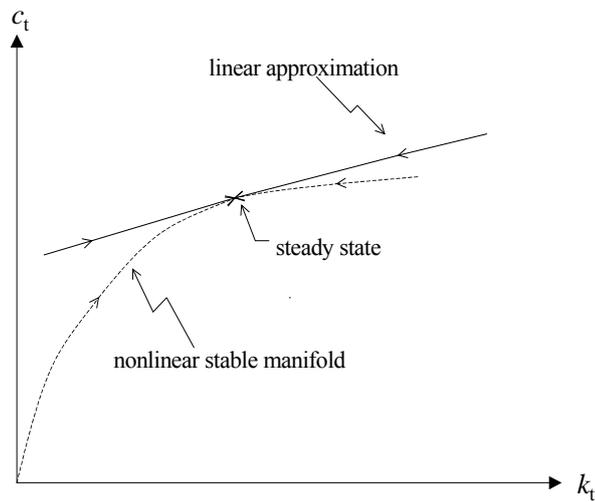


FIGURE 3.3.

3.3.4 Stochastic economies

“Real business cycle theory is the application of general equilibrium theory to the quantitative analysis of business cycle fluctuations.” Edward Prescott (1991, p. 3)

“The Kydland and Prescott model is a complete markets set-up, in which equilibrium and optimal allocations are equivalent. When it was introduced, it seemed to many—myself included—to be much too narrow a framework to be useful in thinking about cyclical issues.” Robert Lucas (1994, p. 184)

In its simplest version, real business cycle theory is an extension of the neoclassical model of Section 3.3.3, in which random productivity shocks are added. The engine of fluctuations, then, comes from the real sphere of the economy. This approach is in contrast with the Lucas approach of the 1970s, based on information and money, where fluctuations arise due to information delays with which agents discover the nature of a shock (real or monetary). As further reviewed in Chapter 9, the Lucas information-theoretic approach has been, instead, more successful in inspiring work on the formation of asset prices, leading to the development of market microstructure theory and, more generally, to information driven explanations of asset prices.

Despite the remarkable switch in the economic motivation, the paradigm underlying real business cycle theory is the same as the information-based approach of Lucas, as it relies on rational expectations: macroeconomic fluctuations and, then, as we shall explain, asset prices fluctuations, stem from the optimal response of the agents vis-à-vis exogenous shocks: agents implement action plans that are state-contingent, i.e. they decide to consume, to work and to invest according to the history of shocks as well as the present shocks they observe.

3.3.4.1 Basic model

We consider an economy with complete markets and no frictions, such that its equilibrium allocations are Pareto-optimal. To characterize these allocations, we implement them through the following program of a social planner:

$$V(k_0, s_0) = \max_{(c_t)_{t=0}^{\infty}} E \left[\sum_{t=0}^{\infty} \beta^t u(c_t) \right], \quad (3.15)$$

subject to a capital accumulation constraint, with capital depreciation. Let I_t denote new investment. It is:

$$I_t = K_{t+1} - (1 - \delta) K_t. \quad (3.16)$$

At time $t - 1$, the available productive capital is K_t . At time t , a portion δK_t of this capital is lost, due to depreciation. Therefore, at time t , the productive system is left with $(1 - \delta) K_t$ units of capital. The capital available at time t , K_{t+1} , equals the capital already in place, $(1 - \delta) K_t$, plus new investments, which is exactly what Eq. (3.16) says.

Next, normalize population normalized to one, such that $K_t = k_t$. The goods market clearing condition is:

$$\tilde{y}(k_t, s_t) = c_t + I_t,$$

where $\tilde{y}(k_t, s_t)$ is the production function, which is \mathcal{F}_t -measurable, and s is the source of randomness—the engine for random fluctuations of the endogenous variables. By replacing Eq. (3.16) into the equilibrium condition,

$$k_{t+1} = \tilde{y}(k_t, s_t) - c_t + (1 - \delta) k_t. \quad (3.17)$$

So the planner maximizes the utility in Eq. (3.15), under the capital accumulation constraint in Eq. (3.17).

We assume that $\tilde{y}(k_t, s_t) \equiv s_t y(k_t)$, where y is as in Section 3.2, and $(s_t)_{t=0}^{\infty}$ is solution to:

$$s_{t+1} = s_t^\rho \epsilon_{t+1}, \quad (3.18)$$

where $\rho \in (0, 1)$, and $(\epsilon_t)_{t=0}^{\infty}$ is a IID sequence with support s.t. $s_t \geq 0$. In this economy, every asset is priced as in the Lucas model of the previous section. Therefore, the gross return on savings $s \cdot y'(k)$ satisfies:

$$u'(c_t) = \beta E_t(u'(c_{t+1})(s_{t+1}y'(k_{t+1}) + 1 - \delta)). \quad (3.19)$$

A rational expectation equilibrium is a stochastic process $(c_t, k_t)_{t=0}^{\infty}$, satisfying Eq. (3.17), the Euler equation in (3.19), for given k_0 and s_0 .

We show the existence of a saddlepoint path for the linearized version of Eqs. (3.17)-(3.18)-(3.19), which implies determinacy of the stochastic (linearized) equilibrium.⁴ We study the behavior of $(c, k, s)_t$ in a neighborhood of $\epsilon \equiv E(\epsilon_t)$. Let (c, k, s) be consumption, capital and productivity shock, corresponding to ϵ , obtained replacing ϵ into Eqs. (3.17)-(3.18)-(3.19), and assuming no uncertainty takes place:

$$c = sy(k) - \delta k, \quad s = \epsilon^{\frac{1}{1-\rho}}, \quad \beta = \frac{1}{sy'(k) + 1 - \delta}.$$

A first-order approximation to Eqs. (3.17)-(3.18)-(3.19) around (k, c, s) , leaves:

$$\hat{z}_{t+1} = \Phi \hat{z}_t + Ru_{t+1}, \quad (3.20)$$

where we have defined $\hat{x}_t \equiv \frac{x_t - x}{x}$, and $\hat{z}_t = (\hat{k}_t, \hat{c}_t, \hat{s}_t)^\top$, $u_t = (u_{c,t}, u_{s,t})^\top$, $u_{c,t} = \hat{c}_t - E_{t-1}(\hat{c}_t)$, $u_{s,t} = \hat{s}_t - E_{t-1}(\hat{s}_t) = \hat{c}_t$, and, finally,

$$\Phi = \begin{pmatrix} \beta^{-1} & & & \\ -\frac{u'(c)}{cu''(c)} s k y''(k) & 1 + \frac{\beta u'(c)}{u''(c)} s y''(k) & -\frac{\beta u'(c)}{cu''(c)} s (s y(k) y''(k) + \rho y'(k)) & \\ 0 & 0 & \rho & \end{pmatrix}, \quad R = \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Let us consider the characteristic equation:

$$0 = \det(\Phi - \lambda I) = (\rho - \lambda) \left[\lambda^2 - \left(\beta^{-1} + 1 + \beta \frac{u'(c)}{u''(c)} s y''(k) \right) \lambda + \beta^{-1} \right].$$

A solution is $\lambda_1 = \rho$. By the same arguments produced for the deterministic case of Section 3.3.3 (see Appendix 1), one finds that $\lambda_2 \in (0, 1)$ and $\lambda_3 > 1$.⁵ As for the deterministic case in Section 3.3.3, we can diagonalize the system by rewriting $\Phi = P\Lambda P^{-1}$, where Λ is a diagonal

⁴A stochastic equilibrium is the situation where there is a stationary measure (definition: $p(+)=\int \pi(+/-)dp(-)$, where π is the transition measure) generating $(c_t, k_t)_{t=1}^{\infty}$.

⁵The linearized model in this section has state variables expressed in growth rates here. However, we can always reformulate this model in terms of first differences, by pre- and post- multiplying Φ by appropriate normalizing matrices. As an example, if G is the 3×3 matrix that has $\frac{1}{k}$, $\frac{1}{c}$ and $\frac{1}{s}$ on its diagonal, (3.20) can be written as: $E(z_{t+1} - z) = G^{-1} \Phi G \cdot (z_t - z)$, where $z_t = (k_t, c_t, s_t)$, and we would arrive at the same conclusions. It is tedious but easy to check that the model in this section collapses to that in Section 3.3.3, once we set $\epsilon_t = 1$, for each t , and $s_0 = 1$.

matrix that has the eigenvalues of Φ on the diagonal, and P is a matrix of the eigenvectors associated to the roots of Φ . The system in (3.20) is, then:

$$\hat{y}_{t+1} = \Lambda \hat{y}_t + w_{t+1}, \quad (3.21)$$

where $\hat{y}_t \equiv P^{-1} \hat{z}_t$ and $w_t \equiv P^{-1} R u_t$. The third equation of this system is:

$$\hat{y}_{3,t+1} = \lambda_3 \hat{y}_{3t} + w_{3,t+1}, \quad (3.22)$$

and \hat{y}_3 explodes unless $\hat{y}_{3t} = 0$ for all t , which is only possible when $w_{3t} = 0$ for all t .⁶

The condition that $\hat{y}_{3t} \equiv 0$ carries an interesting economic interpretation: it tells us that the only sources of uncertainty in this system can stem from shocks to the fundamentals, or that there cannot be extraneous sources of noise, or “sunspots.” The reasons for this are easy to explain. Let $\hat{y}_t = P^{-1} \hat{z}_t \equiv \Pi \hat{z}_t$. We have:

$$0 = \hat{y}_{3t} = \pi_{31} \hat{k}_t + \pi_{32} \hat{c}_t + \pi_{33} \hat{s}_t. \quad (3.23)$$

Eq. (3.23) shows that the three state variables, \hat{k}_t , \hat{c}_t and \hat{s}_t , are mutually linked through a two-dimensional plane. This plane is the saddlepoint of the economy, where the state variables do exhibit a stable behavior, and is formally defined as:

$$\mathcal{S} = \{x \in \mathbb{R}^3 \mid \pi_3 \cdot x = 0\}, \quad \pi_3 = (\pi_{31}, \pi_{32}, \pi_{33}).$$

Furthermore, Eq. (3.23) implies that a linear relation exists between the two expectational errors:

$$\text{For all } t, \quad u_{ct} = -\frac{\pi_{33}}{\pi_{32}} u_{st} \quad (\text{“no-sunspots”}). \quad (3.24)$$

Eq. (3.24) is a “no-sunspots” condition, as it says that the expectational error to consumption cannot be independent of the expectational shock on the fundamentals of the economy, which in this simple economy relates to technological shock. In other words, the source of uncertainty we have assumed in this economy, relates to the technological shock. The remaining expectational errors can only be perfectly correlated to the expectational shock in technology or, there are no sunspots.

The manifold \mathcal{S} brings, mathematically, the same meaning as the stable relation depicted in Figure 3.2, for the deterministic case. In this section, \mathcal{S} is convergent subspace, with $\dim(\mathcal{S}) = 2$, which is the number of roots with modulus less than one. In other words, in this economy with two predetermined variables, \hat{k}_0 and \hat{s}_0 , there exists one, and only one, value of \hat{c}_0 in \mathcal{S} , which ensures stability, and is given by $\hat{c}_0 = -\frac{\pi_{31} \hat{k}_0 + \pi_{33} \hat{s}_0}{\pi_{32}}$. This reasoning generalizes that we made for the deterministic case in Section 3.3.3, and is generalized further in Appendix 1.

The solution to the linearized model can be computed by generalizing the reasoning for the deterministic case. First, by Eq. (3.21) \hat{y} is:

$$\hat{y}_{it} = \lambda_i^t \hat{y}_{i0} + \zeta_{it}, \quad \zeta_{it} \equiv \sum_{j=0}^{t-1} \lambda_i^j w_{i,t-j},$$

⁶In other words, Eq. (3.22) implies that $\hat{y}_{3t} = \lambda_3^{-(T-t)} E_t(\hat{y}_{3,t+T})$, and for all T . Because $\lambda_3 > 1$, this relation holds only when $\hat{y}_{3t} = 0$ for all t .

which implies the solution for \hat{z} is:

$$\hat{z}_t = P\hat{y}_t = (v_1 \ v_2 \ v_3)\hat{y}_t = \sum_{i=1}^3 v_i \hat{y}_{it} = \sum_{i=1}^3 v_i \hat{y}_{i0} \lambda_i^t + \sum_{i=1}^3 v_i \zeta_{it}.$$

To pin down the components of \hat{y}_0 , note that $\hat{z}_0 = P\hat{y}_0 \Rightarrow \hat{y}_0 = P^{-1}\hat{z}_0 \equiv \Pi\hat{z}_0$. The stability condition then requires that the state variables be in \mathcal{S} , or $\hat{y}_0^{(3)} = 0$, which we now use to implement the solution. We have:

$$\hat{z}_t = v_1 \lambda_1^t \hat{y}_{10} + v_2 \lambda_2^t \hat{y}_{20} + v_3 \lambda_3^t \hat{y}_{30} + v_1 \zeta_{1t} + v_2 \zeta_{2t} + v_3 \zeta_{3t}.$$

Moreover, the term $v_3 \lambda_3^t \hat{y}_{30} + v_3 \zeta_{3t}$ needs to be zero, because $\hat{y}_{30} = 0$. Finally, we have that $\zeta_{3t} = \sum_{j=0}^{t-1} \lambda_3^j w_{3,t-j}$, and since $w_{3,t} = 0$, then, then $\zeta_{3t} = 0$ as well. Therefore, the solution for \hat{z}_t is:

$$\hat{z}_t = v_1 \lambda_1^t \hat{y}_{10} + v_2 \lambda_2^t \hat{y}_{20} + v_1 \zeta_{1t} + v_2 \zeta_{2t}.$$

3.3.4.2 Frictions, indeterminacy and sunspots

In the neoclassical model that we are analyzing, the equilibrium is determinate. As explained, this property arises because the number of predetermined variables equals the dimension of the convergent subspace of the economy. If we managed to increase the dimension of the converging subspace, the equilibrium would be indeterminate, as further formalized in Appendix 1. As it turns out, indeterminacy goes hand in hand with sunspots, the expectational shocks extraneous to those in the economic fundamentals, as we discussed earlier, just after Eq. (3.24).

Introducing sunspots in macroeconomics has been an approach pursued in detail by Farmer in a series of articles (see Farmer, 1998, for an introductory account of this approach). The idea is quite interesting, as we know that the basic real business cycle model of this section needs many extensions in order not to be rejected, empirically, as originally shown by Watson (1993). In other words, the basic model in this section offers little room for a rich propagation mechanism, as it entirely relies on impulses, the productivity shocks, which “we hardly read about in the Wall Street Journal,” as provocatively put by King and Rebelo (1999). Sunspots offer an interesting route to enrich the propagation mechanism, although their asset pricing implications in terms of the model analyzed in this section, have not been explored yet.

In a series of articles, David Cass showed that a Pareto-optimal economy cannot harbour sunspots equilibria. On the other hand, any market imperfection has the potential to be a source of sunspots. The typical example is the presence of incomplete markets. The neoclassical model analyzed in this section cannot generate sunspots, as it relies on a system of perfectly competitive markets and absence of any sort of frictions. To introduce sunspots in the economy of this section, we need to think about some deviation from optimality. Two possibilities analyzed in the literature are the presence of imperfect competition and/or externality effects. We provide an example of these effects, by working out the deterministic economy in Section 3.3.3. (Generalizations to the stochastic economy in this section are easy, although more cumbersome.)

How is it that a deterministic economy might generate “stochastic outcomes,” that is, outcomes driven by shocks entirely unrelated to the fundamentals of the economy? Let us imagine this can be possible. Then, both optimal consumption and capital accumulation in Section 3.3.3 are necessarily random processes. The system in (3.13), then, must be rewritten in an expectation format,

$$E_t \begin{pmatrix} \hat{k}_{t+1} \\ \hat{c}_{t+1} \end{pmatrix} = A \begin{pmatrix} \hat{k}_t \\ \hat{c}_t \end{pmatrix}.$$

Next, let us introduce the expectational error process $u_{c,t} \equiv \hat{c}_t - E_{t-1}(\hat{c}_t)$, which we plug back into the previous system, to obtain:

$$\begin{pmatrix} \hat{k}_{t+1} \\ \hat{c}_{t+1} \end{pmatrix} = A \begin{pmatrix} \hat{k}_t \\ \hat{c}_t \end{pmatrix} + \begin{pmatrix} 0 \\ u_{c,t+1} \end{pmatrix}.$$

Naturally, we still have $\lambda_1 \in (0, 1)$ and $\lambda_2 > 1$, as in Section 3.3.3. Therefore, we decompose A as $P\Lambda P^{-1}$, and have:

$$\hat{y}_{t+1} = \Lambda \hat{y}_t + P^{-1} \begin{pmatrix} 0 \\ u_{c,t+1} \end{pmatrix}^\top.$$

Moreover, for $\hat{y}_{2t} = \lambda_2^{-T} E_t(\hat{y}_{2,t+T})$ to hold for all T , we need to have $\hat{y}_{2t} = 0$, for all t . Therefore, the second element of the vector $P^{-1} \begin{pmatrix} 0 \\ u_{c,t+1} \end{pmatrix}^\top$ must be zero, or, for all t ,

$$0 = \pi_{22} u_{c,t} \iff 0 = u_{c,t}.$$

There is no room for expectational errors and, hence, sunspots, in this model. The fact that $\lambda_2 > 1$ implies the dimension of the saddlepoint is less than the number of predetermined variables. So a viable route to pursue here, is to look for economies such that the saddlepoint has a dimension larger than one, i.e. such that $\lambda_2 < 1$. In these economies, indeterminacy and sunspots will be two facets of the same coin. As shown in the appendix, the reasons for which $\lambda_2 > 1$ relate to the classical assumptions about the shape of the utility function u and the production function y . We now modify the production function, to see the effect on the eigenvalues of A .

[Economy with increasing returns]

[Asset pricing implications in further chapters]

3.4 Production-based asset pricing

3.4.1 Firms

For each firm, capital accumulation does satisfy the identity in Eq. (3.16), reproduced here for convenience:

$$K_{t+1} = (1 - \delta) K_t + I_t. \quad (3.25)$$

The additional assumption we make, is that capital adjustment is costly: investing I_t per unit of capital already in place, K_t , entails a cost $\phi(\frac{I_t}{K_t})$, expressed in terms of the price of the final good, which we take to be the numéraire, thereby allowing the investment goods to differ from the final good the firm produces. An investment of I_t , then, leads to a cost $\phi(\frac{I_t}{K_t})K_t$, such that the profit the firm makes at time t is,

$$D(K_t, I_t) \equiv \tilde{y}(K_t, N(K_t)) - w_t N(K_t) - p_t I_t - \phi\left(\frac{I_t}{K_t}\right) K_t, \quad (3.26)$$

where $\tilde{y}(K_t, N_t)$ is the firm's production at time t , obtained with capital K_t and labor N_t , and subject to the productivity shocks described in Section 3.3.4, w_t is the real wage, $N(K)$ is the labor demand schedule, solution to the optimality condition, $\tilde{y}_N(K_t, N(K_t)) = w_t$ for all t , and p_t is the real price of the investment goods, or uninstalled capital. Finally, the adjustment-cost function satisfies $\phi \geq 0$, $\phi' \geq 0$, $\phi'' \geq 0$. In words, capital adjustment is costly when the adjustment is made fastly. Naturally, ϕ is zero in the absence of adjustment costs.

What is the *value* of the profit, from the perspective of time zero? This question can be answered, by utilizing the Arrow-Debreu state prices introduced in Chapter 2. At time t , and in state s , the profit $D_t(s)$ (say) is worth,

$$\phi_{0,t}(s) D(K_t(s), I_t(s)) = m_{0,t}(s) D_t(K_t(s), I_t(s)) P_{0,t}(s),$$

with the same notation as in Chapter 2.

3.4.1.1 The value of the firm

We assume that in each period, the firm distributes all the profits it makes, and that for a given capital K_0 , it maximizes its *cum-dividend* value,

$$V_c(K_0) = \max_{(K_t, I_{t-1})_{t=1}^{\infty}} \left[D(K_0, I_0) + E \left(\sum_{t=1}^{\infty} m_{0,t} D(K_t, I_t) \right) \right],$$

subject to the capital accumulation law of Eq. (3.25).

The value of the firm at time t , $V_c(K_t)$, can be found recursively, through the Bellman's equation,

$$V_c(K_t) = \max_{I_t} [D(K_t, I_t) + E_t(m_{t+1} V_c(K_{t+1}))],$$

where the expectation is taken with respect to the information set as of time t . The first-order conditions for I_t lead to,

$$-D_I(K_t, I_t) = E_t[m_{t+1} V'_c(K_{t+1})]. \quad (3.27)$$

That is, along the optimal capital accumulation path, the marginal cost of new installed capital at time t , $-D_I$, must equal the expected marginal return on the investment, i.e. the expected value of the marginal contribution of capital to the value of the firm at time $t+1$, $V'_c(K_{t+1})$.

By Eq. (3.27), optimal investment is a function $I(K_t)$, and the value of the firm satisfies,

$$V_c(K_t) = D(K_t, I(K_t)) + E_t[m_{t+1} V_c(K_{t+1})].$$

Differentiating the value function in the previous equation, with respect to K_t , and using Eq. (3.27), yields the following envelope condition:

$$\begin{aligned} V'_c(K_t) &= D_K(K_t, I(K_t)) + D_I(K_t, I(K_t)) I'(K_t) + E_t[m_{t+1} V'_c(K_{t+1}) ((1-\delta) + I'(K_t))] \\ &= D_K(K_t, I(K_t)) - (1-\delta) D_I(K_t, I(K_t)). \end{aligned}$$

By replacing this expression for the value function back into Eq. (3.27), leaves:

$$-D_I(K_t, I(K_t)) = E_t[m_{t+1} (D_K(K_{t+1}, I(K_{t+1})) - (1-\delta) D_I(K_{t+1}, I(K_{t+1})))]. \quad (3.28)$$

Along the optimal capital accumulation path, the marginal cost of new installed capital at time t , which by Eq. (3.27) is the expected marginal return on the investment, equals the expected value of (i) the very same marginal cost at time $t+1$, corrected for capital depreciation, $(1-\delta)$, and (ii) capital productivity, net of adjustment costs. Analytically,

$$\begin{aligned} D_K(K_t, I_t) &= \tilde{y}_K(K_t, N(K_t)) + \tilde{y}_N(K_t, N(K_t)) N'(K_t) - w_t N'(K_t) - \frac{\partial}{\partial K} \left(\phi \left(\frac{I_t}{K_t} \right) K_t \right) \\ &= \tilde{y}_K(K_t, N(K_t)) - \frac{\partial}{\partial K} \left(\phi \left(\frac{I_t}{K_t} \right) K_t \right), \\ -D_I(K_t, I(K_t)) &= p_t + \phi' \left(\frac{I_t}{K_t} \right). \end{aligned}$$

We now introduce a fundamental concept in investment theory.

3.4.1.2 q theory

The Tobin's *marginal* q is defined as the ratio of the expected marginal value of an additional unit of capital over its replacement cost:

$$\text{TQ}_t \equiv \text{Tobin's marginal q} \equiv \frac{E[m_{t+1}V'_c(K_{t+1})]}{p_t}.$$

We show that the numerator, $E[m_{t+1}V'_c(K_{t+1})]$, is, simply, the shadow price of installed capital. Consider the Lagrangian at time t ,

$$\mathcal{L}(K_t) = \max_{I_t, K_{t+1}, q_t} [D(K_t, I_t) - q_t(K_{t+1} - (1 - \delta)K_t - I_t) + E_t(m_{t+1}\mathcal{L}(K_{t+1}))], \quad (3.29)$$

which, integrated, gives rise to the value of the firm:

$$\mathcal{L}(K_0) = \max_{(I_t, K_{t+1}, q_t)_{t=0}^{\infty}} E \left[\sum_{t=0}^{\infty} m_{0,t} (D(K_t, I_t) - q_t(K_{t+1} - (1 - \delta)K_t - I_t)) \right].$$

The first-order condition for investment, I_t , is, $q_t = -D_I(K_t, I_t)$, and that for capital, K_{t+1} , is $q_t = E(m_{t+1}\mathcal{L}'(K_{t+1}))$. By Eq. (3.27), then, $\mathcal{L}'(K_t) = V'_c(K_{t+1})$ and, therefore, q_t is the expected marginal return on the investment, that is, the shadow price of installed capital. Therefore, Tobin's marginal q is the ratio of the shadow price of installed capital to its replacement cost:

$$\text{TQ}_t = \frac{q_t}{p_t}.$$

Next, replace the first-order condition for q_t , i.e. Eq. (3.25), into Eq. (3.29), differentiate $\mathcal{L}(K_t)$ with respect to K_t , and use the first-order condition for K_{t+1} , obtaining, $\mathcal{L}'(K_t) = D_K(K_t, I_t) + q_t(1 - \delta)$. These conditions imply that q_t satisfies the valuation equation (3.28):

$$q_t = E_t[m_{t+1}(D_K(K_{t+1}, I_{t+1}) + (1 - \delta)q_{t+1})], \quad (3.30)$$

and therefore that:

$$q_t = p_t + \phi' \left(\frac{I_t}{K_t} \right). \quad (3.31)$$

The shadow price of installed capital, q_t , has to equal the marginal cost of new installed capital, and is larger than the price of uninstalled capital, p_t . It is natural: to install new capital requires some (marginal) adjustment costs, which add to the "row" price of uninstalled capital, p_t . Therefore, in the presence of adjustment costs, Tobin's marginal q is larger than one.

Eq. (3.30) can be solved forward, leaving:

$$q_t = E \left[\sum_{s=1}^{\infty} (1 - \delta)^{s-1} m_{0,t+s} D_K(K_{t+s}, I_{t+s}) \right].$$

The shadow price of installed capital is worth the sum of all its future marginal net productivity, discounted at the depreciation rate. Moreover, Eq. (3.31) can be inverted for I_t/K_t , to deliver:

$$\frac{I_t}{K_t} = \phi'^{-1}(q_t - p_t), \quad (3.32)$$

where ϕ'^{-1} denotes the inverse of ϕ' , and is increasing, since ϕ' is increasing. Given K_t , and the fact that K_{t+1} is predetermined, the firm evaluates q_t through Eq. (3.30), and then determines the level of new investments through Eq. (3.32). These investments are increasing in the difference between the shadow price of installed capital, q_t , and that of uninstalled capital, p_t , as originally assumed by Tobin (1969).

In the absence of adjustment costs, when $q_t = p_t$, Eq. (3.30) delivers the condition,

$$1 = E_t [m_{t+1} (\tilde{y}_K (K_{t+1}, N (K_{t+1})) + (1 - \delta))],$$

where we have set $p_t \equiv 1$ for all t , meaning that the firm's production is just the uninstalled capital. Empirically, however, the marginal productivity of capital, $\tilde{y}_K (K_t, N (K_t))$, is not volatile enough, to rationalize asset returns, as explained in more detail in Chapter 8. Moreover, as we argue in a moment, Tobin's marginal q can be approximated by market-to-book ratios, which are typically time-varying. Therefore, adjustment costs are important for asset pricing.

A difficulty with Tobin's marginal q is that it is quite difficult to estimate. Yet in the special case we are analyzing in this section, where firms act competitively and have access to an homogeneous production function and adjustment costs, Tobin's marginal q can be proxied by the market-to-book ratio of a given firm. Let $V (K_t)$ denote the ex-dividend value of the firm, which is its stock market value, since it nets out the dividend it pays to its holder in the current period. It is:

$$V (K_t) \equiv V_c (K_t) - D (K_t, I (K_t)) = E_t [m_{t+1} V_c (K_{t+1})].$$

The Tobin's *average* q is defined as the ratio of the stock market value of the firm over the replacement cost of the capital:

$$\text{Tobin's average q} \equiv \frac{\text{Stock Mkt Value of the Firm}}{\text{Replacement Cost of Capital}} = \frac{V (K_t)}{p_t K_{t+1}}.$$

The next result was originally obtained by Hayashi (1982) in a continuous-time setting.

THEOREM 3.2. *Tobin's marginal q and average q coincide. That is, we have,*

$$V (K_t) = q_t K_{t+1}.$$

PROOF. By the homogeneity properties of the production function and the adjustment costs,

$$D (K_t, I_t) = D_K (K_t, I_t) K_t + D_I (K_t, I_t) I_t.$$

Therefore, the ex-dividend value of the firm is:

$$\begin{aligned} V (K_0, I_0) &= E \left[\sum_{t=1}^{\infty} m_{0,t} D (K_t, I_t) \right] \\ &= E \left[\sum_{t=1}^{\infty} m_{0,t} (D_K (K_t, I_t) - (1 - \delta) D_I (K_t, I_t)) K_t \right] + E \left[\sum_{t=1}^{\infty} m_{0,t} D_I (K_t, I_t) K_{t+1} \right], \end{aligned}$$

where the second line follows by Eq. (3.25). By Eq. (3.28), and the law of iterated expectations,

$$E \left[\sum_{t=1}^{\infty} m_{0,t} (D_K (K_t, I_t) - (1 - \delta) D_I (K_t, I_t)) K_t \right] = -D_I (K_0, I_0) K_1 - E \left[\sum_{t=1}^{\infty} m_{0,t} K_{t+1} D_I (K_t, I_t) \right].$$

Hence, $V(K_0, I_0) = -D_I(K_0, I_0) K_1 = q_0 K_1$. \parallel

This result, in conjunction with that in Eq. (3.31), provides a simple rule of thumb for investment decisions. Consider, for example, the case of quadratic adjustment costs, where $\phi(x) = \frac{1}{2}\kappa^{-1}x^2$, for some $\kappa > 0$. Then, Eq. (3.32) is:

$$I_t = \kappa (q_t - p_t) K_t = \kappa \left(\frac{\text{Stock Mkt Value of the Firm}}{\text{Replacement Cost of Capital}} - 1 \right) p_t K_t,$$

where the second equality follows by Theorem 3.2. Thus, according to q theory, we expect firms with a market value larger than the cost of reproducing their capital to grow, and firms which are not worth the cost of reproducing their capital to shrink. This basic observation constitutes a first assessment that we can use to assess developments of firms.

3.4.2 Consumers

We now generalize the budget constraint obtained in the program [3.P3], to the uncertainty case. We claim that in this case, the relevant budget constraint is,

$$V_0 = c_0 + E \left[\sum_{t=1}^{\infty} m_{0,t} (c_t - w_t N_t) \right]. \quad (3.33)$$

We have: $c_t + S_t \theta_{t+1} = (S_t + D_t) \theta_t + w_t N_t$ and, then:

$$\begin{aligned} E \left[\sum_{t=1}^{\infty} m_{0,t} (c_t - w_t N_t) \right] &= E \left[\sum_{t=1}^{\infty} m_{0,t} (S_t + D_t) \theta_t \right] - E \left[\sum_{t=1}^{\infty} m_{0,t} S_t \theta_{t+1} \right] \\ &= E \left[\sum_{t=1}^{\infty} E_{t-1} \left(\frac{m_{0,t}}{m_{0,t-1}} m_{0,t-1,t} (S_t + D_t) \theta_t \right) \right] - E \left[\sum_{t=2}^{\infty} m_{0,t-1} S_{t-1} \theta_t \right] \\ &= E \left[\sum_{t=1}^{\infty} m_{0,t-1} S_{t-1} \theta_t \right] - E \left[\sum_{t=2}^{\infty} m_{0,t-1} S_{t-1} \theta_t \right] \\ &= S_0 \theta_1 = V_0 - c_0. \end{aligned}$$

where the third line follows by the properties of the discount factor, $\frac{m_{0,t}}{m_{0,t-1}} = m_{0,t-1}$ and $m_t \equiv m_{t-1,t}$.

Therefore, the program consumers solve is:

$$\max_{(c_t)_{t=0}^{\infty}} E \left[\sum_{t=1}^{\infty} \beta^t u(c_t) \right], \quad \text{s.t. Eq. (3.33)}.$$

We now have two optimality conditions, one intertemporal and another, intratemporal:

$$m_{t+1} = \beta \frac{u_1(c_{t+1}, N_{t+1})}{u_1(c_t, N_t)} \quad (\text{intertemporal}); \quad w_t = -\frac{u_2(c_t, N_t)}{u_1(c_t, N_t)} \quad (\text{intra-temporal}).$$

3.4.3 Equilibrium

For all t ,

$$\tilde{y}(K_t, N_t) = c_t + p_t I_t + \phi \left(\frac{I_t}{K_t} \right) K_t. \quad (3.34)$$

It is easily seen that the condition $\theta_t = 1$ in the financial market, implies that $c_t = D_t + w_t N_t$, which, upon substitution of the profits in Eq. (3.26), delivers the equilibrium condition in Eq. (3.34). Implicit in this reasoning, is the idea the adjustment costs are not paid to anyone. They represent, so to speak, capital losses incurred along the way of growth.

3.5 Money, production and asset prices in overlapping generations models

3.5.1 Introduction: endowment economies

3.5.1.1 A deterministic model

We initially assume the population is constant, and made up of one young and one old. The young agent maximizes his intertemporal utility subject to his budget constraint:

$$\max_{(c_{1t}, c_{2,t+1})} [u(c_{1t}) + \beta u(c_{2,t+1})] \quad \text{subject to} \quad \begin{cases} \text{sav}_t + c_{1t} = w_{1t} \\ c_{2,t+1} = \text{sav}_t R_{t+1} + w_{2,t+1} \end{cases} \quad [3.P5]$$

where w_{1t} and $w_{2,t+1}$ are the endowments the agent receives at his young and old age.

The agent born at time $t - 1$, then, faces the constraints: $\text{sav}_{t-1} + c_{1,t-1} = w_{1,t-1}$ and $c_{2t} = \text{sav}_{t-1} R_t + w_{2t}$. By combining his second period constraint with the first period constraint of the agent born at time t ,

$$\text{sav}_{t-1} R_t + w_t = \text{sav}_t + c_{1t} + c_{2t}, \quad w_t \equiv w_{1t} + w_{2t}. \quad (3.35)$$

The equilibrium in the intergenerational lending market is, naturally:

$$\text{sav}_t = 0, \quad (3.36)$$

and implies that the goods market is also in equilibrium, in that $w_t = \sum_{i=1}^2 c_{i,t}$, and for all t . Therefore, we can analyze the model, by just analyzing the autarkic equilibrium.

As Figure 3.4 illustrates, the first-order condition for the program [3.P5] requires that the slope of the indifference curve be equal to the slope of the lifetime budget constraint, $c_{2,t+1} = -R_{t+1} c_{1,t} + R_{t+1} w_{1t} + w_{2,t+1}$, and leads to:

$$\beta \frac{u'(c_{2,t+1})}{u'(c_{1,t})} = \frac{1}{R_{t+1}}. \quad (3.37)$$

The equilibrium, then, is a sequence of gross returns R_t satisfying Eqs. (3.35), (3.36) and (3.37), or:

$$b_t \equiv \frac{1}{R_{t+1}} = \beta \frac{u'(w_{2,t+1})}{u'(w_{1t})}. \quad (3.38)$$

In this relation, b_t is the shadow price of a bond issued at t , and promising one unit of numéraire at $t + 1$: the sequence of prices, b_t , satisfying Eq. (3.38), is such that agents are happy with not being able to lend and borrow, intergenerationally.

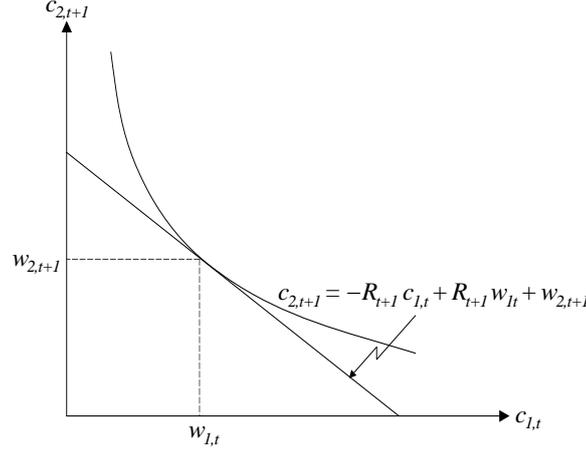


FIGURE 3.4.

The previous model is easy to extend to the case where agents are heterogeneous. The program each agent j solves is, now:

$$\max_{(c_{1j,t}, c_{2j,t+1})} [u_j(c_{1j,t}) + \beta_j u_j(c_{2j,t+1})] \quad \text{subject to} \quad \begin{cases} \text{sav}_{j,t} + c_{1j,t} = w_{1j,t} \\ c_{2j,t+1} = \text{sav}_{j,t} R_{t+1} + w_{2j,t+1} \end{cases}$$

with obvious notation. The first-order condition is, for all time t and agent j ,

$$\beta_j \frac{u'_j(c_{2j,t+1})}{u'_j(c_{1j,t})} = \frac{1}{R_{t+1}} \equiv b_t,$$

and the equilibrium is a sequence of bond prices b_t satisfying the previous relation and the equilibrium in the intrageneration lending market:

$$\sum_{j=1}^J \text{sav}_{j,t} = 0, \quad (3.39)$$

where J denotes the constant number of agents in each generation.

To illustrate, suppose agents have all the same utility, of the CRRA class, with CRRA coefficient equal to η , and the same discount rate, $\beta_j = \beta$. In this case,

$$\text{sav}_{j,t} = \frac{(\beta R_{t+1})^{\frac{1}{\eta}} w_{1t} - w_{2,t+1}}{R_{t+1} + (\beta R_{t+1})^{\frac{1}{\eta}}}.$$

The first term in the numerator reflects an income effect, while the second is a substitution effect. The coefficient $\frac{1}{\eta}$ is the elasticity of intertemporal substitution, as explained in Section 3.2.3. Consider, for example, the logarithmic case, where $\eta = 1$, and:

$$c_{1j,t} = \frac{1}{1+\beta} \left(w_{1j,t} + \frac{w_{2j,t+1}}{R_{t+1}} \right), \quad c_{2j,t+1} = \frac{\beta}{1+\beta} (R_{t+1} w_{1j,t} + w_{2j,t+1}), \quad \text{sav}_{j,t} = \frac{1}{1+\beta} \left(\beta w_{1j,t} - \frac{w_{2j,t+1}}{R_{t+1}} \right), \quad (3.40)$$

and using the equilibrium condition in Eq. (3.39),

$$b_t = \frac{1}{R_{t+1}} = \frac{\beta \sum_{j=1}^J w_{1j,t}}{\sum_{j=1}^J w_{2j,t+1}}. \quad (3.41)$$

3.5.1.2 A tree in a stochastic economy

Suppose, next, that we introduce a tree, which yields a stochastic dividend D_t in each period. Each agent solves the following program:

$$\max_{(c_{1t}, c_{2,t+1})} [u(c_{1t}) + \beta E(u(c_{2,t+1}) | \mathcal{F}_t)] \quad \text{subject to} \quad \begin{cases} S_t \theta_t + c_{1t} = w_{1t} \\ c_{2,t+1} = (S_{t+1} + D_{t+1}) \theta_t + w_{2,t+1} \end{cases} \quad [3.P6]$$

where S_t denotes the asset price and θ the units of the asset the agent chooses in his young age. The agent born at time $t - 1$ faces the constraints $S_{t-1} \theta_{t-1} + c_{1,t-1} = w_{1,t-1}$ and $w_{2t} + (S_t + D_t) \theta_{t-1} = c_{2,t}$. By combining the second period constraint of the agent born at time $t - 1$ with the first period constraint of the agent born at time t ,

$$(S_t + D_t) \theta_{t-1} - S_t \theta_t + w_t = c_{1,t} + c_{2,t}.$$

The clearing condition in the asset market, $\theta_t = 1$, implies that the market for goods also clears, for all t : $D_t + w_{1t} + w_{2t} = c_{1,t} + c_{2,t}$. A characterization of the solution to the program [3.P6] can be obtained by eliminating c from the constraint,

$$\max_{\theta} [u(w_{1t} - S_t \theta) + \beta E(u((S_{t+1} + D_{t+1}) \theta) | \mathcal{F}_t)].$$

The equilibrium is one where $\theta_t = 1$, implying that (i) $c_{1t} = w_{1t} - S_t$ and (ii) $c_{2,t+1} = S_{t+1} + D_{t+1} + w_{2,t+1}$. Using (i) and (ii), the first-order condition for the program [3.P6] leads to:

$$u'(w_{1t} - S_t) S_t = \beta E[u'(S_{t+1} + D_{t+1} + w_{2,t+1}) (S_{t+1} + D_{t+1}) | \mathcal{F}_t].$$

Consider, for example, the case where $u(c) = \ln c$, and set $\tilde{R}_{t+1} = (S_{t+1} + D_{t+1}) / S_t$. We have:

$$\frac{1}{w_{1t} - \text{sav}_t^*} = \beta E \left[\frac{1}{\text{sav}_t^* \tilde{R}_{t+1} + w_{2,t+1}} \tilde{R}_{t+1} \middle| \mathcal{F}_t \right], \quad \text{where } \text{sav}_t^* \equiv S_t \theta_t, \quad \theta_t = 1. \quad (3.42)$$

In a deterministic setting,

$$\frac{1}{w_{1t} - \text{sav}_t} = \beta \frac{1}{\text{sav}_t R_{t+1} + w_{2,t+1}} R_{t+1}, \quad \text{where } \text{sav}_t = 0, \quad (3.43)$$

which leads to the equilibrium bond price in Eq. (3.41). Eqs. (3.42) and (3.43) are formally equivalent. Their fundamental difference is that in the tree economy, savings have to stay positive, as the tree must be held by the young agent, in equilibrium: $\text{sav}_t^* \equiv S_t \geq 0$. In an economy without a tree, instead, the interest rate, R_t , has to be such that savings are zero for all t , $\text{sav}_t = 0$.

Eq. (3.42) can be solved explicitly for the price of the tree, S_t , once we assume $w_{2t} = 0$ for all t . In the absence of a tree, we cannot assume endowments are zero in the old age, since the autarkic economy in this case would be such that the old generation would not consume anything. In the presence of a tree, instead, this assumption is innocuous, conceptually, as the autarkic equilibrium in this case is such that the old generation could consume the fruits of the tree, as well as the proceedings arising from selling the tree to the young generation. Solving Eq. (3.42) for S_t when $w_{2t} = 0$, then, leads to a price for the tree, equal to:

$$S_t = \frac{\beta}{1 + \beta} w_{1t}.$$

3.5.2 *Diamond's model*

$$K_{t+1} = N_t S_t, \quad S_t = S(r(K_t), w(K_t)).$$

[Bubbles]

3.5.3 *Money*

We consider a version of the previous model with endowment (not with capital), and assume that agents can now transfer value through a piece of paper, interpreted as money. The young agent, then, maximizes his intertemporal utility, subject to a new budget constraint:

$$\max_{(c_{1t}, c_{2,t+1})} [u(c_{1t}) + \beta u(c_{2,t+1})] \quad \text{subject to} \quad \begin{cases} \frac{m_t}{p_t} + c_{1t} = w_{1t} \\ c_{2,t+1} = \frac{m_t}{p_{t+1}} + w_{2,t+1} \end{cases} \quad [3.P7]$$

where m_t is the amount of money he holds at time t , and p_t is the price of the consumption good as of time t .

Let

$$\text{sav}_t \equiv \frac{m_t}{p_t}, \quad R_{t+1} \equiv \frac{p_t}{p_{t+1}}. \quad (3.44)$$

Then, the budget constraint for program [3.P7] is *formally* identical to that for program [3.P5]. The difference is that in the monetary economy of this section, the young agent may wish to transfer value over time, by saving money, earning a gross “interest rate” equal to the rate of deflation: the lower the price level the next period, the higher the purchasing power of the money he transfers from the young to the old age. Naturally, then, by aggregating the budget constraints of the young and the old generation, we obtain, formally, Eq. (3.35), where now, sav_t and R_{t+1} are as in (3.44). However, in the setting of this section, sav_t is not necessarily zero, as money can be transferred from a generation to another one. In equilibrium, $\text{sav}_t = \frac{\bar{m}_t}{p_t}$, where \bar{m}_t denotes money supply. Therefore, the real value of money is strictly positive, if the equilibrium price p_t stays bounded over time, which might actually occur, as we shall study below. As we see, the role of money as a medium for transferring value, is, in this context, similar to that of a tree in the stochastic overlapping generations economy of Section 3.5.1.2.

Substituting the equilibrium savings $\text{sav}_t = \frac{\bar{m}_t}{p_t}$ and $R_{t+1} = \frac{p_t}{p_{t+1}}$ into Eq. (3.35), we obtain, $\bar{m}_{t-1} = \bar{m}_t + p_t(c_{1t} + c_{2t} - w_t)$, which used again in Eq. (3.35), delivers,

$$\text{sav}_{t-1} R_t = \text{sav}_t - \frac{\Delta \bar{m}_t}{p_t}. \quad (3.45)$$

We need a law of movement for money creation. We assume that:⁷ $\frac{\Delta \bar{m}_t}{\bar{m}_{t-1}} = \mu_t$, for some bounded sequence μ_t . Replacing this into Eq. (3.45), leaves:

$$(1 + \mu_t) \text{sav}_{t-1} R_t = \text{sav}_t. \quad (3.46)$$

The last relation can be obtained even more simply, noting that by definition, $(1 + \mu_t) \frac{\bar{m}_{t-1}}{p_{t-1}} \frac{p_{t-1}}{p_t} = \frac{\bar{m}_t}{p_t}$. The previous relation can be generalized when population grows. Suppose that at time t ,

⁷In this section, we assume that money transfers are made to the young generation: the money the young generation has to absorb is that from the old generation, \bar{m}_{t-1} , and that created by the “central bank,” $\mu_t \bar{m}_{t-1}$. One might consider an alternative model in which transfers are made to old.

N_t individuals are born, and that $\frac{N_t}{N_{t-1}} = (1+n)$, for some constant n . Let money supply be given by $M_t \equiv N_t \bar{m}_t$, and assume that for all t , $\frac{\Delta M_t}{M_{t-1}} = \mu_t$. Then, by a reasoning similar to that leading to Eq. (3.46),

$$\frac{1 + \mu_t}{1 + n} \text{sav}(R_t) R_t = \text{sav}(R_{t+1}), \quad (3.47)$$

where now, we have set the real savings equal to a function of the interest rate, $\text{sav}_{t-1} \equiv \text{sav}(R_t)$, as it should be, by the solution to the program [3.P7].

Next, suppose that μ_t is independent of R , and that $\lim_{t \rightarrow \infty} \mu_t = \mu$, say, a constant. Eq. (3.47) leads to two stationary equilibria:

- (a) $R = \frac{1+n}{1+\mu}$. This stationary equilibrium relates to the ‘‘Golden Rule,’’ once we set $\mu = 0$, as we shall say in Section 3.6.2. For $\mu \neq 0$, the price is, in this stationary equilibrium, $p_t = \left(\frac{1+\mu}{1+n}\right)^t p_0$. Then, we have: (i) $\frac{\bar{m}_t}{p_t} = \frac{M_t}{N_t p_t} = \frac{M_0}{N_0 p_0}$, and (ii) $\frac{\bar{m}_t}{p_{t+1}} = \frac{M_0}{N_0 p_0} \frac{1+n}{1+\mu}$. All in all, the agents’ budget constraints are bounded and the real value of money is strictly positive. In this stationary equilibrium, agents ‘‘trust’’ money.
- (b) $R_a : \text{sav}(R_a) = 0$. This stationary equilibrium relates to an autarkic state. Generally, we have that $R_a < R$: prices increase more rapidly than per-capita money stocks. Analytically, $\text{sav}(R_a) = 0$ implies that $\lim_{t \rightarrow \infty} \frac{\bar{m}_t}{p_t} \rightarrow 0$, which, in turn, implies that for large t , $\frac{\bar{m}_{t+1}}{p_{t+1}} < \frac{\bar{m}_t}{p_t} \iff \frac{\bar{m}_{t+1}}{\bar{m}_t} = \frac{M_{t+1}/M_t}{N_{t+1}/N_t} = \frac{1+\mu}{1+n} < \frac{p_{t+1}}{p_t} \iff R_a < R$. As for $\frac{\bar{m}_t}{p_{t+1}}$, we have that $\frac{\bar{m}_t}{p_{t+1}} = \frac{\bar{m}_t}{p_t} R_a < \frac{\bar{m}_t}{p_t} R = \frac{\bar{m}_t}{p_t} \frac{1+n}{1+\mu}$, and since $\lim_{t \rightarrow \infty} \frac{\bar{m}_t}{p_t} \rightarrow 0$, then $\lim_{t \rightarrow \infty} \frac{\bar{m}_t}{p_{t+1}} \rightarrow 0$. In this stationary equilibrium, agents do not ‘‘trust’’ money.

If $\text{sav}(\cdot)$ is differentiable and $\text{sav}'(\cdot) \neq 0$, the dynamics of $(R_t)_{t=0}^\infty$ can be analyzed through the slope,

$$\frac{dR_{t+1}}{dR_t} = \frac{\text{sav}'(R_t)R_t + \text{sav}(R_t)}{\text{sav}'(R_{t+1})} \frac{1 + \mu_t}{1 + n}. \quad (3.48)$$

There are three cases:

- (i) $\text{sav}'(R) > 0$. Gross substitutability: the substitution effect dominates the income effect.
- (ii) $\text{sav}'(R) = 0$. Income and substitution effects compensate with each other.
- (iii) $\text{sav}'(R) < 0$. Complementarity: the income effect dominates the substitution effect.

The introductory example of this section leads to an instance of gross substitutability (see Eq. (3.40)). Note that an equilibrium cannot exist in that economy, once we assume agents do not have endowments in the second period, $w_{2,t+1} = 0$, as in this case, savings would be strictly positive, such that the equilibrium condition in Eq. (3.39) would not hold. These issues do not arise in the monetary setting of this section, where savings have to be positive and equal to $\frac{\bar{m}_t}{p_t}$, in order to sustain a monetary equilibrium. Assume, for example, the Cobb-Douglas utility function, $u(c_{1t}, c_{2,t+1}) = c_{1t}^{l_1} \cdot c_{2,t+1}^{l_2}$, which leads to a real saving function equal to $\text{sav}(R_{t+1}) = \frac{\frac{l_2}{l_1} w_{1t} - \frac{w_{2,t+1}}{R_{t+1}}}{1 + \frac{l_2}{l_1}}$. If $w_{2,t+1} = 0$, then, $\text{sav}(R_{t+1}) = \frac{\bar{m}_t}{p_t} = \frac{1}{\nu} w_{1t}$, $\nu \equiv \frac{l_1 + l_2}{l_2}$ and, by reorganizing,

$$\bar{m}_t \nu = p_t w_{1t},$$

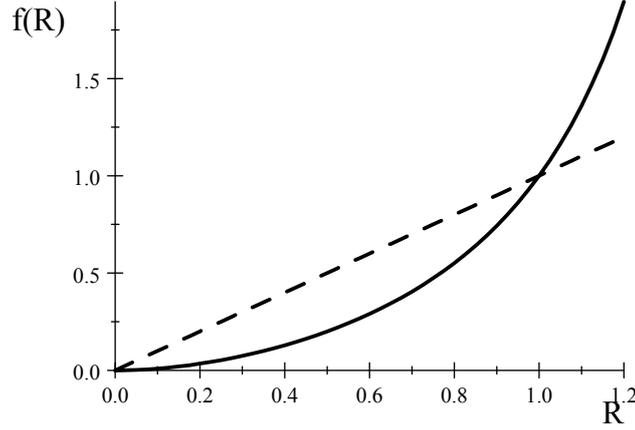


FIGURE 3.5. $f(R) = (R^{-\eta} + R^{-1} - 1)^{-1}$, with $\eta = 2$.

an equation supporting the Quantitative Theory of money. In this economy, the sequence of gross returns satisfies, $R_{t+1} = \frac{p_t}{p_{t+1}} = \frac{\bar{m}_t}{\bar{m}_{t+1}} \frac{w_{1,t+1}}{w_{1,t}}$, or

$$R_{t+1} = \frac{(1+n) \cdot (1+g_{t+1})}{1+\mu_{t+1}}, \quad g_{t+1} \equiv \frac{w_{1,t+1}}{w_{1,t}} - 1.$$

Gross inflation, R_t^{-1} , equals the monetary creation factor, corrected for the growth rate of the economy as measured by g_{t+1} , the youngs' endowments growth rate.

As a final example, consider the utility function $u(c_{1t}, c_{2,t+1}) = \left(l c_{1t}^{(\eta-1)/\eta} + (1-l) c_{2,t+1}^{(\eta-1)/\eta} \right)^{\eta/(\eta-1)}$, which collapses to Cobb-Douglas once $\eta \rightarrow 1$. We have:

$$c_{1t} = \frac{R_{t+1} w_{1t} + w_{2,t+1}}{R_{t+1} + K^\eta R_{t+1}^\eta}, \quad c_{2,t+1} = \frac{R_{t+1} w_{1t} + w_{2,t+1}}{1 + K^{-\eta} R_{t+1}^{1-\eta}}, \quad \text{sav}(R_{t+1}) = \frac{K^\eta R_{t+1}^\eta w_{1t} - w_{2,t+1}}{R_{t+1} + K^\eta R_{t+1}^\eta},$$

where $K \equiv \frac{1-l}{l}$. To simplify, set (i) $K = 1$, (ii) $w_{2t} = \mu_t = n = 0$, and (iii) $w_{1t} = w_{1,t+1}$. It can be shown that in this case, $\text{sign}(\text{sav}'(R)) = \text{sign}(\eta - 1)$. Moreover, the dynamics of the gross interest rate, R , are given by:

$$R_{t+1} = f(R_t) \equiv (R_t^{-\eta} + R_t^{-1} - 1)^{1/(1-\eta)}. \quad (3.49)$$

The stationary equilibria are solutions to $R = f(R)$, and it is easily seen that one of them is $R = 1$, and corresponds to the monetary steady state.

When $\eta > 1$, the slope in Eq. (3.48) has always the same sign, and the mapping f in Eq. (3.49) has two fixed points, $R_a = 0$ and $R = 1$, with R_a being stable and R being unstable, as illustrated by Figure 3.5 when $\eta = 2$.

When $\eta < 1$, the situation is quite delicate. In this case, R_a is not well-defined, and $R = 1$ is not necessarily unstable. We may have sequences of gross interest rates, R_t , converging towards R , or even the emergence of cycles. Mathematically, these properties can be understood by examining the slope of the map in Eq. (3.48), for $R = 1$, $\left. \frac{dR_{t+1}}{dR_t} \right|_{R_{t+1}=R_t=1} = \frac{\eta}{\eta-1}$.

In the general case, Figure 3.6 depicts an hypothetical shape of the map $R_t \mapsto R_{t+1}$, which is that we might expect to arise in the presence of gross substitutability or, in fact, even in the case of complementary, provided $\frac{\text{sav}'(R)R}{\text{sav}(R)} < -1$ for all R . In both cases, the slope, $\frac{dR_{t+1}}{dR_t} > 0$.

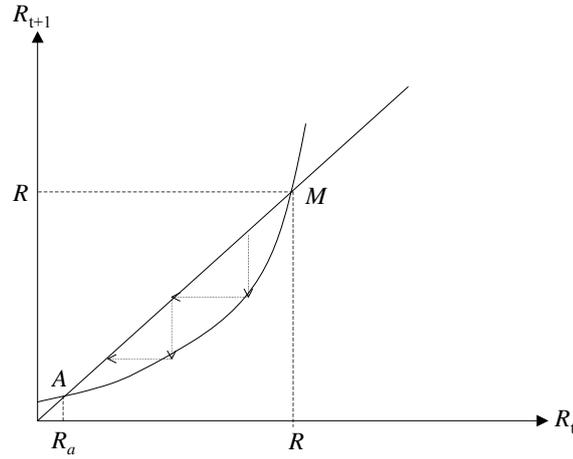


FIGURE 3.6. Gross substitutability

Moreover, the slope at the monetary state, $R = \frac{1+n}{1+\mu}$, is $\left. \frac{dR_{t+1}}{dR_t} \right|_{R_{t+1}=R_t=R} = 1 + \frac{\text{sav}(R)}{\text{sav}'(R)R} > 0$. The slope at the monetary state, the point M in Figure 3.6, is greater than one, provided $\text{sav}'(\cdot) > 0$. In this case, the monetary state M is unstable, while the autarkic state, the point A in Figure 3.6, is stable. Note that any path beginning from the right of the monetary state leads to explosive dynamics for R . These dynamics cannot be part of any equilibrium because they would imply a decreasing sequence of prices, p , thereby tilting the agents' budget constraints in such a way to rule out the existence of a solution to the agents' programs. Therefore, the economy needs to start anywhere between the point A and the point M , although then, we do not have any other piece of information: there exists, in fact, a continuum of points $R_1 \in [R_a, R)$ that are equally likely candidates to the beginning of the equilibrium sequence. Contrary to the representative agent models in the previous sections, the model of this section leads to an indeterminacy of the equilibrium, parametrized by the initial price p_0 .

Would an autarkic equilibrium be the only possible stable steady state? The answer is in the negative. Consider the case where the map $R_t \mapsto R_{t+1}$ bends backwards and is such that $\left. \frac{dR_{t+1}}{dR_t} \right|_{R_{t+1}=R_t=R} < -1$, such that the monetary steady state M is stable. A condition for the map $R_t \mapsto R_{t+1}$ to bend backward is that $\frac{\text{sav}'(R)R}{\text{sav}(R)} > -1$, and the condition for $\left. \frac{dR_{t+1}}{dR_t} \right|_{R_{t+1}=R_t=R} < -1$ to hold is that $\frac{\text{sav}'(R)R}{\text{sav}(R)} > -\frac{1}{2}$. In this case, the point M is reached from any sufficiently neighborhood of M . Figure 3.7 shows a cycle of order two, where $R_{**} = \frac{R^2}{R_*}$.⁸ Note that to analyze the behavior of the gross interest rate, we are needing to make reference to backward-looking dynamics, as there exists an indeterminacy of forward-looking dynamics. Finally, there might exist more complex situations where cycles of order 3 exist, giving rise to what is known as a “chaotic” system. Note that these complex dynamics, including those in Figure 3.7, rely on the assumption that $\text{sav}'(R) < 0$, which might be somehow unappealing.

⁸For the proof, note that by Eq. (3.47), we have, we have that for a cycle of order 2, (i) $\frac{1+\mu}{1+n} R_* s(R_*) = s(R_{**})$, and (ii) $\frac{1+\mu}{1+n} R_{**} s(R_{**}) = s(R_*)$. Multiplying the two equations side by side leaves the result that $R_{**} = R^2/R_*$.

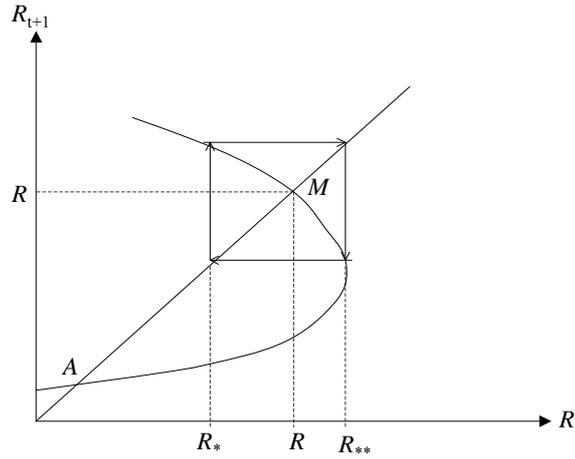


FIGURE 3.7.

3.5.4 Money in a model with real shocks

Lucas (1972) is the first attempt to address issues relating the neutrality of money in contexts with overlapping generations and uncertainty. This section is a simplified version of Lucas model as explained by Stokey et Lucas (1989) (p. 504). Every agent works when young, so as to produce a consumption good, and consumes when he is old, and experiences a disutility of work equal to $-v(n_t)$, where n_t is his labor supply, and v is assumed to satisfy $v', v'' > 0$. Utility drawn from second period consumption is denoted with $u(c_{t+1})$, and has the standard properties. The agent faces the following program:

$$\max_{\{n, c\}} [-v(n_t) + \beta E(u(c_{t+1}) | F_t)] \quad \text{subject to} \quad \begin{cases} m = p_t y_t, & y_t = \epsilon_t n_t \\ p_{t+1} c_{t+1} = m \end{cases}$$

where F_t denotes the information set as of time t , m is money holdings; y_t is the agent's production, obtained through his labor supply n_t , and $(\epsilon_t)_{t=0,1,\dots}$ is a sequence of positive shocks affecting his productivity. Finally, p_t is the price of the consumption good as of time t . By replacing the first constraint into the second leaves, $c_{t+1} = \epsilon_t n_t R_{t+1}$, where $R_{t+1} \equiv \frac{p_t}{p_{t+1}}$. Therefore, the program the agent solves is to $\max_n [-v(n_t) + \beta E(u(\epsilon_t n_t R_{t+1}) | F_t)]$. The first-order condition leads to,

$$v'(n_t) = \beta E[u'(\epsilon_t n_t R_{t+1}) \epsilon_t R_{t+1} | F_t].$$

We have, $c_{t+1} = \epsilon_{t+1} n_{t+1} = \epsilon_t n_t R_{t+1}$, where the first equality follows by the equilibrium in the good market. Replacing this relation into the previous equation leaves,

$$v'(n_t) n_t = \beta E[u'(\epsilon_t n_t R_{t+1}) \epsilon_{t+1} n_{t+1} | F_t]. \quad (3.50)$$

A rational expectation equilibrium is one where $n_t = \eta(\epsilon_t)$, with η satisfying Eq. (3.50), i.e.

$$v'(\eta(\epsilon) \eta(\epsilon)) = \beta \int_{\mathcal{E}} u'(\epsilon^+ \eta(\epsilon^+)) \epsilon^+ \eta(\epsilon^+) dP(\epsilon^+ | \epsilon),$$

where \mathcal{E} denotes the support of ϵ^+ . This equation simplifies as soon as productivity shocks are IID, $P(\epsilon^+ | \epsilon) = P(\epsilon^+)$, in which case, $v'(\eta(\epsilon) \eta(\epsilon))$ is independent of ϵ^+ , and n is a constant

\bar{n} .⁹ This is a result about the neutrality of money, at least provided such a constant \bar{n} exists. Precisely, we have that $v'(\bar{n}) = \beta \int_{\mathcal{E}} u'(\epsilon + \bar{n}) \epsilon^+ dP(\epsilon^+)$. For example, consider $v(x) = \frac{1}{2}x^2$ and $u(x) = \ln x$, in which case $\bar{n} = \sqrt{\beta}$, $y(\epsilon) = \epsilon\sqrt{\beta}$ and $p(\epsilon) = \frac{m}{\epsilon\sqrt{\beta}}$.

3.6 Optimality

3.6.1 Models with productive capital

Consider the usual law of capital accumulation: $K_{t+1} = S_t = Y(K_t, N_t) - C_t$, for K_0 given. Dividing both sides of this equation by N_t leaves:

$$k_{t+1} = \frac{1}{1+n} (y(k_t) - c_t), \quad \text{for } k_0 \text{ given.} \quad (3.51)$$

The stationary state of the economy is achieved when $k_{t+1} = k_t \equiv k$ and $c_{t+1} = c_t \equiv c$, such that:

$$c = y(k) - (1+n)k.$$

In steady-state, per-capita consumption attains its maximum at:

$$\bar{k} : y'(\bar{k}) = 1+n. \quad (3.52)$$

The steady state per-capita capital satisfying Eq. (3.52) is said to satisfy the *Golden Rule*. A social planner would be able to increase per-capita consumption at the stationary state, provided $y'(k) < 1+n$. Indeed, because $y(k)$ is given, we can lower k and have $dc = -(1+n)dk > 0$, immediately, and $dc = (y'(k) - (1+n))dk > 0$, in the next periods. In fact, this outcome would apply along the entire capital accumulation path of the economy, not only in steady state, as we now illustrate. First, a definition. We say that a path $(k, c)_{t=0}^{\infty}$ is consumption-inefficient if there exists another path $(\tilde{k}, \tilde{c})_{t=0}^{\infty}$ satisfying Eq. (3.51), and such that $\tilde{c}_t \geq c_t$ for all t , with at least a strict inequality for one t . The following is a slightly less general version of Theorem 1 in Tirole (p. 161):

THEOREM 3.3 (Cass-Malinvaud theory). *A path $(k, c)_{t=0}^{\infty}$ is: (i) consumption efficient if $\frac{y'(k_t)}{1+n} \geq 1$ for all t , and (ii) consumption inefficient if $\frac{y'(k_t)}{1+n} < 1$ for all t .*

PROOF. As for Part (i), suppose k_t is consumption efficient, and let $\tilde{k}_t = k_t + \epsilon_t$ be an alternative consumption efficient path. Since k_0 is given, $\epsilon_t = 0$. Moreover, by Eq. (3.51),

$$(1+n) \cdot (\tilde{k}_{t+1} - k_{t+1}) = y(\tilde{k}_t) - y(k_t) - (\tilde{c}_t - c_t),$$

and because \tilde{k} is consumption-efficient, $\tilde{c}_t \geq c_t$, with at least one strictly equality for some t . Therefore, by concavity of y , and the definition of \tilde{k}_t ,

$$0 \leq y(\tilde{k}_t) - y(k_t) - (1+n)(\tilde{k}_{t+1} - k_{t+1}) < y(k_t) + y'(k_t)\epsilon_t - y(k_t) - (1+n)\epsilon_{t+1},$$

⁹The proof that $\eta(\epsilon) = \bar{n}$ relies on the following argument. Suppose the contrary, i.e. there exists a point ϵ_0 and a neighborhood of ϵ_0 such that either (i) $\eta(\epsilon_0 + A) > \eta(\epsilon_0)$ or (ii) $\eta(\epsilon_0 + A) < \eta(\epsilon_0)$, for some strictly positive constant A . We deal with the proof of (i) as the proof of (ii) is nearly identical. Since $v'(\eta(\epsilon))\eta(\epsilon)$ is constant, and $v'' > 0$, we have that $v'(\eta(\epsilon_0 + A))\eta(\epsilon_0 + A) = v'(\eta(\epsilon_0))\eta(\epsilon_0) \leq v'(\eta(\epsilon_0 + A))\eta(\epsilon_0)$. Therefore, $v'(\eta(\epsilon_0 + A))[\eta(\epsilon_0 + A) - \eta(\epsilon_0)] < 0$. Next, note that $v' > 0$, such that $\eta(\epsilon_0 + A) < \eta(\epsilon_0)$, contradicting that $\eta(\epsilon_0 + A) > \eta(\epsilon_0)$.

or $\epsilon_{t+1} < \frac{y'(k_t)}{1+n} \epsilon_t$. Evaluating this inequality at $t = 0$ yields $\epsilon_1 < \frac{y'(k_0)}{1+n} \epsilon_0$, and since $\epsilon_0 = 0$, one has that $\epsilon_1 < 0$. Since $\frac{y'(k_t)}{1+n} \geq 1$ for all t , then $\epsilon_t \rightarrow -\infty$ as $t \rightarrow \infty$, which contradicts k_t has bounded trajectories. The proof of Part (ii) is nearly identical, except that, obviously, in this case, $\liminf \epsilon_t \gg -\infty$. Note, in general, there are infinitely many sequences that allow for efficiency improvements. \parallel

The reasoning in this section holds independently of whether the economy has a finite number of agents living forever, or overlapping generations. For example, in the case of overlapping generations, Eq. (3.51) is the capital accumulation path for Diamond's model, once we set $c_t \equiv \frac{C_t}{N_t} = c_{1t} + \frac{c_{2,t+1}}{1+n}$. An important issue is to establish whether actual economies are dynamically efficient? Abel, Mankiw, Summers and Zeckhauser (1989) provide a framework to address this question, which includes uncertainty, and conclude that the US economy does satisfy dynamic efficiency requirements.

3.6.2 Models with money

We wish to find first-best optima, that is, equilibria that a social planner may choose, by acting directly on agents' consumption, without needing to force the agents to make use of money.¹⁰ Let us analyze, first, the stationary state, $R = \frac{1+n}{1+\mu}$. We show that this state corresponds to the stationary state where consumptions and endowments are constants, and that the agents' utility is maximized when $\mu = 0$. Indeed, since the social planner allocates resources without having regard to money, the only constraint is: $w_n \equiv w_1 + \frac{w_2}{1+n} = c_1 + \frac{c_2}{1+n}$, such that the utility of the "stationary agent" is:

$$u(c_1, c_2) = u\left(w_n - \frac{c_2}{1+n}, c_2\right).$$

The first-order condition is $\frac{u_{c2}}{u_{c1}} = \frac{1}{1+n}$. Instead, the first-order condition in the market equilibrium is $\frac{u_{c2}}{u_{c1}} = \frac{1}{R}$. Therefore, the Golden Rule is attained in the market equilibrium, if and only if $\mu = 0$. The social planner policy converges towards the Golden Rule. Indeed, the social planner solves:

$$\max \sum_{t=0}^{\infty} \vartheta^t u^{(t)}(c_{1t}, c_{2,t+1}), \quad \text{subject to } w_{nt} \equiv w_{1t} + \frac{w_{2t}}{1+n} = c_{1t} + \frac{c_{2t}}{1+n},$$

or $\max \sum_{t=0}^{\infty} \vartheta^t u^{(t)}\left(w_{nt} - \frac{c_{2t}}{1+n}, c_{2,t+1}\right)$, where ϑ is the weight the planner gives to the generation as of time t , and the notation $u^{(t)}$ is meant to emphasize that that endowments may change from one generation to another. The first-order conditions, $\frac{u_{c2}^{(t-1)}}{u_{c1}^{(t)}} = \frac{\vartheta}{1+n}$, lead to the "modified" Golden Rule in steady-state state (modified by the weight ϑ).

¹⁰In a second-best equilibrium, a social planner would let the market "play" first, by allowing the agents to use money and, then, would parametrize such virtual equilibria by μ_t . The indirect utility functions that arise as a result would then be expressed in terms of these growth rates μ_t . The social planner would then maximize an aggregator of these utilities with respect to μ_t .

3.7 Appendix 1: Finite difference equations, with economic applications

Let $z_0 \in \mathbb{R}^d$, and consider the following linear system of finite difference equations:

$$z_{t+1} = A \cdot z_t, \quad t = 0, 1, \dots, \quad (3A.1)$$

for some matrix A . The solution to Eq. (3A.1) is:

$$z_t = v_1 \kappa_1 \lambda_1^t + \dots + v_d \kappa_d \lambda_d^t, \quad (3A.2)$$

where λ_i and v_i are eigenvalues and eigenvectors of A , and κ_i are constants, which will be determined below. The standard proof of this result relies on the so-called diagonalization of Eq. (3A.1). Let us consider the system of characteristic equations for A , $(A - \lambda_i I) v_i = \mathbf{0}_{d \times 1}$, where λ_i is scalar and v_i a $d \times 1$ column vector, for $i = 1, \dots, n$, or, in matrix form, $AP = P\Lambda$, where $P = (v_1, \dots, v_d)$ and Λ is a diagonal matrix with λ_i on its diagonal. We assume that $P^\top = P^{-1}$. By post-multiplying by P^{-1} leaves the *spectral decomposition* of A :

$$A = P\Lambda P^{-1}. \quad (3A.3)$$

By replacing Eq. (3A.3) into Eq. (3A.1), and rearranging terms,

$$y_{t+1} = \Lambda \cdot y_t, \quad \text{where } y_t \equiv P^{-1} z_t.$$

The solution for y is $y_{it} = \kappa_i \lambda_i^t$, and the solution for z is: $z_t = P y_t = (v_1, \dots, v_d) y_t = \sum_{i=1}^d v_i y_{it} = \sum_{i=1}^d v_i \kappa_i \lambda_i^t$, which is Eq. (3A.2).

To determine the vector of constants $\kappa = (\kappa_1, \dots, \kappa_d)^\top$, we first evaluate the solution at $t = 0$,

$$z_0 = (v_1, \dots, v_d) \kappa = P \kappa,$$

whence

$$\hat{\kappa} \equiv \kappa(P) = P^{-1} z_0, \quad (3A.4)$$

where the columns of P are vectors belonging to the space of the eigenvectors. Naturally, there is an infinity of these vectors. However, the previous formula shows how the constants $\kappa(P)$ need to “adjust” so as to guarantee the stability of the solution with respect to changes in P .

3A.1 EXAMPLE. Let $d = 2$, and suppose that $\lambda_1 \in (0, 1)$, $\lambda_2 > 1$. The resulting system is unstable for any initial condition, except perhaps for a set with measure zero. This set of measure zero gives rise to the so-called saddlepoint path. We can calculate the coordinates of such a set. We wish to find the set of initial conditions such that $\kappa_2 = 0$, so as to rule out an explosive behavior related to the unstable root $\lambda_2 > 1$. The solution at $t = 0$ is:

$$\begin{pmatrix} x_0 \\ y_0 \end{pmatrix} = z_0 = P \kappa = (v_1, v_2) \begin{pmatrix} \kappa_1 \\ \kappa_2 \end{pmatrix} = \begin{pmatrix} v_{11} \kappa_1 + v_{12} \kappa_2 \\ v_{21} \kappa_1 + v_{22} \kappa_2 \end{pmatrix},$$

where we have set $z = (x, y)^\top$. By replacing the second equation into the first, and solving for κ_2 , yields:

$$\kappa_2 = \frac{v_{11} y_0 - v_{21} x_0}{v_{11} v_{22} - v_{12} v_{21}},$$

which is zero when

$$y_0 = \frac{v_{21}}{v_{11}} x_0.$$

For this system, the saddlepoint is a line with a slope equal to the ratio of the two components of eigenvector for λ_1 —the stable root. Figure 3A.1 depicts the phase diagram for this system, with the “divergent” line satisfying the equation $y_0 = \frac{v_{22}}{v_{12}} x_0$.

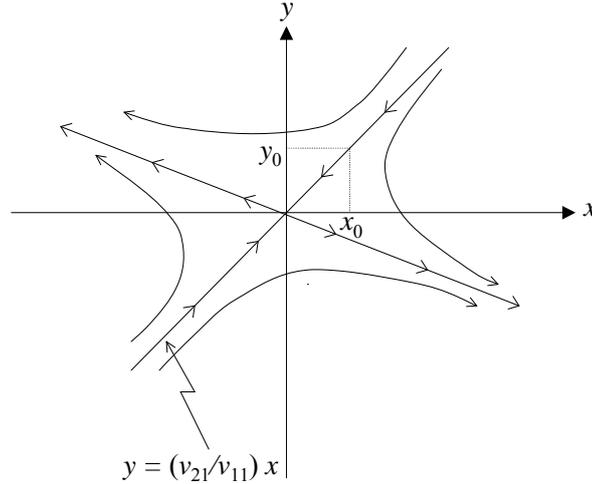


FIGURE 3A.1.

A saddlepoint path brings the following economic content. If x is a *predetermined* variable, y must “jump” to the saddlepoint $y_0 = \frac{v_{21}}{v_{11}}x_0$, so as to ensure the system does not explode. Note, then, that a conceptual difficulty arises should the system include two predetermined variables, as in this case, there are no stable solutions, generically. However, this possibility is unusual in economics. Consider the next example.

3A.2 EXAMPLE. The system of Example 3A.1 is exactly the one for the neoclassic growth model, as we now demonstrate. Section 3.3.3 shows that in a small neighborhood of the stationary values (k, c) , the deviations $(\hat{k}_t, \hat{c}_t)_t$ of capital and consumption from (k, c) , satisfy Eq. (3.13), which is reported here for convenience:

$$\begin{pmatrix} \hat{k}_{t+1} \\ \hat{c}_{t+1} \end{pmatrix} = A \begin{pmatrix} \hat{k}_t \\ \hat{c}_t \end{pmatrix}, \quad A \equiv \begin{pmatrix} y'(k) & -1 \\ -\frac{u'(c)}{u''(c)}y''(k) & 1 + \beta\frac{u'(c)}{u''(c)}y''(k) \end{pmatrix}.$$

By using the relation, $\beta y'(k) = 1$, and the standard conditions on utility and production, u and y , we have that the two eigenvalues of A are: $\lambda_{1/2} = \frac{\text{tr}(A) \pm \sqrt{\text{tr}(A)^2 - 4 \det(A)}}{2}$, where (i) $\det(A) = y'(k) = \beta^{-1} > 1$, and (ii) $\text{tr}(A) = \beta^{-1} + 1 + \beta\frac{u'(c)}{u''(c)}y''(k) > 1 + \det(A)$. Next, note that:

$$a \equiv \text{tr}(A)^2 - 4 \det(A) = \left(\beta^{-1} + 1 + \beta\frac{u'(c)}{u''(c)}y''(k) \right)^2 - 4\beta^{-1} > (\beta^{-1} + 1)^2 - 4\beta^{-1} = (1 - \beta^{-1})^2 > 0.$$

It follows that $\lambda_2 = \frac{1}{2}(\text{tr}(A) + \sqrt{a}) > \frac{1}{2}(1 + \det(A) + \sqrt{a}) > 1 + \frac{1}{2}\sqrt{a} > 1$. Finally, to show that $\lambda_1 \in (0, 1)$, note that since $\det(A) > 0$, one has $2\lambda_1 = \text{tr}(A) - \sqrt{\text{tr}(A)^2 - 4 \det(A)} > 0$; moreover, $\lambda_1 < 1 \Leftrightarrow \text{tr}(A) - \sqrt{\text{tr}(A)^2 - 4 \det(A)} < 2$, or $(\text{tr}(A) - 2)^2 < \text{tr}(A)^2 - 4 \det(A)$, which is true, by simple computations.

We generalize the previous examples to the case where $d > 2$. The counterpart of the saddlepoint for $d = 2$, is called *convergent*, or *stable subspace*. It is the locus of points such that z_t in Eq. (3A.3) does not explode. (In the case of nonlinear systems, this convergent subspace is termed *convergent*, or *stable manifold*. In this appendix we only study linear systems.) Let $\Pi \equiv P^{-1}$, and rewrite Eq. (3A.4), i.e. the system determining the solution for κ , as follows:

$$\hat{\kappa} = \Pi z_0.$$

We assume the elements of z and A are ordered in such a way that $\exists s : |\lambda_i| < 1$, for $i = 1, \dots, s$ and $|\lambda_i| > 1$ for $i = s + 1, \dots, d$. Then, we partition Π as follows:

$$\hat{\kappa} = \begin{pmatrix} \Pi_{\bar{s}} \\ s \times d \\ \Pi_u \\ (d-s) \times d \end{pmatrix} z_0.$$

Proceeding similarly as in Example 3A.1, we aim to make sure the system stays “trapped” in the convergent space and, accordingly, require that: $\hat{\kappa}_{s+1} = \dots = \hat{\kappa}_d = 0$, or,

$$\begin{pmatrix} \hat{\kappa}_{s+1} \\ \vdots \\ \hat{\kappa}_d \end{pmatrix} = \begin{pmatrix} \Pi_u \\ (d-s) \times d \end{pmatrix} z_0 = \mathbf{0}_{(d-s) \times 1}.$$

Let $d \equiv k + k^*$, where k is the number of free variables and k^* is the number of predetermined variables. Partition Π_u and z_0 in such a way to disentangle free from predetermined variables, as follows:

$$\mathbf{0}_{(d-s) \times 1} = \begin{pmatrix} \Pi_u \\ (d-s) \times d \end{pmatrix} z_0 = \begin{pmatrix} \Pi_u^{(1)} & \Pi_u^{(2)} \\ (d-s) \times k & (d-s) \times k^* \end{pmatrix} \begin{pmatrix} z_0^{\text{free}} \\ k \times 1 \\ z_0^{\text{pre}} \\ k^* \times 1 \end{pmatrix} = \begin{pmatrix} \Pi_u^{(1)} & \Pi_u^{(2)} \\ (d-s) \times k & (d-s) \times k^* \end{pmatrix} \begin{pmatrix} z_0^{\text{free}} \\ z_0^{\text{pre}} \end{pmatrix},$$

or,

$$\begin{pmatrix} \Pi_u^{(1)} \\ (d-s) \times k \end{pmatrix} z_0^{\text{free}} = - \begin{pmatrix} \Pi_u^{(2)} \\ (d-s) \times k^* \end{pmatrix} z_0^{\text{pre}}.$$

This system has $d - s$ equations with k unknowns, the components of z_0^{free} : indeed, z_0^{pre} is known, as it the k^* -dimensional vector of predetermined variables, and $\Pi_u^{(1)}, \Pi_u^{(2)}$ depend on the primitive data of the economy, through their relation with A . We assume that $\begin{pmatrix} \Pi_u^{(1)} \\ (d-s) \times k \end{pmatrix}$ has full rank.

We shall refer to s as the dimension of the convergent subspace, \mathcal{S} say. The reason for this terminology is the following. Consider the solution for z_t ,

$$z_t = v_1 \hat{\kappa}_1 \lambda_1^t + \dots + v_s \hat{\kappa}_s \lambda_s^t + v_{s+1} \hat{\kappa}_{s+1} \lambda_{s+1}^t + \dots + v_d \hat{\kappa}_d \lambda_d^t.$$

For z_t to remain stuck in \mathcal{S} , it must be the case that

$$\hat{\kappa}_{s+1} = \dots = \hat{\kappa}_d = 0,$$

in which case,

$$z_t = \begin{pmatrix} v_1 \hat{\kappa}_1 \lambda_1^t + \dots + v_s \hat{\kappa}_s \lambda_s^t \\ d \times 1 \end{pmatrix} = (v_1 \hat{\kappa}_1, \dots, v_s \hat{\kappa}_s) \cdot (\lambda_1^t, \dots, \lambda_s^t)^\top,$$

i.e.,

$$z_t = \begin{pmatrix} \hat{V} \\ d \times s \end{pmatrix} \cdot \begin{pmatrix} \hat{\lambda}_t \\ s \times 1 \end{pmatrix},$$

where $\hat{V} \equiv (v_1 \hat{\kappa}_1, \dots, v_s \hat{\kappa}_s)$ and $\hat{\lambda}_t \equiv (\lambda_1^t, \dots, \lambda_s^t)^\top$. Finally, for each t , introduce the vector subspace:

$$\langle \hat{V} \rangle_t \equiv \{z_t \in \mathbb{R}^d : z_t = \begin{pmatrix} \hat{V} \\ d \times s \end{pmatrix} \cdot \begin{pmatrix} \hat{\lambda}_t \\ s \times 1 \end{pmatrix}, \hat{\lambda}_t \in \mathbb{R}^s\}.$$

Clearly, for each t , $\dim \langle \hat{V} \rangle_t = \text{rank}(\hat{V}) = s$.

There are three cases to consider:

- (i) $d - s = k$, or $s = k^*$. The dimension of the divergent subspace is equal to the number of the free variables or, equivalently, the dimension of the convergent subspace is equal to the number of predetermined variables. In this case, the system is determined. The previous conditions are interpreted as follows. The predetermined variables identify one and only one point in the convergent space, which gives rise to only one possible jump that the free variables can make to ensure the system remain in the convergent space: $z_0^{\text{free}} = -\Pi_u^{(1)-1} \Pi_u^{(2)} z_0^{\text{pre}}$. This case is exactly as in Example 3A.1, where $d = 2$, $k = 1$, and the predetermined variable is x . In this example, x_0 identifies one and only one point in the saddlepoint path, such that starting from that point, there is one and only one value of y_0 guaranteeing that the system does not explode.
- (ii) $d - s > k$, or $s < k^*$. There are generically no solutions lying in the convergent space—a case mentioned just before Example 3A.2.
- (iii) $d - s < k$, or $s > k^*$. There are infinitely many solutions lying in the convergent space, a phenomenon typically referred to as *indeterminacy*. Note that in this case, *sunspots* equilibria may arise. In Example 3A.1, $s = 1$, and so in order for this case to emerge when $d = 2$, we might need to rule out the existence of predetermined variables.

3.8 Appendix 2: Neoclassic growth in continuous-time

3.8.1 Convergence from discrete-time

[To be revised]

Consider chopping time in the law of population growth as follows:

$$N_{hk} - N_{h(k-1)} = \bar{n}N_{h(k-1)}h, \quad k = 1, \dots, \ell,$$

where \bar{n} is an *instantaneous rate*, and $\ell = \frac{t}{h}$ is the number of subperiods in which we have chopped a given time period t . The solution is $N_{h\ell} = (1 + \bar{n}h)^\ell N_0$, or $N_t = (1 + \bar{n}h)^{t/h} N_0$. By taking limits leaves:

$$N(t) = \lim_{h \downarrow 0} (1 + \bar{n}h)^{t/h} N(0) = e^{\bar{n}t} N(0).$$

On the other hand, an exact discretization yields: $N(t - \Delta) = e^{\bar{n}(t-\Delta)} N(0)$, or $\frac{N(t)}{N(t-\Delta)} = e^{\bar{n}\Delta} \equiv 1 + n_\Delta \Leftrightarrow \bar{n} = \frac{1}{\Delta} \ln(1 + n_\Delta)$. Take, for example, $\Delta = 1$, $n_\Delta = n_1 \equiv n : \bar{n} = \ln(1 + n)$.

We apply the same discretization scheme to the law of capital accumulation:

$$K_{h(k+1)} = (1 - \bar{\delta}h) K_{hk} + I_{h(k+1)}h, \quad k = 0, \dots, \ell - 1,$$

where $\bar{\delta}$ is an instantaneous rate, and $\ell = \frac{t}{h}$. By iterating,

$$K_t = (1 - \bar{\delta}h)^{t/h} K_0 + \sum_{j=1}^{t/h} (1 - \bar{\delta}h)^{t/h-j} I_{hj}h.$$

Taking the limits for $h \downarrow 0$ yields:

$$K(t) = e^{-\bar{\delta}t} K_0 + e^{-\bar{\delta}t} \int_0^t e^{\bar{\delta}u} I(u) du,$$

or in differential form:

$$\dot{K}(t) = -\bar{\delta}K(t) + I(t). \quad (3A.5)$$

By replacing the IS equation,

$$Y(t) = F(K(t), N(t)) = C(t) + I(t), \quad (3A.6)$$

into Eq. (3A.5), we obtain the law of capital accumulation:

$$\dot{K}(t) = F(K(t), N(t)) - C(t) - \bar{\delta}K(t).$$

There are a few discretization issues to discuss. First, an exact discretization gives:

$$K(t+1) = e^{-\bar{\delta}} K(t) + e^{-\bar{\delta}(t+1)} \int_t^{t+1} e^{\bar{\delta}u} I(u) du. \quad (3A.7)$$

By identifying with the standard capital accumulation law in the discrete time setting:

$$K_{t+1} = (1 - \delta) K_t + I_t,$$

we get: $\bar{\delta} = \ln \frac{1}{1-\delta}$. It follows that $\delta \in (0, 1) \Rightarrow \bar{\delta} > 0$ and $\delta = 0 \Rightarrow \bar{\delta} = 0$. Hence, while δ can take on only values on $[0, 1)$, $\bar{\delta}$ can take on values on the entire real line. In the continuous time model, then, $\lim_{\delta \rightarrow 1^-} \ln \frac{1}{1-\delta} = \infty$. In continuous time, we cannot imagine a “maximal rate of capital depreciation,” as this would be infinite!

Let us replace δ into the exact discretization in Eq. (3A.7):

$$K(t+1) = (1 - \delta) K(t) + e^{-\bar{\delta}(t+1)} \int_t^{t+1} e^{\bar{\delta}u} I(u) du,$$

such that investments at $t+1$ are simply, $e^{-\bar{\delta}(t+1)} \int_t^{t+1} e^{\bar{\delta}u} I(u) du$.

Finally, we derive per-capita dynamics. Consider dividing both sides of the capital accumulation equation (3A.6) by $N(t)$:

$$\frac{\dot{K}(t)}{N(t)} = \frac{F(K(t), N(t))}{N(t)} - c(t) - \bar{\delta}k(t) = y(k(t)) - c(t) - \bar{\delta}k(t).$$

By using the relation $\dot{k} = d\left(\frac{K(t)}{N(t)}\right) = \frac{\dot{K}(t)}{N(t)} - \bar{n}k(t)$ into the previous equation leads to:

$$\dot{k} = y(k(t)) - c(t) - (\bar{\delta} + \bar{n}) \cdot k(t).$$

It is the capital accumulation constraint used to solve the program of the next section.

3.8.2 The model

Consider the following social planner problem:

$$\begin{cases} \max_c \int_0^{\infty} e^{-\rho t} u(c(t)) dt \\ \text{s.t. } \dot{k}(t) = y(k(t)) - c(t) - (\bar{\delta} + \bar{n}) \cdot k(t) \end{cases} \quad [3A.P1]$$

where all variables are per-capita. We assume there is no capital depreciation. (Note that for the discrete time model, we assumed, instead, a total capital depreciation.) The Hamiltonian is,

$$H(t) = u(c(t)) + \lambda(t) [y(k(t)) - c(t) - (\bar{\delta} + \bar{n}) \cdot k(t)],$$

where λ is a co-state variable. As explained in Appendix 3 of this chapter, the first-order conditions for this problem are:

$$\begin{cases} 0 = \frac{\partial H}{\partial c}(t) & \Leftrightarrow & \lambda(t) = u'(c(t)) \\ \frac{\partial H}{\partial \lambda}(t) = \dot{k}(t) & & \\ \frac{\partial H}{\partial k}(t) = -\dot{\lambda}(t) + \rho\lambda(t) & \Leftrightarrow & \dot{\lambda}(t) = [\rho + \bar{\delta} + \bar{n} - y'(k(t))] \lambda(t) \end{cases} \quad (3A.8)$$

By differentiating the first of equations (3A.8) with respect to time:

$$\dot{\lambda}(t) = \left(\frac{u''(c(t))}{u'(c(t))} \dot{c}(t) \right) \lambda(t).$$

By identifying through the third of equations (3A.8),

$$\dot{c}(t) = \frac{u'(c(t))}{u''(c(t))} (\rho + \bar{\delta} + \bar{n} - y'(k(t))). \quad (3A.9)$$

The equilibrium is the solution of the system consisting of the constraint of the program [3A.P1], and Eq. (3A.9). Similarly as in Section 3.3.3, we analyze the dynamics of the system in a small

neighborhood of the stationary state, defined as the solution (c, k) of the constraint of the program [3A.P1], and Eq. (3A.9), when $\dot{c}(t) = \dot{k}(t) = 0$,

$$\begin{cases} c & = y(k) - (\bar{\delta} + \bar{n})k \\ \rho + \bar{\delta} + \bar{n} & = y'(k) \end{cases}$$

Warning! these are instantaneous figures, so its okay if they are not such that $y'(k) \geq 1+n!$. A first-order approximation of both sides of the constraint of the program [3A.P1], and Eq. (3A.9), near (c, k) , yields:

$$\begin{cases} \dot{c}(t) & = -\frac{u'(c)}{u''(c)}y''(k)(k(t) - k) \\ \dot{k}(t) & = \rho \cdot (k(t) - k) - (c(t) - c) \end{cases}$$

where we used the equality $\rho + \bar{\delta} + \bar{n} = y'(k)$. By setting $x(t) \equiv c(t) - c$ and $y(t) \equiv k(t) - k$ the previous system can be rewritten as:

$$\dot{z}(t) = A \cdot z(t), \quad A \equiv \begin{pmatrix} 0 & -\frac{u'(c)}{u''(c)}y''(k) \\ -1 & \rho \end{pmatrix},$$

where $z \equiv (x, y)^\top$.

Warning! There must be some mistakes somewhere. Let us diagonalize this system by setting $A = P\Lambda P^{-1}$, where P and Λ are as in Appendix 1. We have:

$$\dot{\nu}(t) = \Lambda \cdot \nu(t),$$

where $\nu \equiv P^{-1}z$. The eigenvalues are solutions of the following quadratic equation:

$$0 = \lambda^2 - \rho\lambda - \frac{u'(c)}{u''(c)}y''(k).$$

We see that $\lambda_1 < 0 < \lambda_2$, and $\lambda_1 \equiv \frac{\rho}{2} - \frac{1}{2}\sqrt{\rho^2 + 4\frac{u'(c)}{u''(c)}y''(k)}$. The solution for $\nu(t)$ is:

$$\nu_i(t) = \kappa_i e^{\lambda_i t}, \quad i = 1, 2,$$

whence

$$z(t) = P \cdot \nu(t) = v_1 \kappa_1 e^{\lambda_1 t} + v_2 \kappa_2 e^{\lambda_2 t},$$

where the v_i s are 2×1 vectors. We have,

$$\begin{cases} x(t) & = v_{11}\kappa_1 e^{\lambda_1 t} + v_{12}\kappa_2 e^{\lambda_2 t} \\ y(t) & = v_{21}\kappa_1 e^{\lambda_1 t} + v_{22}\kappa_2 e^{\lambda_2 t} \end{cases}$$

Let us evaluate this solution in $t = 0$,

$$\begin{pmatrix} x(0) \\ y(0) \end{pmatrix} = P\kappa = \begin{pmatrix} v_1 & v_2 \end{pmatrix} \begin{pmatrix} \kappa_1 \\ \kappa_2 \end{pmatrix}.$$

By repeating the reasoning of the previous appendix,

$$\kappa_2 = 0 \Leftrightarrow \frac{y(0)}{x(0)} = \frac{v_{21}}{v_{11}}.$$

As in the discrete time model, the saddlepoint path is located along a line that has as a slope the ratio of the components of the eigenvector associated with the negative root. We can explicitly compute such ratio. By definition, $A \cdot v_1 = \lambda_1 v_1 \Leftrightarrow$

$$\begin{cases} -\frac{u'(c)}{u''(c)}y''(k) & = \lambda_1 v_{11} \\ -v_{11} + \rho v_{21} & = \lambda_1 v_{21} \end{cases}$$

i.e., $\frac{v_{21}}{v_{11}} = -\frac{\lambda_1}{\frac{u'(c)}{u''(c)}y''(k)}$ and simultaneously, $\frac{v_{21}}{v_{11}} = \frac{1}{\rho - \lambda_1}$.

3.9 Appendix 3: Notes on optimization of continuous time systems

Consider the following optimization problem: for $k(t)$ given,

$$\begin{cases} J(k(t), t, T) \equiv \max_{(v(\tau))_{\tau=0}^T} E \left(\int_t^T e^{-\rho(\tau-t)} u(k(\tau), v(\tau)) d\tau + e^{-\rho(T-t)} B(k(T)) \right) \\ \text{s.t. } dk(\tau) = T(k(\tau), v(\tau)) d\tau + \sigma(k(\tau), v(\tau)) dW(\tau) \end{cases} \quad (3A.10)$$

where $W(\tau)$ is a standard Brownian Motion; T and σ are the drift and diffusion functions; u and B are, respectively, an instantaneous payoff, or reward, function, so to speak, and a bequest function; ρ is a discount rate; k is a state variable, and v is a control. Heuristically, the observation of k would provide us with information that can only be used to improve how we choose v . Naturally, controls cannot depend on future observation of k —they can only depend on past values of k . We confine attention to controls known as *feedbacks*, $v(\tau; \omega) \equiv \hat{v}(\tau, k(\tau; \omega))$, for each sample path ω of the state variable $k(\tau; \omega)$. The function \hat{v} makes the control and, then, the state variable k , Markovian.

We can, heuristically, apply a stochastic programming principle: we maximize up to an intermediate point in time $t + \Delta t$, say, assuming the maximization for the remaining time period $[t + \Delta t, T]$ holds. We have:

$$\begin{aligned} J(k(t), t, T) &\equiv \max_{(v(\tau))_{\tau=0}^T} E_t \left(\int_t^T e^{-\rho(\tau-t)} u(k(\tau), v(\tau)) d\tau + e^{-\rho(T-t)} B(k(T)) \right) \\ &= \max_{(v(\tau))_{\tau=0}^T} \left[E_t \left(\int_t^{t+\Delta t} e^{-\rho(\tau-t)} u(k(\tau), v(\tau)) d\tau \right) \right. \\ &\quad \left. + e^{-\rho\Delta t} E_{t+\Delta t} \left(\int_{t+\Delta t}^T e^{-\rho(T-t-\Delta t)} u(k(\tau), v(\tau)) d\tau + e^{-\rho(T-t-\Delta t)} B(k(T)) \right) \right] \\ &= \max_{(v(\tau))_{\tau=0}^T} E_t \left(\int_t^{t+\Delta t} e^{-\rho(\tau-t)} u(k(\tau), v(\tau)) d\tau + e^{-\rho\Delta t} J(k(t+\Delta t), t+\Delta t, T) \right), \end{aligned}$$

where the last two lines follow by the law of iterated expectations, and the last line is the stochastic programming principle. Rearranging terms,

$$\begin{aligned} &\left(\frac{1 - e^{-\rho\Delta t}}{\Delta t} \right) J(k(t), t, T) \\ &= \max_{(v(\tau))_{\tau=0}^T} E_t \left(\frac{1}{\Delta t} \int_t^{t+\Delta t} e^{-\rho(\tau-t)} u(k(\tau), v(\tau)) d\tau + e^{-\rho\Delta t} \frac{1}{\Delta t} (J(k(t+\Delta t), t+\Delta t, T) - J(k(t), t, T)) \right). \end{aligned}$$

For small Δt , we have, after applying Itô's lemma, and assuming enough regularity conditions,

$$0 = \max_v \left(u(k, v) + \left(\frac{\partial}{\partial \tau} + L \right) J(k, \tau, T) - \rho J(k, \tau, T) \right), \quad (3A.11)$$

subject to the boundary condition $J(k, \tau, T) = B(k)$, where $LJ \equiv \frac{\partial J}{\partial k} T + \frac{1}{2} \frac{\partial^2 J}{\partial k^2} \sigma^2$.

The first order condition for Eq. (3A.11) is:

$$0 = u_v(k, \hat{v}) + \frac{\partial J(k, \tau, T)}{\partial k} T_v(k, \hat{v}) + \frac{1}{2} \frac{\partial^2 J(k, \tau, T)}{\partial k^2} \frac{\partial \sigma^2(k, \hat{v})}{\partial v},$$

and implicitly defines the solution $\hat{v} \equiv \hat{v}(\tau, k)$, such that Eq. (3A.11) can be written as:

$$\begin{aligned} \rho J(k, \tau, T) - \frac{\partial J(k, \tau, T)}{\partial \tau} &= u(k, \hat{v}(\tau, k)) + \frac{\partial J(k, \tau, T)}{\partial k} T(k, \hat{v}(\tau, k)) + \frac{1}{2} \frac{\partial J^2(k, \tau, T)}{\partial k^2} \sigma^2(k, \hat{v}(\tau, k)) \\ &\equiv u(k, \hat{v}(\tau, k)) + \lambda(k, \tau, T) T(k, \hat{v}(\tau, k)) + \frac{1}{2} \mu(k, \tau, T) \sigma^2(k, \hat{v}(\tau, k)) \\ &\equiv \hat{H}(\tau, k, \lambda(k, \tau, T), \mu(k, \tau, T)), \end{aligned} \quad (3A.12)$$

where $\lambda(k, \tau, T) \equiv \frac{\partial J(k, \tau, T)}{\partial k}$, i.e. the marginal indirect utility with respect to k in (3A.10), and $\mu(k, \tau, T) \equiv \frac{\partial \lambda(k, \tau, T)}{\partial k} = \frac{\partial^2 J(k, \tau, T)}{\partial k^2}$. The function \hat{H} is usually referred to as the optimized *Hamiltonian*, and Eq. (3A.12) is the *Bellman Equation* for diffusion processes.

By Itô's lemma:

$$d\lambda = \left(\frac{\partial \lambda}{\partial \tau} + \frac{\partial \lambda}{\partial k} T + \frac{1}{2} \frac{\partial^2 \lambda}{\partial k^2} \sigma^2 \right) d\tau + \frac{\partial \lambda}{\partial k} \sigma dW. \quad (3A.13)$$

On the other hand, by differentiating both sides of the Bellman Equation (3A.12) with respect to k ,

$$\begin{aligned} \rho \frac{\partial J}{\partial k} - \frac{\partial^2 J}{\partial k \partial \tau} &= \rho \lambda - \frac{\partial \lambda}{\partial \tau} \\ &= \frac{\partial \hat{H}}{\partial k} + \frac{\partial \hat{H}}{\partial \lambda} \frac{\partial \lambda}{\partial k} + \frac{\partial \hat{H}}{\partial \mu} \frac{\partial \mu}{\partial k} \\ &= \frac{\partial \hat{H}}{\partial k} + T \frac{\partial \lambda}{\partial k} + \frac{1}{2} \sigma^2 \frac{\partial \mu}{\partial k}, \end{aligned} \quad (3A.14)$$

where the first equality follows by the definition of λ , and the third equality holds by the definition of the optimized Hamiltonian function, $\hat{H}(\tau, k, \lambda, \mu)$. Plugging Eq. (3A.14) into Eq. (3A.13) leaves:

$$d\lambda = \left(\rho \lambda - \frac{\partial \hat{H}}{\partial k} \right) d\tau + \frac{\partial \lambda}{\partial k} \sigma dW,$$

which shows that:

$$\rho \lambda - \frac{\partial \hat{H}}{\partial k} = \frac{E(d\lambda)}{d\tau} = \frac{\partial \lambda}{\partial \tau} + \frac{\partial \lambda}{\partial k} T + \frac{1}{2} \frac{\partial^2 \lambda}{\partial k^2} \sigma^2. \quad (3A.15)$$

To summarize, we solve the problem in three steps:

1. Maximize the Hamiltonian function,

$$H(\tau, k, \lambda, \mu) \equiv u(k, v) + \lambda \cdot T(k, v) + \frac{1}{2} \cdot \mu \cdot \sigma(k, v)^2,$$

with respect to the control v . The result is the optimized Hamiltonian function, $\hat{H}(\tau, k, \lambda, \mu)$.

2. Impose the condition in Eq. (3A.15).
3. Solve the partial differential equation (3A.11) for J . Note that for the infinite horizon case, i.e. the problem in (3A.10), for $T = \infty$, the counterpart to Eq. (3A.11) is:

$$0 = \max_v (u(k, v) + LJ(k, \tau, T) - \rho J(k, \tau, T)),$$

subject to the transversality condition,

$$\lim_{T \rightarrow \infty} e^{-\rho T} E[J(k(T))].$$

Succinctly,

$$\begin{cases} \frac{\partial \hat{H}}{\partial v} = 0 \\ \frac{\partial \hat{H}}{\partial \lambda} = T \\ \rho \lambda - \frac{\partial \hat{H}}{\partial k} = \frac{E(d\lambda)}{d\tau} = \frac{\partial \lambda}{\partial \tau} + \frac{\partial \lambda}{\partial k} T + \frac{1}{2} \frac{\partial^2 \lambda}{\partial k^2} \sigma^2 \end{cases} \quad (3A.16)$$

Consider, for example, the deterministic macroeconomic model of Appendix 2. It is easy to see that Eqs. (3A.8) are a special case of Eqs. (3A.16), namely when $\sigma \equiv 0$.

References

- Abel, A.B., N.G. Mankiw, L.H. Summers and R.J. Zeckhauser (1989): "Assessing Dynamic Efficiency: Theory and Evidence." *Review of Economic Studies* 56, 1-20.
- Cochrane, J. H., F. A. Longstaff, and P. Santa-Clara (2008): "Two Trees." *Review of Financial Studies* 21, 347-385.
- Farmer, R. (1998): *The Macroeconomics of Self-Fulfilling Prophecies*. Boston: MIT Press.
- Hayashi, F. (1982): "Tobin's Marginal q and Average q : A Neoclassical Interpretation." *Econometrica* 50, 213-224.
- Kamihigashi, T. (1996): "Real Business Cycles and Sunspot Fluctuations are Observationally Equivalent." *Journal of Monetary Economics* 37, 105-117.
- King, R. G. and S. T. Rebelo (1999): "Resuscitating Real Business Cycles." In: J. B. Taylor and M. Woodford (Editors): *Handbook of Macroeconomics*, Elsevier.
- Lucas, R. E. (1972): "Expectations and the Neutrality of Money." *Journal of Economic Theory* 4, 103-124.
- Lucas, R. E. (1978): "Asset Prices in an Exchange Economy." *Econometrica* 46, 1429-1445.
- Lucas, R. E. (1994): "Money and Macroeconomics." In: *General Equilibrium 40th Anniversary Conference*, CORE DP no. 9482, 184-187.
- Martin, I. (2011): "The Lucas Orchard." Working Paper Stanford University.
- Menzly, L., Santos, T., and P. Veronesi (2004): "Understanding Predictability." *Journal of Political Economy* 112, 1-47.
- Pavlova, A. and R. Rigobon (2008): "The Role of Portfolio Constraints in the International Propagation of Shocks." *Review of Economic Studies* 75, 1215-1256.
- Prescott, E. (1991): "Real Business Cycle Theory: What Have We Learned?" *Revista de Analisis Economico* 6, 3-19.
- Stokey, N. L. and R. E. Lucas, (with E.C. Prescott) (1989): *Recursive Methods in Economic Dynamics*. Harvard University Press.
- Tirole, J. (1988): "Efficacité intertemporelle, transferts intergénérationnels et formation du prix des actifs: une introduction." *Melanges économiques. Essais en l'honneur de Edmond Malinvaud*. Paris: Editions Economica & Editions EHESS, 157-185.
- Tobin, J. (1969): "A General Equilibrium Approach to Monetary Policy." *Journal of Money, Credit and Banking* 1, 15-29.
- Watson, M. (1993): "Measures of Fit for Calibrated Models." *Journal of Political Economy* 101, 1011-1041.

4

Continuous time models

4.1 Introduction

This chapter is an introduction to asset pricing models cast in continuous time. As such, it does not introduce anything conceptually really new, economically, against what we have already learned in previous chapters. Nevertheless, continuous time methods are powerful when it comes to deal with issues arising in economies and markets more complex than those in the previous chapters. Moreover, on an applied perspective, continuous time methods are extremely useful to evaluate derivative instruments that draw value from complex events, such as those relating to baskets of credit events, capital market volatility, or history-dependent developments in fixed income security markets, to name just a few, as we shall see in Part III of these lectures. Continuous time models pose challenges to econometricians—we only observe a discrete realization of an idealized continuous time data generating process. The next chapter develops tools based on simulations, which we can use to cope with these challenges.

This chapter aims to two scopes. The first is to explain in detail how the principle of absence of arbitrage works in continuous time: how do asset prices need to “drift” to ensure that there is no arbitrage? How many of these possible “drifts” do we expect to see in arbitrage-free markets? Our second objective is to develop technical details about the properties of asset prices in continuous time. For example, we shall see that asset prices, once restricted by absence of arbitrage, satisfy partial differential equations, under technical and, sometimes, simplifying assumptions. Yet asset prices are discounted expectations of their future payoffs, taken under the risk-neutral probability. How are these properties tied together? We shall make the connection between partial differential equations and conditional expectations, through the celebrated Feynman-Kac probabilistic representation of the solution to a partial differential equation. Moreover, what is the relation between the risk-neutral probability and the “physical” probability? How do we need to tilt the physical probability to determine the risk-neutral? How many risk-neutral probabilities exist, in incomplete markets or in markets with frictions? The Girsanov theorem is the starting point that we need to deal with these fundamental questions.

The class of models we consider in this chapter are known as diffusion models, along with their extensions accommodating for jumps. and are the workhorse in finance. Diffusion models are, so to speak, those where the variations of a variable of interest are driven by a determin-

istic component (the “drift”) and a stochastic one (the “diffusion”). Heuristically, the diffusion component is normally distributed over an infinitesimal amount of time, being equal to the variations of what is known as a “Brownian motion.” We typically assume that the fundamentals of the economy follow diffusion processes, and that asset prices are rational, in that they are a function of these fundamentals. Absence of arbitrage restricts the set of all possible pricing functions. The fundamental tool with which we link asset prices to fundamentals is Itô’s lemma, a device we need to build new processes (in our case, the asset prices) from old ones (the fundamentals of the economy). The complication in finance is that these new processes are a function of the fundamentals, which we are not given in advance, being, instead, the focus of research. The technical appendixes provide explanations about Itô’s lemma, as well as additional notions of stochastic calculus, and are self-contained. The reader already acquainted with these notions can safely read the chapter without those details.

The chapter is organized as follows. The next section is an introduction to a few and basic properties applying to the price of a long-lived asset, obtained by considering a continuous time limit of models seen in previous chapters. It derives: (i) the fundamental relations that link expected returns, volatilities (the “betas”) and risk-premiums (the “lambdas”); and (ii) a representation of the price-dividend ratio in terms of certain possibly varying discount rates—the risk-adjusted discount rates. These derivations turn out to be quite useful while discussing the properties of equity markets in Chapter 7. Section 4.3 is an introduction to methods. It deals with details leading to the birth of modern continuous time finance—the Black & Scholes formula of evaluation of European options. Section 4.4 ... [In progress]

4.2 On lambdas and betas

4.2.1 Prices

Let S_t be the price of a long-lived asset as of time t , and $D_{t+h} \cdot h$ the dividend paid by this asset at t over a small trading period h . We know that in the absence of arbitrage opportunities, there exists a positive process m_t , known as the *stochastic discount factor*, such that the price of any asset is the expectation of its future payoff, weighted with m_{t+h} ,

$$S_t = E_t [m_{t+h} (S_{t+h} + D_{t+h} \cdot h)], \quad (4.1)$$

where E_t is the conditional expectation given the information set at time t . For example, in a representative economy with risk-neutral agents, we would have that $m_{t+h} = e^{-rth}$, where r is the risk-free rate per unit of time.

Given the stochastic discount factor, we define, as usual, the *pricing kernel*, or state-price, process, ξ_t , as the process that grows by the stochastic discount factor:

$$\xi_{t+h} = m_{t+h} \xi_t, \quad \xi_0 = 1.$$

In terms of the pricing kernel, Eq. (4.1) is, $0 = E_t (\xi_{t+h} S_{t+h} - S_t \xi_t) + E_t (\xi_{t+h} D_{t+h} h)$, which in the limit case of h that tends to zero, is:

$$0 = E_t [d(\xi(t) S(t))] + \xi(t) D(t) dt. \quad (4.2)$$

Eq. (4.2) can be integrated to yield, $\xi(t) S(t) = E_t [\int_t^T \xi(u) D(u) du] + E_t [\xi(T) S(T)]$. Assuming that $\lim_{T \rightarrow \infty} E_t [\xi(T) S(T)] = 0$, the asset price, then, satisfies,

$$\xi(t) S(t) = E_t \left[\int_t^\infty \xi(u) D(u) du \right]. \quad (4.3)$$

Note that, $E_t(m_{t+h}) - 1 = e^{-r_t h} - 1 \approx -r_t h$, whence,

$$\frac{\xi_{t+h} - \xi_t}{\xi_t} \approx -r_t h + [m_{t+h} - E_t(m_{t+h})]. \quad (4.4)$$

In the presence of risk, and risk-averse agents, the innovations to the stochastic discount factor will drive fluctuations of the pricing kernel.

4.2.2 Expected returns

We can elaborate on Eq. (4.2). Assuming pricing kernels are driven by Brownian motions,

$$d(\xi S) = S d\xi + \xi dS + d\xi dS = \xi S \left(\frac{d\xi}{\xi} + \frac{dS}{S} + \frac{d\xi dS}{\xi S} \right).$$

By replacing this expansion into Eq. (4.2) leaves:

$$E_t \left(\frac{dS}{S} \right) + \frac{D}{S} dt = -E_t \left(\frac{d\xi}{\xi} \right) - E_t \left(\frac{d\xi dS}{\xi S} \right). \quad (4.5)$$

This evaluation equation holds for any asset and, hence, for the assets that do not distribute dividends and are locally riskless, i.e. $D = 0$ and $E_t \left(\frac{d\xi dS_0}{\xi S_0} \right) = 0$, where $S_0(t)$ is the price of these locally riskless assets, supposed to satisfy $\frac{dS_0(t)}{S_0(t)} = r(t) dt$, where $r(t)$ is the short term rate process, such that, by Eq. (4.5), $E_t \left(\frac{d\xi}{\xi} \right) = -r(t) dt$, consistently with the discrete time counterpart of the pricing kernel in Eq. (4.4). By Eq. (4.5), then, the expected returns satisfy:

$$E_t \left(\frac{dS}{S} \right) + \frac{D}{S} dt = r dt - E_t \left(\frac{d\xi dS}{\xi S} \right). \quad (4.6)$$

In a diffusion setting, the expectations in Eq. (4.6) can be written in terms of partial derivatives, implying that the asset price solves a certain partial differential equation. Furthermore, in a diffusion setting, the stochastic discount factor satisfies the continuous time limit to Eq. (4.4),

$$\frac{d\xi(t)}{\xi(t)} = -r(t) dt - \lambda(t) \cdot dW(t), \quad (4.7)$$

where W is a vector Brownian motion, supposed to drive fluctuations of the asset prices, and λ is the vector of unit risk-premiums. The interpretation is simple. In the absence of risk, $\xi(\tau) = \xi(t) e^{-r(\tau-t)}$ —the stochastic discount factor is simply the usual discount factor. However, in the presence of risk, the discount factor varies stochastically, driven by the same sources of variations that affect asset prices, $dW(t)$, unless of course some of these sources do not receive compensation, in which case some components of $\lambda(t)$ are zero.

To further our intuitive interpretation of λ as a vector of unit risk-premiums, note that the asset price, S , is obviously driven by the same Brownian motions driving r and ξ . Therefore, we have, $E_t \left(\frac{d\xi dS}{\xi S} \right) = -\text{Vol} \left(\frac{dS}{S} \right) \cdot \lambda dt$, where $\text{Vol} \left(\frac{dS}{S} \right)$ denotes the instantaneous volatility of asset returns. Therefore, the expression for the expected returns in Eq. (4.6) is:

$$E_t \left(\frac{dS}{S} \right) + \frac{D}{S} dt = r dt + \underbrace{\text{Vol} \left(\frac{dS}{S} \right)}_{\text{“betas”}} \cdot \underbrace{\lambda}_{\text{“lambdas”}} dt. \quad (4.8)$$

Expected returns are equal to the short-term rate, and a risk-premium related to the stochastic fluctuations of the asset prices, which is the product of the instantaneous risks related to the asset price fluctuations, $\text{Vol}\left(\frac{dS}{S}\right)$, times the unit risk-premiums that compensate for each individual source of these instantaneous risks. Eq. (4.8) is an APT relation, the continuous time counterpart to those developed in Chapter 1 of these Lectures. The only assumption made to achieve it was absence of arbitrage, via the assumption that a positive stochastic discounting factor or, alternatively, a pricing kernel exists, as in Eq. (4.4) and, then, in the continuous time case, as in Eq. (4.7). The next section aims to a further decomposition of the expected returns.

4.2.3 Risk-adjusted discount rates

How to discount future cash-flows? Do we have to rely on expected returns, such as those predicted by Eq. (4.8)? Expected returns are not necessarily risk-adjusted discount rates. It is a subtle point. Once dividends and asset prices are driven by a single factor, expected returns and risk-adjusted discount rates are the same, as we shall show. In general, the two concepts differ, however. To illustrate, we make a simplification, assuming that the price-dividend ratio, p say, is independent of the dividends D , and driven by a vector of state variables, y , such that:

$$S(y, D) = D \cdot p(y). \quad (4.9)$$

Such a “scale-invariant” property of the asset price arises in many economies, as discussed in more detail in Part II of these Lectures. For example, it arises if the state variables, y , do not depend on D , and if the dividends are a Geometric Brownian Motion with parameters g_0 and σ_D . In this case, the price function in Eq. (4.9) satisfies: $\frac{dS}{S} = \frac{dp}{p} + \frac{dD}{D}$ and, by Eq. (4.6), we have:

$$E_t \left(\frac{dS}{S} \right) + \frac{D}{S} dt = \text{Disc} dt - E_t \left(\frac{dp}{p} \frac{d\xi}{\xi} \right), \quad (4.10)$$

where Disc are defined as follows,

$$\begin{aligned} \text{Disc} &\equiv r - \frac{E_t \left(\frac{dD}{D} \frac{d\xi}{\xi} \right)}{dt} \\ &= r + \underbrace{\text{Vol} \left(\frac{dD}{D} \right)}_{\text{“cash-flow beta”}} \cdot \underbrace{\lambda_{\text{CF}}}_{\text{“cash-flow lambda”}}, \end{aligned}$$

with $\text{Vol}\left(\frac{dD}{D}\right) \equiv \sigma_D$ and λ_{CF} denoting the unit-risk premium required to compensate for the stochastic fluctuations of the dividends.

We refer to Disc as the *risk-adjusted discount rates*. They equal the safe interest rate r , plus the premium, $-E_t \left(\frac{dD}{D} \frac{d\xi}{\xi} \right)$, arising to compensate agents for the stochastic fluctuations of dividends. Eq. (4.10) tells us that if the price-dividend ratio is constant, $E_t \left(\frac{dp}{p} \frac{d\xi}{\xi} \right) = 0$, the risk adjusted discount rates are the same as the expected returns, just as in the one-factor Lucas economy discussed in Chapter 3.

If we assume that the aggregate dividend of this model, D , is consumption, then, σ_0 would be too small, alone, to make the expected returns predicted by this model in line with the data. It is part of the equity premium puzzle, as we shall explain in Part II of these Lectures. But Eq. (4.10) reveals that expected returns might potentially be inflated, once the price-dividend

ratio is driven by additional state variables, and provided of course the circumstance that the risk inherent in these variables is compensated,

$$E_t \left(\frac{dp}{p} \frac{d\xi}{\xi} \right) = - \underbrace{\text{Vol} \left(\frac{dp}{p} \right)}_{\text{"price-betas"}} \cdot \underbrace{\lambda_p}_{\text{"price lambdas"}} dt. \quad (4.11)$$

This term introduces a wedge between expected returns and the risk-adjusted discount rates of Eq. (4.10), and brings the potential to mitigate the equity premium puzzle. Time variation in risk-adjusted discount rates is quite important anyway, especially when it comes to the explanation of return *volatility*, as further explained in Chapter 7. To anticipate, note that the return volatility relates to the term labeled “price-betas” in Eq. (4.11) and as a result, to the sensitivity of the price-dividend ratio to changes in the state variables. We can illustrate this property by re-expressing the price-dividend ratio in Eq. (4.3), in terms of the risk-adjusted discount rates,

$$p(y(t)) = \mathbb{E} \left[\int_t^\infty \frac{D_*(\tau)}{D(t)} \cdot e^{-\int_t^\tau \text{Disc}(y(u)) du} \middle| y(t) \right], \quad \frac{D_*(\tau)}{D(t)} = e^{(g_0 - \frac{1}{2}\sigma_D^2)(\tau-t) + \sigma_D(\hat{W}(\tau) - \hat{W}(t))}, \quad (4.12)$$

where \mathbb{E} and $\hat{W}(\tau)$ denote the expectation and a Brownian motion under the risk-neutral probability. Eq. (4.12) is a present value formula where a fictitious risk-unadjusted dividend growth, $D_*(\tau)/D(t)$, is discounted using the risk-adjusted discount rates, $\text{Disc}(y(t))$. Therefore, the sensitivity of the price-dividend ratio to the state variables depends on that of the risk-adjusted discount rates. If risk-adjusted rates exhibit large swings, the return volatility predicted by Eq. (4.12) is likely to exhibit some of the interesting countercyclical statistics that we see in the data, as we shall explain in Chapter 7.

4.3 An introduction to methods or, the origins: Black & Scholes

4.3.1 Time

By definition, long-lived assets do not have an expiration date. Therefore, as Eq. (4.12) reveals, their prices do not explicitly depend on calendar time, once we assume that the state variables driving them do not explicitly depend on time. Naturally, it does not mean that these prices cannot be time-varying, or stochastic. On the contrary, they may well be driven by stochastic variables, again as in Eq. (4.12). Instead, derivatives have an expiration date, and their current value needs to reflect the time left to maturity, when they will be worth the terminal payoff. The implication of an expiration date is that derivative prices are solutions to partial differential equations that involve the sensitivity of prices to calendar time. In the absence of any expiration date, asset prices are still solutions to partial differential equations, but their sensitivity to time is zero, once we assume that the state variables driving them do not explicitly depend on time. In the special case of a single state variable, these long-lived asset prices are, then, solutions to ordinary differential equations, as in the Lucas model with a single state variable of Section 4.3.

The next section develops the simplest possible setting where we can illustrate the main tools of continuous time finance, which is the Black and Scholes (1973) market leading to the evaluation formula for a European option. In the absence of arbitrage, the derivative price is solution to a certain partial differential equation, the solution of which, we can represent as a

conditional expectation taken under the risk-neutral probability. In Section 4.3.3, we provide the link between this risk-neutral probability and the original probability.

4.3.2 Asset prices as Feynman-Kac representations

4.3.2.1 Black & Scholes partial differential equations

Why are partial differential equations so important in finance? Suppose that the price of a stock follows a geometric Brownian motion:

$$\frac{dS(t)}{S(t)} = \mu dt + \sigma dW(t), \quad \mu, \sigma > 0,$$

and that there exists a riskless accounting technology making spare money evolve as:

$$\frac{dB(t)}{B(t)} = r dt,$$

where $r > 0$. Finally, suppose that there exists another asset, a “call option,” which gives rise to a payoff equal to $(S(T) - K)^+$ at some future date T , where K is the “strike,” or exercise price of the option. Let $c(t)$, $t \in [0, T]$, be the price process of the option. We wish to figure out what this price looks like by formulating as few assumptions as possible. We ignore dividend issues, and assume there are no transaction costs, and rule out any other forms of frictions. We assume *rational expectations*, that is, there exists a function $f \in \mathcal{C}^{1,2}([0, T] \times \mathbb{R}_{++}) : c(t) = f(t, S(t))$. By the previous assumption and Itô’s lemma,

$$dc = (Lf) dt + f_S \sigma S dW,$$

where Lf denotes the so called “infinitesimal generator,” and is defined as $Lf = \frac{\partial f}{\partial t} + \frac{1}{2} \sigma^2 S^2 f_{SS} + \mu S f_S$ and subscripts denote partial derivatives. Next, we create the following portfolio: α units of the risky asset and β units of the riskless accounting technology. Section 4.3 explains that for any portfolio strategy to be “self-financed,” the value V of the resulting portfolio, $V(t) = \alpha(t) S(t) + \beta(t) B(t)$, must be such that:

$$\begin{aligned} dV(t) &= \alpha(t) dS(t) + \beta(t) dB(t) \\ &= (\alpha(t) \mu S(t) + r \beta(t) B(t)) dt + \alpha(t) \sigma S(t) dW(t). \end{aligned}$$

Next, set $V_0 = c_0$ and find $\hat{\alpha}, \hat{\beta}$ such that drift and diffusion terms of V and c are the same. This is done with $\hat{\alpha} = f_S$. Replace this into the previous stochastic differential equation. we have:

$$dV(t) = (\mu S f_S + r \beta B) dt + f_S \sigma S dW.$$

Now find $\hat{\beta} : \text{drift}(V) = \text{drift}(c)$, which after simple calculations is:

$$\hat{\beta} = \frac{Lf - \mu f_S S}{r B}.$$

Since $V_0 = c_0$ and the previous $\hat{\alpha}, \hat{\beta}$ make drifts and diffusion terms of V and c the same, then, by the *unique decomposition property* applying to stochastic differential equations stated in Appendix 1, we have that $V(t) = c(t)$, or:

$$f = c = \hat{V} \equiv \hat{\alpha} S + \hat{\beta} B = f_S S + \frac{Lf - \mu f_S S}{r}.$$

By the definition of L and rearranging,

$$\frac{\partial f}{\partial t} + \frac{1}{2}\sigma^2 S^2 f_{SS} + rSf_S - rf = 0 \quad \forall (t, S) \in [0, T) \times \mathbb{R}_{++}, \quad (4.13)$$

with the “boundary condition” $f(T, S) = (S - K)^+$, $\forall S \in \mathbb{R}_{++}$. This is an example of a Partial Differential Equation. The “unknown” is a function f , which has to be such that it and its partial derivatives are plugged into the left hand side of the first line, we obtain zero. Moreover, the same functions must pick up the boundary condition. The solution to this is the celebrated Black and Scholes (1973) formula.

4.3.2.2 A digression on more general partial differential equations

The Black-Scholes equation (4.13) is a typical (in fact the first) example of partial differential equations in finance. It leads to an equation of the so-called parabolic type, as we shall explain soon. More generally, let us be given,

$$a_0 + a_1 F_t + a_2 F_S + a_3 F_{SS} + a_4 F_{tt} + a_5 F_{St} = 0, \quad (4.14)$$

subject to some boundary condition. This partial differential equation is called: (i) elliptic, if $a_5^2 - 4a_3a_4 < 0$; (ii) parabolic, if $a_5^2 - 4a_3a_4 = 0$; (iii) hyperbolic, if $a_5^2 - 4a_3a_4 > 0$. The typical partial differential equations arising in finance are of the parabolic type. For example, the Black-Scholes function $F = e^{rt}f$ is parabolic. The following section explains how to provide a probabilistic representation to these parabolic partial differential equations.

4.3.2.3 Absence of arbitrage opportunities

Suppose that $c_0 > f_0 = \hat{\alpha}_0 S_0 + \hat{\beta}_0 B_0$. Then sell the option for c_0 , invest $\hat{\alpha}_0 S_0 + \hat{\beta}_0 B_0$, follow the $(\hat{\alpha}(t), \hat{\beta}(t))$ -trading strategy until time T and at time T , obtain $(S(T) - K)^+$ from the portfolio - which is exactly what is due to the buyer of the option. This generates a riskless profit $= c_0 - f_0$ without further expenses in the future - Recall, the $\hat{\alpha}, \hat{\beta}$ strategy is self-financing and there are no transaction costs). This is an arbitrage opportunity.

Suppose then the opposite, i.e. that $c_0 < f_0 = \hat{\alpha}_0 S_0 + \hat{\beta}_0 B_0$. Then buy the option for c_0 and hence claim for $(S(T) - K)^+$ at T . At the same time, “short-sell” the portfolio for $f_0 = \hat{\alpha}_0 S_0 + \hat{\beta}_0 B_0$, and subsequently update the short-selling through the strategy $\hat{\alpha}, \hat{\beta}$. The short-selling position is $(-\hat{\alpha}(t))S(t) + (-\hat{\beta}(t))B(t)$ for all t . At time T , the same short-selling position is $(-\hat{\alpha}(T))S(T) + (-\hat{\beta}(T))B(T) = -(\hat{\alpha}(T)S(T) + \hat{\beta}(T)B(T)) = -(S(T) - K)^+$. This amount of money is exactly the payoff of the option purchased at time zero. Thus use the option payoff to close the short selling position. The whole strategy generate a riskless profit $= f_0 - c_0$ without further expenses in the future. This is an arbitrage opportunity.

Therefore, absence of arbitrage opportunities are ruled out with $c_0 = f_0$. Note, the previous argument does not hinge upon the existence of a market for the option during the life of the option.

4.3.2.4 Feynman-Kac solutions

The typical situation that we encounter in finance is that the asset price is a function F that solves a parabolic partial differential equation, i.e. a special case of Eq. (4.14):

$$-r(x, t)F(x, t) + F_t(x, t) + \mu(x, t)F_x(x, t) + \frac{1}{2}\sigma^2(x, t)F_{xx}(x, t) = 0, \quad \forall (t, x) \in [0, T) \times \mathbb{R}, \quad (4.15)$$

with the boundary condition, $F(x, T) = g(x, T)$ for all x , and the function g is the final payoff. Somehow surprisingly, define, now, a stochastic differential equation, with drift and diffusion μ and σ in Eq. (10.24),

$$dZ(t) = \mu(Z(t), t) dt + \sigma(Z(t), t) dW(t), \quad Z_0 = x. \quad (4.16)$$

where $W(t)$ is a Brownian motion. Under regularity conditions on μ, σ, r , the solution F to Eq. (10.24) is:

$$F(x, t) = E \left[e^{-\int_t^T r(Z(s), s) ds} g(Z(T), T) \Big| z(t) = x \right], \quad (4.17)$$

where Z is solution to Eq. (4.16), and the expectation is taken with respect to the distribution of Z in Eq. (4.16). Note that the existence of the Feynman-Kac representation does not ensure per se the existence of a solution to a given partial differential equation.

Eq. (4.17) can be used to represent the solution to the Black & Scholes partial differential equation (4.13), with “auxiliary” stochastic differential equation (4.16) collapsing to,

$$\frac{dS(t)}{S(t)} = r dt + \sigma d\tilde{W}(t),$$

where $\tilde{W}(t)$ is a Brownian motion, necessarily defined under the risk-neutral probability, due to the drift of $\frac{dS}{S}$ being equal to the risk-free rate, r . That is, by Eq. (4.17), the price of an option in a Black & Scholes market is the risk-neutral expectation of the final payoff, discounted at the risk-free rate.

The Feynman-Kac representation of the solution to partial differential equations is quite useful. First, computing expectations is generally both easier and more intuitive than finding a solution to partial differential equations through guess and trial. Second, except for specific cases, the solution to asset prices is unknown, and a natural way to cope with this problem is to go for Monte-Carlo methods—approximation of the expectation in Eq. (4.17) through simulations and use of the law of large numbers. Finally, the Feynman-Kac representation theorem is useful for some theoretical reasons we shall see later in this chapter.

4.3.2.5 A few heuristic proofs

It is well-beyond the purpose of this appendix to develop detailed proofs of the Feynman-Kac representation theorem. In addition to Karatzas and Shreve (1991, p.366), an excellent source of reference is still Friedman (1975), which relaxes many sufficient conditions given in Karatzas and Shreve through opportune localizations of linear and growth conditions. The heuristic proof provided below covers the slightly more general case in which

$$Lf + q - rf = 0, \quad (4.18)$$

with some boundary condition. Here q is some function $q(x, t) \equiv q_t$. The typical role of q is the one of instantaneous dividend rate promised by the asset. As usual, $Lf = \frac{\partial f}{\partial t} + \mu f_x + \frac{1}{2}\sigma^2 f_{xx}$.

So suppose there exists a solution to Eq. (12.31). To see what a Feynman-Kac representation of such a solution looks like in this case, define

$$y(t) \equiv e^{-\int_0^t r(u) du} f(t) + \int_0^t e^{-\int_0^u r(s) ds} q(u) du,$$

where again, $f(t) = f(t, z(t))$, and:

$$dz(t) = \mu(t) dt + \sigma(t) dW(t), \quad z_0 = x.$$

By Itô's lemma,

$$\begin{aligned} dy &= e^{-\int_0^t r(u) du} q dt - r e^{-\int_0^t r(u) du} f dt + e^{-\int_0^t r(u) du} df \\ &= e^{-\int_0^t r(u) du} q dt - r e^{-\int_0^t r(u) du} f dt + e^{-\int_0^t r(u) du} [(Lf) dt + f_z \sigma dW] \\ &= e^{-\int_0^t r(u) du} \underbrace{[Lf + q - rf]}_{=0} dt + f_z \sigma dW \\ &= e^{-\int_0^t r(u) du} f_z \sigma dW. \end{aligned}$$

Therefore $y(T) = y_0 + \int_0^T e^{-\int_0^t r(u) du} f_z \sigma dW$. Assuming $\sigma f_z \in \mathcal{H}^2$, then, y is a martingale, viz $y_0 = E(y(T))$. We have,

$$y(T) = e^{-\int_0^T r(t) dt} f(T) + \int_0^T e^{-\int_0^u r(s) ds} q(u) du, \quad \text{and } y_0 = f_0.$$

Hence,

$$f_0 = y_0 = E(y(T)) = E \left[e^{-\int_0^T r(t) dt} f(T) \right] + E \left[\int_0^T e^{-\int_0^u r(s) ds} q(u) du \right].$$

4.3.3 Girsanov theorem

4.3.3.1 Motivation again

Consider the Black & Scholes partial differential equation:

$$\frac{\partial f}{\partial t} + Lf - rf = 0, \quad \forall (t, S) \in [0, T] \times \mathbb{R}_{++},$$

with boundary condition, $f(T, S) = (S - K)^+$ for all $S \in \mathbb{R}_{++}$, where $Lf = \frac{\partial f}{\partial t} + \frac{1}{2} \sigma^2 S^2 f_{SS} + rSf_S$, and subscripts denote partial derivatives. By the Feynman-Kac theorem,

$$f(t) = f(t, S(t)) = e^{-r(T-t)} E_Q (S(T) - K)^+,$$

where the expectation E_Q is taken with respect to the probability Q , say, on the σ -field generated by $S(t)$, and $S(t)$ is solution to

$$\frac{dS(t)}{S(t)} = r dt + \sigma d\tilde{W}(t),$$

where \tilde{W} is a Brownian motion defined under Q . For obvious reasons, this probability is usually referred to as the *risk-neutral probability*.

As it turns out, the probability Q is related to the physical probability B on the σ -field generated by $S(t)$, where $S(t)$ is solution to:

$$\frac{dS(t)}{S(t)} = \mu dt + \sigma dW(t),$$

and W is now a Brownian motion defined under B . To see heuristically that this is true, let us consider the following equalities:

$$\begin{aligned} E_Q (S(T) - K)^+ &= \int [S(T)(\omega) - K]^+ Q(d\omega) \\ &= \int \zeta(T)(\omega) [S(T)(\omega) - K]^+ B(d\omega) \\ &\equiv E [\zeta(T)(\omega) (S(T) - K)^+], \end{aligned}$$

where $\zeta(T) \equiv \frac{Q(d\omega)}{P(d\omega)}$, and $\zeta(t)$ is solution to:

$$\frac{d\zeta(t)}{\zeta(t)} = -\lambda dW(t), \quad \zeta_0 = 1,$$

and λ is necessarily equal to, $\lambda = \frac{\mu-r}{\sigma}$.

Indeed, let $y(t) \equiv \zeta(t) f(t)$. We have, $f_0 = e^{-rT} E [\zeta(T) (S(T) - K)^+]$. Because $f(T) = (S(T) - K)^+$, we have that:

$$y_0 = f_0 = e^{-rT} E (y(T)).$$

That is, $e^{-rt}y(t)$ is a P -martingale. Moreover, by Itô's lemma:

$$\begin{aligned} d(e^{-rt}y) &= -re^{-rt}ydt + e^{-rt}dy \\ &= e^{-rt}\zeta \left(\frac{\partial f}{\partial t} + \frac{1}{2}\sigma^2 S^2 f_{SS} + \mu S f_S - \lambda \sigma S f_S - r f \right) dt + e^{-rt}\zeta (\sigma S f_S - \lambda f) dW. \end{aligned}$$

Under the usual pathwise integrability conditions, this is a martingale when,

$$0 = \frac{\partial f}{\partial t} + \frac{1}{2}\sigma^2 S^2 f_{SS} + \mu S f_S - \lambda \sigma S f_S - r f. \quad (4.19)$$

On the other hand, we know that f is solution of the Black-Scholes partial differential equation:

$$0 = \frac{\partial f}{\partial t} + \frac{1}{2}\sigma^2 S^2 f_{SS} + r S f_S - r f. \quad (4.20)$$

Comparing Eq. (4.19) with Eq. (4.20) reveals that the representation $E_P [\zeta(T) (S(T) - K)^+]$ is possible with, $\lambda = \frac{\mu-r}{\sigma}$, as originally claimed. λ has the simple interpretation of *unit risk-premium* for investing in stocks.

The point of the previous computations is that it looks like as if we could start from the original probability space under which

$$\frac{dS}{S} = \mu dt + \sigma dW, \quad (4.21)$$

and, then, we could define a new Brownian motion $d\tilde{W} = dW + \lambda dt$, such that Eq. (4.21) can be written as, $\frac{dS}{S} = (\mu - \lambda\sigma) dt + \sigma d\tilde{W} = r dt + \sigma d\tilde{W}$, under some new probability space. And vice-versa. We formalize this idea in the next subsection, although the following clarification is in order. The definition of Brownian motion you were originally given obviously depends on the underlying probability measure P . As an example, for the definition of the independent, stationary increments of a Brownian motion and for its Gaussian distribution, we must know the probability measure on the σ -field \mathcal{F} . Usually, we do not pay attention to this fact, although this very same fact can be crucial as the previous example demonstrates.

4.3.3.2 The theorem

Consider two probability measures Q_1 and Q_2 , where $Q_1(A) = \int_A \zeta(\omega) Q_2(d\omega)$, for $A \in F$, and the Radon-Nikodym derivative, $\frac{dQ_1}{dQ_2}$. Let $W(t)$ be a Brownian motion under the probability P , and λ be some process satisfying the so-called Novikov's condition: $E[\exp(\frac{1}{2} \int_0^T \|\lambda(t)\|^2 dt)] < \infty$. Under regularity conditions, there exists another probability Q , equivalent to P , displaying the following properties:

- (i) The Radon-Nikodym derivative is $\frac{dQ}{dP} = \zeta(T) = \exp\left(-\frac{1}{2} \int_0^T \|\lambda(t)\|^2 dt - \int_0^T \lambda(t) dW(t)\right)$.
- (ii) $\tilde{W}(t) = W(t) + \int_0^t \lambda(s) ds$ is a Brownian motion under the probability Q .

To develop some intuition, consider the following example. Suppose that some random variable, x , is standard normal: $P(dx) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}x^2) dx$, and that, next, we tilt that distribution by a factor $\zeta(x) = \exp(-\frac{1}{2}\lambda^2 - \lambda x)$. Precisely, the new random variable has distribution, $Q(dx) = \zeta(x) P(dx) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}x^2 - \frac{1}{2}\lambda^2 - \lambda x) dx = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}\tilde{x}^2) d\tilde{x}$, where $\tilde{x} = x + \lambda$. Note that the new density Q is still normal with unit variance. Yet under this new probability, it is $\tilde{x} = x + \lambda$ to have zero expectation. In other words, we have that under Q , \tilde{x} is standard normal, or that alternatively, x is normal with unit variance but drift $-\lambda$.

The fact that changing probability does not lead to change volatility is a well known fact in continuous-time finance where asset prices are driven by Brownian motions. This property does not need to hold in other models, or even in discrete time settings. The typical counterexample is that of a binomial distribution, as in the infinite horizon tree model of Chapter 7, and in all the trees dealt with in Chapter 13.

We conclude this section by discussing a few technical details. The Novikov condition is needed for a variety of reasons. Technically, we need it to ensure that $E\left(\frac{dQ}{dP}\right) = E(\zeta(T)) = E[\exp(-\frac{1}{2} \int_0^T \|\lambda(t)\|^2 dt - \int_0^T \lambda(t) dW(t))] = 1 \Leftrightarrow \int dQ = 1$. This condition rules out extremely ill-behaved λ s which could not allow the equality $\int dQ = 1$ to hold. Thus, it ensures Q is indeed a probability. We may also define the Radon-Nikodym density process of $\frac{dQ}{dP}$. First, some intuition. Suppose we have a claim at time T . Heuristically, we have

$$E_Q[X(T)] = \int X(T) dQ = \int X(T) \left(\frac{dQ}{dP}\right) dP = E_P\left[\left(\frac{dQ}{dP}\right) X(T)\right] = E_P[\zeta(T) X(T)].$$

Similarly, we can “update” the previous formula as time unfolds. The formula to use is,

$$E_Q[X(T) | \mathcal{F}(t)] = \zeta(t)^{-1} E_P[\zeta(T) X(T)],$$

where

$$\frac{d\zeta(t)}{\zeta(t)} = -\lambda(t) dW(t), \quad \zeta_0 = 1.$$

4.4 An introduction to no-arbitrage and equilibrium

4.4.1 Self-financed strategies

A self-financed portfolio leads to a situation where the change in value of the portfolio between two instants t and $t + dt$ is computed as a mark-to-market P&L: the change in the asset prices

times the quantities of the same assets held at time t : there is no injection or withdrawal of funds between any two instants. For example, let θ_1 and S be the number of shares and the price of some risky asset, and θ_2 and b be the number of some riskless assets and its price. Then, the value of a self-financed portfolio, $V = S\theta_1 + \theta_2 b$, satisfies:

$$dV = \theta_1 dS + \theta_2 db = \pi \left(\frac{dS}{S} - \frac{db}{b} \right) + \frac{db}{b} V,$$

where $\pi \equiv S\theta_1$ and the second equality follows by simple computations. If the portfolio strategy involves risky assets distributing a dividend process, and consumption, then, in Appendix 1, we show that value of the self-financed portfolio satisfies:

$$dV(\tau) = \left(\frac{dS(\tau)}{S(\tau)} + \frac{D(\tau)}{S(\tau)} d\tau - r d\tau \right) \pi(\tau) + rV d\tau - c(\tau) d\tau, \quad (4.22)$$

where $D(\tau)$ is the dividend process.

4.4.2 No-arbitrage in Lucas tree

Let us consider the Lucas (1978) model with one tree and a perishable good taken as the numéraire. We assume that the dividend process is solution to:

$$\frac{dD}{D} = \mu_D d\tau + \sigma_D dW,$$

for two positive constants μ_D and σ_D . We assume no-sunspots, and denote the rational pricing function with $S \equiv S(D)$. By Itô's lemma,

$$\frac{dS}{S} = \mu_S d\tau + \sigma_S dW,$$

where

$$\mu_S = \frac{\mu_D D S'(D) + \frac{1}{2} \sigma_D^2 D^2 S''(D)}{S(D)}; \quad \sigma_S = \frac{\sigma_D D S'(D)}{S(D)}.$$

Then, by Eq. (4.22), the value of wealth satisfies,

$$dV = \left[\pi \left(\mu_S + \frac{D}{S} - r \right) + rV - c \right] d\tau + \pi \sigma_S dW.$$

Below, we shall show that in the absence of arbitrage, there must be some process λ , the “unit risk-premium”, such that,

$$\mu_S + \frac{D}{S} - r = \lambda \sigma_S. \quad (4.23)$$

Let us assume that the short-term rate, r , and the risk-premium, λ , are both constant. Below, we shall show that such an assumption is compatible with a general equilibrium economy. By the definition of μ_S and σ_S , Eq. (4.23) can be written as,

$$0 = \frac{1}{2} \sigma_D^2 D^2 S''(D) + (\mu_D - \lambda \sigma_D) D S'(D) - r S(D) + D. \quad (4.24)$$

Eq. (4.24) is a second order differential equation. Its solution, provided it exists, is the rational price of the asset. To solve Eq. (4.24), we initially assume that the solution, S_F say, takes the following simple form,

$$S_F(D) = K \cdot D, \quad (4.25)$$

where K is a constant to be determined. Next, we verify that this is indeed *one* solution to Eq. (4.24). Indeed, if Eq. (4.25) holds, then, by plugging this guess and its derivatives into Eq. (4.24) leaves, $K = (r - \mu_D + \lambda\sigma_D)^{-1}$ and, hence,

$$S_F(D) = \frac{1}{r + \lambda\sigma_D - \mu_D} D. \quad (4.26)$$

This is a Gordon-type formula. It merely states that prices are risk-adjusted expectations of future expected dividends, where the risk-adjusted discount rate is given by $r + \lambda\sigma_D$. Hence, in a comparative statics sense, stock prices are inversely related to the risk-premium, a quite intuitive conclusion.

Eq. (4.26) can be thought to be the Feynman-Kac representation to Eq. (4.24), viz

$$S_F(D(t)) = \mathbb{E}_t \left[\int_t^\infty e^{-r(\tau-t)} D(\tau) d\tau \right], \quad (4.27)$$

where $\mathbb{E}_t[\cdot]$ is the conditional expectation taken under the risk neutral probability Q (say), the dividend process follows,

$$\frac{dD}{D} = (\mu_D - \lambda\sigma_D) d\tau + \sigma_D d\tilde{W},$$

and $\tilde{W}(\tau) = W(\tau) + \lambda(\tau - t)$ is a another standard Brownian motion defined under Q . Formally, the true probability, P , and the risk-neutral probability, Q , are tied up by the Radon-Nikodym derivative,

$$\zeta = \frac{dQ}{dP} = e^{-\lambda(W(\tau) - W(t)) - \frac{1}{2}\lambda^2(\tau-t)}. \quad (4.28)$$

4.4.3 Equilibrium with CRRA

How do precisely preferences affect asset prices? In Eq. (4.26), the asset price relates to the interest rate, r , and the risk-premium, λ . But in equilibrium, agents preferences affect r and λ . However, such an impact can have a non-linear pattern. For example, when the risk-aversion is low, a small change of risk-aversion can make the interest rate and the risk-premium change in the same direction. If the risk-aversion is high, the effects may be different, as the interest rate reflects a variety of factors, including precautionary motives.

To illustrate these features within the simple case of CRRA preferences, let us rewrite, first, Eq. (4.22) under the risk-neutral probability Q . We have,

$$dV = (rV - c) d\tau + \pi\sigma_S d\tilde{W}. \quad (4.29)$$

We assume that the following transversality condition holds,

$$\lim_{\tau \rightarrow \infty} \mathbb{E}_t [e^{-r(\tau-t)} V(\tau)] = 0. \quad (4.30)$$

By integrating Eq. (4.29), and using the previous transversality condition,

$$V(t) = \mathbb{E}_t \left[\int_t^\infty e^{-r(\tau-t)} c(\tau) d\tau \right]. \quad (4.31)$$

By comparing Eq. (4.27) with Eq. (4.31) reveals that the equilibrium in the real markets, $D = c$, also implies that $S = V$. Next, rewrite (4.31) as,

$$V(t) = \mathbb{E}_t \left[\int_t^\infty e^{-r(\tau-t)} c(\tau) d\tau \right] = E_t \left[\int_t^\infty m_t(\tau) c(\tau) d\tau \right],$$

where

$$m_t(\tau) \equiv \frac{\xi(\tau)}{\xi(t)} = e^{-(r+\frac{1}{2}\lambda^2)(\tau-t)-\lambda(W(\tau)-W(t))}.$$

We assume that a representative agent solves the following intertemporal optimization problem,

$$\max_c E_t \left[\int_t^\infty e^{-\rho(\tau-t)} u(c(\tau)) d\tau \right] \quad \text{s.t.} \quad V(t) = E_t \left[\int_t^\infty m_t(\tau) c(\tau) d\tau \right] \quad [\text{P1}]$$

for some instantaneous utility function $u(c)$ and some subjective discount rate ρ .

To solve the program [P1], we form the Lagrangean

$$\mathbb{L} \equiv E_t \left[\int_t^\infty e^{-\rho(\tau-t)} u(c(\tau)) d\tau \right] + \ell \cdot \left[V(t) - E_t \left(\int_t^\infty m_t(\tau) c(\tau) d\tau \right) \right],$$

where ℓ is a Lagrange multiplier. The first order conditions are,

$$e^{-\rho(\tau-t)} u'(c(\tau)) = \ell \cdot m_t(\tau).$$

Moreover, by the equilibrium condition, $c = D$, and the definition of $m_t(\tau)$,

$$u'(D(\tau)) = \ell \cdot e^{-(r+\frac{1}{2}\lambda^2-\rho)(\tau-t)-\lambda(W(\tau)-W(t))}. \quad (4.32)$$

That is, by Itô's lemma,

$$\frac{du'(D)}{u'(D)} = \left[\frac{u''(D)D}{u'(D)} \mu_D + \frac{1}{2} \sigma_D^2 D^2 \frac{u'''(D)}{u'(D)} \right] d\tau + \frac{u''(D)D}{u'(D)} \sigma_D dW. \quad (4.33)$$

Next, let us define the right hand side of Eq. (8A.14) as $U(\tau) \equiv \ell \cdot e^{-(r+\frac{1}{2}\lambda^2-\rho)(\tau-t)-\lambda(W(\tau)-W(t))}$. By Itô's lemma, again,

$$\frac{dU}{U} = (\rho - r) d\tau - \lambda dW. \quad (4.34)$$

By Eq. (8A.14), drift and volatility components of Eq. (4.33) and Eq. (4.34) have to be the same. This is possible if

$$r = \rho - \frac{u''(D)D}{u'(D)} \mu_D - \frac{1}{2} \sigma_D^2 D^2 \frac{u'''(D)}{u'(D)}; \quad \text{and} \quad \lambda = -\frac{u''(D)D}{u'(D)} \sigma_D.$$

Let us assume that λ is constant. After integrating the second of these relations two times, we obtain that besides some irrelevant integration constant,

$$u(D) = \frac{D^{1-\eta} - 1}{1-\eta}, \quad \eta \equiv \frac{\lambda}{\sigma_D},$$

where η is the CRRA. Hence, under CRRA preferences we have that,

$$r = \rho + \eta \mu_D - \frac{1}{2} \eta (\eta + 1) \sigma_D^2, \quad \lambda = \eta \sigma_D.$$

Finally, by replacing these expressions for the short-term rate and the risk-premium into Eq. (4.26) leaves,

$$S(D) = \frac{1}{\rho - (1 - \eta) \left(\mu_D - \frac{1}{2} \eta \sigma_D^2 \right)} D,$$

provided the following conditions holds true:

$$\rho > (1 - \eta) \left(\mu_D - \frac{1}{2} \eta \sigma_D^2 \right). \quad (4.35)$$

We are only left to check that the transversality condition (4.30) holds at the equilibrium $S = V$. We have that under the previous inequality,

$$\begin{aligned} \lim_{\tau \rightarrow \infty} \mathbb{E}_t \left[e^{-r(\tau-t)} V(\tau) \right] &= \lim_{\tau \rightarrow \infty} \mathbb{E}_t \left[e^{-r(\tau-t)} S(\tau) \right] \\ &= \lim_{\tau \rightarrow \infty} E_t \left[m_t(\tau) S(\tau) \right] \\ &= \lim_{\tau \rightarrow \infty} E_t \left[e^{-(r + \frac{1}{2} \lambda^2)(\tau-t) - \lambda(W(\tau) - W(t))} S(\tau) \right] \\ &= S(t) \lim_{\tau \rightarrow \infty} E_t \left[e^{(\mu_D - \frac{1}{2} \sigma_D^2 - r - \frac{1}{2} \lambda^2)(\tau-t) + (\sigma_D - \lambda)(W(\tau) - W(t))} \right] \\ &= S(t) \lim_{\tau \rightarrow \infty} e^{-(r - \mu_D + \sigma_D \lambda)(\tau-t)} \\ &= S(t) \lim_{\tau \rightarrow \infty} e^{-(\rho - (1 - \eta)(\mu_D - \frac{1}{2} \eta \sigma_D^2))(\tau-t)} \\ &= 0. \end{aligned} \quad (4.36)$$

4.4.4 Bubbles

The transversality condition in Eq. (4.30) is often referred to as a *no-bubble condition*. To illustrate the reasons underlying this definition, note that Eq. (4.24) admits an infinite number of solutions. Each of these solutions takes the following form,

$$S(D) = KD + AD^\delta, \quad K, A, \delta \text{ constants.} \quad (4.37)$$

Indeed, by plugging Eq. (4.37) into Eq. (4.24) reveals that Eq. (4.37) holds if and only if the following conditions holds true:

$$0 = K(r + \lambda \sigma_D - \mu_D) - 1, \quad \text{and} \quad 0 = \delta(\mu_D - \lambda \sigma_D) + \frac{1}{2} \delta(\delta - 1) \sigma_D^2 - r. \quad (4.38)$$

The first condition implies that K equals the price-dividend ratio in Eq. (4.26), i.e. $K = S_F(D)/D$. The second condition leads to a quadratic equation in δ , with the two solutions,

$$\delta_1 < 0 \quad \text{and} \quad \delta_2 > 0.$$

Therefore, the asset price function takes the following form:

$$S(D) = S_F(D) + A_1 D^{\delta_1} + A_2 D^{\delta_2}.$$

It satisfies:

$$\lim_{D \rightarrow 0} S(D) = \mp \infty, \text{ if } A_1 \leq 0, \quad \lim_{D \rightarrow 0} S(D) = 0 \text{ if } A_1 = 0.$$

To rule out an explosive behavior of the price as the dividend level, D , gets small, we must set $A_1 = 0$, which leaves,

$$S(D) = S_F(D) + \mathcal{B}(D), \quad \mathcal{B}(D) \equiv A_2 D^{\delta_2}. \quad (4.39)$$

The component, $S_F(D)$, is the *fundamental value of the asset*, as by Eq. (4.27), it is the risk-adjusted present value of the expected dividends. The second component, $\mathcal{B}(D)$, is simply the difference between the market value of the asset, $S(D)$, and the fundamental value, $S_F(D)$. Hence, it is a bubble.

We seek conditions under which Eq. (4.39) satisfies the transversality condition in Eq. (4.30). We have,

$$\lim_{\tau \rightarrow \infty} \mathbb{E}_t [e^{-r(\tau-t)} S(\tau)] = \lim_{\tau \rightarrow \infty} \mathbb{E}_t [e^{-r(\tau-t)} S_F(D(\tau))] + \lim_{\tau \rightarrow \infty} \mathbb{E}_t [e^{-r(\tau-t)} \mathcal{B}(D(\tau))].$$

By Eq. (4.36), the fundamental value of the asset satisfies the transversality condition, under the condition given in Eq. (4.35). As regards the bubble, we have,

$$\begin{aligned} \lim_{\tau \rightarrow \infty} \mathbb{E}_t [e^{-r(\tau-t)} \mathcal{B}(D(\tau))] &= A_2 \cdot \lim_{\tau \rightarrow \infty} \mathbb{E}_t [e^{-r(\tau-t)} D(\tau)^{\delta_2}] \\ &= A_2 \cdot D(t)^{\delta_2} \cdot \lim_{\tau \rightarrow \infty} \mathbb{E}_t [e^{(\delta_2(\mu_D - \lambda\sigma_D) + \frac{1}{2}\delta_2(\delta_2-1)\sigma_D^2 - r)(\tau-t)}] \\ &= A_2 \cdot D(t)^{\delta_2}, \end{aligned} \quad (4.40)$$

where the last line holds as δ_2 satisfies the second condition in Eq. (4.38). Therefore, the bubble can not satisfy the transversality condition, except in the trivial case in which $A_2 = 0$. In other words, in this economy, the transversality condition in Eq. (4.30) holds if and only if there are no bubbles.

4.4.5 Reflecting barriers and absence of arbitrage

Next, suppose that insofar as the dividend $D(\tau)$ fluctuates above a certain level $\underline{D} > 0$, everything goes as in the previous section but that, as soon as the dividends level hits a “barrier” \underline{D} , it is “reflected” back with probability one. In this case, we say that the dividend follows a process with *reflecting barriers*. How does the price behave in the presence of such a barrier? First, if the dividend is above the barrier, $D > \underline{D}$, the price is still as in Eq. (4.37),

$$S(D) = \frac{1}{r - \mu_D + \lambda\sigma_D} D + A_1 D^{\delta_1} + A_2 D^{\delta_2}.$$

First, and as in the previous section, we need to set $A_2 = 0$ to satisfy the transversality condition in Eq. (4.30) (see Eq. (4.40)). However, in the new context of this section, we do not need to set $A_1 = 0$. Rather, this constant is needed to pin down the behavior of the price function $S(D)$ in the neighborhood of the barrier \underline{D} .

We claim that the following *smooth pasting* condition must hold in the neighborhood of \underline{D} ,

$$S'(\underline{D}) = 0. \quad (4.41)$$

This condition is in fact a no-arbitrage condition. Indeed, after hitting the barrier \underline{D} , the dividend is reflected back for the part exceeding \underline{D} . Since the reflection takes place with probability one, the asset is locally riskless at the barrier \underline{D} . However, the dynamics of the asset price is,

$$\frac{dS}{S} = \mu_S d\tau + \underbrace{\frac{\sigma_D DS'}{S}}_{\sigma_S} dW.$$

Therefore, the local risklessness of the asset at \underline{D} is ensured if $S'(\underline{D}) = 0$. [**Warning: We need to add some local time component here.**] Furthermore, rewrite Eq. (4.23) as,

$$\mu_S + \frac{D}{S} - r = \lambda \sigma_S = \lambda \frac{\sigma_D DS'(\underline{D})}{S(\underline{D})}.$$

If $D = \underline{D}$ then, by Eq. (4.41), $S'(\underline{D}) = 0$. Therefore,

$$\mu_S + \frac{\underline{D}}{S} = r.$$

This relations tells us that holding the asset during the reflection guarantees a total return equal to the short-term rate. This is because during the reflection, the asset is locally riskless and, hence, arbitrage opportunities are ruled out when holding the asset will make us earn no more than the safe interest rate, r . Indeed, by previous relation into the wealth equation (4.22), and using the condition that $\sigma_S = 0$, we obtain that

$$dV = \left[\pi \left(\mu_S + \frac{\underline{D}}{S} - r \right) + rV - c \right] d\tau + \pi \sigma_S dW = (rV - c) d\tau.$$

This example illustrates how the relation in Eq. (4.23) works to preclude arbitrage opportunities.

Finally, we solve the model. We have, $K \equiv S_F(D)/D$, and

$$0 = S'(\underline{D}) = K + \delta_1 A_1 \underline{D}^{\delta_1 - 1}; \quad \underline{Q} \equiv S(\underline{D}) = K \underline{D} + A_1 \underline{D}^{\delta_1},$$

where the second condition is the *value matching condition*, which needs to be imposed to ensure continuity of the pricing function with respect to D and, hence absence of arbitrage. The previous system can be solved to yield¹

$$\underline{Q} = \frac{1 - \delta_1}{-\delta_1} K \underline{D} \quad \text{and} \quad A_1 = \frac{K}{-\delta_1} \underline{D}^{1 - \delta_1}.$$

Note, the price is an increasing and convex function of the fundamentals, D .

4.5 Martingales and arbitrage

4.5.1 The information framework

We still consider a Lucas' type economy, but consider a finite horizon $T < \infty$. The primitives include a probability space (Ω, \mathcal{F}, P) . Let W be a standard Brownian motion in \mathbb{R}^d . Define

¹In this model, we take the barrier \underline{D} as given. In other context, we might be interested in “controlling” the dividend D in such a way that as soon as the price, q , hits a level \underline{Q} , the dividend level D is activate to induce the price q to increase. The solution for \underline{Q} reveals that this situation is possible when $\underline{D} = \frac{-\delta_1}{1 - \delta_1} K^{-1} \underline{Q}$, where \underline{Q} is an exogeneously given constant.

$\mathbb{F} = \{\mathcal{F}(t)\}_{t \in [0, T]}$ as the P -augmentation of the natural filtration $\mathcal{F}^W(\tau) = \sigma(W(s), s \leq \tau)$ generated by W , with $\mathcal{F} = \mathcal{F}(T)$.

We consider m trees and a monet market account. These assets, in addition to further assets in zero net supply, or “inside money” assets, to be introduced later, are exchanged without frictions. The trees entitle to receive the usual fruits, or dividends, $D_i(\tau)$, $i = 1, \dots, m$, which are positive $\mathcal{F}(\tau)$ -adapted bounded processes. Fruits are the numéraire. Let $S_+(\tau) = [S_0(\tau), \dots, S_m(\tau)]^\top$ be the positive $\mathcal{F}(\tau)$ -adapted asset price process. The price S_0 is that of a unit money market account, and satisfies: $S_0(\tau) = e^{\int_t^\tau r(u)du}$, where $r(\tau)$ is $\mathcal{F}(\tau)$ -adapted process satisfying $E(\int_t^T r(\tau)du) < \infty$. Moreover, we assume that

$$\frac{dS_i(\tau)}{S_i(\tau)} = \hat{a}_i(\tau)dt + \sigma_i(\tau)dW(\tau), \quad i = 1, \dots, m, \quad (4.42)$$

where $\hat{a}_i(\tau)$ and $\sigma_i(\tau)$ are processes satisfying the same properties as r , with $\sigma_i(\tau) \in \mathbb{R}^d$. We assume that $\text{rank}(\sigma(\tau; \omega)) = m \leq d$ a.s., where $\sigma(\tau) \equiv [\sigma_1(\tau), \dots, \sigma_m(\tau)]^\top$.

We assume that D_i is solution to

$$\frac{dD_i(\tau)}{D_i(\tau)} = a_{D_i}(\tau)d\tau + \sigma_{D_i}(\tau)dW(\tau),$$

where $a_{D_i}(\tau)$ and $\sigma_{D_i}(\tau)$ are $\mathcal{F}(\tau)$ -adapted, with $\sigma_{D_i} \in \mathbb{R}^d$.

A *strategy* is a predictable process in \mathbb{R}^{m+1} , denoted as: $[\theta_0(\tau), \dots, \theta_m(\tau)]^\top$, and satisfying $E(\int_t^T \|\theta(\tau)\|^2 d\tau) < \infty$. The value of a strategy, net of dividends, is: $V \equiv S_+ \cdot \theta$, where S_+ is a row vector. By generalizing Section 4.4.1, we say a strategy is self-financing if its value V , is the solution to:

$$dV = (\pi^\top (a - \mathbf{1}_m r) + Vr - c) dt + \pi^\top \sigma dW, \quad (4.43)$$

where $\mathbf{1}_m$ is a m -dimensional vector of ones, $\pi \equiv (\pi_1, \dots, \pi_m)^\top$, $\pi_i \equiv \theta_i S_i$, $i = 1, \dots, m$, $a \equiv (\hat{a}_1 + \frac{D_1}{S_1}, \dots, \hat{a}_m + \frac{D_m}{S_m})^\top$. The solution to the previous equation is, for each $\tau \in [t, T]$,

$$\frac{V^{x, \pi, c}(\tau)}{S_0(\tau)} = x - \int_t^\tau \frac{c(u)}{S_0(u)} du + \int_t^\tau \frac{\pi^\top(u) (a(u) - \mathbf{1}_m r(u))}{S_0(u)} du + \int_t^\tau \frac{\pi^\top(u) \sigma(u)}{S_0(u)} dW(u), \quad (4.44)$$

where x denotes the initial wealth. We require V to be strictly positive.

4.5.2 Viability

Let $\bar{g}_i = \frac{S_i}{S_0} + \bar{z}_i$, $i = 1, \dots, m$, where $d\bar{z}_i = \frac{1}{S_0} dz_i$ and $z_i(\tau) = \int_t^\tau D_i(u)du$. Let us generalize the definition of the risk-neutral probability in Eq. (4.28), and introduce the set \mathcal{Q} of *risk-neutral, or equivalent martingale, probabilities*, defined as:

$$\mathcal{Q} \equiv \{Q \approx P : \bar{g}_i \text{ is a } Q\text{-martingale}\}.$$

The aim of this section is to show the equivalent of Theorem 2.8 in Chapter 2: \mathcal{Q} is not empty if and only if there are not arbitrage opportunities.

Associated to every $\mathcal{F}(t)$ -adapted process $\lambda(t)$ satisfying some basic regularity conditions (essentially, the Novikov’s condition),

$$W_0(t) = W(t) + \int_t^\tau \lambda(u)du, \quad \tau \in [t, T], \quad (4.45)$$

is a standard Brownian motion under a probability Q which is equivalent to P , with Radon-Nikodym derivative equal to,

$$\zeta(T) \equiv \frac{dQ}{dP} = \exp\left(-\frac{1}{2} \int_t^T \|\lambda(\tau)\|^2 d\tau - \int_t^T \lambda^\top(\tau) dW(\tau)\right). \quad (4.46)$$

The process $(\eta(\tau))_{\tau \in [t, T]}$ is a martingale under P . This result is the celebrated *Girsanov's theorem*.

Now let us rewrite Eq. (4.42) under such a new probability by plugging W_0 in it. Under Q ,

$$\frac{dS_i(\tau)}{S_i(\tau)} = (\hat{a}_i(\tau) - \sigma_i(\tau) \lambda(\tau)) dt + \sigma_i(\tau) dW_0(\tau), \quad i = 1, \dots, m.$$

We also have

$$d\bar{g}_i(\tau) = d\left(\frac{S_i(\tau)}{S_0(\tau)}\right) + d\bar{z}_i(\tau) = \frac{S_i(\tau)}{S_0(\tau)} ((a_i(\tau) - r(\tau)) d\tau + \sigma_i(\tau) dW(\tau)).$$

If \bar{g}_i is a Q -martingale, i.e.

$$S_i(\tau) = E_\tau^Q \left[\frac{S_0(\tau)}{S_0(T)} S_i(T) + \int_\tau^T \frac{S_0(\tau)}{S_0(s)} D_i(s) ds \middle| \mathcal{F}(\tau) \right], \quad i = 1, \dots, m, \quad (4.47)$$

it is necessary and sufficient that $a_i - \sigma_i \lambda = r$, $i = 1, \dots, m$, or

$$a(\tau) - \mathbf{1}_m r(\tau) = \sigma(\tau) \lambda(\tau). \quad (4.48)$$

Therefore, by Eqs. (4.43), (4.45) and (4.48), we have that, for $\tau \in [t, T]$,

$$\frac{V^{x, \pi, c}(\tau)}{S_0(\tau)} = x - \int_t^\tau \frac{c(u)}{S_0(u)} du + \int_t^\tau \frac{\pi^\top(u) \sigma(u)}{S_0(u)} dW_0(u). \quad (4.49)$$

Consider the following definition:

DEFINITION 4.1 (Arbitrage opportunity). *A portfolio π is an arbitrage opportunity if $V^{x, \pi, 0}(t) \leq S_0^{-1}(T) V^{x, \pi, 0}(T)$ and $\Pr(S_0^{-1}(T) V^{x, \pi, 0}(T) - x > 0) > 0$.*

We have:

THEOREM 4.2. *There are no arbitrage opportunities if and only if \mathcal{Q} is not empty.*

A proof of this theorem is in the Appendix. The if part follows easily, by Eq. (4.49). The only if part is more elaborated, but its basic structure can be understood as follows. By the Girsanov's theorem, the statement "absence of arbitrage opportunities $\Rightarrow \exists Q \in \mathcal{Q}$ " is equivalent to "absence of arbitrage opportunities $\Rightarrow \exists \lambda$ satisfying Eq. (4.48)." If Eq. (4.48) didn't hold, one could implement an arbitrage, and find a nonzero $\underline{\pi} : \underline{\pi}^\top \sigma = 0$ and $\underline{\pi}^\top (a - \mathbf{1}_m r) \neq 0$. One could then use $\underline{\pi}$ when $a - \mathbf{1}_m r > 0$ and $-\underline{\pi}$ when $a - \mathbf{1}_m r < 0$, and obtain an appreciation rate of V greater than r in spite of having zeroed uncertainty through $\underline{\pi}^\top \sigma = 0$. If Eq. (4.48) holds, such an arbitrage opportunity would never occur, as in this case for each π , $\pi^\top (a - \mathbf{1}_m r) = \pi^\top \sigma \lambda$. Let

$$\langle \sigma^\top \rangle^\perp \equiv \{x \in L_{t, T, m}^2 : \sigma^\top x = \mathbf{0}_d\}$$

and

$$\langle \sigma \rangle \equiv \{z \in L_{t,T,m}^2 : z = \sigma u, \quad \text{for } u \in L_{t,T,d}^2\}.$$

Then, we may formalize the previous reasoning as follows. The excess return vector, $a - \mathbf{1}_m r$, must be orthogonal to all vectors in $\langle \sigma^\top \rangle^\perp$, and since $\langle \sigma \rangle$ and $\langle \sigma^\top \rangle^\perp$ are orthogonal, $a - \mathbf{1}_m r \in \langle \sigma \rangle$, or $\exists \lambda \in L_{t,T,d}^2 : a - \mathbf{1}_m r = \sigma \lambda$.²

4.5.3 Market completeness

Let $Y \in L^2(\Omega, \mathcal{F}, P)$. Consider the following definition:

DEFINITION 4.3 (Market completeness). *Markets are dynamically complete if for each random variable $Y \in L^2(\Omega, \mathcal{F}, P)$, we can find a portfolio process $\pi : V^{x,\pi,0}(T) = Y$ a.s.*

The previous definition is the natural continuous-time counterpart to that we gave in the discrete-time case (see Chapter 2). In analogy with the conclusions in Chapter 2, we shall prove that in continuous-time, markets are dynamically complete if and only if (i) $m = d$ and (ii) the price volatility matrix of the available assets (primitives and derivatives) is nonsingular. We shall provide a sketch of the proof for the sufficiency part of this statement (see, e.g., Karatzas (1997 pp. 8-9) for the converse), which relates to the existence of fully spanning dynamic strategies. So given a $Y \in L^2(\Omega, \mathcal{F}, P)$, let $m = d$ and suppose the volatility matrix σ is nonsingular. Let us consider the Q -martingale:

$$M(\tau) \equiv E^Q(S_0(T)^{-1} \cdot Y | \mathcal{F}(\tau)). \quad (4.50)$$

By the representation theorem of continuous local martingales as stochastic integrals with respect to Brownian motions (e.g., Karatzas and Shreve (1991) (thm. 4.2 p. 170)), there exists $\varphi \in L_{0,T,d}^2(\Omega, \mathcal{F}, Q)$ such that M can be written as:

$$M(\tau) = M(t) + \int_t^\tau \varphi^\top(u) dW_0(u).$$

We wish to find out a portfolio process π such that the discounted wealth process, net of consumption, $S_0^{-1}(\tau) V^{x,\pi,0}(\tau)$ equals $M(\tau)$ under P (or, equivalently, under Q) a.s. By Eq. (4.49),

$$\frac{V^{x,\pi,0}(\tau)}{S_0(\tau)} = x + \int_t^\tau \frac{\pi^\top(u) \sigma(u)}{S_0(u)} dW_0(u),$$

and so, by identifying, the portfolio we are looking for is $\hat{\pi}^\top = S_0 \varphi^\top \sigma^{-1}$. Set, then, $x = M(t)$. Then, $M(\tau) = S_0^{-1}(\tau) V^{M(t), \hat{\pi}, 0}(\tau)$, and in particular, $M(T) = S_0^{-1}(T) V^{M(t), \hat{\pi}, 0}(T)$ a.s. By comparing with Eq. (4.50), $V^{M(t), \hat{\pi}, 0}(T) = Y$.

Armed with this result, we can now easily state:

²To see that $\langle \sigma \rangle$ and $\langle \sigma^\top \rangle^\perp$ are orthogonal spaces, note that:

$$\begin{aligned} \left\{ x \in L_{t,T,m}^2 : x^\top z = 0, \quad z \in \langle \sigma \rangle \right\} &= \left\{ x \in L_{t,T,m}^2 : x^\top \sigma u = 0, \quad u \in L_{t,T,d}^2 \right\} \\ &= \left\{ x \in L_{t,T,m}^2 : x^\top \sigma = \mathbf{0}_d \right\} \\ &= \left\{ x \in L_{t,T,m}^2 : \sigma^\top x = \mathbf{0}_d \right\} \\ &\equiv \langle \sigma^\top \rangle^\perp. \end{aligned}$$

THEOREM 4.4. \mathcal{Q} is a singleton if and only if markets are complete.

PROOF. There exists a unique $\lambda : a - \mathbf{1}_m r = \sigma \lambda \iff m = d$. The result follows by the Girsanov's theorem. \parallel

When markets are incomplete, there is an infinity of risk-neutral probabilities belonging to \mathcal{Q} . Absence of arbitrage does not allow us to “recover” a unique risk-neutral probability, just as in the discrete time model of Chapter 2. One could make use of general equilibrium arguments, but in this case we go beyond the edge of knowledge, although we shall see something in Part II of these lectures on “Applied asset pricing theory.”

The next results, provide a further representation of the set of risk-neutral probabilities \mathcal{Q} , in the incomplete markets case. Let $L_{0,T,d}^2(\Omega, \mathcal{F}, P)$ be the space of all $\mathcal{F}(t)$ -adapted processes x in \mathbb{R}^d satisfying: $0 < \int_0^T \|x(u)\|^2 du < \infty$, and define,

$$\langle \sigma \rangle^\perp \equiv \{x \in L_{0,T,d}^2(\Omega, \mathcal{F}, P) : \sigma(t)x(t) = \mathbf{0}_m \text{ a.s.}\},$$

where $\mathbf{0}_m$ is a vector of zeros in \mathbb{R}^m . Let

$$\hat{\lambda} = \sigma^\top (\sigma \sigma^\top)^{-1} (a - \mathbf{1}_m r).$$

Under the usual regularity conditions, $\hat{\lambda}$ can be interpreted as the process of unit risk-premia. In fact, *all* processes belonging to the set:

$$\mathcal{Z} = \left\{ \lambda : \lambda(t) = \hat{\lambda}(t) + \nu(t), \eta \in \langle \sigma \rangle^\perp \right\}$$

are bounded and, hence, can be interpreted as unit risk-premia processes. More precisely, define the Radon-Nikodym derivative of Q with respect to P on $\mathcal{F}(T)$:

$$\hat{\zeta}(T) \equiv \frac{dQ}{dP} = \exp \left(-\frac{1}{2} \int_0^T \|\hat{\lambda}(t)\|^2 dt - \int_0^T \hat{\lambda}^\top(t) dW(t) \right),$$

and the density process of all $Q \approx P$ on (Ω, \mathcal{F}) ,

$$\zeta(t) = \hat{\zeta}(t) \cdot \exp \left(\frac{1}{2} \int_0^t \|\nu(u)\|^2 du - \int_0^t \nu^\top(u) dW(u) \right), \quad t \in [0, T],$$

a strictly positive P -martingale. We have the following results, which follows for example by He and Pearson (1991, Proposition 1 p. 271) or Shreve (1991, Lemma 3.4 p. 429):

PROPOSITION 4.5. $Q \in \mathcal{Q}$ if and only if it is of the form: $Q(A) = E(1_A \zeta(T)), \forall A \in \mathcal{F}(T)$.

To summarize, we have that $\dim(\langle \sigma \rangle^\perp) = d - m$. The previous result shows quite nitidly that markets incompleteness implies the existence of an infinity of risk-neutral probabilities. Such a result was shown in great generality by Harrison and Pliska (1983).³

³The so-called Föllmer and Schweizer (1991) measure, or minimal equivalent martingale measure, is defined as: $\hat{P}^*(A) \equiv E(1_A \hat{\zeta}(T))$, for each $A \in \mathcal{F}(T)$.

4.6 Equilibrium with a representative agent

4.6.1 Consumption and portfolio choices: martingale approaches

For now, we assume that markets are complete, $m = d$, and that there are no portfolio constraints or any other frictions. We consider the problem of an agent, who maximizes the expected utility from his consumption flows, $u(\cdot)$, plus the expected utility from terminal wealth, $U(\cdot)$, under the constraint in Eq. (4.44):⁴

$$J(0, V_0) = \max_{(\pi, c, v)} E \left[U(V^{x, \pi, c}(T)) + \int_t^T u(c(\tau)) d\tau \right], \quad \text{s.t. Eq. (4.44) holds.}$$

The first approach to solve this problem was introduced by Merton, which we shall see later. We wish to present another approach, which makes use of Arrow-Debreu state prices, similarly as in Chapter 2. Our first task is to derive a budget constraint paralleling the budget constraint in Chapter 2:

$$0 = c^0 - w^0 + E \left[m \cdot (c^1 - w^1) \right], \quad (4.51)$$

where c and w are consumption and endowments, and m is the discount factor m . In Chapter 2, such a budget constraint arises after having multiplied the initial budget constraint by the Arrow-Debreu state prices,

$$\phi_s = m_s \cdot P_s, \quad m_s \equiv (1+r)^{-1} \zeta_s, \quad \zeta_s = \frac{Q_s}{P_s},$$

and after “having taken the sum over all the states of nature”. We wish to apply the same logic here. First, we define Arrow-Debreu state price *densities*:

$$\phi_{t,T} \equiv m_{t,T} \cdot dP, \quad m_{t,T} = S_0(T)^{-1} \zeta(T), \quad \zeta(T) = \frac{dQ}{dP}. \quad (4.52)$$

As in the finite state space of Chapter 2, we multiply the budget constraint in Eq. (4.44) by these Arrow-Debreu densities, and then, we “take the integral over all states of nature.” The original problem, one with an infinity of trajectory constraints, will then be reduced to one with only one constraint, just as for the budget constraint in Eq. (4.51). Accordingly, multiply both sides in Eq. (4.44) by $\phi_{0,T} = S_0(T)^{-1} \cdot dQ$, and rearrange terms, to obtain:

$$0 = \left[\frac{V^{x, \pi, c}(T)}{S_0(T)} + \int_t^T \frac{c(u)}{S_0(u)} du - x \right] dQ - \left[\int_t^T \frac{(\pi^\top (a - \mathbf{1}_m r)) (u) du + (\pi^\top \sigma)(u) dW(u)}{S_0(u)} \right] dQ.$$

Next, take the integral over all states of nature. By the Girsanov’s theorem,

$$0 = \mathbb{E} \left[\frac{V^{x, \pi, c}(T)}{S_0(T)} + \int_t^T \frac{c(u)}{S_0(u)} du - x \right].$$

We can retrieve back the budget constraint under the probability P . We have, by a change of measure and computations in the Appendix, that:

$$x = \mathbb{E} \left[\frac{V^{x, \pi, c}(T)}{S_0(T)} + \int_t^T \frac{c(u)}{S_0(u)} du \right] = E \left[m_{t,T} \cdot V^{x, \pi, c}(T) + \int_t^T m_{t,u} \cdot c(u) du \right]. \quad (4.53)$$

⁴Moreover, we assume that the agent only considers the choice space in which the control functions satisfy the elementary Markov property and belong to $L^2_{0,T,m}(\Omega, \mathcal{F}, P)$ and $L^2_{0,T,1}(\Omega, \mathcal{F}, P)$.

So the program is,

$$\begin{aligned} J(t, x) &= \max_{(c, v)} E \left[e^{-\rho(T-t)} U(V^{x, \pi, c}(T)) + \int_t^T u(\tau, c(\tau)) d\tau \right], \\ \text{s.t. } x &= E \left[m_{t, T} \cdot V^{x, \pi, c}(T) + \int_t^T m_{t, \tau} \cdot c(\tau) d\tau \right]. \end{aligned}$$

Because of its emphasis on the equivalent martingale measure, this approach to solve the original problem is known as relying on *martingale methods*. Critically, market completeness is needed to use these methods, as in this case, there is one and only one Arrow-Debreu density process. However, the same martingale methods can be applied in the presence of portfolio constraints (which include incomplete markets as a special case) too, although in a slightly modified manner, as we shall see in Section 4.6.

To solve the problem, consider the Lagrangean,

$$\max_{(c, v)} E \left[\int_t^T [u(\tau, c(\tau)) - \psi \cdot m_{t, \tau} \cdot c(\tau)] d\tau + U(v) - \psi \cdot m_{t, T} \cdot v + \psi \cdot x \right],$$

where ψ is the constraint's multiplier, and by Eqs. (4.46) and (4.52),

$$m_{t, \tau} = \exp \left(- \int_t^\tau \left(r(u) + \frac{1}{2} \|\lambda(u)\|^2 \right) du - \int_t^\tau \lambda^\top(u) dW(u) \right). \quad (4.54)$$

The first order conditions are:

$$u_c(\tau, c(\tau)) = \psi \cdot m_{t, \tau}, \text{ for } \tau \in [t, T), \quad \text{and } U'(V^{x, \pi, c}(T)) = \psi \cdot m_{t, T}. \quad (4.55)$$

To compute the portfolio-consumption policy, note that for $c(\tau) \equiv 0$, the proof is just that leading to Theorem 4.4. In the general case, define,

$$M(\tau) \equiv E^Q \left[S_0^{-1}(T) \cdot \hat{v} + \int_t^T S_0(u)^{-1} \hat{c}(u) du \middle| \mathcal{F}(\tau) \right].$$

Notice that:

$$M(\tau) = E^Q \left[S_0^{-1}(T) \cdot \hat{v} + \int_t^T S_0(u)^{-1} \hat{c}(u) du \middle| \mathcal{F}(\tau) \right] = E \left[m_{t, T} \cdot \hat{v} + \int_t^T m_{t, u} \cdot \hat{c}(u) du \middle| \mathcal{F}(\tau) \right].$$

By the predictable representation theorem, $\exists \phi$ such that:

$$M(\tau) = M(t) + \int_t^\tau \phi^\top(u) dW(u).$$

Consider the process $\{m_{0, t} V^{x, \pi, c}(\tau)\}_{\tau \in [t, T]}$. By Itô's lemma,

$$m_{0, t} V^{x, \pi, c}(\tau) + \int_t^\tau m_{t, u} \cdot c(u) du = x + \int_t^\tau m_{t, u} \cdot (\pi^\top \sigma - V^{x, \pi, c} \lambda)(u) dW(u).$$

By identifying,

$$\pi^\top(\tau) = \left[V^{x, \pi, c}(\tau) \lambda(\tau) + \frac{\phi^\top(\tau)}{m_{t, \tau}} \right] \sigma^{-1}(\tau), \quad (4.56)$$

where $V^{x,\pi,c}(\tau)$ can be computed from the constraint:

$$V^{x,\pi,c}(\tau) = E \left[m_{\tau,T} \cdot v + \int_{\tau}^T m_{\tau,u} \cdot c(u) du \middle| \mathcal{F}(\tau) \right],$$

once that the optimal trajectory of c has been computed.

As an example, let $U(v) = \ln v$ and $u(x) = \ln x$. By the first order conditions (4.55), $\frac{1}{\hat{c}(\tau)} = \psi \cdot m_{t,\tau}$, $\frac{1}{\hat{v}} = \psi \cdot m_{\tau,T}$. By plugging these conditions into the constraint, one obtains the solution for the Lagrange multiplier: $\psi = \frac{T+1}{x}$. By replacing this back into the previous first order conditions, one eventually obtains: $\hat{c}(t) = \frac{x}{T+1} \frac{1}{m_{t,\tau}}$, and $\hat{v} = \frac{x}{T+1} \frac{1}{m_{t,T}}$. As regards the portfolio process, one has that:

$$M(\tau) = E \left[m_{\tau,T} \cdot \hat{v} + \int_t^T m_{t,u} \hat{c}(u) du \middle| \mathcal{F}(\tau) \right] = x,$$

which shows that $\phi = 0$ in the representation of Eq. (4.56). So by replacing $\phi = 0$ into (4.56),

$$\pi^\top(\tau) = V^{x,\pi,\hat{c}}(\tau) \lambda(\tau) \sigma^{-1}(\tau).$$

We can compute $V^{x,\pi,\hat{c}}$ in (4.14) by using \hat{c} :

$$V^{x,\pi,\hat{c}}(\tau) = \frac{x}{T+1} E \left[\frac{m_{\tau,T}}{m_{t,T}} + \int_{\tau}^T \frac{m_{\tau,u}}{m_{t,u}} du \middle| \mathcal{F}(\tau) \right] = \frac{x}{m_{t,\tau}} \frac{T+1 - (\tau - t)}{T+1},$$

where we used the property that m satisfies: $m_{t,a} \cdot m_{t,b} = m_{t,b}$, $t \leq a \leq b$. The solution is:

$$\pi^\top(\tau) = \frac{x}{m_{t,\tau}} \frac{T+1 - (\tau - t)}{T+1} \lambda(\tau) \sigma^{-1}(\tau)$$

whence, by taking into account the relation: $a - \mathbf{1}_m r = \sigma \lambda$,

$$\pi(\tau) = \frac{x}{m_{t,\tau}} \frac{T+1 - (\tau - t)}{T+1} [(\sigma \sigma^\top)^{-1} (a - \mathbf{1}_m r)](\tau).$$

4.6.2 The older, Merton's approach: dynamic programming

The Merton's approach derives optimal consumption and portfolio through Bellman's dynamic programming. Let us see how it works in the infinite horizon case. The problem the agent faces is:

$$\begin{aligned} J(V(t)) &= \max_c E \left[\int_t^\infty e^{-\rho(\tau-t)} u(c(\tau)) d\tau \right] \\ \text{s.t. } dV &= [\pi^\top (a - \mathbf{1}_m) + rV - c] d\tau + \pi^\top \sigma dW \end{aligned}$$

Under regularity conditions,

$$0 = \max_c E \left[u(c) + J'(V) (\pi^\top (a - \mathbf{1}_m r) + rV - c) + \frac{1}{2} J''(V) \pi^\top \sigma \sigma^\top \pi - \rho J(V) \right]. \quad (4.57)$$

The first order conditions lead to:

$$u'(c) = J'(v) \quad \text{and} \quad \pi = \left(\frac{-J'(V)}{J''(V)} \right) (\sigma \sigma^\top)^{-1} (a - \mathbf{1}_m r). \quad (4.58)$$

By plugging these expressions back to the Bellman's Equation (4.57) leaves:

$$0 = u(c) + J'(V) \left[\frac{-J'(V)}{J''(V)} \cdot Sh + rV - c \right] + \frac{1}{2} J''(V) \left[\frac{-J'(V)}{J''(V)} \right]^2 Sh - \rho J(V), \quad (4.59)$$

where:

$$Sh \equiv (a - \mathbf{1}_m r)^\top (\sigma \sigma^\top)^{-1} (a - \mathbf{1}_m r),$$

with $\lim_{T \rightarrow \infty} e^{-\rho(T-t)} E[J(V(T))] = 0$.

As an example, consider the CRRA utility $u(c) = (c^{1-\eta} - 1) / (1 - \eta)$. Conjecture that:

$$J(x) = A \frac{x^{1-\eta} - B}{1 - \eta},$$

where A, B are constants to be determined. Using the first condition in (4.58), leaves $c = A^{-1/\eta} V$. By plugging this expression into Eq. (4.59), and using the conjectured analytical form of J , we obtain:

$$0 = AV^{1-\eta} \left(\frac{\eta}{1-\eta} A^{-1/\eta} + \frac{1}{2} \frac{Sh}{\eta} + r - \frac{\rho}{1-\eta} \right) - \frac{1}{1-\eta} (1 - \rho AB).$$

This equation must hold for every V . Therefore

$$A = \left(\frac{\rho - r(1-\eta)}{\eta} - \frac{(1-\eta)Sh}{2\eta^2} \right)^{-\eta}, \quad B = \frac{1}{\rho} \left(\frac{\rho - r(1-\eta)}{\eta} - \frac{(1-\eta)Sh}{2\eta^2} \right)^\eta$$

Clearly, $\lim_{\eta \rightarrow 1} J(V) = \rho^{-1} \ln V$.

4.6.3 Equilibrium

In a complete markets setting, an equilibrium is (i) a consumption plan satisfying the first order conditions (4.55); (ii) a portfolio process having the form in Eq. (4.56), and (iii) the following market clearing conditions:

$$c(\tau) = D(\tau) \equiv \sum_{i=1}^m D_i(\tau), \text{ for } \tau \in [t, T], \quad \underline{q}(T) \equiv \sum_{i=1}^m S_i(T) \quad (4.60)$$

$$\theta_0(\tau) = 0, \quad \pi(\tau) = S(\tau), \text{ for } \tau \in [t, T]. \quad (4.61)$$

We now derive equilibrium allocations and Arrow-Debreu state price densities. First, note that the dividend process, D , satisfies:

$$dD(\tau) = a_D(\tau)D(\tau)d\tau + \sigma_D(\tau)D(\tau)dW(\tau),$$

where $a_D D \equiv \sum_{i=1}^m a_{D_i} D_i$ and $\sigma_D D \equiv \sum_{i=1}^m \sigma_{D_i} D_i$.

We have:

$$\begin{aligned} d \ln u_c(\tau, D(\tau)) &= d \ln u_c(\tau, c(\tau)) \\ &= d \ln m_{t,\tau} \\ &= - \left(r(\tau) + \frac{1}{2} \|\lambda(\tau)\|^2 \right) dt - \lambda^\top(\tau) dW(\tau), \end{aligned} \quad (4.62)$$

where the first equality holds in an equilibrium, the second equality follows by the first order conditions in (4.55), and the third equality is true by the definition of $m_{t,\tau}$ in Eq. (4.54).

Finally, by Itô's lemma, $\ln u_c(\tau, D(\tau))$ is solution to:

$$d \ln u_c = \left[\frac{u_{\tau c}}{u_c} + a_D D \frac{u_{cc}}{u_c} + \frac{1}{2} \sigma_D^2 D^2 \left(\frac{u_{ccc}}{u_c} - \left(\frac{u_{cc}}{u_c} \right)^2 \right) \right] dt + \frac{u_{cc}}{u_c} D \sigma_D dW. \quad (4.63)$$

By identifying drifts and diffusion terms in Eqs. (4.62)-(4.63), we obtain, after a few simplifications, the expression for the equilibrium short term rate and the prices of risk:

$$\begin{aligned} r(\tau) &= - \left[\frac{u_{\tau c}(\tau, D(\tau))}{u_c(\tau, D(\tau))} + a_D(\tau) D(\tau) \frac{u_{cc}(\tau, D(\tau))}{u_c(\tau, D(\tau))} + \frac{1}{2} \sigma_D(\tau)^2 D(\tau)^2 \frac{u_{ccc}(\tau, D(\tau))}{u_c(\tau, D(\tau))} \right] \\ \lambda^\top(\tau) &= - \frac{u_{cc}(\tau, D(\tau))}{u_c(\tau, D(\tau))} \sigma_D(\tau) D(\tau). \end{aligned}$$

For example, consider the CRRA utility function, if $u(\tau, c) = e^{-(\tau-t)\rho} (c^{1-\eta} - 1) / (1 - \eta)$, and $m = 1$. Then,

$$r(\tau) = \rho + \eta a_D(\tau) - \frac{1}{2} \eta(\eta + 1) \sigma_D(\tau)^2, \quad \lambda(\tau) = \eta \sigma_D(\tau).$$

Appendix 2 performs Walras's consistency tests: Eq. (4.60) \iff Eq. (4.61).

4.6.4 Continuous time Consumption-CAPM

By Eq. (4.47),

$$\begin{aligned} S_i(\tau) &= E^Q \left[\frac{S_0(\tau)}{S_0(T)} S_i(T) + \int_\tau^T \frac{S_0(\tau)}{S_0(s)} D_i(s) ds \middle| \mathcal{F}(\tau) \right] \\ &= E \left[\frac{m_{t,T}}{m_{t,\tau}} S_i(T) + \int_\tau^T \frac{m_{t,s}}{m_{t,\tau}} D_i(s) ds \middle| \mathcal{F}(\tau) \right], \end{aligned}$$

where the second line follows by the same arguments leading to Eq. (4.53). Replacing the first order condition in (4.55), and the equilibrium conditions in Eq. (4.60), we obtain the consumption CAPM evaluation of each asset:

$$S_i(\tau) = E \left[\frac{u'(q(T))}{u'(D(\tau))} S_i(T) + \int_\tau^T \frac{u'(D(s))}{u'(D(\tau))} D_i(s) ds \middle| \mathcal{F}(\tau) \right], \quad i = 0, 1, \dots, m.$$

As an example, consider a pure discount bond, with price b . We have that its dividend is zero and that $b(T) = 1$. Therefore,

$$b(\tau) = E \left[\frac{u'(q(T))}{u'(D(\tau))} \middle| \mathcal{F}(\tau) \right] = E \left[\frac{m_{t,T}}{m_{t,\tau}} \middle| \mathcal{F}(\tau) \right],$$

where $m_{t,\tau}$ is as in Eq. (4.54).

4.7 Market imperfections and portfolio choice

The setup is as in Section 4.4, where we fix $m = d$. To allow for frictions such as market incompleteness or short sale constraints, we assume that the vector of normalized portfolio shares in the risky assets, $p(t) \equiv \pi(t)/V^{x,\pi,c}(t)$, is constrained to lie in a closed convex set $K \in \mathbb{R}^d$.

We follow the approach put forward by Cvitanić and Karatzas (1992), which consists in “embedding” the *constrained* portfolio choice of the investor in a set of *unconstrained* portfolio optimization problems. Under regularity conditions that we shall not deal with in these lectures, it is shown that in this set of unconstrained problems, there exists one, which happens to be the solution to the original constrained portfolio problem. So the constrained portfolio problem is solved, once we solve for the unconstrained, which we can do through the martingale methods in Section 4.4. This approach is closely related to the discrete time minimax probability mentioned in Chapter 2. It is a systematic approach to consumption and portfolio policies in a context of constrained portfolio choices, and generalizes results from He and Pearson (1991).

The starting point is the definition of the *support function*,

$$\zeta(\nu) = \sup_{p \in K} (-p^\top \nu), \quad \nu \in \mathbb{R}^d, \quad (4.64)$$

and its *effective domain*,

$$\tilde{K} = \{\nu \in \mathbb{R}^d : \zeta(\nu) < \infty\}.$$

The role of the support function ζ is to “tilt” the dynamics of the price system in Section 4.4, as follows:

$$\frac{dS_0(t)}{S_0(t)} = r^\nu(t) dt, \quad \frac{dS_i(t)}{S_i(t)} = \hat{a}_i^\nu(t) dt + \sigma_i(t) dW(t) \quad (i = 1, \dots, d) \quad (4.65)$$

where:

$$r^\nu \equiv r + \zeta(\nu), \quad \hat{a}_i^\nu \equiv \hat{a}_i + \nu + \zeta(\nu),$$

and \hat{a}_i is as in Section 4.4.

The main result is as follows. Denote with $\text{Val}(x; K)$ the value of the problem faced by an investor facing a portfolio constraint $K \in \mathbb{R}^d$, when his initial wealth is x . Let $\text{Val}_\nu(x)$ be the corresponding value of the problem faced by an unconstrained investor in the market (4.65). Clearly, this value is just $\text{Val}_0(x)$ for the market considered in Sections 4.4 and 4.5. Moreover, for each $\nu \in \mathbb{R}^d$, the unconstrained program the investor faces in the market (4.65), can be solved through martingale methods, using the unique risk-neutral probability Q^ν , equivalent to P , with Radon-Nikodym derivative equal to,

$$\zeta^\eta(T) \equiv \frac{dQ^\nu}{dP} = \zeta^0(T) \exp \left(- \int_0^T (\sigma^{-1}(t) \nu(t))^\top dW(t) - \frac{1}{2} \int_0^T \|\sigma^{-1}(t) \nu(t)\|^2 dt \right). \quad (4.66)$$

Then, under regularity conditions, we have that:

$$\text{Val}(x; K) = \inf_{\nu \in \tilde{K}} (\text{Val}_\nu(x)), \quad (4.67)$$

and optimal consumption and portfolio choices for this unconstrained problem are exactly those chosen by the investor constrained to have $p \in K$. Appendix 4 provides an informal sketch of the arguments leading to Eq. (4.67).

Examples of the support function ζ in Eq. (4.64) are the *unconstrained case*: $K = \mathbb{R}^d$, in which case $\tilde{K} = \{0\}$ and $\zeta = 0$ on \tilde{K} ; *prohibition of short-selling*: $K = [0, \infty)^d$, in which case $\tilde{K} = K$ and $\zeta = 0$ on \tilde{K} , or: *incomplete markets*: $K = \{p \in \mathbb{R}^d : p_{M+1} = \dots = p_D = 0\}$ (i.e. the first M assets can only be traded), in which case $\tilde{K} = \{\nu \in \mathbb{R}^d : \nu_1 = \dots = \nu_M = 0\}$ and $\zeta = 0$ on \tilde{K} .

In the context of log-utility functions, we have that,

$$\hat{\nu} = \arg \min_{\nu \in \tilde{K}} \left(2\zeta(\nu) + \|\lambda + \sigma^{-1}\nu\|^2 \right),$$

where $\lambda = \sigma^{-1}(a - \mathbf{1}_d r)$. Applications of this will be worked out in Part II on “Asset pricing and reality.”

4.8 Jumps

Brownian motions are well suited to model the price behavior of liquid assets or assets issued by names or Governments not subject to default risk. There is, however, a fair amount of interest in modeling discontinuous changes in asset prices. Fixed income instruments may undergo liquidity dry-ups, or even default, causing price discontinuities that we wish to model. This section is an introduction to Poisson models, a class of processes that is particularly useful in addressing these issues.

4.8.1 Poisson jumps

Let (t, T) be a given interval, and consider events in that interval which display the following properties:

- (i) The random number of events arrivals on any disjoint time intervals of (t, T) are independent.
- (ii) Given two arbitrary disjoint but equal time intervals in (t, T) , the probability of a given random number of events arrivals is the same in each interval.
- (iii) The probability that at least two events occur simultaneously in any time interval is zero.

Next, let $P_k(\tau - t)$ be the probability that k events arrive during the time interval $\tau - t$. We make use of the previous three properties to determine the functional form of $P_k(\tau - t)$. First, $P_k(\tau - t)$ must satisfy:

$$P_0(\tau + d\tau - t) = P_0(\tau - t) P_0(d\tau), \quad (4.68)$$

and we impose

$$P_0(0) = 1, \quad P_k(0) = 0 \text{ for } k \geq 1. \quad (4.69)$$

Eq. (4.68) and the first condition in (4.69) are satisfied by $P_0(\tau) = e^{-v\tau}$, for some constant v , which we take to be positive, so as to ensure that $P_0 \in [0, 1]$. Furthermore, we have that:

$$\left\{ \begin{array}{l} P_1(\tau + d\tau - t) = P_0(\tau - t) P_1(d\tau) + P_1(\tau - t) P_0(d\tau) \\ \vdots \\ P_k(\tau + d\tau - t) = P_{k-1}(\tau - t) P_1(d\tau) + P_k(\tau - t) P_0(d\tau) \\ \vdots \end{array} \right. \quad (4.70)$$

The first equation in (4.70) can be rearranged as follows:

$$\frac{P_1(\tau + d\tau - t) - P_1(\tau - t)}{d\tau} = -\frac{1 - P_0(d\tau)}{d\tau}P_1(\tau - t) + \frac{P_1(d\tau)}{d\tau}P_0(\tau - t).$$

For small $d\tau$, $P_1(d\tau) \approx 1 - P_0(d\tau)$ and $P_0(d\tau) = 1 - vd\tau + O(d\tau^2) \approx 1 - vd\tau$. Therefore, $P_1'(\tau - t) = -vP_1(\tau - t) + vP_0(\tau - t)$. By a similar reasoning,

$$P_k'(\tau - t) = -vP_k(\tau - t) + vP_{k-1}(\tau - t).$$

The solution to this equation is:

$$P_k(\tau - t) = \frac{v^k(\tau - t)^k}{k!}e^{-v(\tau-t)}.$$

4.8.2 Interpretation

A Poisson model is one of *rare events*. Moreover, by:

$$E(\text{event arrival in } d\tau) = P_1(d\tau) = vd\tau.$$

For this reason, we usually refer to the parameter v as the *intensity* of event arrivals.

To provide additional intuition about the mathematics of rare events, consider the expression for the probability of k “arrivals” in n trials, predicted by a binomial distribution:

$$P_{n,k} = \binom{n}{k} p^k q^{n-k} = \frac{n!}{k!(n-k)!} p^k q^{n-k}, \quad p, q > 0, \quad p + q = 1,$$

where p is the probability of arrival for each trial. We want to model the probability p as a function of n , with the feature that $\lim_{n \rightarrow \infty} p(n) = 0$, so as to make each arrival “rare.” One possible choice is $p(n) = \frac{a}{n}$, for some constant $a > 0$. Under this assumption, we have:

$$\begin{aligned} P_{n,k} &= \frac{n!}{k!(n-k)!} p(n)^k (1 - p(n))^{n-k} \\ &= \frac{n!}{k!(n-k)!} \left(\frac{a}{n}\right)^k \left(1 - \frac{a}{n}\right)^{n-k} \\ &= \frac{n!}{k!(n-k)!} \left(\frac{a}{n}\right)^k \left(1 - \frac{a}{n}\right)^n \left(1 - \frac{a}{n}\right)^{-k} \\ &= \frac{n!}{n^k (n-k)!} \frac{a^k}{k!} \left(1 - \frac{a}{n}\right)^n \left(1 - \frac{a}{n}\right)^{-k} \\ &= \underbrace{\frac{n}{n} \cdot \frac{n-1}{n} \cdots \frac{n-k+1}{n}}_{k \text{ times}} \frac{a^k}{k!} \left(1 - \frac{a}{n}\right)^n \left(1 - \frac{a}{n}\right)^{-k}, \end{aligned}$$

leaving,

$$\lim_{n \rightarrow \infty} P_{n,k} \equiv P_k = \frac{a^k}{k!} e^{-a}.$$

Next, we split the interval $(\tau - t)$ into n subintervals of length $\frac{\tau-t}{n}$, and then make the probability of one arrival in each sub-interval proportional to each sub-interval length, as illustrated in Figure 4.1,

$$p(n) = v \frac{\tau - t}{n} \equiv \frac{a}{n}, \quad a \equiv v(\tau - t).$$

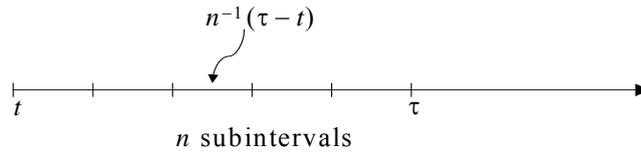


FIGURE 4.1. Heuristic construction of a Poisson process from a binomial distribution.

The Poisson model in the previous section is thus as that we consider here, with $n \rightarrow \infty$, which is continuous-time, as each sub-interval in Figure 4.1 shrinks to $d\tau$. The probability there is one arrival in $d\tau$ is $vd\tau$, which is also the expected number of events in $d\tau$ as shown below:

$$\begin{aligned}
 E(\# \text{ arrivals in } d\tau) &= \Pr(\text{one arrival in } d\tau) \times \text{one arrival} + \Pr(\text{zero arrivals in } d\tau) \times \text{zero arrivals} \\
 &= \Pr(\text{one arrival in } d\tau) \times 1 + \Pr(\text{zero arrivals in } d\tau) \times 0 \\
 &= vd\tau.
 \end{aligned}$$

The heuristic construction in this section opens the way to how we can simulate Poisson processes. We can just simulate a Uniform random variable $U(0, 1)$, with the continuous-time process being approximated by Y , where:

$$Y = \begin{cases} 0 & \text{if } 0 \leq U < 1 - vh \\ 1 & \text{if } 1 - vh \leq U < 1 \end{cases}$$

where h is a discretization interval.

4.8.3 Properties and related distributions

We check that P_k is a probability. We have:

$$\sum_{k=0}^{\infty} P_k = e^{-a} \sum_{k=0}^{\infty} \frac{a^k}{k!} = 1,$$

since $\sum_{k=0}^{\infty} a^k / k!$ is the McLaurin expansion of e^a . Second, we compute the mean,

$$\text{Mean} = \sum_{k=0}^{\infty} k \cdot P_k = e^{-a} \sum_{k=0}^{\infty} k \cdot \frac{a^k}{k!} = a.$$

A related distribution is the *exponential* (or Erlang) distribution. Remember, the probability of zero arrivals in $\tau - t$ predicted by the Poisson model is $P_0(\tau - t) = e^{-v(\tau-t)}$, from which it follows that:

$$G(\tau - t) \equiv 1 - P_0(\tau - t) = 1 - e^{-v(\tau-t)}$$

is the probability of at least one arrival in $\tau - t$. The function G can be also interpreted as the probability the first arrival occurred before τ , starting from t . The density function of G is:

$$g(\tau - t) = \frac{\partial}{\partial \tau} G(\tau - t) = ve^{-v(\tau-t)}.$$

The first two moments of the exponential distribution are:

$$\text{Mean} = \int_0^{\infty} xve^{-vx} dx = v^{-1}, \quad \text{Variance} = \int_0^{\infty} (x - v^{-1})^2 ve^{-vx} dx = v^{-2}.$$

The expected time of the first arrival occurred before τ starting from t equals v^{-1} . More generally, v^{-1} can be interpreted as the average time from an arrival to another.⁵

A more general distribution than the exponential is the Gamma distribution with density:

$$g_\gamma(\tau - t) = v e^{-v(\tau-t)} \frac{(v(\tau-t))^{\gamma-1}}{(\gamma-1)!}.$$

The exponential distribution obtains when $\gamma = 1$.

4.8.4 Asset pricing implications

This section is a short introduction to modeling asset prices as being driven by Brownian motions and jumps processes. We model jumps by interpreting the “arrivals” in the previous sections as those events upon which a certain random variable experiences a jump of size \mathcal{S} , where \mathcal{S} is another random variable with a fixed probability p . A simple model is:

$$dS(\tau) = b(S(\tau))d\tau + \sigma(S(\tau))dW(\tau) + \ell(S(\tau)) \cdot \mathcal{S} \cdot dZ(\tau), \quad (4.71)$$

where b, σ, ℓ are given functions (with $\sigma > 0$), W is a standard Brownian motion, and Z is a Poisson process with intensity equal to v , i.e.

- (i) $\Pr(Z(t)) = 0$.
- (ii) $\forall t \leq \tau_0 < \tau_1 < \dots < \tau_N < \infty$, $Z(\tau_0)$ and $Z(\tau_k) - Z(\tau_{k-1})$ are independent for each $k = 1, \dots, N$.
- (iii) $\forall \tau > t$, $Z(\tau) - Z(t)$ is a random variable with Poisson distribution and expected value $v(\tau - t)$, i.e.:

$$\Pr(Z(\tau) - Z(t) = k) = \frac{v^k (\tau - t)^k}{k!} e^{-v(\tau-t)}.$$

In this framework, k is the number of jumps over the time interval $\tau - t$.⁶ From this, we have that $\Pr(Z(\tau) - Z(t) = 1) = v(\tau - t) e^{-v(\tau-t)}$ and for $\tau - t$ small,

$$\Pr(dZ(\tau) = 1) \equiv \Pr(Z(\tau) - Z(t)|_{\tau \rightarrow t} = 1) = v(\tau - t) e^{-v(\tau-t)}|_{\tau \rightarrow t} \simeq v d\tau.$$

More generally, the process $\{Z(\tau) - v(\tau - t)\}_{\tau \geq t}$ is a martingale.

Armed with these preliminary facts, we can provide a heuristic derivation of Itô’s lemma for jump-diffusion processes. Consider any function f with enough regularity conditions, a rational function of time and S in Eq. (4.71), i.e. $f(\tau) \equiv f(S(\tau), \tau)$. Consider the following expansion of f :

$$df(\tau) = Lf(S(\tau), \tau) d\tau + f_S(S(\tau), \tau) \sigma(S(\tau)) dW(\tau) + [f(S(\tau) + \ell(S(\tau)) \cdot \mathcal{S}, \tau) - f(S(\tau), \tau)] \cdot dZ(\tau).$$

⁵Suppose arrivals are generated by Poisson processes, and consider the random variable “time interval elapsing from one arrival to next one.” Let τ' be the instant at which the last arrival occurred. Then, the probability the time $\tau - \tau'$ which will elapse from the last arrival to the next is less than Δ is the same as the probability that during the time interval $\tau - \tau'$, there is at least one arrival.

⁶For simplicity, we take v to be constant. If v is a deterministic function of time, we have that

$$\Pr(Z(\tau) - Z(t) = k) = \frac{(\int_t^\tau v(u) du)^k}{k!} \exp\left(-\int_t^\tau v(u) du\right), \quad k = 0, 1, \dots$$

and there is also the possibility to model v as a function of the state: $v = v(q)$, for example. Cox processes.

The first two terms in are the usual Itô's lemma terms, with L denoting the usual infinitesimal generator for diffusions. The third term accounts for jumps. If there are no jumps from time τ_- to time τ (where $d\tau = \tau - \tau_-$), then $dZ(\tau) = 0$. If there is a jump then $dZ(\tau) = 1$, and in this case f , as a "rational" function, needs also instantaneously jump to $f(S(\tau) + \ell(S(\tau)) \cdot \mathcal{S}, \tau)$. The jump will be exactly $f(S(\tau) + \ell(S(\tau)) \cdot \mathcal{S}, \tau) - f(S(\tau), \tau)$, where \mathcal{S} is another random variable with a fixed probability measure. Clearly, if $f(S, \tau) = S$, we are back to the initial jump-diffusion model in Eq. (4.71).

To derive the infinitesimal generator for jumps-diffusion, $L^J f$ say, note that:

$$\begin{aligned} E(df) &= (Lf) d\tau + E[(f(S + \ell\mathcal{S}, \tau) - f(S, \tau)) \cdot dZ(\tau)] \\ &= (Lf) d\tau + E[(f(S + \ell\mathcal{S}, \tau) - f(S, \tau)) \cdot v \cdot d\tau], \end{aligned}$$

or

$$L^J f = Lf + v \cdot \int_{\text{supp}(\mathcal{S})} [f(S + \ell\mathcal{S}, \tau) - f(S, \tau)] p(d\mathcal{S}),$$

where $\text{supp}(\mathcal{S})$ denotes the support of \mathcal{S} . Therefore, the infinitesimal generator for jumps-diffusion is, simply, $L^J f$.

4.8.5 An option pricing formula

Merton (1976, JFE), Bates (1988, working paper), Naik and Lee (1990, RFS) are the seminal papers.

4.9 Continuous time Markov chains

Needed to model credit risk.

4.10 Appendix 1: Self-financed strategies

We have,

$$c_t + S_t \theta_{1,t+1} + b_t \theta_{2,t+1} = (S_t + D_t) \theta_{1,t} + b_t \theta_{2,t} \equiv V_t + D_t \theta_{1,t},$$

where $V_t \equiv S_t \theta_{1,t} + b_t \theta_{2,t}$ is wealth net of dividends. We have,

$$\begin{aligned} V_t - V_{t-1} &= S_t \theta_{1,t} + b_t \theta_{2,t} - V_{t-1} \\ &= S_t \theta_{1,t} + b_t \theta_{2,t} - (c_{t-1} + S_{t-1} \theta_{1,t} + b_{t-1} \theta_{2,t} - D_{t-1} \theta_{1,t-1}) \\ &= (S_t - S_{t-1}) \theta_{1,t} + (b_t - b_{t-1}) \theta_{2,t} - c_{t-1} + D_{t-1} \theta_{1,t-1}, \end{aligned}$$

and more generally,

$$V_t - V_{t-\Delta} = (S_t - S_{t-\Delta}) \theta_{1,t} + (b_t - b_{t-\Delta}) \theta_{2,t} - (c_{t-\Delta} \cdot \Delta) + (D_{t-\Delta} \cdot \Delta) \theta_{1,t-\Delta}.$$

Now let $\Delta \downarrow 0$ and assume that θ_1 and θ_2 are constant between t and $t - \Delta$. We have:

$$dV(\tau) = (dS(\tau) + D(\tau)d\tau) \theta_1(\tau) + db(\tau)\theta_2(\tau) - c(\tau)d\tau.$$

Assume that

$$\frac{db(\tau)}{b(\tau)} = rd\tau.$$

The budget constraint can then be written as:

$$\begin{aligned} dV(\tau) &= (dS(\tau) + D(\tau)d\tau) \theta_1(\tau) + rb(\tau)\theta_2(\tau)d\tau - c(\tau)d\tau \\ &= (dS(\tau) + D(\tau)d\tau) \theta_1(\tau) + r(V - S(\tau)\theta_1(\tau)) d\tau - c(\tau)d\tau \\ &= (dS(\tau) + D(\tau)d\tau - rS(\tau)d\tau) \theta_1(\tau) + rV d\tau - c(\tau)d\tau \\ &= \left(\frac{dS(\tau)}{S(\tau)} + \frac{D(\tau)}{S(\tau)} d\tau - rd\tau \right) \theta_1(\tau) S(\tau) + rV d\tau - c(\tau)d\tau \\ &= \left(\frac{dS(\tau)}{S(\tau)} + \frac{D(\tau)}{S(\tau)} d\tau - rd\tau \right) \pi(\tau) + rV d\tau - c(\tau)d\tau. \end{aligned}$$

4.11 Appendix 2: An introduction to stochastic calculus for finance

4.11.1 Stochastic integrals

4.11.1.1 Motivation

Given is a Brownian motion $W(t) \equiv W_t(\omega)$, $t \geq 0$, and the associated natural filtration $\mathcal{F}(t)$. We aim to give a sense to the “integral”

$$I_t(\omega) \equiv \int_0^t f(s) dW_s(\omega), \quad (4A.1)$$

where f is a given function. More generally, this appendix aims to provide explanations about the sense to give to “integrals” which look like:

$$I_t(\omega) \equiv \int_0^t g(s; \omega) dW_s(\omega),$$

where g is now a progressively $\mathcal{F}(t)$ -measurable function.

The motivation for this aim is that we can build up a class of useful processes from Brownian motions. Let us illustrate. Given (Ω, \mathcal{F}, P) on which W is Brownian motion, and let $T < \infty$. Let us write dW for the increment of W over an infinitesimal amount of time. In some sense, $dW(t)$ equals $W(t + \Delta t) - W(t)$ as $\Delta t \rightarrow 0$. We may think of the “increment” $dW(t)$ as being normally distributed: $dW(t) \sim N(0, dt)$. From here, we may consider some richer processes X (say)

$$dX_t(\omega) = \mu_t(\omega) dt + \sigma_t(\omega) dW(t) \quad (4A.2)$$

for some objects $\mu_t(\cdot)$ and $\sigma_t(\cdot)$ to be defined later. These processes are known as Itô’s processes, as further explained below. The intuition on $\mu_t(\cdot)$ and $\sigma_t(\cdot)$ is as follows. Heuristically, we have that $E[dX_t(\omega)] = E[\mu_t(\omega)] dt + \sigma E(dW(t)) = E[\mu_t(\omega)] dt$, such that $\mu_t(\cdot)$ is related to the instantaneous expected changes of dX . So this model is richer than Brownian motions because μ can be different from identically zero. Useful for asset pricing. Think of X as an asset price process. Hard to imagine that we would be willing to invest if the expected variation of X (that is the expected capital gain) over some time horizon is just zero. Following the interpretation of X as an asset price, we now compute the variance of dX . We have, $var(dX(t)) = E[dX(t) - E(dX(t))]^2 = E(\sigma dW(t))^2$ which turns out to equal $\sigma^2 dt$.

A quite important terminology issue. The “process” μ is called the *drift* and the “process” σ is called the *diffusion coefficient*, or the *volatility* of X . Clearly, the drift μ determines the trend, and the volatility determines the noisiness of X around that trend. Both drift and diffusion coefficients need to be adapted processes, as we shall explain. One example of drift and diffusion coefficients. Assume that $\mu \equiv W(t)$ and $\sigma \equiv 0$. In this case, we have that: $dX(t) = W(t) dt$, which shows that $X(t)$ still fluctuates randomly. Here, μ is a stochastic process and so is $X(t)$. Its infinitesimal variations can be predicted. But its further evolution cannot. In finance jargon, we would say that $X(t)$ is *locally riskless* in this example.

Let us proceed with a more delicate example, relating to *strategies and trading gains*. Suppose that a stock price is just a Brownian motion. Assume it does not distribute dividends over some time-horizon of interest, and that we hold $\theta(t)$ units of it at time t . What are our trading gains from 0 to t ? Later, we shall argue that the expression, $\int_0^t \theta(s) dW_s(\omega)$, is the answer to this question: intuitively, this expression is the sum of the instantaneous capital gains on the assets, multiplied by the units of the asset that are held. This expression is what is known as *stochastic integral*.

Why are we insisting in modeling asset prices through Brownian motions? As we shall see, Brownian motions are wild in some sense, i.e. they are of unbounded variation on any interval. So why don’t we go for smoother processes? The answer is that “smoother” processes would give rise to arbitrage opportunities. Harrison, Pitbladdo and Schaefer (1984) showed that in continuous time models, asset

prices must be “wild.” Intuitively, if stock-prices are continuous in time and have finite variation, we could predict them over the immediate future, thus cashing-in the capital gains.

Let us mention a few technicalities. We already know $W(t)$ is nowhere differentiable. So the expression in Eq. (8.20) should be only understood as a shorthand for,

$$X_t(\omega) = X_0 + \int_0^t \mu_s(\omega) ds + \int_0^t \sigma_s(\omega) dW_s(\omega).$$

The question, then, is what does the “stochastic integral” $\int_0^t \sigma_s(\omega) dW_s(\omega)$ mean, and why we would ever need it. In standard calculus, the integral can be defined from its differential. To anticipate, in stochastic calculus this is no longer the case, in that the stochastic integral is the real thing.

In the following sections, we provide short reviews of the ordinary Riemann integral, the Riemann-Stieltjes integral and explain why these two approaches to pathwise integration generically fail to provide a solid foundation to the “expression” $I_t(\omega)$ in Eq. (4A.1). To anticipate, the main issue relates to unboundedness of Brownian motions:

$$\forall \omega \in \Omega, \sup_{\tau} \sum_{i=1}^n |W_{t_i}(\omega) - W_{t_{i-1}}(\omega)| = \infty,$$

where the supremum is taken over all partitions of $[0, T]$. We shall state conditions on “how much bounded” the integrator and integrands in $I_t(\omega)$ should be in order for the Riemann-Stieltjes theory to hold. Unfortunately, these conditions are restrictive within the context of interest here. We shall explain that in general, no Riemann or Riemann-Stieltjes explanation can be given to “expressions” such as $\int_0^t f(s; \omega) dW_s(\omega)$. However, there are still cases where the Riemann-Stieltjes theory works. For example, consider the functions $f(t) = 1$, or $f(t) = t$. But in general, the Riemann-Stieltjes theory doesn’t work, so we have to attack the problem with a more general approach. Intuitively, we can only consider a probabilistic representation of $I_t(\omega)$.

4.11.1.2 Riemann

Given is $x \mapsto f(x)$, $x \in (0, 1)$. We consider two standard definitions. First, we define a *partition* as $\tau_n : 0 = t_0 < t_1 < \dots < t_{n-1} < t_n = 1$ and $\Delta_i = t_i - t_{i-1}$, $i = 1, \dots, n$, as in the following picture.



Second, we define an *intermediate partition* as σ_n : any collection of values y_i satisfying $t_{i-1} \leq y_i < t_i$, $i = 1, \dots, n$. Then, for a given partition τ_n and intermediate partition σ_n , the Riemann sum is defined as:

$$S_n(\tau_n, \sigma_n) \equiv \sum_{i=1}^n f(y_i) \Delta_i.$$

It’s a “weighted average of the values $f(y_i)$.” Next, let $\text{Mesh}(\tau_n) \equiv \max_{i=1, \dots, n} \Delta_i$. Consider letting $\text{Mesh}(\tau_n) \rightarrow 0$ by sending $n \rightarrow \infty$. If the limit, $\lim_{n \rightarrow \infty} S_n(\tau_n, \sigma_n)$, exists, and is independent of τ_n and σ_n , then it is called the *Riemann integral* of f on $(0, 1)$ and it is written:

$$\int_0^1 f(t) dt.$$

Two properties are worth mentioning:

1. *Linearity*: Given two constants c_1 and c_2 , $\int_0^1 (c_1 f_1(t) + c_2 f_2(t)) dt = c_1 \int_0^1 f_1(t) dt + c_2 \int_0^1 f_2(t) dt$.
2. *Linearity on adjacent intervals*: $\int_0^1 f(t) dt = \int_0^a f(t) dt + \int_a^1 f(t) dt$ for every $a \in (0, 1)$.

4.11.1.3 Riemann-Stieltjes

The main idea is to “integrate one function f with respect to another function g .” One standard example relates to the computation of the expectation of a random variable with distribution function g . Heuristically, we have that:

$$\int_0^1 t dg(t) \approx \sum_i t_i [g(t_i) - g(t_{i-1})].$$

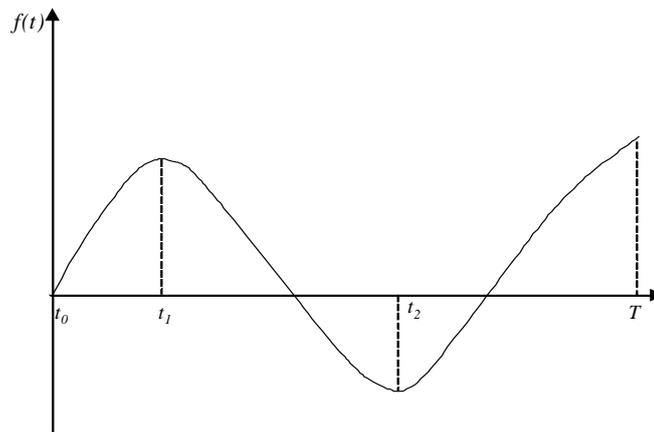
In general, let us be given two functions f and g . Consider, again, the definitions of τ_n, σ_n given earlier, and set: $\Delta g_i = g(t_i) - g(t_{i-1})$, $i = 1, \dots, n$. The Riemann-Stieltjes sum is defined as:

$$S_n(\tau_n, \sigma_n) = \sum_{i=1}^n f(y_i) \Delta g_i.$$

Clearly the Riemann sum is a special case obtained with the identity function $g(t) = t$. Similarly as in the definition of the Riemann sum, here we have that if the limit, $\lim_{n \rightarrow \infty} S_n(\tau_n, \sigma_n)$, exists, and is independent of τ_n and σ_n , then is called the *Riemann-Stieltjes integral* of f with respect to g on $(0, 1)$ and it is written:

$$\int_0^1 f(t) dg(t).$$

The crucial issue is, can we now use Riemann-Stieltjes theory to define integrals of functions w.r.t Brownian motions? That is, can we interpret $\int_0^1 f(t) dW_t(\omega)$ as a Riemann-Stieltjes integral, path by path, i.e. $\forall \omega \in \Omega$? The answer is in the negative, except in very special cases. Indeed, a natural example of an integral of functions with respect to Brownian motion is $I_t(\omega) \equiv \int_0^t f(s) dW_s(\omega)$. But what does this representation mean? We know that a ω - W_t path is not-differentiable. However, the main point, here, does not even relate to differentiability, but to a quite peculiar property, known as *unboundedness of Brownian motions*. To introduce this property, consider a certain function f , such as that depicted in the next picture.



Consider, then, its “first variation,” defined as:

$$\begin{aligned} [f(t_1) - f(t_0)] - [f(t_2) - f(t_1)] + [f(T) - f(t_2)] &= \int_0^{t_1} f'(t) dt + \int_{t_1}^{t_2} (-f'(t)) dt + \int_{t_2}^T f'(t) dt \\ &= \int_{t_0}^T |f'(t)| dt. \end{aligned}$$

We can see that this first variation is a measure of the total amount of up and down motion of the path of the function f . We can formalize this reasoning as follows. Let f be a function of a real variable. Its variation in an interval $[a, b]$ is defined as

$$V_f([a, b]) = \sup \sum_{i=1}^n \left| f(t_i^{(n)}) - f(t_{i-1}^{(n)}) \right|,$$

where the supremum is taken over the partitions $a \equiv t_0^{(n)} < t_1^{(n)} < \dots < t_n^{(n)} \equiv b$. By the triangle inequality, $|f(x) - f(y)| = |f(x) - f(z) + f(z) - f(y)| \leq |f(x) - f(z)| + |f(z) - f(y)|$, the sums in $V_f([a, b])$ can only increase as we add more and more into the partition, such that,

$$V_f([a, b]) = \lim_{\text{mesh} \downarrow 0} \sum_{i=1}^n \left| f(t_i^{(n)}) - f(t_{i-1}^{(n)}) \right|.$$

Next, consider the following definition.

DEFINITION 4A.1. A real function h on $(0, 1)$ has bounded p -variation, $p > 0$, if

$$\sup_{\tau} \sum_{i=1}^n |h(t_i) - h(t_{i-1})|^p < \infty,$$

where the supremum is taken over all partitions of $(0, 1)$.

We have:

THEOREM 4A.2. The Riemann-Stieltjes integral, $\int_0^1 f(t) dg(t)$, exists under the following conditions:

- (i) The functions f and g don't have discontinuities at the same points.
- (ii) f has bounded p -variation and g has bounded q -variation, with $\frac{1}{p} + \frac{1}{q} > 1$, that is, f, g satisfy $\sup_{\tau} \sum_{i=1}^n |f(t_i) - f(t_{i-1})|^p < \infty$ and $\sup_{\tau} \sum_{i=1}^n |g(t_i) - g(t_{i-1})|^q < \infty$ with $\frac{1}{p} + \frac{1}{q} > 1$.

Now, it is well-known that almost every ω - W_t path has bounded p -variation for $p \geq 2$. And, as expected, unbounded p -variation for $p < 2$, as further argued below. Consider, then, the integral, $\int_0^1 f(t) dW_t(\omega)$, and suppose f is differentiable with bounded derivatives. By the mean value theorem, there exists a $K > 0$ such that: $|f(t) - f(s)| \leq K(t - s)$ for $s < t$. Therefore, $\sup_{\tau} \sum_{i=1}^n |f(t_i) - f(t_{i-1})| \leq K \sum_{i=1}^n (t_i - t_{i-1}) = K$. That is, f has bounded p -variation, with $p = 1$.

By Theorem 4A.2, we now have that for almost every ω - W_t path, the Riemann-Stieltjes integral of f with respect to Brownian motions,

$$I_t(\omega) \equiv \int_0^t f(s) dW_s(\omega),$$

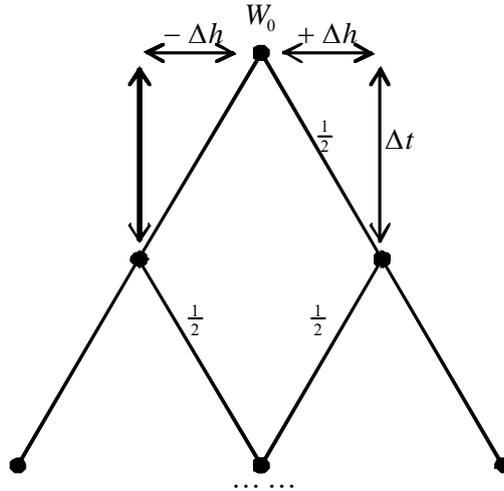
exists for every deterministic function f which is differentiable with bounded first-order derivative. For example, $f(t) = 1$, or $f(t) = t$. We aren't done. Consider $f_t(\omega) = W_t(\omega)$ and, then:

$$I(W)(\omega) = \int_0^1 W_t(\omega) dW_t(\omega).$$

Let $p = 2 + \epsilon$, for some $\epsilon > 0$. Hence $p = q = 2 + \epsilon$, and so $\frac{1}{p} + \frac{1}{q} = \frac{2}{2+\epsilon} < 1$. The Riemann-Stieltjes theory doesn't work even with this simple example. This is where the theory of Itô's stochastic integrals comes in.

4.11.1.4 A digression on unboundedness of Brownian motions

Why do Brownian motions display unbounded variation? Consider the “Brownian tree” in the picture below.



Time is Δt and space is Δh . In the Brownian tree, we must have,

$$\Delta h = \sqrt{\Delta t}. \tag{4A.3}$$

Indeed, and heuristically, we have that $var(\Delta W) = (\Delta h)^2$, which matched to $var(\Delta W) = \Delta t$, leaves precisely Eq. (4A.3). Therefore, $E(|\Delta W|) = \Delta h = \sqrt{\Delta t}$. Next let us chop a time interval of length t in $n \equiv \frac{t}{\Delta t}$ parts. The total expected length traveled by a Brownian motion is,

$$\frac{t}{\Delta t} \Delta h = \frac{t}{\Delta t} \sqrt{\Delta t} \rightarrow \infty \text{ as } \Delta t \rightarrow 0.$$

A more substantive proof is one for example of Corollary 2.5 p. 25 in Revuz and Yor (1999). A sketch of this proof proceeds as follows. We have:

$$\sum_i (W_{t_i} - W_{t_{i-1}})^2 \leq \max_i |W_{t_i} - W_{t_{i-1}}| \cdot \sum_i |W_{t_i} - W_{t_{i-1}}|.$$

Moreover, $\max_i |W_{t_i} - W_{t_{i-1}}|$ converges to zero $\forall \omega \in \Omega$ because W is continuous, and by the Heine-Cantor theorem, continuous functions are uniformly continuous on finite intervals. Then, suppose that W_t has bounded variation, which would imply that

$$\forall \omega \in \Omega, \quad \sum_i (W_{t_i} - W_{t_{i-1}})^2 \rightarrow 0, \quad \text{Mesh} \downarrow 0,$$

which is impossible, as we shall now argue. Indeed, in the next section, we shall establish that $L_n \equiv \sum_i (W_{t_i} - W_{t_{i-1}})^2 \xrightarrow{q.m} t$, implying that $\text{p lim}_n L_n = t$. Therefore, there exists a sequence $n_k : L_{n_k} \rightarrow t$ for all $\omega \in \Omega$. (Convergence in probability does not imply almost sure convergence, yet it implies that there exists a suitable subsequence n_k s.t. \exists a.s. convergence, which is what we just need here.)

4.11.1.5 Itô

Let us begin with a first example, which can help grasp the nature of the issues under study. Consider

$$I(W)(\omega) = \int_0^1 W_t(\omega) dW_t(\omega).$$

Consider, then, the following Riemann-Stieltjes sum:

$$S_n = \sum_{i=1}^n W_{t_{i-1}} \Delta_i W, \quad \Delta_i W = W_{t_i} - W_{t_{i-1}},$$

where the intermediate partition makes simply use of the left-end points $y_i = t_{i-1}$, $i = 1, \dots, n$. Simple computations leave:

$$S_n = \frac{1}{2} [W_t^2 - Q_n(t)], \quad Q_n(t) \equiv \sum_{i=1}^n (\Delta_i W)^2.$$

The quantity $Q_n(t)$ is known as the *Quadratic Variation*, a quite useful concept in financial econometrics. We have

$$E[Q_n(t)] = \sum_{i=1}^n E(\Delta_i W)^2 = \sum_{i=1}^n \Delta_i = t.$$

Moreover, $\text{var}[(\Delta_i W)^2] = \text{var}[(\frac{1}{\sqrt{\Delta_i}} \Delta_i W \sqrt{\Delta_i})^2] = \Delta_i^2 \text{var}[(\frac{\Delta_i W}{\sqrt{\Delta_i}})^2] = 2\Delta_i^2$, where the last equality follows because $\frac{\Delta_i W}{\sqrt{\Delta_i}} \sim N(0, 1)$, which implies that $(\frac{\Delta_i W}{\sqrt{\Delta_i}})^2 \sim \chi^2(1)$. Hence,

$$\text{var}[Q_n(t)] = \sum_{i=1}^n \text{var}[(\Delta_i W)^2] = 2 \sum_{i=1}^n \Delta_i^2 \leq 2 \sum_{i=1}^n \text{Mesh}(\tau_n) \cdot \Delta_i = 2t \cdot \text{Mesh}(\tau_n) \rightarrow 0.$$

But $\text{var}[Q_n(t)] = E[Q_n(t) - E(Q_n(t))]^2 = E[Q_n(t) - t]^2$. Therefore,

$$\text{var}[Q_n(t)] = E[Q_n(t) - t]^2 \rightarrow 0 \quad t\text{-pointwise.}$$

This type of convergence is called *convergence in quadratic mean of $Q_n(t)$ to t* and it is written $Q_n(t) \xrightarrow{q.m.} t$, as we shall explain in the appendix of the next chapter. By the celebrated Chebyshev's inequality, convergence in quadratic mean implies convergence in probability:

$$\forall \delta > 0, \Pr\{|Q_n(t) - t| > \delta\} \leq \frac{E[Q_n(t) - t]^2}{\delta^2}.$$

Issues related to uniform convergence issues will be dealt with later.

To sumup, $\int_0^t W_s(\omega) dW_s(\omega)$ doesn't exist as a Riemann-Stieltjes integral. Nevertheless, the previous facts suggest that a good definition of it could hinge upon the notion of a mean square limit, viz

$$S_n = \sum_{i=1}^n W_{t_{i-1}} \Delta_i W = \frac{1}{2} [W_t^2 - Q_n(t)] \xrightarrow{q.m.} \frac{1}{2} (W_t^2 - t),$$

or, as we shall explain,

$$S_n = \sum_{i=1}^n W_{t_{i-1}} \Delta_i W \xrightarrow{q.m.} \int_0^t W_s dW_s,$$

where $\int_0^t W_s dW_s = \frac{1}{2} (W_t^2 - t)$ has the Itô's sense.

Clearly, $\int_0^t W_s dW_s$ does not satisfy the usual Riemann-Stieltjes rule of integration. (For any smooth function f such that $f(0) = 0$, the Riemann-Stieltjes integral $\int_0^t f(u) df(u) = \frac{1}{2} f(t)^2$.) This doesn't work here because we have yet to see what the chain-rule for functions of ω - W_t is. This will lead us to the celebrated Itô's lemma, which shall confirm that $\int_0^t W_s dW_s = \frac{1}{2} (W_t^2 - t)$. This example vividly illustrated that standard integration methods fails. In fact, the timing of the integrands is quite critical. For example, in Riemann integration, the integrand can be evaluated at any point in the interval. If we apply this to the kind of integrals we are studying here we obtain, $\lim \sum_i f(W_{t_{i-1}}) (W_{t_i} - W_{t_{i-1}})$

(for the left boundary) and $\lim \sum_i f(W_{t_i})(W_{t_i} - W_{t_{i-1}})$ (for the right boundary). But the two limits do not agree. The expectation of the first is zero (by the law of iterated expectations), while the expectation of the second is not necessarily zero. Finally, Riemann integration theory differs from the integration theory underlying the previous example because of the mode of convergence utilized in the two theories.

A short digression is order. The so-called Stratonovich stochastic integral selects as points of the intermediate partition the central ones:

$$\tilde{S}_n = \sum_{i=1}^n f(W_{y_i}) \Delta_i W, \quad y_i = \frac{1}{2}(t_{i-1} + t_i).$$

For the Stratonovich integral, the usual Riemann-Stieltjes rule applies, yet the Stratonovich stochastic integral isn't Riemann-Stieltjes.

4.11.1.6 The Itô's stochastic integral for *simple* processes

Let \mathcal{F} be the P -augmentation of the filtration of W . Consider $[0, T]$ and partitions $\tau_n : 0 = t_0 < t_1 < \dots < t_n = T$, and the following definition:

DEFINITION 4A.3 (SIMPLE PROCESSES). *The process $C = (C_t, t \in [0, T])$ is simple if*

(i) *There exists a partition τ_n and a sequence of r.v. $Z_i, i = 1, \dots, n$, s.t*

$$C_t = \begin{cases} Z_n, & \text{if } t = T \\ Z_i, & \text{if } t_{i-1} \leq t < t_i, \quad i = 1, \dots, n \end{cases}$$

(ii) *The sequence (Z_i) is $F_{t_{i-1}}$ -adapted, $i = 1, \dots, n$.*

(iii) *$E(Z_i^2) < \infty$ all i (L^2).*

As an example, consider $C_t = W_{t_{n-1}}$, if $t = T$, and $C_t = W_{t_{i-1}}$, if $t_{i-1} \leq t < t_i$, $i = 1, \dots, n$. Next, we have:

DEFINITION 4A.4. *The Itô's stochastic integral of a simple process C is,*

$$\begin{aligned} \int_0^T C_s dW_s &= \sum_{i=1}^n C_{t_{i-1}} (W_{t_i} - W_{t_{i-1}}) = \sum_{i=1}^n Z_i (W_{t_i} - W_{t_{i-1}}), \quad \text{on } [0, T] \\ \int_0^t C_s dW_s &= \sum_{i=1}^{k-1} C_{t_{i-1}} (W_{t_i} - W_{t_{i-1}}) + Z_k (W_t - W_{t_{k-1}}), \quad t \in [t_{k-1}, t_k]. \end{aligned}$$

with the notation $\sum_{i=1}^0 m_i \equiv 0$.

It is a Riemann-Stieltjes sum of C with respect to Brownian motions evaluated at *left-end* points. Finally, we proceed with listing a set of useful properties.

PROPERTY 4A.P1. *$I_t(C) = \int_0^t C_s dW_s, t \in [0, T]$ is a F_t -martingale and has expectation equal to zero.*

Proof. Let us check that $I_t(C)$ is a \mathcal{F}_t -martingale. We have to check three conditions: (i) $E|I_t(C)| < \infty$, all $t \in [0, T]$; (ii) $I_t(C)$ is \mathcal{F}_t -adapted; (iii) $E[I_t(C)|\mathcal{F}_s] = I_s(C)$, $s < t$. Condition (i) follows by the

isometry property to be introduced below. Condition (ii) is trivial. To show (iii), suppose, initially, that $s, t \in [t_{k-1}, t_k]$, $s < t$. We have:

$$\begin{aligned} I_t(C) &= \sum_{i=1}^{k-1} Z_i (W_{t_i} - W_{t_{i-1}}) + Z_k (W_t - W_{t_{k-1}}) \\ &= \sum_{i=1}^{k-1} Z_i (W_{t_i} - W_{t_{i-1}}) + Z_k (W_s - W_{t_{k-1}}) + Z_k (W_t - W_s) \\ &= I_s(C) + Z_k (W_t - W_s) \end{aligned}$$

$$\begin{aligned} E[I_t(C) | \mathcal{F}_s] &= E[I_s(C) | \mathcal{F}_s] + E[Z_k (W_t - W_s) | \mathcal{F}_s] \\ &= I_s(C) + Z_k E[(W_t - W_s) | \mathcal{F}_s] = I_s(C). \end{aligned}$$

The case $s \in [t_{l-1}, t_l]$ and $t \in [t_{k-1}, t_k]$, $l < k$ is proven similarly. Finally, $I_t(C)$ has zero expectation because it starts from the origin by the definition: $I_0(C) = 0 \Rightarrow E(I_t(C)) = 0$ all t . That is, $\forall t$, $E[I_t(C)] = E[I_0(C)] = I_0(C) = 0$.

PROPERTY 4A.P2 (ISOMETRY). $E\left(\int_0^t C_s dW_s\right)^2 = \int_0^t E(C_s^2) ds$, for all $t \in [0, T]$.

Proof. Without loss of generality, set $t = t_k$. We have:

$$\begin{aligned} E\left[\int_0^t C_s dW_s\right]^2 &= E\left[\sum_{i=1}^k C_{t_{i-1}} (W_{t_i} - W_{t_{i-1}})\right]^2 \\ &= E\left[\sum_{i=1}^k \sum_{j=1}^k C_{t_{i-1}} (W_{t_i} - W_{t_{i-1}}) C_{t_{j-1}} (W_{t_j} - W_{t_{j-1}})\right] \\ &= E\left[\sum_{i=1}^k C_{t_{i-1}}^2 (W_{t_i} - W_{t_{i-1}})^2\right], \end{aligned}$$

where the last equality follows because $(W_{t_i} - W_{t_{i-1}})$ and $(W_{t_j} - W_{t_{j-1}})$ are independent, for all $i \neq j$. Then,

$$\begin{aligned} E\left[\int_0^t C_s dW_s\right]^2 &= E\left[\sum_{i=1}^k C_{t_{i-1}}^2 (W_{t_i} - W_{t_{i-1}})^2\right] \\ &= E\left[\sum_{i=1}^k E\left(C_{t_{i-1}}^2 (W_{t_i} - W_{t_{i-1}})^2 \mid \mathcal{F}_{t_{i-1}}\right)\right] \\ &= E\left[\sum_{i=1}^k E\left(C_{t_{i-1}}^2 \mid \mathcal{F}_{t_{i-1}}\right) (t_i - t_{i-1})\right] \\ &= \sum_{i=1}^k E\left(C_{t_{i-1}}^2\right) (t_i - t_{i-1}) \\ &= \int_0^t E(C_s^2) ds. \end{aligned}$$

PROPERTY 4A.P3 (LINEARITY AND LINEARITY ON ADJACENT INTERVALS).

PROPERTY 4A.P4. $I_t(C)$ has continuous ω -paths.

4.11.1.7 The general Itô's stochastic integral

We now consider a more general class of integrand \mathcal{F} -adapted processes C_t , $t \in [0, T]$ satisfying $\int_0^T E(C_s^2) ds < \infty$, and $\in L^2(P \otimes dt)$, which is obviously satisfied by simple processes, although now we are now moving to continuous time. Clearly, \mathcal{H}^2 is a closed linear subspace of $L^2(P \otimes dt)$. So let $\|\cdot\|_{L^2(P \otimes dt)}$ be the norm of $L^2(P \otimes dt)$. Let \mathcal{H}_0^2 be the subset of \mathcal{H}^2 consisting of all simple processes. We now outline how to construct the stochastic integral, in four steps.

Step 1: (\mathcal{H}_0^2 is dense in \mathcal{H}^2). It is possible to show that for any $C \in \mathcal{H}^2$, there exists a sequence of simple processes $C^{(n)}$ s.t $\|C - C^{(n)}\|_{L^2(P \otimes dt)} \rightarrow 0$, i.e. $\int_0^T E(C_s - C_s^{(n)})^2 ds \rightarrow 0$.

Step 2: By step 1, $\{C^{(n)}\}$ is a Cauchy sequence in $L^2(P \otimes dt)$. By the isometry property of the Itô's integral for simple processes

$$\|I_T(C^{(n)}) - I_T(C^{(n')})\|_{L^2(P)} = \|C^{(n)} - C^{(n')}\|_{L^2(P \otimes dt)}.$$

Therefore, $I_T(C^{(n)})$ is a Cauchy sequence in $L^2(P)$. Now it is well-known that $L^2(P)$ is complete, and so $I_T(C^{(n)})$ must converge to some element of $L^2(P)$, denoted as $I_T(C)$.

Step 3: This limit is called the *Itô's stochastic integral* of C , and is written as

$$I_T(C) = \int_0^T C_s dW_s.$$

Finally, the limit is well-defined: if there is another $C_*^{(n)}$: $\|C - C_*^{(n)}\|_{L^2(P \otimes dt)} \rightarrow 0$, then $\lim I_T(C_*^{(n)}) = \lim I_T(C^{(n)}) = I_T(C)$ in the $L^2(P)$ norm.

Step 4: (Itô's integral as a process) We wish to create a whole “continuum” of Itô's integrals at a single glance. Step 3 is not enough because we need uniform convergence on $[0, T]$. To show that it's feasible lies beyond the aim of these introductory lectures. The final result is, For any $C \in \mathcal{H}^2$, there exists a process $(I_t, t \in [0, T])$ which is a continuous \mathcal{F}_t -martingale s.t

$$I_t = \int_0^t C_s dW_s, \quad t \in [0, T], \quad P \otimes dt\text{-a.s.}$$

To summarize, then, let us introduce the two spaces,

$$\mathcal{L}^n = \left\{ \theta \in \mathcal{L} : \int_0^T |\theta_t|^n dt < \infty \text{ a.s.} \right\}, \quad n = 1, 2, \quad \mathcal{H}^2 = \left\{ \theta \in \mathcal{L}^2 : E \left(\int_0^T |\theta_t|^2 dt \right) < \infty \text{ a.s.} \right\}$$

where \mathcal{L} denotes the set of all adapted processes. Let $\theta \in \mathcal{H}^2$. The stochastic integral $I_t(\theta) = \int_0^t \theta_s dW_s$ satisfies the following properties: (i) Continuous sample paths, and $I_t(\theta)$ is a \mathcal{F}_t -martingale; (ii) Expectation equal to zero; (iii) Itô's isometry on \mathcal{H}^2 , i.e. $E(\int_0^t \theta_s dW_s)^2 = \int_0^t E(\theta_s^2) ds < \infty$, $t \in [0, T]$, hence $E\left[\int_0^t C_s dW_s\right]^2 \leq E[\int_0^t C_s dW_s]^2 = \int_0^t E(C_s^2) ds < \infty$; (iv) Linearity and linearity on adjacent intervals.

A few remarks are in order. If $\theta \in \mathcal{H}^2$, then X solution to $dX_t = \theta_t dW_t$ is a martingale. If $\theta \notin \mathcal{H}^2$, but $\in \mathcal{L}^2$, X is, instead, called a *local martingale*. The converse is the *Martingale Representation Theorem*. This theorem states that if X is a \mathcal{F}_t -martingale, then there exists a $\theta \in \mathcal{H}^2$: $dX_t = \theta_t dW_t$. This result is utilized in the main text of this chapter, when it helps us tell whether we live in a world with complete or incomplete markets. Moreover, in continuous-time finance, θ is often a portfolio strategy. It must be in \mathcal{H}^2 to avoid doubling strategies, which are a kind of arbitrage opportunities (at least in absence

of frictions such as short-selling constraints). Assume, for example, that an asset price is W , and that this asset does not distribute dividends from 0 to T . Then dW is the instantaneous gain from holding one unit of this asset. The condition $\theta \in \mathcal{H}^2$ implies that these strategies cannot become arbitrarily large according to the \mathcal{H}^2 criterion. Moreover, the previous properties of $I_t(\theta) = \int_0^t \theta_s dW_s$ suggest that the “cumulative” gain process $G_t = G_0 + I_t(\theta)$ is a martingale (not only a “local” martingale). Therefore, no investor expects to make profits from investing in this asset.

4.11.1.8 Itô’s lemma: Introduction

We develop, heuristically, a basic version of Itô’s lemma, with its most general version stated further in this appendix. Let $f : \mathbb{R} \mapsto \mathbb{R}$ be twice continuously differentiable. We have:

$$f(W_t) = f(W_0) + \int_0^t f'(W_s) dW_s + \frac{1}{2} \int_0^t f''(W_s) ds, \quad (4A.4)$$

where the first integral is an Itô’s stochastic integral, and second one is a Riemann’s one. For example, let $f(B) = B^2$. Then,

$$W_t^2 = 2 \int_0^t W_s dW_s + \int_0^t ds \Leftrightarrow \int_0^t W_s dW_s = \frac{1}{2} (W_t^2 - t). \quad (4A.5)$$

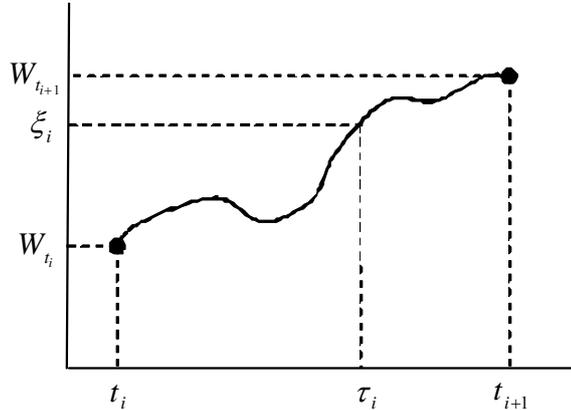
To provide a sketchy proof of Eq. (4A.4), note that:

$$f(W_t) - f(W_0) = \sum_{i=0}^{k-1} [f(W_{t_{i+1}}) - f(W_{t_i})].$$

By Taylor,

$$f(W_{t_{i+1}}) - f(W_{t_i}) = f'(W_{t_i}) (W_{t_{i+1}} - W_{t_i}) + \frac{1}{2} f''(\xi_i) (W_{t_{i+1}} - W_{t_i})^2,$$

where $\min(W_{t_i}, W_{t_{i+1}}) < \xi_i < \max(W_{t_i}, W_{t_{i+1}})$, as in the figure below. Because W is continuous, $\xi_i(\omega) = W_{\tau_i(\omega)}$ for some $\tau_i(\omega) : t_i \leq \tau_i(\omega) \leq t_{i+1}$.



Therefore,

$$f(W_t) - f(W_0) = \sum_{i=0}^{k-1} f'(W_{t_i}) (W_{t_{i+1}} - W_{t_i}) + \frac{1}{2} \sum_{i=0}^{k-1} f''(W_{\tau_i}) (W_{t_{i+1}} - W_{t_i})^2.$$

We have

$$\sum_{i=0}^{k-1} f''(W_{\tau_i}) (W_{t_{i+1}} - W_{t_i})^2 \approx \sum_{i=0}^{k-1} f''(W_{\tau_i}) (t_{i+1} - t_i).$$

Finally,

$$\begin{aligned}\sum_i f'(W_{t_i})(W_{t_{i+1}} - W_{t_i}) &\rightarrow \int f'(W_s) dW_s \\ \sum_i f''(W_{\tau_i})(t_{i+1} - t_i) &\rightarrow \int f''(W_s) ds\end{aligned}$$

More technical details in order of descending difficulty can be found in Karatzas and Shreve (1991), Arnold (1974), Steele (2001) and Mikosch (1998).

Let us reconsider the example in Eq. (4A.4). By the stochastic integral theorem, is a martingale. This is confirmed by Eq. (4A.4). According to Eq. (4A.4),

$$\int_0^t W_s dW_s = \frac{1}{2}(W_t^2 - t)$$

and $(W_t^2 - t)$ is indeed a martingale for $E(W_t^2) = t$ all t .

4.11.2 Stochastic differential equations

4.11.2.1 Background

Consider the differential equation:

$$dx_t = \mu(t, x_t) dt, \quad x_0 = x,$$

for some function μ . Randomness can be introduced via an additional “noise term”:

$$dx_t = \mu(t, x_t) dt + \sigma(t, x_t) dW_t, \quad x_0 = x.$$

We already know that a ω - W_t is not differentiable, so this is only a short-hand notation for:

$$x_t = x_0 + \int_0^t \mu(s, x_s) ds + \int_0^t \sigma(s, x_s) dW_s, \quad (4A.6)$$

where the first integral is Riemann and the second integral is an Itô's stochastic integral.

We have the following definitions. First, we say that an *Itô's process* is,

$$dx_t(\omega) = \mu_t(\omega) dt + \sigma_t(\omega) dW_t, \quad x_0 = x.$$

Moreover, we say that an *Itô's diffusion process* is,

$$dx(t) = \mu(t, x(t)) dt + \sigma(t, x(t)) dW(t), \quad x_0 = x.$$

It is known that an Itô's diffusion process is a Markov process. The previous equation is also called a *stochastic differential equation* (SDE). In a SDE, μ and σ “depend” on ω only through x . Finally, we say that a *time-homogeneous diffusion process* is,

$$dx(t) = \mu(x(t)) dt + \sigma(x(t)) dW(t), \quad x_0 = x.$$

There is a beautiful property that is used to price financial derivatives, using replication arguments, as explained in the main text, called the *unique decomposition property*. Suppose we were given two processes x and y with $x_0 = y_0$, and that:

$$dx_t = \mu_t^x dt + \sigma_t^x dW_t \quad \text{and} \quad dy_t = \mu_t^y dt + \sigma_t^y dW_t.$$

Then $x_t = y_t$ almost surely if and only if $\mu_t^x = \mu_t^y$ and $\sigma_t^x = \sigma_t^y$ almost everywhere, in the sense that $E[\int_0^T |a_t^x - a_t^y| dt = 0] = E[\int_0^T |b_t^x - b_t^y| dt] = 0$.

4.11.2.2 Basic definitions, properties and regularity conditions

How do we know whether the various integrals given before are well-defined. As an example, the Itô's integral representation $\int_0^t \sigma(s, x_s) dW_s$ works if σ is \mathcal{F}_t -adapted and $\int_0^t E[\sigma(s, x_s)^2] ds < \infty$. But how can be sure that these two basic conditions are satisfied if we don't know yet the *solution* of x ? And, above all, what is a solution to a SDE? We have two concepts of such a solution, *strong* and *weak*.

DEFINITION 4A.5. (STRONG SOLUTION TO A SDE) *A strong solution to Eq. (4A.6) is a stochastic process $x = (x_t, t \in [0, T])$ such that:*

- (i) x is \mathcal{F}_t -adapted.
- (ii) The integrals in Eq. (4A.6) are well-defined in the Riemann's and Itô's sense and Eq. (4A.6) holds $P \otimes dt$ -almost surely
- (iii) $E\left(\int_0^T |x_s|^2 ds\right) < \infty$.

In other words, the definition of a *strong* solution requires that a Brownian motion be "given in advance," and that the solution x_t constructed from it be then \mathcal{F}_t -adapted.

Next, suppose, instead, that we were only given x_0 and two functions $\sigma(t, x)$ and $\mu(t, x)$, and that we were asked to find a pair of processes (\tilde{x}, \tilde{W}) on some probability space $(\Omega, \tilde{\mathcal{F}}, P)$ such that Eq. (4A.6) holds with \tilde{x} being $\tilde{\mathcal{F}}_t$ -adapted on some space, not necessarily the one in Eq. (4A.6). (Clearly such a \tilde{x} needs not to be \mathcal{F}_t -adapted.) In this case (\tilde{x}, \tilde{W}) is called a *weak solution* on $(\Omega, \tilde{\mathcal{F}}, P)$. In the case of a weak solution, we are given x, μ, σ and then "we have to find" two things: a Brownian motion \tilde{W} and a $\tilde{\mathcal{F}}_t$ -adapted process \tilde{x} such that $\tilde{x}_t = x_0 + \int_0^t \mu(s, \tilde{x}_s) ds + \int_0^t \sigma(s, \tilde{x}_s) d\tilde{W}_s$ holds $P \otimes dt$ -almost surely. Clearly, a strong solution is also weak, but the converse is not true. Consider the following example.

EXAMPLE 4A.6. (TANAKA EQUATION) Let $x(t)$ satisfy:

$$dx(t) = \text{sign}(x(t))dW(t), \quad x_0 = 0. \quad (4A.7)$$

This equation has no strong solutions, for define

$$y(t) = \int_0^t \text{sign}(\hat{W}(s))d\hat{W}(s), \quad (4A.8)$$

where \hat{W} is a Brownian motion. It can be shown that $y(t)$ is $\mathcal{G}(t)$ -measurable, where $\mathcal{G}(t)$ is the σ -algebra generated by $|\hat{W}(t)|$. Clearly $\mathcal{G}(t) \subset \hat{\mathcal{F}}(t)$, where $\hat{\mathcal{F}}(t)$ is the σ -algebra generated by $\hat{W}(t)$. Therefore, the σ -algebra generated by $y(t)$ is also strictly contained in $\hat{\mathcal{F}}(t)$. Armed with this result, we can easily show that there are no strong solutions to Eq. (4A.7). To show this, suppose the contrary. There is a theorem saying that $x(t)$ would then be a Brownian motion. On the other hand, Eq. (4A.7) can also be written

$$dW(t) = \text{sign}(x(t))dx(t), \quad x_0 = 0,$$

or

$$W(t) = \int_0^t \text{sign}(x(s))dx(s).$$

By the same reasoning produced to show that the σ -algebra generated by $y(t)$ is strictly contained in $\hat{\mathcal{F}}(t)$ in Eq. (4A.8), we conclude that the σ -algebra generated by $W(t)$ is strictly contained in the σ -algebra generated by $x(t)$. But this contradicts that $x(t)$ is a strong solution to Eq. (4A.7).

Clearly, one needs to be able to impose conditions that allow to distinguish between weak and strong solutions. However, the only focus of the following discussion is about the regularity conditions ensuring existence and uniqueness of strong solutions—the case of interest in continuous-time finance. We need two types of restrictions on μ and σ . Consider the following definition. For a given function f , we say that it satisfies a Lipschitz condition in x if there exists a constant L , such that for all $(x, y) \in \mathbb{R}^d \times \mathbb{R}^d$,

$$\|f(x, t) - f(y, t)\| \leq L \|x - y\| \quad \text{uniformly in } t.$$

where $\|A\| \equiv \sqrt{\text{Tr}(AA^\top)}$. In other words, f cannot change too widely. We also say f satisfies a growth condition in x , if there exists a constant G such that for all $(x, y) \in \mathbb{R}^d \times \mathbb{R}^d$,

$$\|f(x, t)\|^2 \leq G \left(1 + \|x\|^2\right) \quad \text{uniformly in } t.$$

That is, f cannot grow too much.

Next, we turn to the concepts of existence and uniqueness of a solution to a stochastic differential equation. We say that if $x_t^{(1)}(\omega)$ and $x_t^{(2)}(\omega)$ are both strong solutions to Eq. (4A.6), then $x_t^{(1)}(\omega) = x_t^{(2)}(\omega)$ $P \otimes dt$ -a.s. We have:

THEOREM 4A.7. *Suppose that μ, σ satisfy Lipschitz and growth conditions in x , then there exists a unique Itô's process x satisfying Eq. (4A.6) which is continuous adapted Markov.*

Consider the following stochastic differential equation:

$$dx(t) = \mu(a - x(t))dt + \sigma\sqrt{x(t)}dW(t),$$

for some constants μ, a, σ . This is the so-called square-root process utilized to model equity volatility (see Chapter 10), the short-term rate (see Chapter 12) or instantaneous probabilities of default of debt issuers (see Chapter 13). The point here, for now, is that the diffusion component does not satisfy the conditions in Theorem 4A.7. Yet it is possible to show that under suitable parameter restrictions there exists a strong solution. Incidentally, the solution to this simple equation is still unknown.

What about *uniqueness* of the solution? It is well-known that if μ, σ are locally Lipschitz continuous in x , then strong uniqueness holds. But even for ordinary differential equations, a local Lipschitz condition is not necessarily enough to guarantee global existence (i.e. for all t) of a solution. For example, consider the following equation:

$$\frac{dx(t)}{dt} = \mu(x(t)) \equiv x^2(t), \quad x_0 = 1,$$

has as unique solution:

$$x(t) = \frac{1}{1-t}, \quad 0 \leq t < 1.$$

Yet is impossible to find a global solution, i.e. one defined for all t . This is exactly the kind of pathology ruled out by linear-growth conditions. More generally, linear-growth conditions ensure that $|x_t(\omega)|$ is unique and doesn't explode in finite time. Naturally, Lipschitz and growth conditions are only *sufficient* conditions to guarantee the previous conclusions.

A final remark. The uniqueness concept used here refers to *strong* or *pathwise* uniqueness. There are also definitions of *weak* uniqueness to mean that any two solutions (weak or strong) have the same finite-dimensional distributions. For example, the Tanaka's equation introduced earlier has no strong solution, yet it can be shown that it has a (weakly) unique weak solution.

4.11.2.3 Itô's lemma

Itô's lemma is a fundamental tool of analysis in continuous-time finance. It helps build up “new” processes from “old” processes. Two examples might clarify.

- (i) A share price is certainly a function of its dividend process. If the dividend process is solution to some SDE, then the asset price is a solution to another SDE. Which SDE? Itô's lemma will give us the answer.
- (ii) Derivative products, reviewed in the third part of these lectures, are financial instruments, with a value depending on some underlying factors, whence, the terminology, “derivative.” In other words, derivative prices are functions of these factors. If factors are solutions to SDE, derivative prices are also solutions to SDE. Once again, Itô's lemma will provide us with the right SDE.

Naturally, the functional form linking the dividend process (or the factors) to the asset prices is unknown. But in situations of interest, no-arbitrage restrictions will help to pin down such a functional form.

Let us proceed with a few preliminary heuristic considerations. A useful heuristic definition is that the increments of a Brownian motion, $dW(t)$, can be thought of as being equal to $W(T + \Delta t) - W(t)$ as $\Delta t \rightarrow 0$. We may think of the “increments” $dW(t)$ as being normally distributed, $dW(t) \sim N(0, dt)$. Heuristically, indeed, $\Delta W(t) \equiv W(t + \Delta t) - W(t) \sim N(0, \Delta t)$. But then, by the previous normality property of $\Delta W(t)$,

$$E[\Delta W(t)] = 0 \text{ and } E[(\Delta W(t))^2] = \Delta t, \text{ hence } \text{var}[\Delta W(t)] = \Delta t, \text{ and } \text{var}[(\Delta W(t))^2] = 2(\Delta t)^2,$$

where the second equality follows by the property χ^2 distributions.

The point of the previous computations is that for small Δt , the variance of $(\Delta W(t))^2$, which is proportional to $(\Delta t)^2$, is negligible if compared to its expectation, which is Δt . Heuristically, $Q_n(t) \xrightarrow{q.m.} t$ and $(dW(t))^2 \equiv Q_n(dt) \xrightarrow{q.m.} dt$. These heuristic considerations lead to the following, celebrated table below.

<i>Itô's multiplication table</i>		
$(dt)^n$	= 0	for $n > 1$
$dt \cdot dW$	= 0	
$(dW)^2$	= dt	
$(dW)^n$	= 0	for $n > 2$
$dW_1 dW_2$	= 0	for two independent Brownian motions

We now use this table, and heuristically derive Itô's lemma. Let $x(t)$ be the solution to,

$$dx(t) = \mu(t) dt + \sigma(t) dW(t),$$

and suppose we are given a function $f(x, t)$, which we assume to be as differentiable in (x, t) as many times as needed below. We expand f as follows:

$$df(x, t) = f_t(x, t) dt + f_x(x, t) dx + \frac{1}{2} f_{xx}(x, t) (dx)^2 + \text{Remainder},$$

where the remainder contains only terms of order higher than $(dx)^2$ and $(dt)^2$. So for reasons which will be clear in one moment we will discard it. We have,

$$\begin{aligned} df &= f_t dt + f_x dx + \frac{1}{2} f_{xx} (dx)^2 \\ &= f_t dt + f_x (\mu dt + \sigma dW) + \frac{1}{2} f_{xx} (\mu dt + \sigma dW)^2 \\ &= f_t dt + f_x \mu dt + f_x \sigma dW + \frac{1}{2} f_{xx} [\mu^2 (dt)^2 + \sigma^2 (dW)^2 + 2\mu\sigma (dt \cdot dW)]. \end{aligned}$$

By the Itô's multiplication table,

$$\begin{aligned} df &= f_t dt + f_x \mu dt + f_x \sigma dW + \frac{1}{2} f_{xx} \left[\mu^2 (dt)^2 + \sigma^2 (dW)^2 + 2\mu\sigma (dt \cdot dW) \right] \\ &= f_t dt + f_x \mu dt + f_x \sigma dW + \frac{1}{2} f_{xx} \left[\begin{array}{ccc} 0 & + & \sigma^2 \cdot dt & + & 0 \end{array} \right] \end{aligned}$$

By rearranging terms,

$$df(x, t) = \left[f_t(x, t) + f_x(x, t) \mu + \frac{1}{2} f_{xx}(x, t) \sigma^2 \right] dt + f_x(x, t) \sigma dW,$$

and the remainder is also zero by the Itô's multiplication table. This is *Itô's lemma*.

Naturally, Itô's lemma also holds when x is a multidimensional process. A heuristic derivation of it can be obtained through the Itô's multiplication table applied to the following expansion:

$$df(x, t) = f_t dt + f_x dx + \frac{1}{2} \sum_{i,j} f_{x_i x_j} dx_i dx_j.$$

Then, we have:

THEOREM 4A.8. (ITÔ'S LEMMA, MULTIDIMENSIONAL) *Let us be given a multidimensional process $x \in R^n$ solution to,*

$$dx(t) = \mu(x(t), t) dt + \sigma(x(t), t) dW(t), \quad (4A.9)$$

where μ is in R^n , σ is in $R^{n \times d}$ and W is a d -dimensional vector of independent Brownian motions. Moreover, let us be given a function $f(x, t)$, which is twice continuously differentiable in x and differentiable in t . Then f is an Itô's process, solution to:

$$df(x(t), t) = \mathcal{L}f(x(t), t) dt + f_x(x(t), t) \sigma(t) dW(t)$$

or more formally,

$$f(x(t), t) = f(x_0, 0) + \int_0^t \mathcal{L}f(x(s), s) ds + \int_0^t f_x(x(s), s) \sigma(x(s), s) dW(s), \quad (4A.10)$$

where

$$\mathcal{L}f(x, t) = f_t(x, t) + f_x(x, t) \mu + \frac{1}{2} \text{Tr} \left[\sigma \sigma^\top f_{xx}(x, t) \right]$$

and $f_x(x, t)$ and $f_{xx}(x, t)$ are the gradient and Hessian of f with respect to x .

Note that by Eq. (4A.10), and provided $f_x \sigma \in \mathcal{H}^2$, f is a martingale whenever $\mathcal{L}f(x, t) = 0$, for all x, t . As a matter of terminology, the operator $\mathcal{A}f(x, t) = f_x(x, t) \mu + \frac{1}{2} \text{Tr} \left[\sigma \sigma^\top f_{xx}(x, t) \right]$ is usually referred to as the *infinitesimal generator* of the diffusion process in Eq. (4A.9).

4.12 Appendix 3: Proof of selected results

4.12.1 Proof of Theorem 4.2

As mentioned in the main text, we have that by the Girsanov's theorem, \mathcal{Q} is non-empty if and only if Eq. (4.48) holds true. Therefore, the proof will rely on Eq. (4.48). *If part.* With $c \equiv 0$, Eq. (4.49) is:

$$\frac{V^{x,\pi,0}(\tau)}{S_0(\tau)} = x + \int_t^\tau \frac{\pi^\top(u) \sigma(u)}{S_0(u)} dW_0(u), \quad \tau \in [t, T],$$

which implies, $x = E_\tau^{\mathcal{Q}}[S_0(T)^{-1} V^{x,\pi,0}(T)]$. An arbitrage opportunity is $V^{x,\pi,0}(t) \leq S_0(T)^{-1} V^{x,\pi,0}(T)$ a.s., which combined with the previous equality leaves: $V^{x,\pi,0}(t) = S_0(T)^{-1} V^{x,\pi,0}(T)$ \mathcal{Q} -a.s. (if a r.v. $\tilde{y} \geq 0$ and $E_t(\tilde{y}) = 0$, this means that $\tilde{y} = 0$ a.s.) and, hence, P -a.s. The last equality is in contradiction with $\Pr(S_0(T)^{-1} V^{x,\pi,0}(T) - x > 0) > 0$, as required by Definition 4.3.

Only if part. We combine portions of proofs in Karatzas (1997, thm. 0.2.4 pp. 6-7) and Øksendal (1998, thm. 12.1.8b, pp. 256-257). We let:

$$\begin{aligned} Z(\tau) &= \{\omega \in \Omega : \text{Eq. (4.48) has no solutions}\} \\ &= \{\omega \in \Omega : a(\tau; \omega) - \mathbf{1}_m r(\tau; \omega) \notin \langle \sigma \rangle\} \\ &= \left\{ \omega \in \Omega : \exists \underline{\pi}(\tau; \omega) : \underline{\pi}(\tau; \omega)^\top \sigma(\tau; \omega) = 0 \text{ and } \underline{\pi}(\tau; \omega)^\top (a(\tau; \omega) - \mathbf{1}_m r(\tau; \omega)) \neq 0 \right\}, \end{aligned}$$

and consider the following portfolio,

$$\hat{\pi}(\tau; \omega) = \begin{cases} k \cdot \text{sign} [\underline{\pi}(\tau; \omega)^\top (a(\tau; \omega) - \mathbf{1}_m r(\tau; \omega))] \cdot \underline{\pi}(\tau; \omega) & \text{for } \omega \in Z(\tau) \\ 0 & \text{for } \omega \notin Z(\tau) \end{cases}$$

Clearly $\hat{\pi}$ is $(\tau; \omega)$ -measurable, and generates, by Eq. (4.44),

$$\begin{aligned} \frac{V^{x,\hat{\pi},0}(\tau)}{S_0(\tau)} &= x + \int_t^\tau \frac{\hat{\pi}^\top(u) (a(u) - \mathbf{1}_m r(u))}{S_0(u)} \mathbb{I}_{Z(u)} du + \int_t^\tau \frac{\hat{\pi}^\top(u) \sigma(u)}{S_0(u)} \mathbb{I}_{Z(u)} dW(u) \\ &= x + \int_t^\tau \left(\frac{\hat{\pi}^\top(u) (a(u) - \mathbf{1}_m r(u))}{S_0(u)} \right) \mathbb{I}_{Z(u)} du \\ &\geq x. \end{aligned}$$

So the market has no arbitrage only if $\mathbb{I}_{Z(u)} = 0$, i.e. only if Eq. (4.48) has at least one solution. \parallel

4.12.2 Proof of Eq. (4.53).

We have:

$$\begin{aligned} x &= \mathbb{E} \left[\frac{V^{x,\pi,c}(T)}{S_0(T)} + \int_t^T \frac{c(u)}{S_0(u)} du \right] \\ &= \zeta(t)^{-1} E \left[\frac{\zeta(T) V^{x,\pi,c}(T)}{S_0(T)} + \int_t^T \frac{\zeta(T) c(u)}{S_0(u)} du \right] \\ &= \zeta(t)^{-1} E \left[\frac{\zeta(T) V^{x,\pi,c}(T)}{S_0(T)} + \int_t^T E \left(\frac{\zeta(T) c(u)}{S_0(u)} \middle| \mathcal{F}(u) \right) du \right] \\ &= \zeta(t)^{-1} E \left[\frac{\zeta(T) V^{x,\pi,c}(T)}{S_0(T)} + \int_t^T \frac{E(\zeta(T) | \mathcal{F}(u)) c(u)}{S_0(u)} du \right] \\ &= \zeta(t)^{-1} E \left[\frac{\zeta(T) V^{x,\pi,c}(T)}{S_0(T)} + \int_t^T \frac{\zeta(u) c(u)}{S_0(u)} du \right] \\ &= E \left[m_{t,T} \cdot V^{x,\pi,c}(T) + \int_t^T m_{t,u} \cdot c(u) du \right], \end{aligned}$$

where we used the fact that c is adapted, the law of iterated expectations, the martingale property of ζ , and the definition of $m_{0,t}$.

4.12.3 Walras's consistency tests

First, we show that Eq. (4.60) \Rightarrow Eq. (4.61). To grasp intuition about the ongoing proof, consider the two-period economy of Chapter 2. In that economy, absence of arbitrage opportunities implies that $\exists \phi \in \mathbb{R}^d : \phi^\top (c^1 - w^1) = S\theta = -(c_0 - w_0)$, whence $c_s = w_s, s = 0, \dots, d \iff \theta = \mathbf{0}_m$. In the model of this chapter, absence of arbitrage opportunities implies that there exists a unique $Q \in \mathcal{Q}$ such that:

$$\frac{V^{x,\pi,c}(\tau)}{S_0(\tau)} \equiv \frac{\theta_0(\tau)S_0(\tau) + \pi^\top(\tau)\mathbf{1}_m}{S_0(\tau)} = \mathbf{1}_m^\top S(t) + \int_t^\tau \frac{\pi^\top \sigma(u)}{S_0(u)} dW_0(u) - \int_t^\tau \frac{c(u)}{S_0(u)} du.$$

That is,

$$\begin{aligned} & \frac{\theta_0(\tau)S_0(\tau) + (\pi^\top(\tau) - S^\top(\tau))\mathbf{1}_m}{S_0(\tau)} + \frac{S^\top(\tau)\mathbf{1}_m}{S_0(\tau)} \\ &= \mathbf{1}_m^\top S(t) + \int_t^\tau \frac{(\pi^\top(u) - S^\top(u))\sigma(u)}{S_0(u)} dW_0(u) - \int_t^\tau \frac{c(u)}{S_0(u)} du + \int_t^\tau \frac{S^\top(u)\sigma(u)}{S_0(u)} dW_0(u). \end{aligned}$$

Plugging the solution $\left(\frac{S_i}{S_0}\right)(\tau) = S_i(t) + \int_t^\tau (S_0^{-1}S_i)(u)\sigma_i(u)dW_0(u) - \int_t^\tau (S_0^{-1}D_i)(u)du$ in the previous relation,

$$\frac{\theta_0(T)S_0(T) + (\pi^\top(T) - S^\top(T))\mathbf{1}_m}{S_0(T)} = \int_t^T \frac{\pi^\top(u) - S^\top(u)}{S_0(u)} \sigma(u) dW_0(u) + \int_t^T \frac{D(u) - c(u)}{S_0(u)} du. \quad (4A.11)$$

When Eq. (4.60) holds, we have that $V^{x,\pi,c}(T) = \theta_0(T)S_0(T) + \pi^\top(T)\mathbf{1}_m = \underline{q}(T) = S^\top(T)\mathbf{1}_m$, and $D = c$, and Eq. (4A.11) becomes:

$$0 = x(T) \equiv \int_t^T \frac{\pi^\top(u) - S^\top(u)}{S_0(u)} \sigma(u) dW_0(u),$$

a martingale starting at zero, satisfying:

$$dx(\tau) = \frac{\pi^\top(\tau) - S^\top(\tau)}{S_0(\tau)} \sigma(\tau) dW_0(\tau) = 0.$$

Since $\ker(\sigma) = \{\emptyset\}$ then, we have that $\pi(\tau) = S(\tau)$ a.s. for $\tau \in [t, T]$ and, hence, $\pi(\tau) = S(\tau)$ a.s. for $\tau \in [t, T]$. It is easily checked that this implies $\theta_0(T) = 0$ P -a.s. and that in fact, $\theta_0(\tau) = 0$ a.s.

Next, we show that Eq. (4.61) \Rightarrow Eq. (4.60). When Eq. (4.61) holds, Eq. (4A.11) becomes:

$$0 = y(T) \equiv \int_t^T \frac{D(u) - c(u)}{S_0(u)} du,$$

a martingale starting at zero. We conclude by the same arguments used in the proof of the previous part. \parallel

4.13 Appendix 4: The Green's function

4.13.1 Setup

In Section 4.6, it is shown that in frictionless markets, the value of a security as of time τ is:

$$V(x(\tau), \tau) = E \left[\frac{m_{t,T}}{m_{t,\tau}} V(x(T), T) + \int_{\tau}^T \frac{m_{t,s}}{m_{t,\tau}} h(x(s), s) ds \right], \quad (4A.12)$$

where $m_{t,\tau}$ is the stochastic discount factor,

$$m_{t,\tau} = \frac{\zeta(t, \tau)}{S_0(\tau)} = \frac{1}{S_0(\tau)} \cdot \frac{dQ}{dP} \Big|_{\mathcal{F}(\tau)}.$$

The Arrow-Debreu state price density is:

$$\phi_{t,T} = m_{t,T} dP = \frac{S_0(t)}{S_0(T)} dQ.$$

Our aim is to characterize this density in terms of partial differential equations. By the same reasoning produced in Section 4.6, Eq. (4A.12) can be rewritten as:

$$V(x(\tau), \tau) = \mathbb{E} \left[a(\tau, T) V(x(T), T) + \int_{\tau}^T a(\tau, s) h(x(s), s) ds \right], \quad a(t', t'') \equiv \frac{S_0(t')}{S_0(t'')}. \quad (4A.13)$$

Next, consider the state vector, $y(u) \equiv (a(\tau, u), x(u))$, $\tau \leq u \leq T$, and let $q(y(t')|y(\tau))$ be the risk-neutral density of y . We have,

$$\begin{aligned} V(x(\tau), \tau) &= \mathbb{E} \left[a(\tau, T) V(x(T), T) + \int_{\tau}^T a(\tau, s) h(x(s), s) ds \right] \\ &= \int a(\tau, T) V(x(T), T) q(y(T)|y(\tau)) dy(T) + \int_{\tau}^T \int a(\tau, s) h(x(s), s) q(y(s)|y(\tau)) dy(s) ds. \end{aligned}$$

If $V(x(T), T)$ and $a(\tau, T)$ are independent,

$$\int a(\tau, T) V(x(T), T) q(y(T)|y(\tau)) dy(T) = \int_X G(\tau, T) V(x(T), T) dx(T)$$

where:

$$G(\tau, T) \equiv \int_A a(\tau, T) q(y(T)|y(\tau)) dy(T).$$

Assuming the same for h ,

$$V(x(\tau), \tau) = \int_X G(\tau, T) V(x(T), T) dx(T) + \int_{\tau}^T \int_X G(\tau, s) h(x(s), s) dx(s) ds.$$

The function G is known as the *Green's function*:

$$G(t, \ell) \equiv G(x, t; \xi, \ell) = \int_A a(t, \ell) q(y(\ell)|y(t)) da.$$

It is the value in state $x \in \mathbb{R}^d$ as of time t of a unit of numéraire at $\ell > t$ if future states lie in a neighborhood (in \mathbb{R}^d) of ξ . It is thus the Arrow-Debreu state-price density.

For example, a pure discount bond has $V(x, T) = 1 \forall x$, and $h(x, s) = 1 \forall x, s$, and

$$V(x(\tau), \tau) = \int_X G(x(\tau), \tau; \xi, T) d\xi, \quad \text{with } \lim_{\tau \uparrow T} G(x(\tau), \tau; \xi, T) = \delta(x(\tau) - \xi),$$

where δ is the Dirac's delta.

4.13.2 The PDE connection

We show the Green's function satisfies the same partial differential equation (PDE) satisfied by the security price, but with a different boundary condition, and with the instantaneous dividend taken out. We have:

$$V(x(t), t) = \int_X G(x(t), t; \xi(T), T) V(\xi(T), T) d\xi(T) + \int_t^T \int_X G(x(t), t; \xi(s), s) h(\xi(s), s) d\xi(s) ds. \quad (4A.14)$$

Consider the scalar case. By Eq. (4A.13), and the Feynman-Kac connection between PDEs and conditional expectations reviewed in Section 4.2, we have that under regularity conditions, V is solution to:

$$0 = V_t + \mu V_x + \frac{1}{2} \sigma^2 V_{xx} - rV + h, \quad (4A.15)$$

where μ is the risk-neutral drift of x . Next, take the following partial derivatives of $V(x, t)$ in Eq. (4A.14):

$$\begin{aligned} V_t &= \int_X G_t V d\xi - \int_X \delta(x - \xi) h d\xi + \int_t^T \int_X G_t h d\xi ds = \int_X G_t V d\xi - h + \int_t^T \int_X G_t h d\xi ds \\ V_x &= \int_X G_x V d\xi + \int_t^T \int_X G_x h d\xi ds \\ V_{xx} &= \int_X G_{xx} V d\xi + \int_t^T \int_X G_{xx} h d\xi ds \end{aligned}$$

and replace them into Eq. (4A.15) to obtain:

$$\begin{aligned} 0 &= \int_X \left[G_t + \mu G_x + \frac{1}{2} \sigma^2 G_{xx} - rG \right] V(\xi(T), T) d\xi(T) \\ &\quad + \int_t^T \int_X \left[G_t + \mu G_x + \frac{1}{2} \sigma^2 G_{xx} - rG \right] h(\xi(s), s) d\xi(s) ds. \end{aligned}$$

This shows that G is solution to

$$0 = G_t + \mu G_x + \frac{1}{2} \sigma^2 G_{xx} - rG, \quad \text{with } \lim_{t \uparrow T} G(x, t; \xi, T) = \delta(x - \xi).$$

4.14 Appendix 5: Portfolio constraints

We are looking for a portfolio-consumption policy $(p_{\hat{\nu}}, c_{\hat{\nu}})$ such that

$$\text{Val}(x; K) = E \left[\int_0^T u(t, c_{\hat{\nu}}(t)) dt + U(V^{x, p_{\hat{\nu}}, c_{\hat{\nu}}}(T)) \right] \equiv \text{Val}_{\hat{\nu}}(x), \quad (4A.16)$$

and $p_{\hat{\nu}}(t) \in K$ for all $t \in [0, T]$.

Note that because K contains the origin, then, the support function ζ in Eq. (4.64) satisfies $\zeta(\nu) \geq 0$ for each $\nu \in \tilde{K}$. Moreover, an intuitive and important property of ζ is that,

$$p \in K \iff \zeta(\nu) + p^\top \nu \geq 0, \quad \forall \nu \in \tilde{K}. \quad (4A.17)$$

Next, define the standard Brownian motion under the probability Q^ν , defined through the Radon-Nikodym in Eq. (4.66):

$$W_\nu(t) = W(t) + \int_0^t (\lambda(u) + \sigma^{-1}(u) \nu(u)) du \equiv W_0(t) + \int_0^t (\sigma^{-1}(u) \nu(u)) du,$$

where $\lambda = \sigma^{-1}(a - \mathbf{1}_d r)$, and W_0 is the usual Brownian under the risk-neutral probability in a market without any frictions. If the price system is as in Eqs. (4.65), then, for *any* unconstrained portfolio-consumption (p, c) , the dynamics of wealth, $V_\nu^{x, p, c}$ say, are easily seen to be:

$$dV_\nu^{x, p, c} = \left(p^\top \nu + \zeta(\nu) V_\nu^{x, p, c} + r V_\nu^{x, p, c} - c \right) dt + p^\top \sigma dW_0.$$

So we have that under Q^0 ,

$$\begin{aligned} & \frac{V_\nu^{x, p, c}(T)}{S_0(T)} + \int_0^T \frac{c(t)}{S_0(t)} dt \\ &= x + \int_0^T \frac{V_\nu^{x, p, c}(t)}{S_0(t)} \left[p^\top(t) \nu(t) + \zeta(\nu(t)) \right] dt + \int_0^T \frac{V_\nu^{x, p, c}(t)}{S_0(t)} p^\top(t) \sigma(t) dW_0(t). \end{aligned}$$

Therefore, for any normalized portfolio-consumption (p, c) , we have that the wealth difference, $\Delta(t) \equiv \frac{V_\nu^{x, \pi, c}(T) - V_\nu^{x, p, c}(T)}{S_0(T)}$, satisfies:

$$d\Delta(t) = \underbrace{\frac{V_\nu^{x, \pi, c}(t)}{S_0(t)} \left[p^\top(t) \nu(t) + \zeta(\nu(t)) \right]}_{\equiv m(t)} dt + \Delta(t) p^\top(t) \sigma(t) dW(t), \quad \Delta(0) = 0.$$

Next, consider, the simpler equation,

$$d\bar{\Delta}(t) = \bar{\Delta}(t) p^\top(t) \sigma(t) dW(t), \quad \bar{\Delta}(0) = 0. \quad (4A.18)$$

Because $m(t) \geq 0$ by Eq. (4A.17), then, by a comparison theorem (e.g., Karatzas and Shreve (1991, p. 291-295)), $\Delta(t) \geq \bar{\Delta}(t) = 0$, where the last equality follows because the solution to Eq. (4A.18) is $\bar{\Delta}(t) = \bar{\Delta}(0) L(t)$, for some positive process $L(t)$. Therefore, we have,

$$V_\nu^{x, p, c}(t) \geq V_\nu^{x, \pi, c}(t), \quad \text{with an equality if } \zeta(\nu(t)) + p^\top \nu(t) = 0 \text{ for all } t. \quad (4A.19)$$

Finally, suppose there is a constrained portfolio-consumption pair $(p_{\hat{\nu}}, c_{\hat{\nu}})$, such that

$$\zeta(\hat{\nu}(t)) + p^\top(t) \hat{\nu}(t) = 0. \quad (4A.20)$$

Naturally, we have that $\text{Val}(x; K) \leq \text{Val}_\nu(x)$ for all ν and, hence,

$$\text{Val}(x; K) \leq \inf_{\nu \in \tilde{K}} (\text{Val}_\nu(x)). \quad (4A.21)$$

Moreover, we have,

$$\begin{aligned} \text{Val}(x; K) &= E \left[\int_0^T u(t, c(t)) dt + U(V^{x, p, c}(T)) \right], \quad p(t) \in K \\ &\geq E \left[\int_0^T u(t, c_{\hat{\nu}}(t)) dt + U(V^{x, p_{\hat{\nu}}, c_{\hat{\nu}}}(T)) \right] \\ &= E \left[\int_0^T u(t, c_{\hat{\nu}}(t)) dt + U(V_{\hat{\nu}}^{x, p_{\hat{\nu}}, c_{\hat{\nu}}}(T)) \right] \\ &= \text{Val}_{\hat{\nu}}(x), \end{aligned} \quad (4A.22)$$

where the second line follows, because the value of the unconstrained problem is, of course, the largest we may have, once we consider any arbitrary *constrained* portfolio-consumption $(p_{\hat{\nu}}, c_{\hat{\nu}})$. The third line follows by Eq. (4A.20) and (4A.19). The fourth line is the definition of $\text{Val}_\nu(x)$. Combining (4A.21) with (4A.22) leaves,

$$\text{Val}(x; K) = \text{Val}_{\hat{\nu}}(x).$$

The converse, namely “if there exists a $\hat{\nu} \in \tilde{K}$ that minimizes $\text{Val}_{\hat{\nu}}(x)$, then, the corresponding portfolio-consumption process $(p_{\hat{\nu}}, c_{\hat{\nu}})$ is optimal for the constrained problem,” is also true, but its arguments (even informal) are omitted here.

4.15 Appendix 6: Models with final consumption only

Sometimes, we may be interested in models with consumption taking place in at the end of the period only. Let $\bar{S} = (S^{(0)}, S)^\top$ and $\bar{\theta} = (\theta^{(0)}, \theta)$, where θ and S are both m -dimensional. Define as usual wealth as of time t as $V_t \equiv \bar{S}_t \bar{\theta}_t$. There are no dividends. A *self-financing* strategy $\bar{\theta}$ satisfies,

$$\bar{S}_t^+ \bar{\theta}_{t+1} = \bar{S}_t \bar{\theta}_t \equiv V_t, \quad t = 1, \dots, T.$$

Therefore,

$$\begin{aligned} V_t &= \bar{S}_t \bar{\theta}_t + \bar{S}_{t-1} \bar{\theta}_{t-1} - \bar{S}_{t-1} \bar{\theta}_{t-1} \\ &= \bar{S}_t \bar{\theta}_t + \bar{S}_{t-1} \bar{\theta}_{t-1} - \bar{S}_{t-1} \bar{\theta}_t \quad (\text{because } \bar{\theta} \text{ is self-financing}) \\ &= V_{t-1} + \Delta \bar{S}_t \bar{\theta}_t, \quad \Delta \bar{S}_t \equiv \bar{S}_t - \bar{S}_{t-1}, \quad t = 1, \dots, T, \end{aligned}$$

or,

$$V_t = V_1 + \sum_{n=1}^t \Delta \bar{S}_n \bar{\theta}_n.$$

Next, suppose that

$$\Delta S_t^{(0)} = r_t S_{t-1}^{(0)}, \quad t = 1, \dots, T,$$

with $\{r_t\}_{t=1}^T$ given and to be defined more precisely below. The term $\Delta \bar{S}_t \bar{\theta}_t^+$ can then be rewritten as:

$$\begin{aligned} \Delta \bar{S}_t \bar{\theta}_t &= \Delta S_t^{(0)} \theta_t^{(0)} + \Delta S_t \theta_t \\ &= r_t S_{t-1}^{(0)} \theta_t^{(0)} + \Delta S_t \theta_t \\ &= r_t S_{t-1}^{(0)} \theta_t^{(0)} + r_t S_{t-1} \theta_t - r_t S_{t-1} \theta_t + \Delta S_t \theta_t \\ &= r_t \bar{S}_{t-1} \bar{\theta}_t - r_t S_{t-1} \theta_t + \Delta S_t \theta_t \\ &= r_t \bar{S}_{t-1} \bar{\theta}_{t-1} - r_t S_{t-1} \theta_t + \Delta S_t \theta_t \quad (\text{because } \bar{\theta} \text{ is self-financing}) \\ &= r_t V_{t-1} - r_t S_{t-1} \theta_t + \Delta S_t \theta_t, \end{aligned}$$

and we obtain

$$V_t = (1 + r_t) V_{t-1} - r_t S_{t-1} \theta_t + \Delta S_t \theta_t,$$

or,

$$V_t = V_1 + \sum_{n=1}^t (r_n V_{n-1} - r_n S_{n-1} \theta_n + \Delta S_n \theta_n).$$

Next, considering “small” time intervals. In the limit we obtain:

$$dV(t) = r(t)V(t)dt - r(t)S(t)\theta(t)dt + dS(t)\theta(t).$$

Such an equation can also be arrived at by noticing that current wealth is nothing but initial wealth plus gains from trade accumulated up to now:

$$V(t) = V(0) + \int_0^t d\bar{S}(u)\bar{\theta}(u).$$

\Leftrightarrow

$$\begin{aligned} dV(t) &= d\bar{S}(t)\bar{\theta}(t)^+ \\ &= dS_0(t)\theta_0(t) + dS(t)\theta(t) \\ &= r(t)S_0(t)\theta_0(t)dt + dS(t)\theta(t) \\ &= r(t)(V(t) - S(t)\theta(t))dt + dS(t)\theta(t) \\ &= r(t)V(t)dt - r(t)S(t)\theta(t)dt + dS(t)\theta(t). \end{aligned}$$

Now consider the sequence of problems of terminal wealth maximization:

$$\text{For } t = 1, \dots, T, \quad \mathcal{P}_t : \begin{cases} \max_{\theta_t} E[u(V(T)) | \mathcal{F}_{t-1}], \\ \text{s.t. } V_t = (1 + r_t) V_{t-1} - r_t S_{t-1} \theta_t + \Delta S_t \theta_t \end{cases}$$

Even if markets are incomplete, agents can solve the sequence of problems $\{\mathcal{P}_t\}_{t=1}^T$ as time unfolds. Each problem can be written as:

$$\max_{\theta_t} E \left[u \left(V_1 + \sum_{t=1}^T (r_t V_{t-1} - r_t S_{t-1} \theta_t + \Delta S_t \theta_t) \right) \middle| \mathcal{F}_{t-1} \right].$$

The FOC for $t = 1$ is:

$$E[u'(V(T)) (S_1 - (1 + r_0) S_0) | \mathcal{F}_0],$$

whence

$$S_0 = (1 + r_0)^{-1} \frac{E[u'(V(T)) \cdot S_1 | \mathcal{F}_0]}{E[u'(V(T)) | \mathcal{F}_0]}.$$

In general

$$S_t = (1 + r_t)^{-1} \frac{E[u'(V(T)) \cdot S_{t+1} | \mathcal{F}_t]}{E[u'(V(T)) | \mathcal{F}_t]}, \quad t = 0, \dots, T - 1.$$

The previous relations suggest that we can define a *martingale measure* Q for the discounted price process by defining

$$\frac{dQ}{dP} \bigg|_{\mathcal{F}_t} = \frac{u'(V(T))}{E[u'(V(T)) | \mathcal{F}_t]}.$$

Connections with the CAPM. It's easy to show that:

$$E(\tilde{r}_{t+1}) - r_t = \text{cov} \left[\frac{u'(V(T))}{E[u'(V(T)) | \mathcal{F}_t]}, \tilde{r}_{t+1} \right],$$

where $\tilde{r}_{t+1} \equiv (S_{t+1} - S_t) / S_t$.

4.16 Appendix 7: Topics on jumps

4.16.1 The Radon-Nikodym derivative

This appendix derives, heuristically, results about Radon-Nikodym derivatives for jump-diffusion processes. Precise mathematical details can be found in Brémaud (1981). Consider the jump times $0 < \tau_1 < \tau_2 < \dots < \tau_n = \hat{T}$. The probability of a jump in a neighborhood of τ_i is $v(\tau_i)d\tau$. To define the same probability under the risk-neutral world, write $v^Q(\tau_i)d\tau$ under Q , and set $v^Q = v\lambda^J$, for some λ^J . The probability that no-jump would occur between any two adjacent random points τ_{i-1} and τ_i and a jump would at time τ_{i-1} is, for $i \geq 2$, proportional to:

$$v(\tau_{i-1})e^{-\int_{\tau_{i-1}}^{\tau_i} v(u)du} \quad \text{under } P,$$

and to

$$v^Q(\tau_{i-1})e^{-\int_{\tau_{i-1}}^{\tau_i} v^Q(u)du} = v(\tau_{i-1})\lambda^J(\tau_{i-1})e^{-\int_{\tau_{i-1}}^{\tau_i} v(u)\lambda^J(u)du} \quad \text{under } Q.$$

As explained in Section 4.7, these are in fact densities of time intervals elapsing from one arrival to the next one.

Next, let A be the event of marks at time $\tau_1, \tau_2, \dots, \tau_n$. The Radon-Nikodym derivative is the likelihood ratio of the two probabilities Q and P of A :

$$\frac{Q(A)}{P(A)} = \frac{e^{-\int_t^{\tau_1} v(u)\lambda^J(u)du} \cdot v(\tau_1)\lambda^J(\tau_1)e^{-\int_{\tau_1}^{\tau_2} v(u)\lambda^J(u)du} \cdot v(\tau_2)\lambda^J(\tau_2)e^{-\int_{\tau_2}^{\tau_3} v(u)\lambda^J(u)du} \cdot \dots}{e^{-\int_t^{\tau_1} v(u)du} \cdot v(\tau_1)e^{-\int_{\tau_1}^{\tau_2} v(u)du} \cdot v(\tau_2)e^{-\int_{\tau_2}^{\tau_3} v(u)du} \cdot \dots},$$

where we have used the fact that given that at $\tau_0 = t$, there are no-jumps, the probability that no-jumps would occur from t to τ_1 is $e^{-\int_t^{\tau_1} v(u)du}$ under P , and $e^{-\int_t^{\tau_1} v(u)\lambda^J(u)du}$ under Q . Simple algebra yields,

$$\begin{aligned} \frac{Q(A)}{P(A)} &= \lambda^J(\tau_1) \cdot \lambda^J(\tau_2) \cdot e^{-\int_t^{\tau_1} v(u)(\lambda^J(u)-1)du} \cdot e^{-\int_{\tau_1}^{\tau_2} v(u)(\lambda^J(u)-1)du} \cdot e^{-\int_{\tau_2}^{\tau_3} v(u)(\lambda^J(u)-1)du} \cdot \dots \\ &= \prod_{i=1}^n \lambda^J(\tau_i) \cdot e^{-\int_t^{\tau_n} v(u)(\lambda^J(u)-1)du} \\ &= \exp \left[\ln \left(\prod_{i=1}^n \lambda^J(\tau_i) \cdot e^{-\int_t^{\tau_n} v(u)(\lambda^J(u)-1)du} \right) \right] \\ &= \exp \left[\sum_{i=1}^n \ln \lambda^J(\tau_i) - \int_t^{\tau_n} v(u) (\lambda^J(u) - 1) du \right] \\ &= \exp \left[\int_t^{\hat{T}} \ln \lambda^J(u) dZ(u) - \int_t^{\hat{T}} v(u) (\lambda^J(u) - 1) du \right], \end{aligned}$$

where the last equality follows from the definition of the Stieltjes integral.

Consider, finally, the following definition. Let M be a martingale. The unique solution to the equation:

$$L(\tau) = 1 + \int_t^\tau L(u) dM(u),$$

is named the *Doléans-Dade exponential semimartingale* and is denoted as $\mathcal{E}(M)$. We now turn to the arbitrage restrictions arising whilst dealing with asset prices driven by jump-diffusion processes.

4.16.2 Arbitrage restrictions

As in the main text, let now S be the price of a primitive asset, solution to:

$$\begin{aligned}\frac{dS}{S} &= b d\tau + \sigma dW + \ell S dZ \\ &= b d\tau + \sigma dW + \ell S (dZ - v d\tau) + \ell S v d\tau \\ &= (b + \ell S v) d\tau + \sigma dW + \ell S (dZ - v d\tau).\end{aligned}$$

Next, define

$$d\tilde{Z} = dZ - v^Q d\tau \quad (v^Q = v\lambda^J); \quad d\tilde{W} = dW + \lambda d\tau.$$

Both \tilde{Z} and \tilde{W} are Q -martingales. We have:

$$\frac{dS}{S} = (b + \ell S v^Q - \sigma \lambda) d\tau + \sigma d\tilde{W} + \ell S d\tilde{Z}.$$

The characterization of the equivalent martingale measure for the discounted price is given by the following Radon-Nikodym density of Q with respect to P :

$$\frac{dQ}{dP} = \mathcal{E} \left(- \int_t^T \lambda(\tau) dW(\tau) + \int_t^T (\lambda^J(\tau) - 1) (dZ(\tau) - v(\tau)) d\tau \right),$$

where $\mathcal{E}(\cdot)$ is the Doléans-Dade exponential semimartingale, and so:

$$b = r + \sigma \lambda - \ell v^Q E_S(S) = r + \sigma \lambda - \ell v \lambda^J E_S(S).$$

Clearly, markets are incomplete here. It is possible to show that if S is deterministic, a representative agent with utility function $u(x) = \frac{x^{1-\eta}-1}{1-\eta}$ makes $\lambda^J(S) = (1 + S)^{-\eta}$.

4.16.3 State price density: introduction

We have:

$$L(T) = \exp \left[- \int_t^T v(\tau) (\lambda^J(\tau) - 1) d\tau + \int_t^T \ln \lambda^J(\tau) dZ(\tau) \right].$$

The objective here is to use Itô's lemma for jump processes to express L in differential form. Define the jump process y as:

$$y(\tau) \equiv - \int_t^\tau v(u) (\lambda^J(u) - 1) du + \int_t^\tau \ln \lambda^J(u) dZ(u).$$

In terms of y , L is $L(\tau) = l(y(\tau))$ with $l(y) = e^y$. We have:

$$\begin{aligned}dL(\tau) &= -e^{y(\tau)} v(\tau) (\lambda^J(\tau) - 1) d\tau + \left(e^{y(\tau) + \text{jump}} - e^{y(\tau)} \right) dZ(\tau) \\ &= -e^{y(\tau)} v(\tau) (\lambda^J(\tau) - 1) d\tau + e^{y(\tau)} \left(e^{\ln \lambda^J(\tau)} - 1 \right) dZ(\tau)\end{aligned}$$

or,

$$\frac{dL(\tau)}{L(\tau)} = -v(\tau) (\lambda^J(\tau) - 1) d\tau + (\lambda^J(\tau) - 1) dZ(\tau) = (\lambda^J(\tau) - 1) (dZ(\tau) - v(\tau) d\tau).$$

The general case (with stochastic distribution) is covered in the following subsection.

4.16.4 State price density: general case

Assume that the primitive is:

$$dx(\tau) = \mu(x(\tau_-))d\tau + \sigma(x(\tau_-))dW(\tau) + dZ(\tau),$$

and let u denote the price of a derivative. Introduce the P -martingale,

$$dM(\tau) = dZ(\tau) - v(x(\tau))d\tau.$$

By Itô's lemma for jump-diffusion processes,

$$\begin{aligned} \frac{du(x(\tau), \tau)}{u(x(\tau_-), \tau)} &= \mu^u(x(\tau_-), \tau)d\tau + \sigma^u(x(\tau_-), \tau)dW(\tau) + J^u(\Delta x, \tau) dZ(\tau) \\ &= (\mu^u(x(\tau_-), \tau) + v(x(\tau_-))J^u(\Delta x, \tau)) d\tau + \sigma^u(x(\tau_-), \tau)dW(\tau) + J^u(\Delta x, \tau) dM(\tau), \end{aligned}$$

where $\mu^u = \frac{1}{u} \left(\frac{\partial}{\partial t} + L \right) u$, $\sigma^u = \frac{1}{u} \left(\frac{\partial u}{\partial x} \sigma \right)$, $\frac{\partial}{\partial t} + L$ is the generator for pure diffusion processes and, finally:

$$J^u(\Delta x, \tau) \equiv \frac{u(x(\tau), \tau) - u(x(\tau_-), \tau)}{u(x(\tau_-), \tau)}.$$

To generalize the steps made to deal with the standard diffusion case, let

$$d\tilde{W} = dW + \lambda d\tau, \quad d\tilde{Z} = dZ - v^Q d\tau.$$

We wish to find restrictions on both λ and v^Q , such that both \tilde{W} and \tilde{Z} are Q -martingales. Let J^ξ be the jump component for the state price density ξ :

$$\frac{d\xi(\tau)}{\xi(\tau_-)} = -\lambda(x(\tau_-))dW(\tau) + J^\xi(\Delta x, \tau) dM(\tau), \quad \xi(t) = 1.$$

We shall show that:

$$v^Q = v \left(1 + J^\xi \right).$$

Note that in this case,

$$\frac{d\xi(\tau)}{\xi(\tau_-)} = -\lambda(x(\tau_-))dW(\tau) + (\lambda^J - 1) dM(\tau), \quad \xi(t) = 1,$$

a clear generalization of the pure diffusion case.

As for the derivative price:

$$\begin{aligned} \frac{du}{u} &= (\mu^u + vJ^u) d\tau + \sigma^u dW + J^u (dZ - v d\tau) \\ &= (\mu^u + v^Q J^u - \sigma^u \lambda) d\tau + \sigma^u d\tilde{W} + J^u d\tilde{Z} \\ &= \left(\mu^u + v(1 + J^\xi)J^u - \sigma^u \lambda \right) d\tau + \sigma^u d\tilde{W} + J^u d\tilde{Z}. \end{aligned}$$

Finally, by the Q -martingale property of the discounted u ,

$$\mu^u - r = \sigma^u \lambda - v^Q \cdot E_{\Delta x}(J^u) = \sigma^u \lambda - v \cdot E_{\Delta x} \left((1 + J^\xi)J^u \right),$$

where $E_{\Delta x}$ is taken with respect to the jump-size distribution, which is the same under Q and P .

PROOF THAT $v^Q = v(1 + J^\xi)$. As usual, the state-price density ξ has to be a P -martingale in order to be able to price bonds (in addition to all other assets). In addition, ξ clearly “depends” on W and Z . Therefore, it satisfies:

$$\frac{d\xi(\tau)}{\xi(\tau_-)} = -\lambda(x(\tau_-))dW(\tau) + J_\xi(\Delta x, \tau) dM(\tau), \quad \xi(t) = 1.$$

We wish to find v^Q in $d\tilde{Z} = dZ - v^Q d\tau$ such that \tilde{Z} is a Q -martingale, viz

$$\tilde{Z}(\tau) = \mathbb{E}[\tilde{Z}(T)],$$

i.e.,

$$\mathbb{E}(\tilde{Z}(t)) = \frac{E\left(\xi(T) \cdot \tilde{Z}(T)\right)}{\xi(t)} = \tilde{Z}(t) \Leftrightarrow \xi(t)\tilde{Z}(t) = E[\xi(T)\tilde{Z}(T)],$$

i.e.,

$$\xi(t)\tilde{Z}(t) \text{ is a } P\text{-martingale.}$$

By Itô’s lemma,

$$\begin{aligned} d(\xi\tilde{Z}) &= d\xi \cdot \tilde{Z} + \xi \cdot d\tilde{Z} + d\xi \cdot d\tilde{Z} \\ &= d\xi \cdot \tilde{Z} + \xi (dZ - v^Q d\tau) + d\xi \cdot d\tilde{Z} \\ &= d\xi \cdot \tilde{Z} + \xi \underbrace{[dZ - v d\tau]}_{dM} + (v - v^Q) d\tau + d\xi \cdot d\tilde{Z} \\ &= d\xi \cdot \tilde{Z} + \xi \cdot dM + \xi (v - v^Q) d\tau + d\xi \cdot d\tilde{Z}. \end{aligned}$$

Because ξ , M and $\xi\tilde{Z}$ are P -martingales,

$$\forall T, \quad 0 = E \left[\int_t^T \xi(\tau) \cdot (v(\tau) - v^Q(\tau)) d\tau + \int_t^T d\xi(\tau) \cdot d\tilde{Z}(\tau) \right].$$

But

$$d\xi \cdot d\tilde{Z} = \xi \left(-\lambda dW + J^\xi dM \right) (dZ - v^Q d\tau) = \xi \left[-\lambda dW + J^\xi (dZ - v d\tau) \right] (dZ - v^Q d\tau),$$

and since $(dZ)^2 = dZ$,

$$E(d\xi \cdot d\tilde{Z}) = \xi \cdot J^\xi v \cdot d\tau,$$

and the previous condition collapses to:

$$\forall T, \quad 0 = E \left[\int_t^T \xi(\tau) \cdot \left(v(\tau) - v^Q(\tau) + J^\xi(\Delta x)v(\tau) \right) d\tau \right],$$

which implies

$$v^Q(\tau) = v(\tau) \left(1 + J^\xi(\Delta x) \right), \quad \text{a.s.}$$

||

References

- Arnold, L. (1974): *Stochastic Differential Equations: Theory and Applications*, New York: Wiley.
- Black, F. and M. Scholes (1973): “The Pricing of Options and Corporate Liabilities.” *Journal of Political Economy* 81, 637-659.
- Brémaud, P. (1981): *Point Processes and Queues: Martingale Dynamics*. Berlin: Springer Verlag.
- Cvitanović, J. and I. Karatzas (1992): “Convex Duality in Constrained Portfolio Optimization.” *Annals of Applied Probability* 2, 767-818.
- Föllmer, H. and M. Schweizer (1991): “Hedging of Contingent Claims under Incomplete Information.” In: Davis, M. and R. Elliott (Editors): *Applied Stochastic Analysis*. New York: Gordon & Breach, 389-414.
- Friedman, A. (1975): *Stochastic Differential Equations and Applications* (Vol. I). New York: Academic Press.
- Harrison, J.M. and S. Pliska (1983): “A Stochastic Calculus Model of Continuous Trading: Complete Markets.” *Stochastic Processes and Their Applications* 15, 313-316.
- Harrison, J.M, R. Pitbladdo and S.M. Schaefer (1984): “Continuous Price Processes in Frictionless Markets Have Infinite Variation.” *Journal of Business* 57, 353-365.
- He, H. and N. Pearson (1991): “Consumption and Portfolio Policies with Incomplete Markets and Short-Sales Constraints: The Infinite Dimensional Case.” *Journal of Economic Theory* 54, 259-304.
- Karatzas, I. and S.E. Shreve (1991): *Brownian Motion and Stochastic Calculus*. New York: Springer Verlag.
- Mikosch, T. (1998): *Elementary Stochastic Calculus with Finance in View*. Singapore: World Scientific.
- Revuz, D. and M. Yor (1999): *Continuous Martingales and Brownian Motion*. New York: Springer Verlag.
- Shreve, S. (1991): “A Control Theorist’s View of Asset Pricing.” In: Davis, M. and R. Elliot (Editors): *Applied Stochastic Analysis*. New York: Gordon & Breach, 415-445.
- Steele, J.M. (2001): *Stochastic Calculus and Financial Applications*. New York: Springer-Verlag.

5

Taking models to data

5.1 Introduction

This chapter surveys methods to estimate and test dynamic models of asset prices. It begins with foundational issues on identification, specification and testing. Then, it surveys classical estimation and testing methodologies such as the Method of Moments, where the number of moment conditions equals the dimension of the parameter vector (Pearson, 1894); Maximum Likelihood (ML) (Gauss, 1816; Fisher, 1912); the Generalized Method of Moments (GMM), where the number of moment conditions exceeds the dimension of the parameter vector, leading to the minimum chi-squared (Neyman and Pearson, 1928; Hansen, 1982); and, finally, the recent developments relying on simulations, which aim to implement ML and GMM estimation for models that are analytically quite complex, but that can be simulated. The chapter concludes with an illustration of how joint estimation of fundamentals and asset prices in arbitrage-free models can lead to statistical efficiency, asymptotically.

5.2 Data generating processes

5.2.1 Basics

Given is a multidimensional stochastic process y_t , a *data generating process* (DGP). While we do not know the probability distribution underlying y_t , we use the available data to get insights into its nature. A few definitions. A DGP is a conditional law, say the law of y_t given the set of past values $\underline{y}_{t-1} = \{y_{t-1}, y_{t-2}, \dots\}$, and some exogenous [define] variable z , with $\underline{z}_t = \{z_t, z_{t-1}, z_{t-2}, \dots\}$,

$$\text{DGP} : \ell_0(y_t | x_t),$$

where $x_t = (\underline{y}_{t-1}, \underline{z}_t)$, and ℓ_0 denotes the conditional density of the data, the true law. Then, we have three basic definitions. First, we define a *parametric model* as a set of conditional laws for y_t , indexed by a parameter vector $\theta \in \Theta \subseteq \mathbb{R}^p$,

$$(M) = \{\ell(y_t | x_t; \theta), \theta \in \Theta \subseteq \mathbb{R}^p\}.$$

Second, we say that the model (M) is *well-specified* if,

$$\exists \theta_0 \in \Theta : \ell(y_t | x_t; \theta_0) = \ell_0(y_t | x_t).$$

Third, we say that the model (M) is *identifiable* if θ_0 is unique. The main goal of this chapter is to review tools aimed at drawing inference about the true parameter θ_0 , given the observations.

5.2.2 Restrictions on the DGP

The previous definition of DGP is too rich to be of practical relevance. This chapter deals with estimation methods applying to DGPs satisfying a few restrictions. Two fundamental restrictions are usually imposed on the DGP:

- Restrictions on the *heterogeneity* of the stochastic process, which lead to *stationary* random processes.
- Restrictions on the *memory* of the stochastic process, which pave the way to *ergodic* processes.

5.2.2.1 Stationarity

Stationary processes describe phenomena leading to long run equilibria, in some statistical sense: as time unfolds, the probability generating the observations settles down to some “long-run” probability density, a time invariant probability. As Chapter 3 explains, in the early 1980s, theorists began to define a long-run equilibrium as a well-defined stationary, probability distribution generating economic outcomes. We have two notions of stationarity: (i) Strong, or strict, stationarity. Definition: Homogeneity in law; (ii) Weak stationarity, or stationarity of order p . Definition: Homogeneity in moments.

Even with stationary DGP, there might be situations where the number of parameters to be estimated increases with the sample size. As an example, consider two stochastic processes: one, for which $cov(y_t, y_{t+\tau}) = \tau^2$; and another, for which $cov(y_t, y_{t+\tau}) = \exp(-|\tau|)$. In both cases, the DGP is stationary. Yet for the first process, the dependence increases with τ , and for the second, the dependence decreases with τ . As this simple example reveals, a stationary stochastic process may have “long memory.” “Ergodicity” further restricts DGP, so as to make this memory play a more limited role.

5.2.2.2 Ergodicity

We shall deal with DGPs where the dependence between y_{t_1} and y_{t_2} decreases with $|t_2 - t_1|$. To introduce some concepts and notation, say two events A and B are independent, when $P(A \cap B) = P(A)P(B)$. A stochastic process is *asymptotically independent* if, for some function β_τ ,

$$\beta_\tau \geq |F(y_{t_1}, \dots, y_{t_n}, y_{t_1+\tau}, \dots, y_{t_n+\tau}) - F(y_{t_1}, \dots, y_{t_n}) F(y_{t_1+\tau}, \dots, y_{t_n+\tau})|,$$

we also have that $\lim_{\tau \rightarrow \infty} \beta_\tau \rightarrow 0$. A stochastic process is *p-dependent* if $\forall \tau \leq p$, $\beta_\tau \neq 0$. A stochastic process is *asymptotically uncorrelated* if there exists ρ_τ such that for all t , $\rho_\tau \geq cov(y_t, y_{t+\tau}) / \sqrt{var(y_t) \cdot var(y_{t+\tau})}$, and that $0 \leq \rho_\tau \leq 1$ with $\sum_{\tau=0}^{\infty} \rho_\tau < \infty$. For example, $\rho_\tau = \tau^{-(1+\delta)}$, $\delta > 0$, in which case $\rho_\tau \downarrow 0$ as $\tau \uparrow \infty$.

Let \mathbb{B}_1^t denote the σ -algebra generated by $\{y_1, \dots, y_t\}$ and $A \in \mathbb{B}_{-\infty}^t$, $B \in \mathbb{B}_{t+\tau}^\infty$, and define:

$$\alpha_\tau = \sup_{\tau} |P(A \cap B) - P(A)P(B)|, \quad \varphi(\tau) = \sup_{\tau} |P(B | A) - P(B)|, \quad P(A) > 0.$$

We say that (i) y is *strongly mixing*, or α -*mixing* if $\lim_{\tau \rightarrow \infty} \alpha_\tau \rightarrow 0$; (ii) y is *uniformly mixing* if $\lim_{\tau \rightarrow \infty} \varphi_\tau \rightarrow 0$. Clearly, a uniformly mixing process is also strongly mixing. A second order stationary process is *ergodic* if $\lim_{T \rightarrow \infty} \sum_{\tau=1}^T \text{cov}(y_t, y_{t+\tau}) < \infty$. If a second order stationary process is strongly mixing, it is also ergodic.

5.2.3 Parameter estimators

Consider an estimator of the parameter vector θ of the model,

$$(M) = \{\ell(y_t | x_t; \theta), \theta \in \Theta \subset \mathbb{R}^p\}.$$

Naturally, any estimator does necessarily depend on the sample size, which we write as $\hat{\theta}_T \equiv t_T(y)$. Of a given estimator $\hat{\theta}_T$, we say that it is:

- *Correct*, or *unbiased*, if $E(\hat{\theta}_T) = \theta_0$. The difference $E(\hat{\theta}_T) - \theta_0$ is called distortion, or bias.
- *Weakly consistent* if $\text{plim} \hat{\theta}_T = \theta_0$. And *strongly consistent* if $\hat{\theta}_T \xrightarrow{\text{a.s.}} \theta_0$.

Finally, an estimator $\hat{\theta}_T^{(1)}$ is more *efficient* than another estimator $\hat{\theta}_T^{(2)}$ if, for any vector of constants c , we have that $c^\top \cdot \text{var}(\hat{\theta}_T^{(1)}) \cdot c < c^\top \cdot \text{var}(\hat{\theta}_T^{(2)}) \cdot c$.

5.2.4 Basic properties of density functions

We have T observations $y_T^1 = \{y_1, \dots, y_T\}$. Suppose these observations are the realization of a T -dimensional random variable with joint density, $f(\tilde{y}_1, \dots, \tilde{y}_T; \theta) = f(\tilde{y}_1^T; \theta)$. We have momentarily put tildes on y_i , to emphasize that we view each \tilde{y}_i as a random variable.¹ However, to ease notation, from now on, we write y_i instead of \tilde{y}_i . By construction, $\int f(y | \theta) dy \equiv \int \dots \int f(y_1^T | \theta) dy_1^T = 1$ or,

$$\forall \theta \in \Theta, \int f(y; \theta) dy = 1.$$

Now suppose that the support of y doesn't depend on θ . Under regularity conditions,

$$\nabla_\theta \int f(y; \theta) dy = \int \nabla_\theta f(y; \theta) dy = \mathbf{0}_p,$$

where $\mathbf{0}_p$ is a column vector of zeros in \mathbb{R}^p . Moreover, for all $\theta \in \Theta$,

$$\mathbf{0}_p = \int \nabla_\theta f(y; \theta) dy = E_\theta [\nabla_\theta \ln f(y; \theta)]. \quad (5.1)$$

Finally, we have,

$$\begin{aligned} \mathbf{0}_{p \times p} &= \nabla_\theta \int [\nabla_\theta \ln f(y; \theta)] f(y; \theta) dy \\ &= \int [\nabla_{\theta\theta} \ln f(y; \theta)] f(y; \theta) dy + \int |\nabla_\theta \ln f(y; \theta)|_2 f(y; \theta) dy, \end{aligned}$$

where $|x|_2$ denotes the outer product, i.e. $|x|_2 = x \cdot x^\top$. Hence, by Eq. (5.1),

$$E_\theta [\nabla_{\theta\theta} \ln f(y; \theta)] = -E_\theta |\nabla_\theta \ln f(y; \theta)|_2 = -\text{var}_\theta [\nabla_\theta \ln f(y; \theta)] \equiv -\mathcal{J}(\theta), \quad \forall \theta \in \Theta.$$

The matrix \mathcal{J} is known as the Fisher's information matrix.

¹Therefore, we follow a classical perspective. A Bayesian statistician would view the sample as *given*. We do not review Bayesian methods in this chapter.

5.2.5 The Cramer-Rao lower bound

Let $t(y)$ some unbiased estimator of θ , and set the dimension of the parameter space to $p = 1$. We have,

$$E[t(y)] = \int t(y) f(y; \theta) dy.$$

Under regularity conditions,

$$\nabla_{\theta} E[t(y)] = \int t(y) [\nabla_{\theta} \ln f(y; \theta)] f(y; \theta) dy = \text{cov}(t(y), \nabla_{\theta} \ln f(y; \theta)).$$

By Cauchy-Schwartz inequality, $[\text{cov}(t(y), \nabla_{\theta} \ln f(y; \theta))]^2 \leq \text{var}[t(y)] \cdot \text{var}[\nabla_{\theta} \ln f(y; \theta)]$. Therefore,

$$[\nabla_{\theta} E(t(y))]^2 \leq \text{var}[t(y)] \cdot \text{var}[\nabla_{\theta} \ln f(y; \theta)] = -\text{var}[t(y)] \cdot E[\nabla_{\theta\theta} \ln f(y; \theta)].$$

But if $t(y)$ is unbiased, or $E[t(y)] = \theta$,

$$\text{var}[t(y)] \geq [-E(\nabla_{\theta} \ln f(y; \theta))]^{-1} \equiv \mathcal{J}(\theta)^{-1}.$$

This is the celebrated Cramer-Rao bound. The same results holds in the multidimensional case, through a mere change in notation (see, e.g., Amemiya, 1985, p. 14-17).

5.3 Maximum likelihood estimation

5.3.1 Basics

The density of the data, $f(y_1^T | \theta)$, maps every possible sample and parameter values of θ on to positive numbers, the “likelihood” of occurrence of any given sample, given the parameter $\theta: \mathbb{R}^{nT} \times \Theta \mapsto \mathbb{R}_+$. We trace the joint density of the entire sample through a thought experiment, in which we change the sample y_1^T . So the sample is viewed as the realization of a random variable, a view opposite to the Bayesian perspective. We ask: Which value of θ makes the sample we observed the most likely to have occurred? We introduce the “likelihood function,” $L(\theta | y_1^T) \equiv f(y_1^T; \theta)$. It is the function $\theta \mapsto f(y; \theta)$ for y_1^T given and equal to \bar{y} , say:

$$L(\theta | \bar{y}) \equiv f(\bar{y}; \theta).$$

Then, we maximize $L(\theta | y_1^T)$ with respect to θ . That is, we look for the value of θ , which maximizes the probability to observe the sample we have effectively observed. The resulting estimator is called *maximum likelihood estimator* (MLE). As we shall see, the MLE attains the Cramer-Rao lower bound, provided the model is not misspecified.

5.3.2 Factorizations

Consider a series of events $\{A_i\}$. In the Appendix, we show that,

$$\Pr\left(\bigcap_{i=1}^n A_i\right) = \prod_{i=1}^n \Pr\left(A_i \left| \bigcap_{j=1}^{i-1} A_j\right.\right). \quad (5.2)$$

By Eq. (5.2), then, the MLE satisfies:

$$\hat{\theta}_T = \arg \max_{\theta \in \Theta} L_T(\theta) = \arg \max_{\theta \in \Theta} \left(\frac{1}{T} \ln L_T(\theta) \right),$$

where, assuming IID data,

$$\ln L_T(\theta) \equiv \ln \prod_{t=1}^T f(y_t | y_1^{t-1}; \theta) = \sum_{t=1}^T \ln f(y_t | y_1^{t-1}; \theta) \equiv \sum_{t=1}^T \ln f(y_t; \theta) \equiv \sum_{t=1}^T \ell_t(\theta), \quad (5.3)$$

and $\ell_t(\theta)$ is the “log-likelihood” of a single observation.

5.3.3 Asymptotic properties

We consider the i.i.d. case only, as in Eq. (5.3). Moreover, we provide heuristic arguments, leaving more rigorous proofs and general results in the Appendix.

5.3.3.1 The limiting problem

The MLE satisfies the following first order conditions,

$$\mathbf{0}_p = \nabla_{\theta} \ln L_T(\theta) |_{\theta = \hat{\theta}_T} \equiv \nabla_{\theta} \ln L_T(\hat{\theta}_T).$$

Consider a Taylor expansion of the first order conditions around θ_0 ,

$$\mathbf{0}_p = \nabla_{\theta} \ln L_T(\hat{\theta}_T) \stackrel{d}{=} \nabla_{\theta} \ln L_T(\theta_0) + \nabla_{\theta\theta} \ln L_T(\theta_0)(\hat{\theta}_T - \theta_0), \quad (5.4)$$

where the notation $x_T \stackrel{d}{=} y_T$ means that the difference $x_T - y_T = o_p(1)$, and θ_0 is defined as the solution to the limiting problem,

$$\theta_0 = \arg \max_{\theta \in \Theta} \left[\lim_{T \rightarrow \infty} \left(\frac{1}{T} \ln L_T(\theta) \right) \right] = \arg \max_{\theta \in \Theta} [E(\ell(\theta))],$$

and, finally, ℓ satisfies regularity conditions needed to ensure that,

$$\theta_0 : E[\nabla_{\theta} \ell(\theta_0)] = \mathbf{0}_p.$$

To show that this is indeed the solution, suppose θ_0 is identified; that is, $\theta \neq \theta_0$ and $\theta, \theta_0 \in \Theta \Leftrightarrow f(y|\theta) \neq f(y|\theta_0)$. Suppose, further, that for each $\theta \in \Theta$, $E_{\theta}[\ln f(y|\theta)] < \infty$. Then, we have that $\theta_0 = \arg \max_{\theta \in \Theta} E_{\theta}[\ln f(y|\theta)]$, and this value of θ is unique. The proof is, indeed, very simple. We have,

$$\begin{aligned} E_{\theta_0} \left[-\ln \left(\frac{f(y|\theta)}{f(y|\theta_0)} \right) \right] &> -\ln E_{\theta_0} \left(\frac{f(y|\theta)}{f(y|\theta_0)} \right) \\ &= -\ln \int \frac{f(y|\theta)}{f(y|\theta_0)} f(y|\theta_0) dy \\ &= -\ln \int f(y|\theta) dy = 0. \end{aligned}$$

5.3.3.2 Consistency and asymptotic normality

Provided the model is well-specified, we have that $\hat{\theta}_T \xrightarrow{p} \theta_0$ and even $\hat{\theta}_T \xrightarrow{a.s.} \theta_0$, under regularity conditions. One example of conditions required to obtain weak consistency is that the following uniform weak law of large numbers holds,

$$\lim_{T \rightarrow \infty} \Pr \left[\sup_{\theta \in \Theta} |\ell_T(\theta) - E(\ell(\theta))| \right] \rightarrow 0.$$

Next, consider again the asymptotic expansion in Eq. (5.4), which can be elaborated, so as to have,

$$\begin{aligned} \sqrt{T}(\hat{\theta}_T - \theta_0) &\stackrel{d}{=} - \left[\frac{1}{T} \nabla_{\theta\theta} \ln L_T(\theta_0) \right]^{-1} \frac{1}{\sqrt{T}} \nabla_{\theta} \ln L_T(\theta_0) \\ &= - \left[\frac{1}{T} \sum_{t=1}^T \nabla_{\theta\theta} \ell_t(\theta_0) \right]^{-1} \frac{1}{\sqrt{T}} \sum_{t=1}^T \nabla_{\theta} \ell_t(\theta_0). \end{aligned}$$

By the law of large numbers reviewed in the Appendix (weak law no. 1),

$$\frac{1}{T} \sum_{t=1}^T \nabla_{\theta\theta} \ell_t(\theta_0) \xrightarrow{p} E_{\theta_0} [\nabla_{\theta\theta} \ell_t(\theta_0)] = -\mathcal{J}(\theta_0).$$

Therefore, asymptotically,

$$\sqrt{T}(\hat{\theta}_T - \theta_0) \stackrel{d}{=} \mathcal{J}(\theta_0)^{-1} \frac{1}{\sqrt{T}} \sum_{t=1}^T \nabla_{\theta} \ell_t(\theta_0).$$

We also have,

$$\frac{1}{\sqrt{T}} \sum_{t=1}^T \nabla_{\theta} \ell_t(\theta_0) \xrightarrow{d} N(0, \mathcal{J}(\theta_0)).$$

Indeed, let $\overline{\nabla_{\theta} \ell(\theta_0)}_T = \frac{1}{T} \sum_{t=1}^T \nabla_{\theta} \ell_t(\theta_0)$, and note that $E(\nabla_{\theta} \ell_t(\theta_0)) = 0$. Then, by the central limit theorem reviewed in the Appendix:

$$\frac{1}{\sqrt{T}} \frac{\sum_{t=1}^T \nabla_{\theta} \ell_t(\theta_0)}{\sqrt{\text{var}[\nabla_{\theta} \ell_t(\theta_0)]}} = \frac{\sqrt{T} \left(\overline{\nabla_{\theta} \ell(\theta_0)}_T - E(\nabla_{\theta} \ell_t(\theta_0)) \right)}{\sqrt{\text{var}[\nabla_{\theta} \ell_t(\theta_0)]}},$$

where, for each t , $\text{var}[\nabla_{\theta} \ell_t(\theta_0)] = \mathcal{J}(\theta_0)$.

Finally, by the Slutsky's theorem reviewed in the Appendix,

$$\sqrt{T}(\hat{\theta}_T - \theta_0) \xrightarrow{d} N(0, \mathcal{J}(\theta_0)^{-1}).$$

Therefore, the ML estimator attains the Cramer-Rao lower bound.

5.4 M-estimators

Consider a function g of the unknown parameters θ . Given a function Ψ , a *M-estimator* of the function $g(\theta)$ is the solution to,

$$\max_{g \in G} \sum_{t=1}^T \Psi(x_t, y_t; g),$$

where y and x are as in Section 5.2.1. We assume that a solution to this problem exists, that it is interior and that it is unique. Let us denote the M-estimator with $\hat{g}_T(x_1^T, y_1^T)$. Naturally, the M-estimator satisfies the following first order conditions,

$$0 = \frac{1}{T} \sum_{t=1}^T \nabla_g \Psi(y_t, x_t; \hat{g}_T(x_1^T, y_1^T)).$$

To simplify the presentation, we assume that (x, y) are independent in time, and that they have the same law. By the law of large numbers,

$$\frac{1}{T} \sum_{t=1}^T \Psi(y_t, x_t; g) \xrightarrow{p} \iint \Psi(y, x; g) dF(x, y) = \iint \Psi(y, x; g) dF(y|x) dZ(x) \equiv E_x E_0[\Psi(y, x; g)],$$

where E_0 is the expectation operator taken with respect to the true conditional law of y given x and E_x is the expectation operator taken with respect to the true marginal law of x . The limit problem is,

$$g_\infty = g_\infty(\theta_0) = \arg \max_{g \in G} E_x E_0[\Psi(y, x; g)].$$

Under standard regularity conditions,² there exists a sequence of M-estimators $\hat{g}_T(x, y)$ converging a.s. to $g_\infty = g_\infty(\theta_0)$. Under additional regularity conditions, the M-estimator is also asymptotic normal:

THEOREM 5.1: *Let $\mathcal{I} \equiv E_x E_0 \left(\nabla_g \Psi(y, x; g_\infty(\theta_0)) [\nabla_g \Psi(y, x; g_\infty(\theta_0))]^\top \right)$ and assume that the matrix $\mathcal{J} \equiv E_x E_0 [-\nabla_{gg} \Psi(y, x; g)]$ exists and has an inverse. We have,*

$$\sqrt{T}(\hat{g}_T - g_\infty(\theta_0)) \xrightarrow{d} N(0, \mathcal{J}^{-1} \mathcal{I} \mathcal{J}^{-1}).$$

SKETCH OF THE PROOF. The M-estimator satisfies the following first order conditions,

$$\begin{aligned} 0 &= \frac{1}{\sqrt{T}} \sum_{t=1}^T \nabla_g \Psi(y_t, x_t; \hat{g}_T) \\ &\stackrel{d}{=} \frac{1}{\sqrt{T}} \sum_{t=1}^T \nabla_g \Psi(y_t, x_t; g_\infty) + \sqrt{T} \left[\frac{1}{T} \sum_{t=1}^T \nabla_{gg} \Psi(y_t, x_t; g_\infty) \right] \cdot (\hat{g}_T - g_\infty). \end{aligned}$$

² G is compact; Ψ is continuous with respect to g and integrable with respect to the true law, for each g ; $\frac{1}{T} \sum_{t=1}^T \Psi(y_t, x_t; g) \xrightarrow{a.s.} E_x E_0[\Psi(y, x; g)]$ uniformly on G ; the limit problem has a unique solution $g_\infty = g_\infty(\theta_0)$.

By rearranging terms,

$$\begin{aligned}\sqrt{T}(\hat{g}_T - g_\infty) &\stackrel{d}{=} \left[-\frac{1}{T} \sum_{t=1}^T \nabla_{gg} \Psi(y_t, x_t; g_\infty) \right]^{-1} \cdot \frac{1}{\sqrt{T}} \sum_{t=1}^T \nabla_g \Psi(y_t, x_t; g_\infty) \\ &\stackrel{d}{=} [E_x E_0 (-\nabla_{gg} \Psi(y, x; g))]^{-1} \cdot \frac{1}{\sqrt{T}} \sum_{t=1}^T \nabla_g \Psi(y_t, x_t; g_\infty) \\ &\stackrel{d}{=} \mathcal{J}^{-1} \cdot \frac{1}{\sqrt{T}} \sum_{t=1}^T \nabla_g \Psi(y_t, x_t; g_\infty).\end{aligned}$$

By the limiting problem, $E_x E_0 [\nabla_g \Psi(y, x; g_\infty)] = 0$. Then, $\text{var}(\nabla_g \Psi) = E(\nabla_g \Psi \cdot [\nabla_g \Psi]^\top) = \mathcal{I}$, and, then,

$$\frac{1}{\sqrt{T}} \sum_{t=1}^T \nabla_g \Psi(y_t, x_t; g_\infty) \xrightarrow{d} N(0, \mathcal{I}).$$

The result follows by the Slutsky's theorem and the symmetry of \mathcal{J} . \parallel

One simple example of M-estimator is the Nonlinear Least Squares estimator,

$$\hat{\theta}_T = \arg \min_{\theta \in \Theta} \sum_{t=1}^T [y_t - m(x_t; \theta)]^2,$$

for some function m . In this case, $\Psi(x, y; \theta) = [y - m(x; \theta)]^2$.

5.5 Pseudo, or quasi, maximum likelihood

The maximum likelihood estimator is an M-estimator: set $\Psi = \ln L$, the log-likelihood function. Indeed, assume the model is well-specified, in which case $\mathcal{J} = \mathcal{I}$, which confirms we are back to the MLE.

Next, suppose that we implement the MLE to estimate a model, when in fact the model is *misspecified* in that the true DGP $\ell_0(y_t | x_t)$ does not belong to the family of laws spanned by our model,

$$\ell_0(y_t | x_t) \notin (M) = \{f(y_t | x_t; \theta), \theta \in \Theta\}.$$

Suppose we insist in maximizing $\Psi = \ln L$, where $L = \sum_t f(y_t | x_t; \theta)$. In this case,

$$\sqrt{T}(\hat{\theta}_T - \theta_0^*) \xrightarrow{d} N(0, \mathcal{J}^{-1} \mathcal{I} \mathcal{J}^{-1}),$$

where θ_0^* is the “pseudo-true” value,³ and

$$\mathcal{J} = -E_x E_0 \left[\nabla_{\theta\theta} \ln f(y_t | \underline{y}_{t-1}; \theta_0^*) \right], \quad \mathcal{I} = E_x E_0 \left(\nabla_\theta \ln f(y_t | \underline{y}_{t-1}; \theta_0^*) \cdot \left[\nabla_\theta \ln f(y_t | \underline{y}_{t-1}; \theta_0^*) \right]^\top \right).$$

In the presence of specification errors, $\mathcal{J} \neq \mathcal{I}$. By comparing the two estimated matrices leads to detect specification errors. Finally, note that in this general case, the variance-covariance

³That is, θ_0^* is, clearly, the solution to some misspecified limiting problem. This θ_0^* has an appealing interpretation in terms of some entropy distance minimizer.

matrix $\mathcal{J}^{-1}\mathcal{I}\mathcal{J}^{-1}$ depends on the unknown law of (y_t, x_t) . To assess the precision of the estimates of \hat{g}_T , one needs to estimate such a variance-covariance matrix. A common practice is to use the following a.s. consistent estimators,

$$\hat{J} = -\frac{1}{T} \sum_{t=1}^T \nabla_{gg} \Psi(y_t, x_t; \hat{g}_T), \quad \text{and} \quad \hat{\mathcal{I}} = -\frac{1}{T} \sum_{t=1}^T (\nabla_g \Psi(y_t, x_t; \hat{g}_T) [\nabla_g \Psi(y_t, x_t; \hat{g}_T)^\top]).$$

5.6 GMM

Economic theory often places restrictions on models that have the following format,

$$E[h(y_t; \theta_0)] = \mathbf{0}_q, \quad (5.5)$$

where $h : \mathbb{R}^n \times \Theta \mapsto \mathbb{R}^q$, θ_0 is the true parameter vector, y_t is the n -dimensional vector of the observable variables and $\Theta \subseteq \mathbb{R}^p$. Typically, then, the MLE cannot be used to estimate θ_0 . Moreover, MLE requires specifying a density function. Hansen (1982) proposed the following Generalized Method of Moments (GMM) estimation procedure. Consider the sample counterpart to the population in Eq. (5.5),

$$\bar{h}(y_1^T; \theta) = \frac{1}{T} \sum_{t=1}^T h(y_t; \theta), \quad (5.6)$$

where we have rewritten h as a function of the parameter vector $\theta \in \Theta$. The basic idea of GMM is to find a θ which makes $\bar{h}(y_1^T; \theta)$ as close as possible to zero. Precisely, we have,

DEFINITION (GMM estimator): The GMM estimator *is the sequence* $\hat{\theta}_T$ *satisfying,*

$$\hat{\theta}_T = \arg \min_{\theta \in \Theta \subseteq \mathbb{R}^p} \bar{h}(y_1^T; \theta) \begin{matrix} \top \\ 1 \times q \end{matrix} \cdot \begin{matrix} W_T \\ q \times q \end{matrix} \cdot \bar{h}(y_1^T; \theta) \begin{matrix} \\ q \times 1 \end{matrix},$$

where $\{W_T\}$ is a sequence of weighting matrices, with elements that may depend on the observations.

When $p = q$, we say the GMM is *just-identified*, and is, simply, the MM, satisfying:

$$\hat{\theta}_T : \bar{h}(y_1^T; \hat{\theta}_T) = \mathbf{0}_q.$$

When $p < q$, we say the GMM estimator imposes *overidentifying* restrictions.

We analyze the i.i.d. case only. Under regularity conditions, there exists a matrix W_T that minimizes the asymptotic variance of the GMM estimator, which satisfies asymptotically,

$$W = \left[\lim_{T \rightarrow \infty} T \cdot E \left(\bar{h}(y_1^T; \hat{\theta}_T) \cdot \bar{h}(y_1^T; \hat{\theta}_T)^\top \right) \right]^{-1} \equiv \Sigma_0^{-1}. \quad (5.7)$$

An estimator of Σ_0 can be:

$$\Sigma_T = \frac{1}{T} \sum_{t=1}^T \left[h(y_t; \hat{\theta}_T) \cdot h(y_t; \hat{\theta}_T)^\top \right].$$

Note that $\hat{\theta}_T$ depends on the weighting matrix Σ_T , and the weighting matrix Σ_T depends on $\hat{\theta}_T$. Therefore, we need to implement an iterative procedure. The more one iterates, the less likely the final outcome depends on the initial weighting matrix $\Sigma_T^{(0)}$. For example, one can start with $\Sigma_T^{(0)} = \mathbf{I}_q$.

We have:

THEOREM 5.2: *Suppose to be given a sequence of GMM estimators $\hat{\theta}_T$ with weighing matrix as in Eq. (5.7), and such that: $\hat{\theta}_T \xrightarrow{p} \theta_0$. We have,*

$$\sqrt{T}(\hat{\theta}_T - \theta_0) \xrightarrow{d} N\left(\mathbf{0}_p, \left[E(h_\theta) \Sigma_0^{-1} E(h_\theta)^\top\right]^{-1}\right), \quad \text{where } h_\theta \equiv \nabla_\theta h(y; \theta_0).$$

SKETCH OF THE PROOF: The assumption that $\hat{\theta}_T \xrightarrow{p} \theta_0$ is easy to check under mild regularity conditions. Moreover, the GMM satisfies,

$$\mathbf{0}_p = \nabla_\theta \bar{h}(y_1^T; \hat{\theta}_T) \Sigma_T^{-1} \bar{h}(y_1^T; \hat{\theta}_T). \quad (5.8)$$

Eq. (5.8) confirms that if $p = q$, the GMM satisfies $\hat{\theta}_T : \bar{h}(y_1^T; \hat{\theta}_T) = 0$. Indeed, $\nabla_\theta h \Sigma_T^{-1}$ is full-rank with $p = q$, and Eq. (5.8) can only be satisfied with $\bar{h} = 0$. In the general case, $q > p$, we have,

$$\sqrt{T} \bar{h}(y_1^T; \hat{\theta}_T) = \sqrt{T} \bar{h}(y_1^T; \theta_0) + \left[\nabla_\theta \bar{h}(y_1^T; \theta_0)\right]^\top \sqrt{T}(\hat{\theta}_T - \theta_0) + o_p(1).$$

By premultiplying both sides of the previous equality by $\nabla_\theta \bar{h}(y_1^T; \hat{\theta}_T) \Sigma_T^{-1}$,

$$\begin{aligned} & \sqrt{T} \nabla_\theta \bar{h}(y_1^T; \hat{\theta}_T) \Sigma_T^{-1} \cdot \bar{h}(y_1^T; \hat{\theta}_T) \\ &= \sqrt{T} \nabla_\theta \bar{h}(y_1^T; \hat{\theta}_T) \Sigma_T^{-1} \cdot \bar{h}(y_1^T; \theta_0) + \nabla_\theta \bar{h}(y_1^T; \hat{\theta}_T) \Sigma_T^{-1} \cdot \left[\nabla_\theta \bar{h}(y_1^T; \theta_0)\right]^\top \sqrt{T}(\hat{\theta}_T - \theta_0) + o_p(1). \end{aligned}$$

The l.h.s. of this equality is zero by the first order conditions in Eq. (5.8). By rearranging terms,

$$\begin{aligned} & \sqrt{T}(\hat{\theta}_T - \theta_0) \stackrel{d}{=} - \left(\nabla_\theta \bar{h}(y_1^T; \theta_0) \Sigma_T^{-1} \left[\nabla_\theta \bar{h}(y_1^T; \theta_0)\right]^\top\right)^{-1} \nabla_\theta \bar{h}(y_1^T; \hat{\theta}_T) \Sigma_T^{-1} \cdot \sqrt{T} \bar{h}(y_1^T; \theta_0) \\ &= - \left(\frac{1}{T} \sum_{t=1}^T \nabla_\theta h(y_t; \hat{\theta}_T) \Sigma_T^{-1} \frac{1}{T} \sum_{t=1}^T \left[\nabla_\theta h(y_t; \hat{\theta}_T)\right]^\top\right)^{-1} \frac{1}{T} \sum_{t=1}^T \nabla_\theta h(y_t; \hat{\theta}_T) \Sigma_T^{-1} \sqrt{T} \bar{h}(y_1^T; \theta_0) \\ &\stackrel{d}{=} - \left(E(h_\theta) \Sigma_0^{-1} E(h_\theta)^\top\right)^{-1} E(h_\theta) \Sigma_0^{-1} \cdot \frac{1}{\sqrt{T}} \sum_{t=1}^T h(y_t; \theta_0). \end{aligned}$$

We have: $\frac{1}{\sqrt{T}} \sum_{t=1}^T h(y_t; \theta_0) \xrightarrow{d} N(E(h), \text{var}(h))$, where, by Eq. (5.5), $E(h) = 0$, and $\text{var}(h) = E(h \cdot h^\top) = \Sigma_0$. Hence:

$$\frac{1}{\sqrt{T}} \sum_{t=1}^T h(y_t; \theta_0) \xrightarrow{d} N(0, \Sigma_0).$$

Therefore, $\sqrt{T}(\hat{\theta}_T - \theta_0)$ is asymptotically normal with expectation $\mathbf{0}_p$, and variance,

$$\left(E(h_\theta) \Sigma_0^{-1} E(h_\theta)^\top\right)^{-1} E(h_\theta) \Sigma_0^{-1} \Sigma_0 \Sigma_0^{-1} E(h_\theta)^\top \left(E(h_\theta) \Sigma_0^{-1} E(h_\theta)^\top\right)^{\top-1} = \left(E(h_\theta) \Sigma_0^{-1} E(h_\theta)^\top\right)^{-1}.$$

||

A widely used global specification test is that of the celebrated “overidentifying restrictions.” Consider the following intuitive result:

$$\sqrt{T}\bar{h}(y_1^T; \theta_0)^\top \Sigma_0^{-1} \sqrt{T}\bar{h}(y_1^T; \theta_0) \xrightarrow{d} \chi^2(q).$$

Would we be expecting the same, if we were to replace the true parameter θ_0 with the GMM estimator $\hat{\theta}_T$, which is, anyway, a consistent estimator for θ_0 ? The answer is no. Define:

$$\mathcal{C}_T = \sqrt{T}\bar{h}(y_1^T; \hat{\theta}_T)^\top \Sigma_T^{-1} \cdot \sqrt{T}\bar{h}(y_1^T; \hat{\theta}_T).$$

We have,

$$\begin{aligned} \sqrt{T}\bar{h}(y_1^T; \hat{\theta}_T) &\stackrel{d}{=} \sqrt{T}\bar{h}(y_1^T; \theta_0) + \nabla_{\theta}\bar{h}(y_1^T; \theta_0) \sqrt{T}(\hat{\theta}_T - \theta_0) \\ &\stackrel{d}{=} \sqrt{T}\bar{h}(y_1^T; \theta_0) - [\nabla_{\theta}\bar{h}(y_1^T; \theta_0)]^\top \left[E(h_\theta) \Sigma_0^{-1} E(h_\theta)^\top \right]^{-1} E(h_\theta) \Sigma_0^{-1} \cdot \sqrt{T}\bar{h}(y_1^T; \theta_0) \\ &\stackrel{d}{=} \sqrt{T}\bar{h}(y_1^T; \theta_0) - E(h_\theta)^\top \left[E(h_\theta) \Sigma_0^{-1} E(h_\theta)^\top \right]^{-1} E(h_\theta) \Sigma_0^{-1} \cdot \sqrt{T}\bar{h}(y_1^T; \theta_0) \\ &= \underbrace{(\mathbf{I}_q - \mathbf{P}_q)}_{q \times q} \underbrace{\sqrt{T}\bar{h}(y_1^T; \theta_0)}_{q \times 1}, \end{aligned}$$

and

$$\mathbf{P}_q \equiv E(h_\theta)^\top \left[E(h_\theta) \Sigma_0^{-1} E(h_\theta)^\top \right]^{-1} E(h_\theta) \Sigma_0^{-1}$$

is the orthogonal projector in the space generated by the columns of $E(h_\theta)$ by the inner product Σ_0^{-1} . Thus, we have shown that,

$$\mathcal{C}_T \stackrel{d}{=} \sqrt{T}\bar{h}(y_1^T; \theta_0)^\top (\mathbf{I}_q - \mathbf{P}_q)^\top \Sigma_T^{-1} (\mathbf{I}_q - \mathbf{P}_q) \sqrt{T}\bar{h}(y_1^T; \theta_0).$$

But,

$$\sqrt{T}\bar{h}(y_1^T; \theta_0) \xrightarrow{d} N(0, \Sigma_0),$$

and, by a classical result,

$$\mathcal{C}_T \xrightarrow{d} \chi^2(q - p).$$

Hansen and Singleton (1982, 1983) started the literature on the estimation and testing of dynamic asset pricing models within a fully articulated rational expectations framework. Consider the classical system of Euler equations arising in the Lucas tree,

$$E \left[\beta \frac{u'(c_{t+1})}{u'(c_t)} (1 + r_{i,t+1}) - 1 \middle| \mathcal{F}_t \right] = 0, \quad i = 1, \dots, m,$$

where u is the utility function of the representative agent, r_i is the return on asset i , β is the time-discount factor, \mathcal{F}_t is the information set as of time t , and m is the number of assets. Consider the CRRA utility function, $u(x) = x^{1-\eta}/(1-\eta)$. If the model is well-specified, then, there exist some β_0 and η_0 such that:

$$E \left[\beta_0 \left(\frac{c_{t+1}}{c_t} \right)^{-\eta_0} (1 + r_{i,t+1}) - 1 \middle| \mathcal{F}_t \right] = 0, \quad i = 1, \dots, m.$$

To sum up, the dimension of the parameter vector is $p = 2$. To estimate the true parameter vector $\theta_0 \equiv (\beta_0, \eta_0)$, we may build up a system of orthogonality conditions. This system can be based on projecting observable variables predicted by the model onto other variables, some “instruments” included in the information set \mathcal{F}_t :

$$E[h(y_t; \theta_0)] = 0,$$

where, for some vector of z instruments, say, $\text{In}_t = [i_{1,t}, \dots, i_{z,t}]^\top$,

$$h(y_t; \theta) = \begin{pmatrix} \left[\beta \left(\frac{c_{t+1}}{c_t} \right)^{-\eta} (1 + r_{1,t+1}) - 1 \right] \cdot \text{In}_t \\ \vdots \\ \left[\beta \left(\frac{c_{t+1}}{c_t} \right)^{-\eta} (1 + r_{m,t+1}) - 1 \right] \cdot \text{In}_t \end{pmatrix}, \quad q = m \cdot z. \quad (5.9)$$

The instruments used to produce the orthogonality restrictions, may include constants, past values of consumption growth, $\frac{c_{t+1}}{c_t}$, or even past returns.

5.7 Simulation-based estimators

Ideally, MLE should be the preferred estimation method of parametric Markov models, as it leads to first-order efficiency. Yet economic theory places restrictions that make these models problematic to estimate through maximum ML. In these cases, GMM is a natural estimation method. But GMM can be unfeasible as well, in situations of interest. Assume, for example, that the data generating process is not i.i.d. Instead, data are generated by the transition function,

$$y_{t+1} = H(y_t, \epsilon_{t+1}; \theta_0), \quad (5.10)$$

where $H : \mathbb{R}^n \times \mathbb{R}^d \times \Theta \mapsto \mathbb{R}^n$, and ϵ_t is a vector of i.i.d. disturbances in \mathbb{R}^d . Assume the econometrician knows the function H . Let $z_t = (y_t, y_{t-1}, \dots, y_{t-l+1})$, $l < \infty$. In many cases of interest, the function \bar{h} in Eq. (5.6) can be written as,

$$\bar{h}(y_1^T; \theta) = \frac{1}{T} \sum_{t=1}^T \underbrace{[f_t^* - E(f(z_t, \theta))]}_{\equiv h(z_t, \theta)}, \quad (5.11)$$

where,

$$f_t^* = f(z_t, \theta_0),$$

is a vector-valued moment function, or “observation function,” a function that summarizes satisfactorily the data, so to speak. Consider, for example, Eq. (5.9) without the instruments In_t , where $f_t^* = (1 + r_{i,t+1})^{-1}$ and $E(f(z_t, \theta)) = \beta E\left(\left(\frac{c_{t+1}}{c_t}\right)^{-\eta}\right)$. Once we identify consumption growth with y_{t+1} , $y_{t+1} = \ln \frac{c_{t+1}}{c_t}$, and take the transition law in Eq. (5.10) to be log-normally distributed, as in some basic models we shall see in Part II of these lectures, we can compute $E(f(z_t, \theta))$ in closed form. Needless to say, the GMM estimator is unfeasible, if we are not able to compute the expectation $E(f(z_t, \theta))$ in closed form, for each θ . Simulation-based methods can make the method of moments feasible in this case.

5.7.1 Three simulation-based estimators

The basic idea underlying simulation-based methods is quite simple. While the moment conditions are too complex to be evaluated analytically, the model in Eq. (5.10) can be simulated. Accordingly, draw ϵ_t from its distribution, and save the simulated values $\hat{\epsilon}_t$. Compute recursively,

$$y_{t+1}^\theta = H(y_t^\theta, \hat{\epsilon}_{t+1}, \theta),$$

and create simulated moment functions as follows,

$$f_t^\theta \equiv f(z_t^\theta, \theta).$$

Consider the following parameter estimator,

$$\theta_T = \arg \min_{\theta \in \Theta} G_T(\theta)^\top W_T G_T(\theta), \quad (5.12)$$

where W_T is some weighing matrix, $G_T(\theta)$ is the simulated counterpart to \bar{h} in Eq. (5.11),

$$G_T(\theta) = \frac{1}{T} \sum_{t=1}^T \left(f_t^* - \frac{1}{S(T)} \sum_{s=1}^{S(T)} f_s^\theta \right),$$

and $S(T)$ is the simulated sample size, which we write as a function of the sample size T , for the purpose of the asymptotic theory.

The estimator θ_T , also known as the *Simulated Method of Moments* (SMM) estimator, aims to match the sample properties of the actual and simulated processes f_t^* and f_t^θ . It was introduced in a series of works, by McFadden (1989), Pakes and Pollard (1989), Lee and Ingram (1991) and Duffie and Singleton (1993). The simulated pseudo-maximum likelihood method of Laroque and Salanié (1989, 1993, 1994) can also be interpreted as a SMM estimator.

A second simulation-based estimator relies on the *indirect inference* principle (IIP), and was proposed by Gouriéroux, Monfort and Renault (1993) and Smith (1993). Instead of minimizing the distance of some moment conditions, the IIP relies on minimizing the parameters of an auxiliary, possibly misspecified model. For example, consider the following auxiliary parameter estimator,

$$\beta_T = \arg \max_{\beta} \ln L(y_1^T; \beta), \quad (5.13)$$

where L is the likelihood of some possibly misspecified model. Consider simulating S times the process y_t in Eq. (5.10), and computing,

$$\beta_T^s(\theta) = \arg \max_{\beta} \ln L(y_s(\theta)_1^T; \beta), \quad s = 1, \dots, S,$$

where $y_s(\theta)_1^T = (y_t^{\theta,s})_{t=1}^T$ are the simulated variables (for $s = 1, \dots, S$) when the parameter vector is θ . The IIP-based estimator is defined similarly as θ_T in Eq. (5.12), but with the function G_T given by,

$$G_T(\theta) = \beta_T - \frac{1}{S} \sum_{s=1}^S \beta_T^s(\theta). \quad (5.14)$$

The diagram in Figure 5.1 illustrates the main ideas underlying the IIP.

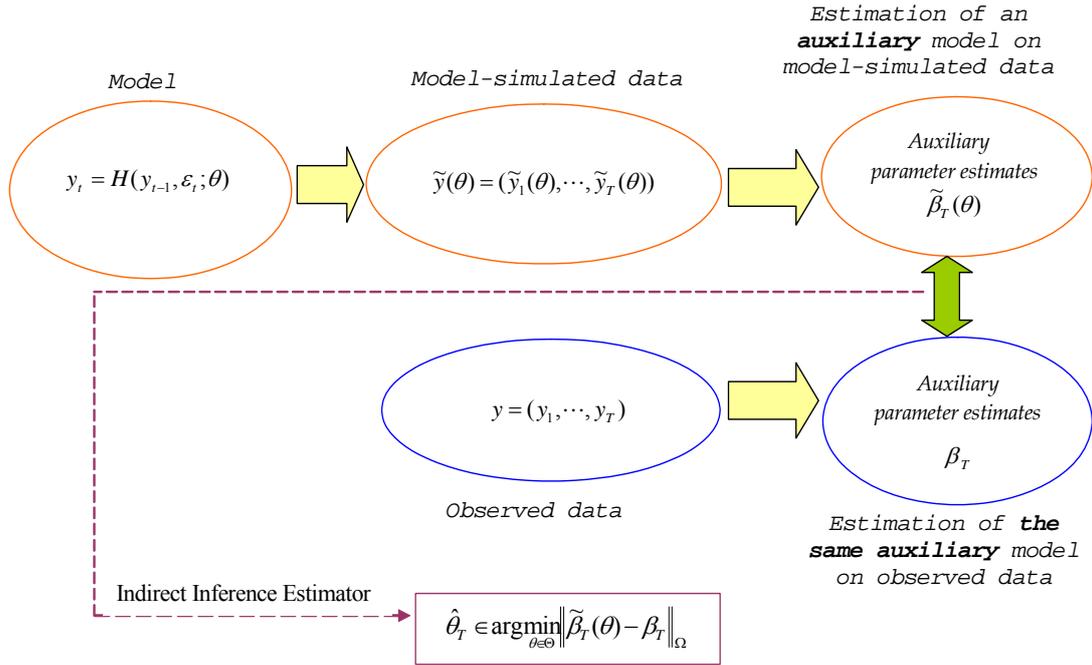


FIGURE 5.1. *The Indirect Inference principle.* Given the true model $y_t = H(y_{t-1}, \varepsilon_t; \theta)$, an estimator of θ based on the indirect inference principle ($\hat{\theta}_T$ say) makes the parameters of some *auxiliary* model $\tilde{\beta}_T(\hat{\theta}_T)$ as close as possible to the parameters β_T of *the same auxiliary* model estimated on the observations. That is, $\hat{\theta}_T = \operatorname{argmin}_{\theta \in \Theta} \|\tilde{\beta}_T(\theta) - \beta_T\|_{\Omega}$, for some norm Ω .

Finally, Gallant and Tauchen (1996) propose a simulation-based estimation method they label *efficient method of moments* (EMM). Their estimator sets,

$$G_T(\theta, \beta_T) = \frac{1}{N} \sum_{n=1}^N \frac{\partial}{\partial \beta} \ln f(y_n^\theta | z_{n-1}^\theta; \beta_T),$$

where $\frac{\partial}{\partial \beta} \ln f(y | z; \beta)$ is the score of some auxiliary model f , also known as the score generator, β_T is the Pseudo ML estimator of the auxiliary model, and $(y_n^\theta)_{n=1}^N$ is a long simulation (i.e. N is very large) of Eq. (5.10), with parameter vector set equal to θ . Finally, the weighting matrix W_T in Eq. (5.12) is taken to be any matrix \mathcal{I}_T^{-1} converging in probability to:

$$\mathcal{I} = E \left[\left[\frac{\partial}{\partial \beta} \ln f(y_2 | z_1; \beta) \right]_2 \right]. \tag{5.15}$$

To motivate this choice of $G_T(\theta)$, note that the auxiliary score, $\frac{\partial}{\partial \beta} \ln f(y_t | z_{t-1}; \beta_T)$, satisfies the following first order conditions:

$$\frac{1}{T} \sum_{n=1}^T \frac{\partial}{\partial \beta} \ln f(y_t | z_{t-1}; \beta_T) = 0,$$

which is the sample equivalent of

$$E \left[\frac{\partial}{\partial \beta} \ln f(y_2 | z_1; \beta^*) \right] = 0,$$

for some β^* . Likewise, we must have that with $\theta = \theta_0$, $G_T(\theta_0, \beta_T) = 0$, for large N . All in all, we want to find a stochastic process $H(y_t, \epsilon_{t+1}; \cdot)$ in Eq. (5.10), or a parameter vector θ such that the expectation of the score of the auxiliary model is zero, a very property of the score, arising even when the model is misspecified.

5.7.2 Asymptotic normality

We show, heuristically, how asymptotic normality obtains for the three estimators of Section 5.7.1, and then, define conditions under which asymptotic efficiency might obtain for the EMM.

5.7.2.1 SMM

Let,

$$\Sigma_0 = \sum_{j=-\infty}^{\infty} E \left[(f_t^* - E(f_t^*)) (f_{t-j}^* - E(f_{t-j}^*))^\top \right],$$

and suppose that

$$W_T \xrightarrow{p} W_0 = \Sigma_0^{-1}.$$

We now demonstrate that under this condition, as $T \rightarrow \infty$ and $S(T) \rightarrow \infty$,

$$\sqrt{T}(\theta_T - \theta_0) \xrightarrow{d} N \left(\mathbf{0}_p, (1 + \tau) (D_0^\top \Sigma_0^{-1} D_0)^{-1} \right), \quad (5.16)$$

where $\tau = \lim_{T \rightarrow \infty} \frac{T}{S(T)}$, $D_0 = E(\nabla_\theta G_\infty(\theta_0)) = E(\nabla_\theta f_\infty^{\theta_0})$, and the notation G_∞ means that G is drawn from its stationary distribution.

Indeed, the first order conditions satisfied by the SMM in Eq. (5.12) are,

$$\mathbf{0}_p = [\nabla_\theta G_T(\theta_T)]^\top W_T G_T(\theta_T) = [\nabla_\theta G_T(\theta_T)]^\top W_T \cdot [G_T(\theta_0) + \nabla_\theta G_T(\theta_0)(\theta_T - \theta_0)] + o_p(1).$$

That is,

$$\begin{aligned} \sqrt{T}(\theta_T - \theta_0) &\stackrel{d}{=} - \left([\nabla_\theta G_T(\theta_T)]^\top W_T \nabla_\theta G_T(\theta_0) \right)^{-1} [\nabla_\theta G_T(\theta_T)]^\top W_T \cdot \sqrt{T} G_T(\theta_0) \\ &\stackrel{d}{=} - (D_0^\top W_0 D_0)^{-1} D_0^\top W_0 \cdot \sqrt{T} G_T(\theta_T) \\ &= - (D_0^\top \Sigma_0^{-1} D_0)^{-1} D_0^\top \Sigma_0^{-1} \cdot \sqrt{T} G_T(\theta_0). \end{aligned} \quad (5.17)$$

We have,

$$\begin{aligned} \sqrt{T} G_T(\theta_0) &= \sqrt{T} \cdot \frac{1}{T} \sum_{t=1}^T \left(f_t^* - \frac{1}{S(T)} \sum_{s=1}^{S(T)} f_s^{\theta_0} \right) \\ &= \frac{1}{\sqrt{T}} \sum_{t=1}^T (f_t^* - E(f_\infty^*)) - \frac{\sqrt{T}}{\sqrt{S(T)}} \cdot \frac{1}{\sqrt{S(T)}} \sum_{s=1}^{S(T)} (f_s^{\theta_0} - E(f_\infty^{\theta_0})) \\ &\stackrel{d}{\rightarrow} N(0, (1 + \tau) \Sigma_0), \end{aligned}$$

where we have used the fact that $E(f_\infty^*) = E(f_\infty^{\theta_0})$. By using this result into Eq. (5.17) produces the convergence in Eq. (5.16). If $\tau = \lim_{T \rightarrow \infty} \frac{T}{S(T)} = 0$ (i.e. if the number of simulations grows faster than the sample size), the SMM estimator is as efficient as the GMM estimator. Finally, and obviously, we need that $\tau = \lim_{T \rightarrow \infty} \frac{T}{S(T)} < \infty$: the number of simulations $S(T)$ cannot grow more slowly than the sample size.

5.7.2.2 Indirect inference

The IIP-based estimator works slightly differently. For this estimator, even if the number of simulations S is fixed, asymptotic normality obtains without requiring S to go to infinity faster than the sample size. Basically, what really matters here is that ST goes to infinity.

By Eq. (5.17), and the discussion in Section 5.7.1, we know that asymptotically, the first order conditions satisfied by the IIP-based estimator are,

$$\sqrt{T}(\theta_T - \theta_0) \stackrel{d}{=} - (D_0^\top W_0 D_0)^{-1} D_0^\top W_0 \cdot \sqrt{T} G_T(\theta_0),$$

where G_T is as in Eq. (5.14), $D_0 = \nabla_\theta b(\theta)$, and $b(\theta)$ is solution to the limiting problem corresponding to the estimator in Eq. (5.13), viz

$$\beta(\theta) = \arg \max_{\beta} \left(\lim_{T \rightarrow \infty} \frac{1}{T} \ln L(y_1^T; \beta) \right).$$

We need to find the distribution of G_T in Eq. (5.14). We have,

$$\begin{aligned} \sqrt{T} G_T(\theta_0) &= \frac{1}{S} \sum_{s=1}^S \sqrt{T} (\beta_T - \beta_T^s(\theta_0)) \\ &= \frac{1}{S} \sum_{s=1}^S \sqrt{T} [(\beta_T - \beta_0) - (\beta_T^s(\theta_0) - \beta_0)] \\ &= \sqrt{T} (\beta_T - \beta_0) - \frac{1}{S} \sum_{s=1}^S \sqrt{T} (\beta_T^s(\theta_0) - \beta_0), \end{aligned}$$

where $\beta_0 = \beta(\theta_0)$. Hence, given the independence of the sample and the simulations,

$$\sqrt{T} G_T(\theta_0) \stackrel{d}{\rightarrow} N \left(0, \left(1 + \frac{1}{S} \right) \cdot \text{Asy.Var} \left(\sqrt{T} \beta_T \right) \right).$$

That is, asymptotically S can be fixed with respect to T .

5.7.2.3 Efficient method of moments

We have,

$$\theta_T = \arg \min_{\theta} G_T(\theta, \beta_T)^\top W_T G_T(\theta, \beta_T), \quad G_T(\theta, \beta_T) = \frac{1}{N} \sum_{n=1}^N \frac{\partial}{\partial \beta} \ln f(y_n^\theta | z_{n-1}^\theta; \beta_T).$$

The first order conditions are:

$$\begin{aligned} 0 &= \nabla_\theta G_T(\theta_T, \beta_T)^\top W_T G_T(\theta_T, \beta_T) \\ &\stackrel{d}{=} \nabla_\theta G_T(\theta_0, \beta_T)^\top W_T (G_T(\theta_0, \beta_T) + \nabla_\theta G_T(\theta_0, \beta_T) (\theta_T - \theta_0)), \end{aligned}$$

or

$$\sqrt{T}(\theta_T - \theta_0) \stackrel{d}{=} - \left(\nabla_\theta G_T(\theta_0, \beta_T)^\top W_T \nabla_\theta G_T(\theta_0, \beta_T) \right)^{-1} \nabla_\theta G_T(\theta_0, \beta_T)^\top W_T \sqrt{T} G_T(\theta_0, \beta_T).$$

We have, for some β^* ,

$$\sqrt{T}G_T(\theta_0, \beta_T) \stackrel{d}{=} \mathcal{J}\sqrt{T}(\beta_T - \beta^*) \stackrel{d}{\rightarrow} N(0, \mathcal{I}),$$

where $\mathcal{J} = E\left(\frac{\partial}{\partial\beta\partial\beta^\top} \ln f(y_2|y_1; \beta)\right)$ and \mathcal{I} is as in Eq. (5.15). Hence,

$$\sqrt{T}(\theta_T - \theta_0) \stackrel{d}{\rightarrow} N(0, V),$$

where,

$$V = (\nabla_\theta G^\top W \nabla_\theta G)^{-1} \nabla_\theta G^\top W \mathcal{I} W^\top \nabla_\theta G (\nabla_\theta G^\top W \nabla_\theta G)^{-1}.$$

With $W = \mathcal{I}^{-1}$, this variance collapses to,

$$V = (\nabla_\theta G^\top \mathcal{I}^{-1} \nabla_\theta G)^{-1}. \quad (5.18)$$

5.7.2.4 Spanning scores

This section provides a heuristic discussion of the conditions under which the EMM achieves the Cramer-Rao lower bound. Consider the following definition, which is similar to that in Tauchen (1997). Of a given span of moment conditions s_f , say that of the EMM, we say that it also spans the true score if,

$$\text{var}(s|s_f) = 0, \quad (5.19)$$

where s denotes the true score. From Eq. (5.18), we know that the asymptotic variance of the EMM, say var_{EMM} , satisfies:

$$\text{var}_{\text{EMM}}^{-1} \equiv V^{-1} = \nabla_\theta G^\top \text{var}(s_f)^{-1} \nabla_\theta G.$$

By the linear projection,

$$s = Bs_f + \epsilon, \quad B = \text{cov}(s, s_f) \text{var}(s_f)^{-1},$$

we have,

$$\text{var}_{\text{MLE}}^{-1} = \text{var}(s) = B \text{var}(s_f) B^\top + \text{var}(s|s_f) = \text{cov}(s, s_f) \text{var}(s_f)^{-1} \text{cov}(s, s_f)^\top + \text{var}(s|s_f), \quad (5.20)$$

where var_{MLE} denotes the asymptotic variance of the MLE. We claim that:

$$\text{cov}(s, s_f)^\top = \nabla_\theta G. \quad (5.21)$$

Indeed, under regularity conditions,

$$\begin{aligned} \nabla_\theta G(\theta_0, \beta^*) &= \left[\frac{\partial}{\partial\theta} \left(\int \frac{\partial}{\partial\beta} \ln f(y; \beta^*) p(y, \theta) dy \right) \right]_{\theta=\theta_0} \\ &= \int \frac{\partial}{\partial\beta} \ln f(y; \beta^*) \frac{\partial}{\partial\theta} p(y, \theta_0) dy \\ &= \int \left(\frac{\partial}{\partial\beta} \ln f(y; \beta^*) \frac{\partial}{\partial\theta} \ln p(y, \theta_0) \right) p(y, \theta_0) dy \\ &= \text{cov}(s, s_f)^\top, \end{aligned}$$

where $p(y, \theta_0)$ is the true density. Next, replace Eq. (5.21) into Eq. (5.20),

$$\text{var}_{\text{MLE}}^{-1} = \nabla_\theta G^\top \text{var}(s_f)^{-1} \nabla_\theta G + \text{var}(s|s_f) = \text{var}_{\text{EMM}}^{-1} + \text{var}(s|s_f).$$

Therefore, the EMM estimator achieves the Cramer-Rao lower bound under the spanning condition in Eq. (5.19).

5.7.3 A fourth simulation-based estimator: Simulated maximum likelihood

Estimating the parameters of stochastic differential equations is a recurrent theme in empirical finance. Consider a continuous time model,

$$dy(\tau) = b(y(\tau); \theta) d\tau + \Sigma(y(\tau); \theta) dW(\tau), \quad (5.22)$$

where $W(\tau)$ is a Brownian motion and b and Σ are two functions guaranteeing a strong solution to Eq. (5.22). Except in special cases (e.g., the affine models reviewed in Chapter 12), the likelihood function of the data generated by this process is unknown. We can then use one of the three estimators we have presented in section 5.7.1. Alternatively, we might use simulated maximum likelihood, a method introduced in finance by Santa-Clara (1995) (see, also, Brandt and Santa-Clara, 2002). We only provide the idea of the method, not the asymptotic theory.

Suppose, then, that we observe discretely sample data generated by Eq. (5.22): $y_0, y_1, \dots, y_t, \dots, y_T$, where T is the sample size. We need to know the transition density, say $p(y_{t+1}|y_t; \theta)$, to implement maximum likelihood, which we assume we do not know. Consider, then, the Euler approximation to Eq. (5.22),

$$y_{(k+1)/n} = y_{k/n} + b(y_{k/n}; \theta) \frac{1}{n} + \Sigma(y_{k/n}; \theta) \sqrt{\frac{1}{n}} \epsilon_{k+1}, \quad (5.23)$$

where ϵ_k is a sequence of i.i.d. random variables with expectation zero and unit variance. This stochastic process is defined at the dates $\frac{k}{n}$, for k integer. Let $[Tn]$ denote the integer part of Tn , and for $k = 1, \dots, [Tn]$, set

$$\hat{y}_\tau^{(n)} = y_{k/n}, \quad \text{if } \frac{k}{n} \leq \tau \leq \frac{k+1}{n}.$$

In other words, we are chopping the time interval between two observations, $[t, t+1]$, in n pieces, and then take n to be large. We know that as $n \rightarrow \infty$, $\hat{y}_t^{(n)} \Rightarrow y(t)$ as $n \rightarrow \infty$, where \Rightarrow denotes “weak convergence,” or “convergence in distribution,” meaning that all finite dimensional distributions of $\hat{y}_t^{(n)}$ converge to those of $y(t)$ as $n \rightarrow \infty$. The idea underlying simulated maximum likelihood, then, is to estimate the transition density, $p(y_{t+1}|y_t; \theta)$, through simulations of Eq. (5.23), performed using a large value of n . Note, we cannot guarantee the transition density is recovered by simulating Eq. (5.23), not even for a large value of n . We can only perform an imperfect simulation of Eq. (5.23).

The likelihood function is,

$$L = p(y_0; \theta) \prod_{t=0}^{T-1} p(y_{t+1}|y_t; \theta),$$

where $p(y_0; \theta)$ denotes the marginal density of the first observation, y_0 .

Let $p^n(y'|y; \theta)$ the transition density of the data generated by Eq. (5.23). Then, if ϵ is normally distributed,

$$p^n(y_{(k+1)/n}|y_{k/n}; \theta) = \varphi\left(y_{(k+1)/n}; y_{k/n} + b(y_{k/n}; \theta) \frac{1}{n}; \Sigma^2(y_{k/n}; \theta) \frac{1}{n}\right),$$

where $\varphi(u; \mu; \sigma^2)$ denotes the Gaussian density with mean μ and variance σ^2 . Moreover, we have, approximately,

$$\begin{aligned} p^n(y_{t+1}|y_t; \theta) &= \int p^n(y_{t+1}|x; \theta) p^n(x|y_t; \theta) dx \\ &= \int \varphi\left(y_{t+1}; x + b(x; \theta) \frac{1}{n}; \Sigma^2(x; \theta) \frac{1}{n}\right) p^n(x|y_t; \theta) dx, \end{aligned}$$

where we have set $x = y_{t+1 - \frac{1}{n}}$. We may, now, draw values of x from $p^n(x|y_t; \theta)$, as explained in a moment, and estimate $p^n(y_{t+1}|y_t; \theta)$ through:

$$p^{n,S}(y_{t+1}|y_t; \theta) \equiv \frac{1}{S} \sum_{j=1}^S \varphi\left(y_{(k+1)/n}; \tilde{x}^j + b(\tilde{x}^j; \theta) \frac{1}{n}; \Sigma^2(\tilde{x}^j; \theta) \frac{1}{n}\right),$$

where \tilde{x}^j is obtained by iterating Eq. (5.23) from time t to time $t + 1 - \frac{1}{n}$. Under regularity conditions, we have that for all $\theta \in \Theta$, $\sup_{y', y} |p^{n,S}(y'|y; \theta) - p(y'|y; \theta)| \rightarrow 0$ as n and S get large, with $\frac{\sqrt{S}}{n} \rightarrow 0$.

5.7.4 Advances

The three estimators that we have examined in Sections 5.7.1-5.7.2, are general-purpose, but in general, they do not lead to asymptotic efficiency, unless the true score belongs to the span of the moment conditions, as explained in Section 5.7.2.4. There exist other simulation-based methods, which aim to approximate the likelihood function through simulations (e.g., Lee, 1995; Hajivassiliou and McFadden, 1998): for example, the simulated maximum likelihood estimator in Section 5.7.2.3 can be used to estimate the parameters of stochastic differential equations. While methods based on simulated likelihood lead to asymptotically efficient estimators, they address specific estimation problems, just as the example of Section 5.7.2.3 illustrates.

There exist estimators that are both general purpose and that can lead to asymptotic efficiency. Fermanian and Salanié (2004) consider an estimator that relies on approximating the likelihood function through kernel estimates obtained simulating the model of interest. Carrasco, Chernov, Florens and Ghysels (2007) rely on a “continuum of moment conditions” matching model-based (simulated) characteristic functions to data-based characteristic functions. Altissimo and Mele (2009) propose an estimator based on a continuum of moment conditions, which minimizes a certain distance between conditional densities estimated with the true data and conditional densities estimated with data simulated from the model, where both conditional densities are estimated through kernel methods.

5.7.5 In practice? Latent factors and identification

The estimation theory of this section does not rule out the situation where some of the variables in Eq. (5.10) are unobservable. The principle to follow is very simple, one applies any of the methods we have discussed to those variables simulated out of Eq. (5.10), which correspond to the observed ones. For example, we may want to estimate the following model of the short-term rate $r(\tau)$, discussed at length in Chapter 12:

$$\begin{aligned} dr(\tau) &= \kappa_r (\bar{r} - r(\tau)) d\tau + \sqrt{v(\tau)} dW_1(\tau) \\ dv(\tau) &= \kappa_v (\bar{v} - v(\tau)) d\tau + \xi \sqrt{v(\tau)} dW_2(\tau) \end{aligned} \tag{5.24}$$

where $v(\tau)$ is the short-term rate instantaneous, stochastic variance, W_1 and W_2 are two standard Brownian motions, and the parameter vector of interest is $\theta = [\kappa_r \bar{r} \kappa_v \bar{v} \xi]$. Let us consider one of the methods discussed so far, say indirect inference. The logical steps to follow, then, are (i) to simulate Eqs. (5.24), and (ii) to calibrate an auxiliary model to the short term rate data simulated out of Eqs. (5.24) which is as close as possible to the very same auxiliary model fitted on true data. Note, in doing so, we just have to neglect the volatility data simulated out of Eqs. (5.24), as these data are obviously unobservable.

The question arises, therefore, as to whether the auxiliary model one chooses is rich enough to allow identifying the model's parameter vector θ . There might be many combinations of unobserved random processes $v(\tau)$ that are consistent with the likelihood of any given auxiliary model. So which auxiliary model to fit, in practice? Gallant and Tauchen (1996) asked this question long time ago. Needless to mention, there are no general answers to this question. Very simply, one requires the model to be identifiable, which is likely to happen once the auxiliary model is "rich enough." In an impressive series of applied work, Gallant and Tauchen and their co-authors have proposed semi-nonparametric score generators, as a way to get as close as possible to a "rich" model. Intuitively, by increasing the order of Hermite expansions, semi-nonparametric scores might converge to the true ones. Alternatively, one might use a continuum of moment conditions, as explained in Section 5.7.4. For example, the nonparametric density estimates in Altissimo and Mele (2009) converge to the true ones, once the bandwidth parameters used to smooth out these estimates gets smaller and smaller. In the next section, we provide a discussion of how asset prices might help convey information about unobserved processes and lead to statistical efficiency.

5.8 Asset pricing, prediction functions, and statistical inference

We develop conditions, which ensure the feasibility of estimation methods in a context where an unobservable multidimensional process is estimated in conjunction with prediction functions suggested by asset pricing models.⁴ We assume that the data generating process is a multidimensional partially observed diffusion process solution to,

$$dy(\tau) = b(y(\tau); \theta) d\tau + \Sigma(y(\tau); \theta) dW(\tau), \quad (5.25)$$

where W is a multidimensional process and (b, Σ) satisfy some regularity conditions we single out below. We analyze situations where the original partially observed system in Eq. (5.25) can be estimated by augmenting it with a number of observable deterministic functions of the state. In many situations, such deterministic functions are suggested by asset pricing theories in a natural way. Typical examples include the price of derivatives or in general, any functional of asset prices (such as asset returns, bond yields, implied volatilities).

The idea to use asset pricing predictions to improve the fit of models with unobservable factors has been explored at least by, e.g., Christensen (1992), Pastorello, Renault and Touzi (2000), Chernov and Ghysels (2000), Singleton (2001), and Pastorello, Patilea and Renault (2003).

We consider a standard Markov pricing setting. For fixed $t \geq 0$, we let M be the expiration date of a contingent claim with rational price process $c = \{c(y(\tau), M - \tau)\}_{\tau \in [t, M]}$, and let $\{z(y(\tau))\}_{\tau \in [t, M]}$ and $\Pi(y)$ be the associated intermediate payoff process and final payoff function,

⁴This section is based on an unpublished appendix of Altissimo and Mele (2009).

respectively. Let $\partial/\partial\tau + L$ be the usual infinitesimal generator of the system in Eq. (5.25), taken under the risk-neutral probability. Then, as we saw in Chapter 4, we have that in a frictionless, arbitrage-free market, c is the solution to the following partial differential equation:

$$\begin{cases} 0 = \left(\frac{\partial}{\partial\tau} + L - R \right) c(y, M - \tau) + z(y), \quad \forall (y, \tau) \in Y \times [t, M] \\ c(y, 0) = \Pi(y), \quad \forall y \in Y \end{cases} \quad (5.26)$$

where $R \equiv R(y)$ is the short-term rate. We call *prediction function* any continuous and twice differentiable function $c(y; M - \tau)$ solution to the partial differential equation and boundary condition in (5.26). Examples of contingent claims with prices satisfying (5.26) are derivatives, typically.

Next, we augment the system in Eq. (5.25) with $d - q$ prediction functions, where q denotes the number of the observable variables in Eq. (5.25). Precisely, we let:

$$C(\tau) \equiv (c(y(\tau), M_1 - \tau), \dots, c(y(\tau), M_{d-q} - \tau)), \quad \tau \in [t, M_1]$$

where $\{M_i\}_{i=1}^{d-q}$ is an increasing sequence of fixed maturity dates. Furthermore, we define the measurable vector valued function:

$$\phi(y(\tau); \theta, \gamma) \equiv (y^o(\tau), C(y(\tau))), \quad \tau \in [t, M_1], \quad (\theta, \gamma) \in \Theta \times \Gamma, \quad (5.27)$$

where $y^o(\tau)$ denotes the vector of observable variables in Eq. (5.25), and $\Gamma \subset \mathbb{R}^{p_\gamma}$ is a compact parameter set containing additional parameters. These new parameters arise from the change of measure leading to the pricing model in Eq. (5.27), and are now part of our estimation problem.

We assume that the pricing model in Eq. (5.27) is correctly specified. That is, all contingent claim prices in the economy are taken to be generated by the prediction function $c(y, M - \tau)$ for some $(\theta_0, \gamma_0) \in \Theta \times \Gamma$. For simplicity, we also consider a stylized situation in which all contingent claims have the same contractual characteristics specified by $\mathcal{C} \equiv (z, \Pi)$. More generally, one may define a series of classes of contingent claims $\{\mathcal{C}_j\}_{j=1}^J$, where the class of contingent claims j has characteristics specified by $\mathcal{C}_j \equiv (z_j, \Pi_j)$. As an example, assets belonging to the class \mathcal{C}_1 can be European options, assets belonging to the class \mathcal{C}_2 can be bonds. The number of prediction functions that we would introduce in this case would be equal to $d - q = \sum_{j=1}^J M^j$, where M^j is the number of prediction functions within class of assets j . To keep the presentation simple, we do not consider such a more general situation.

The objective is to define estimators of the parameter vector (θ_0, γ_0) , under which observations were generated. We want to use any of the simulation methods reviewed in Section 5.7 to produce an estimator of (θ_0, γ_0) . The idea, as usual, is to make the finite dimensional distributions of ϕ implied by the pricing model in Eq. (5.27) and the fundamentals in Eq. (5.26) as close as possible to the sample counterparts of ϕ . Let $\Phi \subseteq \mathbb{R}^d$ be the domain on which ϕ takes values. As illustrated by Figure 5.2, we want to move from the unfeasible domain Y of the original state variables in Eq. (5.25) (observables and not) to the domain Φ on which only observable variables take value. Ideally, we would like to implement such a change in domain in order to recover as much information as possible about the original unobserved process in (5.25). Clearly, ϕ is fully revealing whenever it is globally invertible. However, we will show that estimation is feasible even when ϕ is only locally one-to-one.

An important feature of the theory in this section is that it does not hinge upon the availability of contingent prices data covering the same sample period covered by the observables

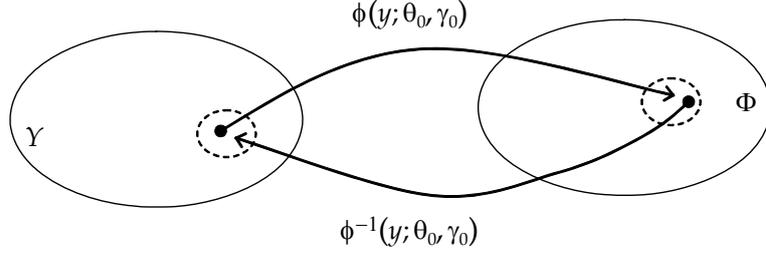


FIGURE 5.2. *Asset pricing, the Markov property, and statistical efficiency.* Y is the domain on which the partially observed primitive state process $y \equiv (y^o \ y^u)^\top$ takes values, Φ is the domain on which the observed system $\phi \equiv (y^o \ C(y))^\top$ takes values in Markovian economies, and $C(y)$ is a contingent claim price process in \mathbb{R}^{d-q^*} . Let $\phi^c = (y^o, c(y, \ell_1), \dots, c(y, \ell_{d-q^*}))$, where $\{c(y, \ell_j)\}_{j=1}^{d-q^*}$ forms an intertemporal cohort of contingent claim prices, as in Definition 5.3. If the *local restrictions* of ϕ are one-to-one and onto, statistical inference about θ and γ can be made, using information about the price of derivative contracts, ϕ^c . If ϕ is also *globally* invertible, statistical inference can lead to first-order asymptotic efficiency, once conditioned upon ϕ^c .

in Eq. (5.25). First, the price of a given contingent claim is typically not available for a long sample period. As an example, available option data often include option prices with a life span smaller than the usual sample span of the underlying asset prices. By contrast, it is common to observe long time series of option prices having the same maturity. Second, the price of a single contingent claim depends on the time-to-maturity of the claim; therefore, it does not satisfy the stationarity assumptions maintained in this paper. To address these issues, we deal with data on assets having the same characteristics at each point in time. Precisely, consider the data generated by the following random processes:

DEFINITION 5.3. (Intertemporal (ℓ, N) -cohort of contingent claim prices) *Given a prediction function $c(y; M - \tau)$ and a N -dimensional vector $\ell \equiv (\ell_1, \dots, \ell_N)$ of fixed time-to-maturity, an intertemporal (ℓ, N) -cohort of contingent claim prices is any collection of contingent claim price processes $c(\tau, \ell) \equiv (c(y(\tau), \ell_1), \dots, c(y(\tau), \ell_N))$ ($\tau \geq 0$) generated by the pricing model (5.27).*

Consider for example a sample realization of three-months at-the-money option prices, or a sample realization of six-months zero-coupon bond prices. Long sequences such as the ones in these examples are common to observe. If these sequences were generated by the pricing model in Eq. (5.27), as in Definition 5.3, they would be deterministic functions of y , and hence stationary. We now develop conditions ensuring both feasibility and first-order efficiency of the class of simulation-based estimators, as applied to this kind of data. Let \bar{a} denote the matrix having the first q rows of Σ , the diffusion matrix in Eq. (5.25). Let ∇C denote the Jacobian of C with respect to y . We have:

THEOREM 5.4. (Asset pricing and Cramer-Rao lower bound) *Suppose to observe an intertemporal $(\ell, d-q)$ -cohort of contingent claim prices $c(\tau, \ell)$, and that there exist prediction functions C in \mathbb{R}^{d-q} with the property that for $\theta = \theta_0$ and $\gamma = \gamma_0$,*

$$\begin{pmatrix} \bar{a}(\tau) \cdot \Sigma(\tau)^{-1} \\ \nabla C(\tau) \end{pmatrix} \neq 0, \quad P \otimes d\tau\text{-a.s.} \quad \text{all } \tau \in [t, t+1], \quad (5.28)$$

where C satisfies the initial condition $C(t) = c(t, \ell) \equiv (c(y(t), \ell_1), \dots, c(y(t), \ell_{d-q}))$. Let $\phi_t^c = (y^o(t), c(y(t), \ell_1), \dots, c(y(t), \ell_{d-q}))$. Then, any simulation-based estimator applied to ϕ_t^c is feasible. Moreover, assume ϕ_t^c is also Markov. Then, any estimator with a span of moment conditions for ϕ_t^c that also spans the true score, attains the Cramer-Rao lower bound, with respect to the fields generated by ϕ_t^c .

According to Theorem 5.4, any estimator is feasible, whenever ϕ is locally invertible for a time span equal to the sampling interval. As Figure 5.2 illustrates, condition (5.28) is satisfied whenever ϕ is locally one-to-one and onto.⁵ If ϕ is also globally invertible for the same time span, ϕ^c is Markov. The last part of this theorem says that in this case, any estimator is asymptotically efficient. We emphasize that this conclusion is about first-order efficiency in the joint estimation of θ and γ given the observations on ϕ^c .

Naturally, condition (5.28) does not ensure that ϕ is globally one-to-one and onto: ϕ might have many locally invertible restrictions.⁶ In practice, ϕ might fail being globally invertible because monotonicity properties of ϕ may break down in multidimensional diffusion models. For example, in models with stochastic volatility, option prices can be decreasing in the underlying asset price (see Bergman, Grundy and Wiener, 1996). In models of the yield curve with stochastic volatility, to cite a second example, medium-long term bond prices can be increasing in the short-term rate (see Mele, 2003). These cases might arise as there is no guarantee that the solution to a stochastic differential system is nondecreasing in the initial condition of one if its components, which is, instead, always true in the scalar case.

When all components of vector y^o represent the prices of assets actively traded in frictionless markets, (5.28) corresponds to a condition ensuring market completeness in the sense of Harrison and Pliska (1983). As an example, condition (5.28) for Heston's (1993) model is $\partial c / \partial \sigma \neq 0$ $P \otimes dt$ -a.s., where σ denotes instantaneous volatility of the price process. This condition is satisfied by the Heston's model. In fact, Romano and Touzi (1997) showed that within a fairly general class of stochastic volatility models, option prices are *always* strictly increasing in σ whenever they are convex in Q . Theorem 5.4 can be used to implement efficient estimators in other complex multidimensional models. Consider for example a three-factor model of the yield curve. Consider a state-vector (r, σ, ℓ) , where r is the short-term rate and σ, ℓ are additional factors (such as, say, instantaneous short-term rate volatility and a central tendency factor). Let $u^{(i)} = u(r(\tau), \sigma(\tau), \ell(\tau); M_i - \tau)$ be the time τ rational price of a pure discount bond expiring at $M_i \geq \tau$, $i = 1, 2$, and take $M_1 < M_2$. Let $\phi \equiv (r, u^{(1)}, u^{(2)})$. Condition (5.28) for this model is then,

$$u_\sigma^{(1)} u_\ell^{(2)} - u_\ell^{(1)} u_\sigma^{(2)} \neq 0, \quad P \otimes dt\text{-a.s.} \quad \tau \in [t, t + 1], \quad (5.29)$$

where subscripts denote partial derivatives. It is easily checked that this same condition must be satisfied by models with correlated Brownian motions and by yet more general models. Classes of models of the short-term rate for which condition (5.29) holds are more intricate to identify than in the European option pricing case seen above (see Mele, 2003).

⁵Local invertibility of ϕ means that for every $y \in Y$, there exists an open set Y_* containing y such that the restriction of ϕ to Y_* is invertible. Let $J\phi$ denote the Jacobian of ϕ . Then, we have that ϕ is locally invertible on Y_* if $\det J\phi \neq 0$ on Y_* , which is condition (5.28).

⁶As an example, consider the mapping $\mathbb{R}^2 \mapsto \mathbb{R}^2$ defined as $\phi(y_1, y_2) = (e^{y_1} \cos y_2, e^{y_1} \sin y_2)$. The Jacobian satisfies $\det J\phi(y_1, y_2) = e^{2y_1}$, yet ϕ is 2π -periodic with respect to y_2 . For example, $\phi(0, 2\pi) = \phi(0, 0)$.

5.9 Appendix 1: Proof of selected results

PROOF OF EQ. (5.2). We have: $P(A_1 \cap A_2) = P(A_1) \cdot P(A_2 | A_1)$. Consider the event $E \equiv A_1 \cap A_2$. We still have,

$$\Pr(A_3 | A_1 \cap A_2) = \Pr(A_3 | E) = \frac{\Pr(A_3 \cap E)}{\Pr(E)} = \frac{\Pr(A_3 \cap A_1 \cap A_2)}{\Pr(A_1 \cap A_2)}.$$

That is,

$$\Pr\left(\bigcap_{i=1}^3 A_i\right) = \Pr(A_1 \cap A_2) \cdot \Pr(A_3 | A_1 \cap A_2) = \Pr(A_1) \cdot \Pr(A_2 | A_1) \cdot \Pr(A_3 | A_1 \cap A_2).$$

Continuing, we obtain Eq. (5.2). \parallel

5.10 Appendix 2: Collected notions and results

CONVERGENCE IN PROBABILITY. A sequence of random vectors $\{x_T\}$ converges in probability to the random vector \tilde{x} if for each $\epsilon > 0$, $\delta > 0$ and each $i = 1, 2, \dots, N$, there exists a $T_{\epsilon, \delta}$ such that for every $T \geq T_{\epsilon, \delta}$,

$$\Pr(|x_{Ti} - \tilde{x}_i| > \delta) < \epsilon.$$

This is succinctly written as $x_T \xrightarrow{p} \tilde{x}$, or $\text{plim } x_T = \tilde{x}$, if $\tilde{x} \equiv \bar{x}$, a constant.

Convergence in probability generalizes the standard notion of a limit of a deterministic sequence. Of a deterministic sequence x_T , we say it converges to some limit \bar{x} if, for $\kappa > 0$, there exists a T_κ : for each $T \geq T_\kappa$ we have that $|x_T - \bar{x}| < \kappa$. Convergence in probability can also be restated as saying that:

$$\lim_{T \rightarrow \infty} \Pr(|x_{Ti} - \tilde{x}_i| > \delta) = 0.$$

The following is a stronger notion of convergence:

ALMOST SURE CONVERGENCE. A sequence of random vectors $\{x_T\}$ converges almost surely to the random vector \tilde{x} if, for each $i = 1, 2, \dots, N$, we have:

$$\Pr(\omega : x_{Ti}(\omega) \rightarrow \tilde{x}_i) = 1,$$

where ω denotes the entire random sequence x_{Ti} . This is succinctly written as $x_T \xrightarrow{a.s.} \tilde{x}$.

Almost sure convergence implies convergence in probability. Convergence in probability means that for each $\epsilon > 0$, $\lim_{T \rightarrow \infty} \Pr(\omega : |x_{Ti}(\omega) - \tilde{x}_i| < \epsilon) = 1$. Almost sure convergence requires that $\Pr(\lim_{T \rightarrow \infty} x_{Ti} \rightarrow \tilde{x}_i) = 1$ or that

$$\lim_{T' \rightarrow \infty} \Pr\left(\sup_{T \geq T'} |x_{Ti} - \tilde{x}_i| > \delta\right) = \lim_{T' \rightarrow \infty} \Pr\left(\bigcup_{T \geq T'} |x_{Ti} - \tilde{x}_i| > \delta\right) = 0.$$

Next, assume that the second order moments of all x_i are finite. We have:

CONVERGENCE IN QUADRATIC MEAN. A sequence of random vectors $\{x_T\}$ converges in quadratic mean to the random vector \tilde{x} if for each $i = 1, 2, \dots, N$, we have:

$$\lim_{T' \rightarrow \infty} E[(x_{Ti} - \tilde{x}_i)^2] \rightarrow 0.$$

This is succinctly written as $x_T \xrightarrow{q.m.} \tilde{x}$.

REMARK. By Chebyshev's inequality,

$$\Pr(|x_{Ti} - \tilde{x}_i| > \delta) \leq \frac{E[(x_{Ti} - \tilde{x}_i)^2]}{\delta^2},$$

which shows that convergence in quadratic mean implies convergence in probability.

We now turn to a weaker notion of convergence:

CONVERGENCE IN DISTRIBUTION. Let $\{f_T(\cdot)\}_T$ be the sequence of probability distributions (that is, $f_T(x) = \text{pr}(x_T \leq x)$) of the sequence of the random vectors $\{x_T\}$. Let \tilde{x} be a random vector with probability distribution $f(x)$. A sequence $\{x_T\}$ converges in distribution to \tilde{x} if, for each $i = 1, 2, \dots, N$, we have:

$$\lim_{T \rightarrow \infty} f_T(x) = f(x).$$

This is succinctly written as $x_T \xrightarrow{d} \tilde{x}$.

The following two results are useful to the purpose of this chapter:

SLUTZKY'S THEOREM. If $y_T \xrightarrow{p} \bar{y}$ and $x_T \xrightarrow{d} \tilde{x}$, then:

$$y_T \cdot x_T \xrightarrow{d} \bar{y} \cdot \tilde{x}.$$

CRAMER-WOLD DEVICE. Let λ be a N -dimensional vector of constants. We have:

$$x_T \xrightarrow{d} \tilde{x} \Leftrightarrow \lambda^\top \cdot x_T \xrightarrow{d} \lambda^\top \cdot \tilde{x}.$$

The following example illustrates the Cramer-Wold device. If $\lambda^\top \cdot x_T \xrightarrow{d} N(0; \lambda^\top \Sigma \lambda)$, then $x_T \xrightarrow{d} N(0; \Sigma)$.

We now state two laws about convergence in probability.

WEAK LAW (NO. 1) (Khinchine). Let $\{x_T\}$ be a i.i.d. sequence satisfying $E(x_T) = \mu < \infty \forall T$. We have:

$$\bar{x}_T \equiv \frac{1}{T} \sum_{t=1}^T x_t \xrightarrow{p} \mu.$$

WEAK LAW (NO. 2) (Chebyshev). Let $\{x_T\}$ be a sequence independent but not identically distributed, satisfying $E(x_T) = \mu_T < \infty$ and $E[(x_T - \mu_T)^2] = \sigma_T^2 < \infty$. If $\lim_{T \rightarrow \infty} \frac{1}{T^2} \sum_{t=1}^T \sigma_t^2 \rightarrow 0$, then:

$$\bar{x}_T \equiv \frac{1}{T} \sum_{t=1}^T x_t \xrightarrow{p} \bar{\mu}_T \equiv \frac{1}{T} \sum_{t=1}^T \mu_t.$$

We now state and provide a proof of the central limit theorem in a simple setting.

CENTRAL LIMIT THEOREM. Let $\{x_T\}$ be a i.i.d. sequence, satisfying $E(x_T) = \mu < \infty$ and $E[(x_T - \mu)^2] = \sigma^2 < \infty \forall T$. Let $\bar{x}_T \equiv \frac{1}{T} \sum_{t=1}^T x_t$. We have,

$$\frac{\sqrt{T}(\bar{x}_T - \mu)}{\sigma} \xrightarrow{d} N(0, 1).$$

The multidimensional version of this theorem requires a mere change in notation. For the proof, the classic method relies on the characteristic functions. Let:

$$\varphi(t) \equiv E(e^{itx}) = \int e^{itx} f(x) dx, \quad \mathbf{i} \equiv \sqrt{-1}.$$

We have $\frac{\partial^r}{\partial t^r} \varphi(t) \Big|_{t=0} = \mathbf{i}^r m^{(r)}$, where $m^{(r)}$ is the r -th order moment. By a Taylor's expansion,

$$\varphi(t) = \varphi(0) + \frac{\partial}{\partial t} \varphi(t) \Big|_{t=0} t + \frac{1}{2} \frac{\partial^2}{\partial t^2} \varphi(t) \Big|_{t=0} t^2 + \dots = 1 + \mathbf{i} m^{(1)} t - m^{(2)} \frac{1}{2} t^2 + \dots$$

Next, let $\bar{x}_T = \frac{1}{T} \sum_{t=1}^T x_t$, and consider the random variable,

$$Y_T \equiv \frac{\sqrt{T}(\bar{x}_T - \mu)}{\sigma} = \frac{1}{\sqrt{T}} \sum_{t=1}^T \frac{x_t - \mu}{\sigma}.$$

The characteristic function of Y_T is the product of the characteristic functions of $a_t \equiv \frac{x_t - \mu}{\sqrt{T}\sigma}$, which are all the same: $\varphi_{Y_T}(t) = (\varphi_a(t))^T$, where $\varphi_a(t) = 1 - \frac{t^2}{2T} + \dots$. Therefore,

$$\varphi_{Y_T}(t) = \varphi\left(\frac{t}{\sqrt{T}}\right)^T = \left(1 - \frac{1}{2} \frac{t^2}{T} + o(T^{-1})\right)^T.$$

Clearly, $\lim_{T \rightarrow \infty} \varphi_{Y_T}(t) = e^{-\frac{1}{2}t^2}$, which is the characteristic function of a standard Gaussian variable.

5.11 Appendix 3: Theory for maximum likelihood estimation

Assume that $\hat{\theta}_T \xrightarrow{a.s.} \theta_0$, and that $H(y, \theta) \equiv \nabla_{\theta\theta} \ln L(\theta|y)$ exists, it is continuous in θ uniformly in y and that we can differentiate twice inside the integral $\int L(\theta|y)dy = 1$. We have:

$$s_T(\theta) = \frac{1}{T} \sum_{t=1}^T \nabla_{\theta} \ln L(\theta|y_t).$$

Consider the c -parametrized curves $\theta(c) = c\Box(\theta_0 - \hat{\theta}_T) + \hat{\theta}_T$ where, for all $c \in (0, 1)^p$ and $\theta \in \Theta$, $c\Box\theta$ denotes a vector in Θ where the i th element is $c^{(i)}\theta^{(i)}$. By the intermediate value theorem, there exists then a c^* in $(0, 1)^p$ such that we have almost surely:

$$s_T(\hat{\theta}_T) = s_T(\theta_0) + H_T(\theta^*) \cdot (\hat{\theta}_T - \theta_0),$$

where $\theta^* \equiv \theta(c^*)$ and:

$$H_T(\theta) = \frac{1}{T} \sum_{t=1}^T H(\theta|y_t).$$

The first order conditions tell us that $s_T(\hat{\theta}_T) = 0$. Hence,

$$0 = s_T(\theta_0) + H_T(\theta^*) \cdot (\hat{\theta}_T - \theta_0).$$

We also have that:

$$|H_T(\theta^*) - H_T(\theta_0)| \leq \frac{1}{T} \sum_{t=1}^T |H(\theta^*|y_t) - H(\theta_0|y_t)| \leq \sup |H(\theta^*) - H(\theta_0)|, \quad (5A.1)$$

where the supremum is taken over the set of all the observations. Since $\hat{\theta}_T \xrightarrow{a.s.} \theta_0$, we also have that $\theta_T^* \xrightarrow{a.s.} \theta_0$. Moreover, by the law of large numbers,

$$H_T(\theta_0) = \frac{1}{T} \sum_{t=1}^T H(\theta_0|y_t) \xrightarrow{p} E[H(\theta_0|y_t)] = -\mathcal{J}(\theta_0). \quad (5A.2)$$

Since H is continuous in θ uniformly in y , the inequality in (5A.1), and (5A.2) both imply that:

$$H_T(\theta_T^*) \xrightarrow{a.s.} -\mathcal{J}(\theta_0).$$

Therefore, as $T \rightarrow \infty$,

$$\sqrt{T}(\hat{\theta}_T - \theta_0) = -H_T^{-1}(\theta_0) \cdot s_T(\theta_0)\sqrt{T} = \mathcal{J}^{-1} \cdot \sqrt{T}s_T(\theta_0).$$

By the central limit theorem, and $E(s_T) = 0$, the score, $s_T(\theta_0) = \frac{1}{T} \sum_{t=1}^T s(\theta_0, y_t)$, is such that

$$\sqrt{T} \cdot s_T(\theta_0) \xrightarrow{d} N(0, \text{var}(s(\theta_0, y_t))),$$

where

$$\text{var}(s(\theta_0, y_t)) = \mathcal{J}.$$

The result follows by the Slutsky's theorem and the symmetry of \mathcal{J} .

Finally, one should show the existence of a sequence $\hat{\theta}_T$ converging a.s. to θ_0 . Proofs on this type of convergence can be found in Amemiya (1985), or in Newey and McFadden (1994).

5.12 Appendix 4: Dependent processes

5.12.1 Weak dependence

Let $\sigma_T^2 = \text{var}(\sum_{t=1}^T x_t)$, and assume that $\sigma_T^2 = O(T)$, and that $\sigma_T^2 = O(T^{-1})$. If

$$\sigma_T^{-1} \sum_{t=1}^T (x_t - E(x_t)) \xrightarrow{d} N(0, 1),$$

we say that $\{x_t\}$ is *weakly dependent*. Of a process, we say it is “nonergodic,” when it exhibits such a strong dependence that it does not even satisfy the law of large numbers.

- Stationarity
- Weak dependence
- Ergodicity

5.12.2 The central limit theorem for martingale differences

Let x_t be a martingale difference sequence with $E(x_t^2) = \sigma_t^2 < \infty$ for all t , and define $\bar{x}_T \equiv \frac{1}{T} \sum_{t=1}^T x_t$, and $\bar{\sigma}_T^2 \equiv \frac{1}{T} \sum_{t=1}^T \sigma_t^2$. Let,

$$\forall \epsilon > 0, \lim_{T \rightarrow \infty} \frac{1}{T \bar{\sigma}_T^2} \sum_{t=1}^T x_t^2 \mathbb{I}_{|x_t| \geq \epsilon T \cdot \bar{\sigma}_T} = 0, \quad \text{and} \quad \frac{1}{T} \sum_{t=1}^T x_t^2 - \bar{\sigma}_T^2 \xrightarrow{p} 0.$$

Under the previous condition,

$$\frac{\sqrt{T} \cdot \bar{x}_T}{\bar{\sigma}_T} \xrightarrow{d} N(0, 1).$$

5.12.3 Applications to maximum likelihood

We use the central limit theorem for martingale differences to prove asymptotic normality of the MLE, in the case of weakly dependent processes. We have,

$$\ln L_T(\theta) = \sum_{t=1}^T \ell_t(\theta), \quad \ell_t(\theta) \equiv \ell(\theta; y_t | x_t).$$

The MLE satisfies the following first order conditions,

$$\mathbf{0}_p = \nabla_{\theta} \ln L_T(\theta)|_{\theta=\hat{\theta}_T} \stackrel{d}{=} \sum_{t=1}^T \nabla_{\theta} \ell_t(\theta)|_{\theta=\theta_0} + \sum_{t=1}^T \nabla_{\theta\theta} \ell_t(\theta)|_{\theta=\theta_0} (\hat{\theta}_T - \theta_0),$$

whence

$$\sqrt{T}(\hat{\theta}_T - \theta_0) \stackrel{d}{=} - \left[\frac{1}{T} \sum_{t=1}^T \nabla_{\theta\theta} \ell_t(\theta_0) \right]^{-1} \frac{1}{\sqrt{T}} \sum_{t=1}^T \nabla_{\theta} \ell_t(\theta_0). \quad (5A.3)$$

We have:

$$E_{\theta_0} [\nabla_{\theta} \ell_{t+1}(\theta_0) | F_t] = \mathbf{0}_p,$$

which shows that $\frac{\partial \ell_t(\theta_0)}{\partial \theta}$ is a martingale difference. Naturally, here we also have that:

$$E_{\theta_0} (|\nabla_{\theta} \ell_{t+1}(\theta_0)|_2 | F_t) = -E_{\theta_0} (\nabla_{\theta\theta} \ell_{t+1}(\theta_0) | F_t) \equiv \mathcal{J}_t(\theta_0).$$

Next, for a given constant $c \in \mathbb{R}^p$, let:

$$x_t \equiv c^\top \nabla_{\theta} \ell_t(\theta_0).$$

Clearly, x_t is also a martingale difference. Furthermore,

$$E_{\theta_0}(x_{t+1}^2 | F_t) = -c^\top \mathcal{J}_t(\theta_0) c,$$

and because x_t is a martingale difference, $E(x_t x_{t-i}) = E[E(x_t \cdot x_{t-i} | F_{t-i})] = E[E(x_t | F_{t-i}) \cdot x_{t-i}] = 0$, for all i . That is, x_t and x_{t-i} are mutually uncorrelated. It follows that,

$$\begin{aligned} \text{var} \left(\sum_{t=1}^T x_t \right) &= \sum_{t=1}^T E(x_t^2) \\ &= \sum_{t=1}^T c^\top E_{\theta_0} (|\nabla_{\theta} \ell_t(\theta_0)|_2) c \\ &= \sum_{t=1}^T c^\top E_{\theta_0} [E_{\theta_0} (|\nabla_{\theta} \ell_t(\theta_0)|_2 | F_{t-1})] c \\ &= - \sum_{t=1}^T c^\top E_{\theta_0} [\mathcal{J}_{t-1}(\theta_0)] c \\ &= -c^\top \left[\sum_{t=1}^T E_{\theta_0} (\mathcal{J}_{t-1}(\theta_0)) \right] c. \end{aligned}$$

Next, define:

$$\bar{x}_T \equiv \frac{1}{T} \sum_{t=1}^T x_t \quad \text{and} \quad \bar{\sigma}_T^2 \equiv \frac{1}{T} \sum_{t=1}^T E(x_t^2) = -c^\top \left[\frac{1}{T} \sum_{t=1}^T E_{\theta_0} (\mathcal{J}_{t-1}(\theta_0)) \right] c.$$

Under the conditions underlying the central limit theorem for weakly dependent processes provided earlier, to be spelled out below,

$$\frac{\sqrt{T} \bar{x}_T}{\bar{\sigma}_T} \xrightarrow{d} N(0, 1).$$

By the Cramer-Wold device,

$$\left[\frac{1}{T} \sum_{t=1}^T E_{\theta_0} (\mathcal{J}_{t-1}(\theta_0)) \right]^{-1/2} \frac{1}{\sqrt{T}} \sum_{t=1}^T \nabla_{\theta} \ell_t(\theta_0) \xrightarrow{d} N(0, \mathbf{I}_p).$$

The conditions that need to be satisfied are,

$$\frac{1}{T} \sum_{t=1}^T \nabla_{\theta\theta} \ell_t(\theta_0) - \frac{1}{T} \sum_{t=1}^T E_{\theta_0} [\mathcal{J}_{t-1}(\theta_0)] \xrightarrow{p} 0, \quad \text{and} \quad \text{plim} \frac{1}{T} \sum_{t=1}^T E_{\theta_0} [\mathcal{J}_{t-1}(\theta_0)] \equiv \mathcal{J}_{\infty}(\theta_0).$$

Under the previous conditions, it follows from Eq. (5A.3) that,

$$\sqrt{T}(\hat{\theta}_T - \theta_0) \xrightarrow{d} N(\mathbf{0}_p, \mathcal{J}_{\infty}(\theta_0)^{-1}).$$

5.13 Appendix 5: Proof of Theorem 5.4

Let $\pi_t \equiv \pi_t(\phi(y(t+1), \mathbf{M} - (t+1)\mathbf{1}_{d-q}) | \phi(y(t), \mathbf{M} - t\mathbf{1}_{d-q}))$ denote the transition density of

$$\phi(y(t), \mathbf{M} - t\mathbf{1}_{d-q}) \equiv \phi(y(t)) \equiv (y^o(t), c(y(t), M_1 - t), \dots, c(y(t), M_{d-q} - t)),$$

where we have emphasized the dependence of ϕ on the time-to-maturity vector:

$$\mathbf{M} - t\mathbf{1}_{d-q} \equiv (M_1 - t, \dots, M_{d-q} - t).$$

By $\Sigma(\tau)$ full rank $P \otimes d\tau$ -a.s., and Itô's lemma, ϕ satisfies, for $\tau \in [t, t+1]$,

$$\begin{cases} dy^o(\tau) &= b^o(\tau)d\tau + F(\tau)\Sigma(\tau)dW(\tau) \\ dc(\tau) &= b^c(\tau)d\tau + \nabla c(\tau)\Sigma(\tau)dW(\tau) \end{cases}$$

where b^o and b^c are, respectively, q -dimensional and $(d-q)$ -dimensional measurable functions, and $F(\tau) \equiv \bar{a}(\tau) \cdot \Sigma(\tau)^{-1} P \otimes d\tau$ -a.s. Under condition (5.28), π_t is not degenerate. Furthermore, $C(y(t); \ell) \equiv C(t)$ is deterministic in $\ell \equiv (\ell_1, \dots, \ell_{d-q})$. That is, for all $(\bar{c}, \bar{c}^+) \in \mathbb{R}^d \times \mathbb{R}^d$, there exists a function μ such that for any neighbourhood $N(\bar{c}^+)$ of \bar{c}^+ , there exists another neighborhood $N(\mu(\bar{c}^+))$ of $\mu(\bar{c}^+)$ such that,

$$\begin{aligned} & \{\omega \in \Omega : \phi(y(t+1), \mathbf{M} - (t+1)\mathbf{1}_{d-q}) \in N(\bar{c}^+) \mid \phi(y(t), \mathbf{M} - t\mathbf{1}_{d-q}) = \bar{c}\} \\ = & \{\omega \in \Omega : (y^o(t+1), c(y(t+1), M_1 - t), \dots, c(y(t+1), M_{d-q} - t)) \in N(\mu(\bar{c}^+)) \\ & \quad \mid \phi(y(t), \mathbf{M} - t\mathbf{1}_{d-q}) = \bar{c}\} \\ = & \{\omega \in \Omega : (y^o(t+1), c(y(t+1), M_1 - t), \dots, c(y(t+1), M_{d-q} - t)) \in N(\mu(\bar{c}^+)) \\ & \quad \mid (y^o(t), c(y(t), M_1 - t), \dots, c(y(t), M_{d-q} - t)) = \bar{c}\} \end{aligned}$$

where the last equality follows by the definition of ϕ . In particular, the transition laws of ϕ_t^c given ϕ_{t-1}^c are not degenerate; and ϕ_t^c is stationary. The feasibility of simulation based method of moments estimation is proved. The efficiency claim follows by the Markov property of ϕ , and the usual score martingale difference argument. ■

References

- Altissimo, F. and A. Mele (2009): “Simulated Nonparametric Estimation of Dynamic Models.” *Review of Economic Studies* 76, 413-450.
- Amemiya, T. (1985): *Advanced Econometrics*. Cambridge, Mass.: Harvard University Press.
- Bergman, Y. Z., B. D. Grundy, and Z. Wiener (1996): “General Properties of Option Prices.” *Journal of Finance* 51, 1573-1610.
- Brandt, M. and P. Santa-Clara (2002): “Simulated Likelihood Estimation of Diffusions with an Applications to Exchange Rate Dynamics in Incomplete Markets.” *Journal of Financial Economics* 63, 161-210.
- Carrasco, M., M. Chernov, J.-P. Florens and E. Ghysels (2007): “Efficient Estimation of General Dynamic Models with a Continuum of Moment Conditions.” *Journal of Econometrics* 140, 529-573.
- Chernov, M. and E. Ghysels (2000): “A Study towards a Unified Approach to the Joint Estimation of Objective and Risk-Neutral Measures for the Purpose of Options Valuation.” *Journal of Financial Economics* 56, 407-458.
- Christensen, B. J. (1992): “Asset Prices and the Empirical Martingale Model.” Working paper, New York University.
- Duffie, D. and K. J. Singleton (1993): “Simulated Moments Estimation of Markov Models of Asset Prices.” *Econometrica* 61, 929-952.
- Fermanian, J.-D. and B. Salanié (2004): “A Nonparametric Simulated Maximum Likelihood Estimation Method.” *Econometric Theory* 20, 701-734.
- Fisher, R. A. (1912): “On an Absolute Criterion for Fitting Frequency Curves.” *Messages of Mathematics* 41, 155-157.
- Gallant, A. R. and G. Tauchen (1996): “Which Moments to Match?” *Econometric Theory* 12, 657-681.
- Gauss, C. F. (1816): “Bestimmung der Genauigkeit der Beobachtungen.” *Zeitschrift für Astronomie und Verwandte Wissenschaften* 1, 185-196.
- Gouriéroux, C., A. Monfort and E. Renault (1993): “Indirect Inference.” *Journal of Applied Econometrics* 8, S85-S118.
- Hajivassiliou, V. and D. McFadden (1998): “The Method of Simulated Scores for the Estimation of Limited-Dependent Variable Models.” *Econometrica* 66, 863-896.
- Hansen, L. P. (1982): “Large Sample Properties of Generalized Method of Moments Estimators.” *Econometrica* 50, 1029-1054.
- Hansen, L. P. and K. J. Singleton (1982): “Generalized Instrumental Variables Estimation of Nonlinear Rational Expectations Models.” *Econometrica* 50, 1269-1286.

- Hansen, L. P. and K. J. Singleton (1983): “Stochastic Consumption, Risk Aversion, and the Temporal Behavior of Asset Returns.” *Journal of Political Economy* 91, 249-265.
- Harrison, J. M. and S. R. Pliska (1983): “A Stochastic Calculus Model of Continuous Trading: Complete Markets.” *Stochastic Processes and their Applications* 15, 313-316.
- Heston, S. (1993): “A Closed-Form Solution for Options with Stochastic Volatility with Applications to Bond and Currency Options.” *Review of Financial Studies* 6, 327-343.
- Laroque, G. and B. Salanié (1989): “Estimation of Multimarket Fix-Price Models: An Application of Pseudo-Maximum Likelihood Methods.” *Econometrica* 57, 831-860.
- Laroque, G. and B. Salanié (1993): “Simulation-Based Estimation of Models with Lagged Latent Variables.” *Journal of Applied Econometrics* 8, S119-S133.
- Laroque, G. and B. Salanié (1994): “Estimating the Canonical Disequilibrium Model: Asymptotic Theory and Finite Sample Properties.” *Journal of Econometrics* 62, 165-210.
- Lee, B-S. and B. F. Ingram (1991): “Simulation Estimation of Time-Series Models.” *Journal of Econometrics* 47, 197-207.
- Lee, L. F. (1995): “Asymptotic Bias in Simulated Maximum Likelihood Estimation of Discrete Choice Models.” *Econometric Theory* 11, 437-483.
- McFadden, D. (1989): “A Method of Simulated Moments for Estimation of Discrete Response Models without Numerical Integration.” *Econometrica* 57, 995-1026.
- Mele, A. (2003): “Fundamental Properties of Bond Prices in Models of the Short-Term Rate.” *Review of Financial Studies* 16, 679-716.
- Newey, W. K. and D. L. McFadden (1994): “Large Sample Estimation and Hypothesis Testing.” In: Engle, R. F. and D. L. McFadden (Editors): *Handbook of Econometrics*, Vol. 4, Chapter 36, 2111-2245. Amsterdam: Elsevier.
- Neyman, J. and E. S. Pearson (1928): “On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference.” *Biometrika* 20A, 175-240, 263-294.
- Pakes, A. and D. Pollard (1989): “Simulation and the Asymptotics of Optimization Estimators.” *Econometrica* 57, 1027-1057.
- Pastorello, S., E. Renault and N. Touzi (2000): “Statistical Inference for Random-Variance Option Pricing.” *Journal of Business and Economic Statistics* 18, 358-367.
- Pastorello, S., V. Patilea, and E. Renault (2003): “Iterative and Recursive Estimation in Structural Non Adaptive Models.” *Journal of Business and Economic Statistics* 21, 449-509.
- Pearson, K. (1894): “Contributions to the Mathematical Theory of Evolution.” *Philosophical Transactions of the Royal Society of London*, Series A 185, 71-78.
- Romano, M. and N. Touzi (1997): “Contingent Claims and Market Completeness in a Stochastic Volatility Model.” *Mathematical Finance* 7, 399-412.

-
- Santa-Clara, P. (1995): “Simulated Likelihood Estimation of Diffusions With an Application to the Short Term Interest Rate.” Ph.D. dissertation, INSEAD.
- Singleton, K. J. (2001): “Estimation of Affine Asset Pricing Models Using the Empirical Characteristic Function.” *Journal of Econometrics* 102, 111-141.
- Smith, A. (1993): “Estimating Nonlinear Time Series Models Using Simulated Vector Autoregressions.” *Journal of Applied Econometrics* 8, S63-S84.
- Tauchen, G. (1997): “New Minimum Chi-Square Methods in Empirical Finance.” In D. Kreps and K. Wallis (Editors): *Advances in Econometrics*, 7th World Congress, Econometrics Society Monographs, Vol. III. Cambridge UK: Cambridge University Press, 279-317.

Part II

Applied asset pricing theory

6

Neo-classical kernels and puzzles

6.1 Introduction

This chapter discusses methods of statistical inference about the salient implications of asset pricing models. These models generate a quite high number of potential restrictions on security returns. A key observation is that the single engine of these restrictions is the pricing kernel. A data-reduction device to test for these restrictions relies on the assessment of the quantitative restrictions any pricing kernel should satisfy, to be consistent with the data. This approach leads to the celebrated Hansen and Jagannathan (1991) bounds to the pricing kernel, which we illustrate through the simplest version of the Lucas' tree model, and some of its variants. There are issues about this methodology, arising due to the finite sample performance of the bounds, which have been tackled over the course of research, and which we shall survey. Nevertheless, there are no models that might ever escape from a level of scrutiny, such as that relating to the bounds we illustrate in this chapter.

The next section develops a simple model, which we use as a benchmark to explain the equity premium puzzle. Section 6.3 develops the central tool of analysis, a non-parametric bound on the conditional variance of the pricing kernels (i.e. the risk-premium), for a given level of the short-term interest rate. Section 6.4 considers multifactor extensions, with closed-form solutions, arising under a number of analytically convenient assumptions made on the stochastic discounting factor. One of the striking points of Section 6.4 is that in spite of these added dimensionalities, the resulting models might spectacularly fail explain the dynamics of asset prices—pumping up volatility is not enough, if this added volatility is not accompanied by time-varying countercyclical statistics. Section 6.5 develops a link between stochastic discount factors and Sharpe ratios, and Section 6.6 develops dynamic versions of the core bounds at the heart of this chapter.

6.2 The equity premium puzzle

6.2.1 A single-factor model

We consider an economy with a single agent endowed with a CRRA utility, $u(x) = x^{1-\eta}/(1-\eta)$, and a constant discount factor β . We assume cum-dividends gross returns, $(S_t + D_t)/S_{t-1}$, are generated by the following model:

$$\begin{cases} \ln(S_t + D_t) &= \ln S_{t-1} + \mu_S - \frac{1}{2}\sigma_S^2 + \epsilon_{S,t} \\ \ln D_t &= \ln D_{t-1} + \mu_D - \frac{1}{2}\sigma_D^2 + \epsilon_{D,t} \end{cases} \quad (6.1)$$

where

$$\begin{bmatrix} \epsilon_{S,t} \\ \epsilon_{D,t} \end{bmatrix} \sim NID \left(\mathbf{0}_2; \begin{bmatrix} \sigma_S^2 & \sigma_{SD} \\ \sigma_{SD} & \sigma_D^2 \end{bmatrix} \right).$$

Naturally, the coefficients μ_S , μ_D , σ_S^2 , σ_D^2 and σ_{SD} need to satisfy restrictions compatible with an optimizing behavior of the agent, and an equilibrium. The key intertemporal restriction applying to the asset price is the standard Euler equation, which is, by results in Part I of these lectures,

$$1 = E(m_{t+1} \cdot e^{Q_{t+1}} | F_t), \quad Q_{t+1} = \ln \left(\frac{S_{t+1} + D_{t+1}}{S_t} \right), \quad m_{t+1} \equiv e^{Z_{t+1}} \equiv \beta \left(\frac{D_{t+1}}{D_t} \right)^{-\eta}, \quad (6.2)$$

where F_t is the information set as of time t , and m_{t+1} is the stochastic discounting factor. Naturally, Eq. (6.2) holds for any asset. In particular, it holds for a one-period bond with price $S_t^b \equiv b_t$, $S_{t+1}^b \equiv 1$ and $D_{t+1}^b \equiv 0$. Define, then, $Q_{t+1}^b \equiv \ln(b_t^{-1}) \equiv \ln R_t$. By replacing R_t into Eq. (6.2), one gets $R_t^{-1} = E(e^{Z_{t+1}} | F_t)$, such that we are left with the following system:

$$\frac{1}{R_t} = E(m_{t+1} | F_t), \quad 1 = E(m_{t+1} \cdot e^{Q_{t+1}} | F_t). \quad (6.3)$$

The following result helps solve analytically the two equations (6.3).

LEMMA 6.1: *Let Z be conditionally normally distributed. Then, for any $\gamma \in \mathbb{R}$,*

$$\begin{aligned} E(e^{-\gamma Z_{t+1}} | F_t) &= e^{-\gamma E(Z_{t+1}|F_t) + \frac{1}{2}\gamma^2 \text{var}(Z_{t+1}|F_t)} \\ \sqrt{\text{var}(e^{-\gamma Z_{t+1}} | F_t)} &= e^{-\gamma E(Z_{t+1}|F_t) + \gamma^2 \text{var}(Z_{t+1}|F_t)} \sqrt{1 - e^{-\gamma^2 \text{var}(Z_{t+1}|F_t)}} \end{aligned}$$

By the definition of Z , Eq. (6.1), and Lemma 6.1,

$$\frac{1}{R_t} = E(e^{Z_{t+1}} | F_t) = e^{E(Z_{t+1}|F_t) + \frac{1}{2}\text{var}(Z_{t+1}|F_t)} = e^{\ln \beta - \eta(\mu_D - \frac{1}{2}\sigma_D^2) + \frac{1}{2}\eta^2 \sigma_D^2}.$$

Therefore, the equilibrium interest rate is constant, and its expression is given in the second of Eqs. (6.4) below.

The second of equations (6.3) can be written as,

$$1 = E(\exp(Z_{t+1} + Q_{t+1}) | F_t) = e^{\ln \beta - \eta(\mu_D - \frac{1}{2}\sigma_D^2) + \mu_S - \frac{1}{2}\sigma_S^2} \cdot E(e^{\tilde{n}_{t+1}} | F_t),$$

where $\tilde{n}_{t+1} \equiv \epsilon_{S,t+1} - \eta\epsilon_{D,t+1} \sim N(0, \sigma_S^2 + \eta^2\sigma_D^2 - 2\eta\sigma_{SD})$. The expectation in the above equation can be computed using Lemma 6.1. The result is,

$$0 = \underbrace{\ln \beta - \eta\mu_D + \frac{1}{2}\eta(\eta+1)\sigma_D^2}_{-\ln R_t} + \mu_S - \eta\sigma_{SD}.$$

By defining $R_t \equiv e^{r_t}$, and rearranging terms,

$$\underbrace{\mu_S - r}_{\text{risk premium}} = \eta\sigma_{SD}.$$

To sum up,

$$\mu_S = r + \eta\sigma_{SD}, \quad r = -\ln \beta + \eta\mu_D - \frac{1}{2}\eta(\eta+1)\sigma_D^2, \quad (6.4)$$

and the expected gross return on the risky asset is,

$$E \left[\frac{S_{t+1} + D_{t+1}}{S_t} \middle| F_t \right] = e^{\mu_S - \frac{1}{2}\sigma_S^2} \cdot E [e^{\epsilon_{S,t+1}} | F_t] = e^{\mu_S} = e^{r + \eta\sigma_{SD}}.$$

Therefore, if $\sigma_{SD} > 0$, then $E((S_{t+1} + D_{t+1})/S_t | F_t) > E(b_t^{-1} | F_t)$, as expected.

The expressions for the equity premium and the short-term rate are the discrete-time counterpart of those derived in Chapter 4. Consider, for example, the interest rate. The second term, $\eta\mu_D$, reflects “intertemporal substitution” effects: consumption endowment increases, on average, as μ_D increases, which reduces the demand for bonds, thereby increasing the interest rate. The last term, instead, relates to “precautionary” motives: an increase in the uncertainty related to consumption endowment, σ_D , raises concerns with the our representative agent, who then increases his demand for bonds, thereby leading to a drop in the equilibrium interest rate.

We finally check the internal consistency of the model. The coefficients of the model satisfy some restrictions. In particular, the asset price volatility must be determined endogeneously. Let us conjecture, first, that the following “no-sunspots” condition holds, for each period t :

$$\epsilon_{S,t} = \epsilon_{D,t}. \quad (6.5)$$

Below, we shall show this condition does hold. By Eq. (6.5), then,

$$\mu_S - r = \lambda\sigma_D, \quad \lambda \equiv \eta\sigma_D, \quad (6.6)$$

and

$$Z_{t+1} = - \left(r + \frac{1}{2}\lambda^2 \right) - \lambda u_{D,t+1}, \quad u_{D,t+1} \equiv \frac{\epsilon_{D,t+1}}{\sigma_D},$$

such that we can define the pricing kernel ξ , from the stochastic discounting factor, recursively, as follows:

$$\frac{\xi_{t+1}}{\xi_t} = m_{t+1} \equiv e^{Z_{t+1}}, \quad \xi_0 = 1.$$

It is the discrete-time counterpart to the continuous-time representation of the Arrow-Debreu state price density given in Chapter 4.

Let us iterate the asset price equation (6.2),

$$\begin{aligned} S_t &= E \left[\left(\prod_{j=1}^n e^{Z_{t+j}} \right) \cdot S_{t+n} \middle| F_t \right] + \sum_{i=1}^n E \left[\left(\prod_{j=1}^i e^{Z_{t+j}} \right) \cdot D_{t+i} \middle| F_t \right] \\ &= E \left(\frac{\xi_{t+n}}{\xi_t} \cdot S_{t+n} \middle| F_t \right) + \sum_{i=1}^n E \left(\frac{\xi_{t+i}}{\xi_t} \cdot D_{t+i} \middle| F_t \right). \end{aligned}$$

By letting $n \rightarrow \infty$ and assuming the first term in the previous equation goes to zero,

$$S_t = \sum_{i=1}^{\infty} E \left(\frac{\xi_{t+i}}{\xi_t} \cdot D_{t+i} \middle| F_t \right). \quad (6.7)$$

Eq. (6.7) holds, as just mentioned, under a transversality condition, similar to that analyzed in Chapter 4, Section 4.3.3, which always holds, under the inequalities given in (6.8) below.

The expectations in Eq. (6.7) are, by Lemma 6.1,

$$E \left(\frac{\xi_{t+i}}{\xi_t} \cdot D_{t+i} \middle| F_t \right) = E \left(e^{\sum_{j=1}^i Z_{t+j}} \cdot D_{t+i} \middle| F_t \right) = D_t e^{(\mu_D - r - \sigma_D \lambda)i}.$$

Suppose the “risk-adjusted” discount rate $r + \sigma_D \lambda$ is higher than the growth rate of the economy, viz

$$r + \sigma_D \lambda > \mu_D \Leftrightarrow k \equiv e^{\mu_D - r - \sigma_D \lambda} < 1. \quad (6.8)$$

Under this condition, the summation in Eq. (6.7) converges, leaving:

$$\frac{S_t}{D_t} = \frac{k}{1 - k}. \quad (6.9)$$

This pricing equation relates to the celebrated Gordon’s formula (Gordon, 1962). The price-dividend ratio increases with the expected dividend growth, μ_D , and decreases with the (risk-adjusted) discount rate, $r + \sigma_D \lambda$. It predicts that price-dividend ratios are constant, a counterfactual prediction addressed in the next two chapters. Finally, the solution for the price-dividend ratio in Eq. (6.9) is, of course, consistent with that of the Lucas model in Chapter 3, Section 3.2.4, as shown below.

We now check that the no-sunspots condition in Eq. (6.5) holds, and derive the variance of the asset price. Note then, that Eq. (6.9) and the second equation in (6.1) imply that:

$$\ln(S_t + D_t) - \ln S_{t-1} = -\ln k + \mu_D - \frac{1}{2} \sigma_D^2 + \epsilon_{D,t}.$$

By the first equation in (6.1),

$$\mu_S - \frac{1}{2} \sigma_S^2 = \mu_D - \frac{1}{2} \sigma_D^2 - \ln k, \quad \epsilon_{S,t} = \epsilon_{D,t}, \text{ for each } t. \quad (6.10)$$

The second condition confirms the no-sunspots condition in Eq. (6.5) holds. It also informs us that, $\sigma_S^2 = \sigma_{SD} = \sigma_D^2$. By replacing this into the first condition, delivers back $\mu_S = \mu_D - \ln k = r + \sigma_D \lambda$.

Note, finally, that by replacing the expression for the interest rate in the second of Eqs. (6.4) and the equity premium in Eq. (6.6) into Eq. (6.8), the constant k simplifies to $k \equiv \beta \exp^{(\eta-1)(-\mu_D + \frac{1}{2} \eta \sigma_D^2)}$, such that the price-dividend ratio in the log-utility, $\eta = 1$, collapses to $\beta / (1 - \beta)$, as established in Chapter 3. This section provides a solution to the general CRRA case, under the additional assumption that dividends are normally distributed.

6.2.2 Extensions

Chapter 3 shows that within a IID environment, prices are convex (resp. concave) in the dividend rate whenever $\eta > 1$ (resp. $\eta < 1$). The pricing formula in Eq. (6.9) reveals that this property may be lost in a dynamic environment. In Eq. (6.9), prices are always linear in the dividends' rate. By using techniques developed in the next chapter, we may show that within a dynamic context, convexity properties of the price function are inherited by those of the dividend process in the following sense: if the expected dividend growth under the risk-neutral measure is a convex (resp. concave) function of the initial dividend rate, then prices are convex (resp. concave) in the initial dividend rate. In the model analyzed here, the expected dividend growth under the risk-neutral measure is linear in the dividends' rate, which explains the linear property in Eq. (6.9).

6.2.3 The puzzles

“Average excess returns on the US stock market [the equity premium] is too high to be easily explained by standard asset pricing models.” Mehra and Prescott (1985)

Mehra and Prescott (1985) noted the following difficulty with the Lucas model, which gave rise to what is widely known as the *equity premium puzzle*. To be consistent with US data, the equity premium in Eq. (6.6), $\mu_S - r = \lambda\sigma_D^2$, must be an approximate 6% annualized, as explained in the next chapter. If the asset we are trying to price is literally a *consumption claim*, then, σ_D would be consumption volatility, which is very low, approximately 3.3%. For the equity premium to be high, we would need, by reverse-engineering the two equations in (6.6), a quite high value of the relative risk-aversion, say $\eta \approx \frac{0.06}{0.033^2} \approx 55$. Section 6.4 explains this number, 55, can be slightly improved to 35, once we also condition on the volatility of short-term bonds.

One assumption underlying the previous calculations is that the aggregate dividend equals aggregate consumption, which is obviously not the case, in the real world. Note, then, that dividend growth volatility is around 6%, which implies the implied η is $17 \approx \frac{0.06}{0.06^2}$, thereby mitigating the premium puzzle. Still, the model would fail deliver realistic predictions about return volatility, as in this case, return volatility would be just 6%, by Eq. (6.10), which is less than a half of what we see in the data, as explained in the next chapter. Moreover, the model would fail predict countercyclical statistics, such as countercyclical expected returns or dividend yields. We shall return to these topics in the next chapter.

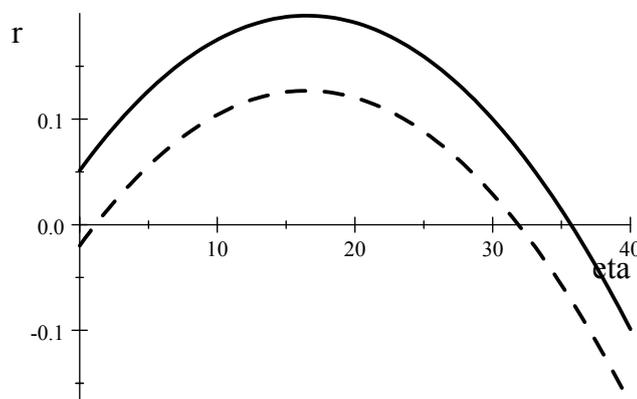
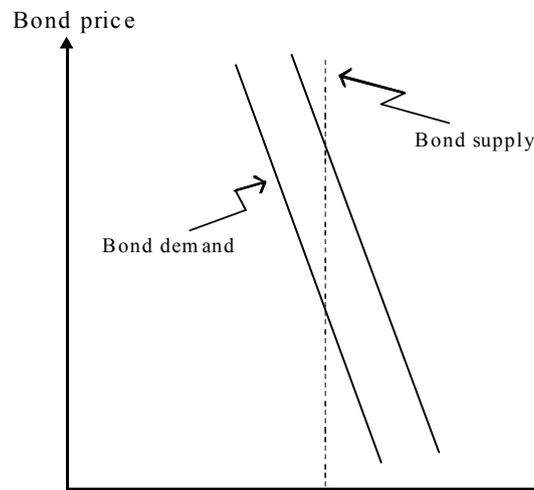


FIGURE 6.1. The risk-free rate puzzle: the two curves depict the graph $\eta \mapsto r(\eta) = -\ln \beta + 0.0183 \cdot \eta - \frac{1}{2} (0.0328)^2 \cdot \eta (\eta + 1)$, with $\beta = 0.95$ (solid line) and $\beta = 1.05$ (dashed line). Even if risk aversion were to be as high as $\eta = 30$, the equilibrium short-term rate would behave counterfactually, reaching a level as high as 10%. In order for r to be lower when η is high, it might be required that $\beta > 1$.

The equity premium is not the only puzzle. Even if we are willing to consider that a CRRA as large as $\eta = 30$ is plausible, another puzzle arises—an interest rate puzzle. As the expression for r in equations (6.4) shows, a large value of η can lead the interest rate to take very high values, as illustrated by Figure 6.1. Finally, related to the interest rate puzzle is an interest rate volatility puzzle. In the model of this chapter, the safe rate is constant. However, in models where both the equity premium and interest rates change over time, driven by state variables related to, say, preference shocks or market imperfections, the short-term rate is too volatile. For example, in the presence of time-varying expected dividend growth, the expression for the short-term rate is the same as in Eq. (6.4), but with $\mu_{D,t}$ replacing the constant μ_D , as explained in the next chapters. It is easily seen, then, that the interest rate is quite volatile for high values of η .

This interest rate volatility puzzle relates to the assumption of a representative agent. Chapter 3 (Section 3.2.3) explains that agents with low elasticity of intertemporal substitution (EIS) have an inelastic demands for bonds. In the context of CRRA utility functions, a low EIS corresponds to a high CRRA, as $\text{EIS} = \frac{1}{\eta}$, as explained in Chapter 3. So now suppose there is a wide-economy shock that shifts the demand for bonds, as in the following picture—for example, a shock that makes $\mu_{D,t}$ change.



An economy with a representative agent is one where the supply of bonds is fixed. The combination of a representative agent with a low EIS, then, implies a high volatility of the short-term rate, which is counterfactual. To mitigate this issue, one may consider preferences that disentangle the EIS from risk-aversion, as such as those relying on non-expected utility (Epstein and Zin, 1989, 1991; Weil, 1989), or a framework with multiple agents, where bond supply is positively sloped, as in the limited participation model of Guvenen (2009). These models are examined in Chapter 8.

6.3 Hansen-Jagannathan cup

Suppose there are n risky assets. The n asset pricing equations for these assets are,

$$1 = E[m_{t+1}(1 + R_{j,t+1}) | F_t], \quad j = 1, \dots, n.$$

Assuming $R_{j,t+1}$ is stationary, and taking the unconditional expectation of both sides of the previous equation, leaves,

$$\mathbf{1}_n = E[m_t(\mathbf{1}_n + R_t)], \quad R_t = (R_{1,t}, \dots, R_{n,t})^\top.$$

Next, let $\bar{m} \equiv E(m_t)$, and create a family of stochastic discount factors m_t^* , parametrized by \bar{m} , by projecting m on to the asset returns, as follows:

$$\text{Proj}(m | \mathbf{1}_n + R_t) \equiv m_t^*(\bar{m}) = \bar{m} + [R_t - E(R_t)]_{1 \times n}^\top \beta_{\bar{m}},_{n \times 1}$$

where¹

$$\beta_{\bar{m}} = \frac{\sum_{n \times 1}^{-1} \text{cov}(m, \mathbf{1}_n + R_t)}{\sum_{n \times n}} = \Sigma^{-1} [\mathbf{1}_n - \bar{m}E(\mathbf{1}_n + R_t)],$$

and $\Sigma \equiv E[(R_t - E(R_t))(R_t - E(R_t))^\top]$. As shown in the Appendix, we also have that,

$$\mathbf{1}_n = E[m_t^*(\bar{m}) \cdot (\mathbf{1}_n + R_t)]. \quad (6.11)$$

We have,

$$\sqrt{\text{var}(m_t^*(\bar{m}))} = \sqrt{\beta_{\bar{m}}^\top \Sigma \beta_{\bar{m}}} = \sqrt{(\mathbf{1}_n - \bar{m}E(\mathbf{1}_n + R_t))^\top \Sigma^{-1} (\mathbf{1}_n - \bar{m}E(\mathbf{1}_n + R_t))}. \quad (6.12)$$

Eq. (6.12), provides the expression for the celebrated *Hansen-Jagannathan “cup”*—after the work of Hansen and Jagannathan (1991). It leads to an important tool of analysis, as the following theorem shows.

THEOREM 6.2: *Among all stochastic discount factors with fixed expectation \bar{m} , $m_t^*(\bar{m})$ is the one with the smallest variance.*

PROOF: Consider another discount factor indexed by \bar{m} , i.e. $m_t(\bar{m})$. Naturally, $m_t(\bar{m})$ satisfies $\mathbf{1}_n = E[m_t(\bar{m})(\mathbf{1}_n + R_t)]$. Moreover, by Eq. (6.11),

$$\begin{aligned} \mathbf{0}_n &= E[(m_t(\bar{m}) - m_t^*(\bar{m}))(\mathbf{1}_n + R_t)] \\ &= E[(m_t(\bar{m}) - m_t^*(\bar{m}))((\mathbf{1}_n + E(R_t)) + (R_t - E(R_t)))] \\ &= E[(m_t(\bar{m}) - m_t^*(\bar{m}))(R_t - E(R_t))] \\ &= \text{cov}[m_t(\bar{m}) - m_t^*(\bar{m}), R_t] \end{aligned}$$

where the third line follows because $E[m_t(\bar{m})] = E[m_t^*(\bar{m})] = \bar{m}$, and the fourth line holds by $E[(m_t(\bar{m}) - m_t^*(\bar{m}))] = 0$. But $m_t^*(\bar{m})$ is a linear combination of R_t . By the previous equation,

¹We have, $\text{cov}(m, \mathbf{1}_n + R_t) = E[m(\mathbf{1}_n + R)] - E(m)E(\mathbf{1}_n + R_t) = \mathbf{1}_n - \bar{m}E(\mathbf{1}_n + R_t)$.

then, it must be that $0 = \text{cov} [m_t(\bar{m}) - m_t^*(\bar{m}), m_t^*(\bar{m})]$. Therefore:

$$\begin{aligned} \text{var} [m_t(\bar{m})] &= \text{var} [m_t^*(\bar{m}) + m_t(\bar{m}) - m_t^*(\bar{m})] \\ &= \text{var} [m_t^*(\bar{m})] + \text{var} [m_t(\bar{m}) - m_t^*(\bar{m})] + 2 \cdot \text{cov} [m_t(\bar{m}) - m_t^*(\bar{m}), m_t^*(\bar{m})] \\ &= \text{var} [m_t^*(\bar{m})] + \text{var} [m_t(\bar{m}) - m_t^*(\bar{m})] \\ &\geq \text{var} [m_t^*(\bar{m})]. \end{aligned}$$

||

Hansen and Jagannathan (1991) consider an extension of this result, where the stochastic discount factor satisfies the non-negativity constraint, $m > 0$.

Consider, then, the space, $(\bar{m}, \text{var} [m_t^*(\bar{m})])$, and any model giving rise to a pair $(\bar{m}, \text{var} [m_t(\bar{m})])$. By Theorem 6.1, the pair $(\bar{m}, \text{var} [m_t(\bar{m})])$ has to lie above the cup $(\bar{m}, \text{var} [m_t^*(\bar{m})])$, for each possible \bar{m} . As an example, apply the Hansen-Jagannathan bounds in Eq. (6.12) to the neo-classical model of Section 6.2. The stochastic discount factor for this model is:

$$m_{t+1} = \frac{\xi_{t+1}}{\xi_t} = \exp(Z_{t+1}), \quad Z_{t+1} = -\left(r + \frac{1}{2}\lambda^2\right) - \lambda u_{D,t+1}, \quad u_{D,t+1} \equiv \frac{\epsilon_{D,t+1}}{\sigma_D}.$$

By Lemma 6.1, the first two moments of this stochastic discount factor are:

$$\bar{m} = E(m_t) = e^{-r} \quad \text{and} \quad \bar{\sigma}_m = \sqrt{\text{var}(m_t(\bar{m}))} = e^{-r + \frac{1}{2}\lambda^2} \sqrt{1 - e^{-\lambda^2}}, \quad (6.13)$$

where r is as in Eq. (6.4), and $\lambda = \eta\sigma_D$, as usual. For given μ_D and σ_D^2 , these two equations in (6.13) form a η -parametrized curve in the space $(\bar{m}, \bar{\sigma}_m)$. The issue is to check whether this curve enters the Hansen-Jagannathan cup for plausible values of η . It is not the case. Rather, we have the situation depicted in Figure 6.2. The reason the circles bend back is easily explained. When η is low, an increase in η leads to an increase in $\bar{\sigma}_m$ and a decrease in \bar{m} , because r increases with η when η is small. Yet as soon as η gets sufficiently large, the interest rate decreases, due to precautionary motives, which make both \bar{m} and $\bar{\sigma}_m$ increase. In general, we expect that for any model where the intertemporal elasticity of substitution is somehow confounded with relative risk-aversion, a bending pattern such that in Figure 6.2 would arise, being indicative of conflicts between intertemporal elasticity of substitution and precautionary effects relating to the demand of riskless assets.

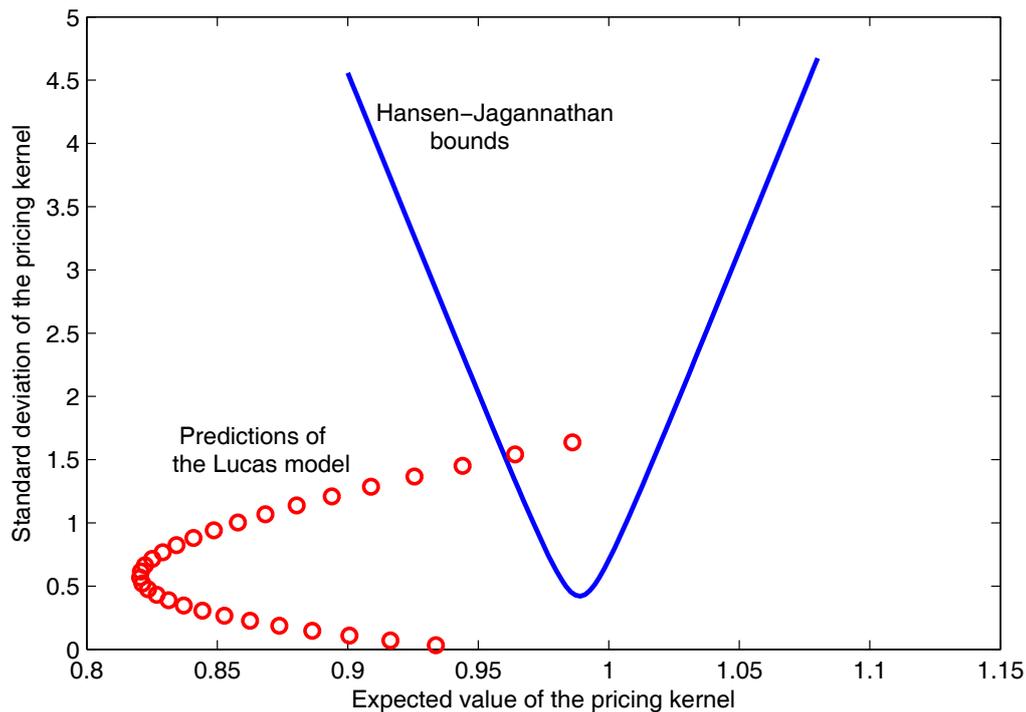


FIGURE 6.2. The solid line depicts the Hansen-Jagannathan bounds, obtained through Eq. (6.12), through aggregate stock market data and the short-term rate. The average return and standard deviation of the stock market are taken to be 0.07 and 0.14. The average short-term rate (three-month bill) and its volatility are, instead, 0.01 and 0.02. These estimates relate to the sample period from January 1948 to December 2002. The circles are predictions of the Lucas model in Eq. (6.13), with $\beta = 0.95$, $\mu_D = 0.0183$, $\sigma_D = 0.0328$ and η ranging from 1 to 35. The two circles inside the cup are the pairs $(\bar{m}, \bar{\sigma}_m)$ in Eq. (6.13) obtained with $\eta = 35$ and 33. Progressively lower values of η lead the pairs $(\bar{m}, \bar{\sigma}_m)$ to lie outside the cup, nonlinearly.

The Lucas model predicts that the pricing kernel is quite moderately volatile. The following chapters discuss models with both heterogeneous agents or more general preferences, which can help boost the volatility of the pricing kernel.

6.4 Multifactor extensions

A natural way to increase the variance of the pricing kernel is to increase the number of factors. We consider two possibilities: one in which returns are normally distributed, and one in which returns are lognormally distributed.

6.4.1 Exponential affine pricing kernels

Consider again the simple model in Section 6.2. In this section, we shall make a different assumption regarding the returns distributions. But we shall maintain the hypothesis that the

stochastic discount factor satisfies an exponential-Gaussian type structure,

$$m_{t+1} = \exp(Z_{t+1}), \quad Z_{t+1} = -\left(r + \frac{1}{2}\lambda^2\right) - \lambda u_{D,t+1}, \quad u_{D,t+1} \sim \text{NID}(0, 1),$$

where r and λ are some constants. We have,

$$1 = E(m_{t+1} \cdot \tilde{R}_{t+1}) = E(m_{t+1}) E(\tilde{R}_{t+1}) + \text{cov}(m_{t+1}, \tilde{R}_{t+1}), \quad \tilde{R}_{t+1} \equiv \frac{S_{t+1} + D_{t+1}}{S_t}.$$

By rearranging terms,² and using the fact that $E(m_{t+1}) = R^{-1}$,

$$E(\tilde{R}_{t+1}) - R = -R \cdot \text{cov}(m_{t+1}, \tilde{R}_{t+1}). \quad (6.14)$$

Consider the following result, which we shall use later:

LEMMA 6.3 (Stein's lemma): *Suppose that two random variables x and y are jointly normal. Then,*

$$\text{cov}[g(x), y] = E[g'(x)] \cdot \text{cov}(x, y),$$

for any function $g : E(|g'(x)|) < \infty$.

Next, suppose \tilde{R} is normally distributed. This assumption is inconsistent with the model in Section 6.2, where \tilde{R} is lognormally distributed, in equilibrium, being equal to $\ln \tilde{R} = \mu_D - \frac{1}{2}\sigma_D^2 + \epsilon_S$, where ϵ_S is normal. Let us explore, however, the asset pricing implications of the assumption \tilde{R} is normally distributed. Because \tilde{R}_{t+1} and Z_{t+1} are both normal, and $m_{t+1} = m(Z_{t+1}) = \exp(Z_{t+1})$, we may apply Lemma 6.3 and obtain,

$$\text{cov}(m_{t+1}, \tilde{R}_{t+1}) = E[m'(Z_{t+1})] \cdot \text{cov}(Z_{t+1}, \tilde{R}_{t+1}) = -\lambda R^{-1} \cdot \text{cov}(u_{D,t+1}, \tilde{R}_{t+1}).$$

Replacing this expression for the covariance, $\text{cov}(m_{t+1}, \tilde{R}_{t+1})$, into Eq. (6.14), leaves:

$$E(\tilde{R}_{t+1}) - R = \lambda \cdot \text{cov}(u_{D,t+1}, \tilde{R}_{t+1}).$$

We wish to extend the previous observations to a more general setup. Clearly, the stochastic discount factor is some function of K factors $m(\epsilon_{1t}, \dots, \epsilon_{Kt})$. A particularly convenient analytical assumption is to make m exponential-affine and the factors $(\epsilon_{i,t})_{i=1}^K$ normal, as in the following definition:

DEFINITION 6.4 (EAPK: Exponential Affine Pricing Kernel): *Let,*

$$Z_t \equiv \phi_0 + \sum_{i=1}^K \phi_i \epsilon_{i,t}.$$

A EAPK is a function,

$$m_t = m(Z_t) = \exp(Z_t). \quad (6.15)$$

²With a portfolio return that is perfectly correlated with m , we have:

$$E_t(\tilde{R}_{t+1}^M) - \frac{1}{E_t(m_{t+1})} = -\frac{\sigma_t(m_{t+1})}{E_t(m_{t+1})} \sigma_t(\tilde{R}_{t+1}^M).$$

In more general setups than the ones considered in this introductory example, both $\frac{\sigma_t(m_{t+1})}{E_t(m_{t+1})}$ and $\sigma_t(\tilde{R}_{t+1}^M)$ should be time-varying.

If $(\epsilon_{i,t})_{i=1}^K$ are jointly normal, and each $\epsilon_{i,t}$ has mean zero and variance σ_i^2 , $i = 1, \dots, K$, the EAPK is called a Normal EAPK (NEAPK).

In the previous definition, we assumed each $\epsilon_{i,t}$ has a mean equal to zero, which entails no loss of generality insofar as $\phi_0 \neq 0$. Next, suppose \tilde{R} is normally distributed. By Lemma 6.3 and the NEAPK assumption,

$$\text{cov}(m_{t+1}, \tilde{R}_{t+1}) = \text{cov}[\exp(Z_{t+1}), \tilde{R}_{t+1}] = R^{-1} \text{cov}(Z_{t+1}, \tilde{R}_{t+1}) = R^{-1} \sum_{i=1}^K \phi_i \text{cov}(\epsilon_{i,t+1}, \tilde{R}_{t+1}).$$

By replacing this into Eq. (6.14) leaves the linear factor representation,

$$E(\tilde{R}_{t+1}) - R = - \sum_{i=1}^K \phi_i \underbrace{\text{cov}(\epsilon_{i,t+1}, \tilde{R}_{t+1})}_{\text{“betas”}}. \quad (6.16)$$

We have thus shown the following result:

PROPOSITION 6.5: *Suppose that \tilde{R} is normally distributed. Then, NEAPK \Rightarrow linear factor representation for asset returns.*

The APT representation in Eq. (6.16), is similar to a result in Cochrane (1996).³ Cochrane (1996) assumes that m is affine, i.e. $m(Z_t) = Z_t$ where Z_t is as in Definition 6.1. This assumption implies that $\text{cov}(m_{t+1}, \tilde{R}_{t+1}) = \sum_{i=1}^K \phi_i \text{cov}(\epsilon_{i,t+1}, \tilde{R}_{t+1})$. By replacing this expression for the covariance, $\text{cov}(m_{t+1}, \tilde{R}_{t+1})$, into Eq. (6.14), leaves

$$E(\tilde{R}_{t+1}) - R = -R \sum_{i=1}^K \phi_i \text{cov}(\epsilon_{i,t+1}, \tilde{R}_{t+1}), \quad \text{where } R = \frac{1}{E(m)} = \frac{1}{\phi_0}.$$

The NEAPK assumption, compared to Cochrane’s, carries the obvious advantage to guarantee the stochastic discount factor is strictly positive—a theoretical condition we need to rule out arbitrage opportunities.

6.4.2 Lognormal returns

Next, we assume that \tilde{R} is lognormally distributed, and that the NEAPK holds. We have,

$$1 = E(m_{t+1} \cdot \tilde{R}_{t+1}) \iff e^{-\phi_0} = E\left(e^{\sum_{i=1}^K \phi_i \epsilon_{i,t+1}} \cdot \tilde{R}_{t+1}\right). \quad (6.17)$$

Consider, first, the case $K = 1$, and let $y_t = \ln \tilde{R}_t$ be normally distributed. The previous equation can be written as,

$$e^{-\phi_0} = E\left[e^{\phi_1 \epsilon_{t+1} + y_{t+1}}\right] = e^{E(y_{t+1}) + \frac{1}{2}(\phi_1^2 \sigma_\epsilon^2 + \sigma_y^2 + 2\phi_1 \sigma_{\epsilon y})}.$$

³To recall why eq. (6.16) is indeed a APT equation, suppose that \tilde{R} is a n -(column) vector of returns and that $\tilde{R} = a + bf$, where f is K -(column) vector with zero mean and unit variance and a, b are some given vector and matrix with appropriate dimension. Then clearly, $b = \text{cov}(\tilde{R}, f)$. A portfolio π delivers $\pi^\top \tilde{R} = \pi^\top a + \pi^\top \text{cov}(\tilde{R}, f)f$. Arbitrage opportunity is: $\exists \pi : \pi^\top \text{cov}(\tilde{R}, f) = 0$ and $\pi^\top a \neq r$. To rule that out, we may show as in Part I of these *Lectures* that there must exist a K -(column) vector λ s.t. $a = \text{cov}(\tilde{R}, f)\lambda + r$. This implies $\tilde{R} = a + bf = r + \text{cov}(\tilde{R}, f)\lambda + bf$. That is, $E(\tilde{R}) = r + \text{cov}(\tilde{R}, f)\lambda$.

This is,

$$E(y_{t+1}) = - \left[\phi_0 + \frac{1}{2}(\phi_1^2 \sigma_\epsilon^2 + \sigma_y^2 + 2\phi_1 \sigma_{\epsilon y}) \right].$$

By applying the pricing equation (6.17) to a zero coupon bond,

$$e^{-\phi_0} = E(e^{\phi_1 \epsilon_{t+1}}) e^{\ln R_{t+1}} = e^{\ln R_{t+1} + \frac{1}{2} \phi_1^2 \sigma_\epsilon^2},$$

which we can solve for R_{t+1} :

$$\ln R_{t+1} = - \left(\phi_0 + \frac{1}{2} \phi_1^2 \sigma_\epsilon^2 \right).$$

The expected excess return is,

$$E(y_{t+1}) - \ln R_{t+1} + \frac{1}{2} \sigma_y^2 = -\phi_1 \sigma_{\epsilon y}. \quad (6.18)$$

Eq. (6.18) shows that the theory in Section 6.2 through a different angle. Apart from Jensen's inequality effects ($\frac{1}{2} \sigma_y^2$), this is indeed the Lucas model of Section 6.2 once $\phi_1 = -\eta$. As is clear, this is a poor model, as it is bound to explain returns with only one "stochastic discount-factor parameter," i.e. ϕ_1 .

Next consider the general case. Assume as usual that dividends are as in (6.1). To find the price function in terms of the state variable ϵ , we may proceed as in Section 6.2. In the absence of bubbles,

$$S_t = \sum_{i=1}^{\infty} E \left[\frac{\xi_{t+i}}{\xi_t} \cdot D_{t+i} \right] = D_t \cdot \sum_{i=1}^{\infty} e^{(\mu_D + \phi_0 + \frac{1}{2} \sum_{i=1}^K \phi_i (\phi_i \sigma_i^2 + 2\sigma_{i,D})) \cdot i}, \quad \sigma_{i,D} \equiv \text{cov}(\epsilon_i, \epsilon_D).$$

Thus, if

$$\hat{k} \equiv \mu_D + \phi_0 + \frac{1}{2} \sum_{i=1}^K \phi_i (\phi_i \sigma_i^2 + 2\sigma_{i,D}) < 0,$$

then,

$$\frac{S_t}{D_t} = \frac{\hat{k}}{1 - \hat{k}}.$$

Even in this multi-factor setting, price-dividend ratios are constant, a counterfactual prediction, as explained in the next chapter. Note, then, the following facts. The first two moments of the stochastic discount factor satisfying a NEAPK structure can be easily found, by Eq. (6.15), and an application of Lemma 6.1:

$$E(m_t) = e^{E(Z_t) + \frac{1}{2} \text{var}(Z_t)}, \quad \sqrt{\text{var}(m_t)} = e^{E(Z_t) + \text{var}(Z_t)} \sqrt{1 - e^{-\text{var}(Z_t)}}.$$

We can always calibrate the parameters of this model, so as to make sure the first two moments of the pricing kernel enter into the Hansen-Jagannathan cup. However, remember, the model still predicts price-dividend ratios to be constant: the model does not really work, even if we are able to arbitrarily increase the variance of the pricing kernel underlying it. A model that satisfies the Hansen-Jagannathan bounds is not necessarily a good one. We need further theoretical test conditions. The next chapter illustrates theoretical test conditions addressing these concerns, and attempting to answer questions such as: (i) when are price-dividend ratios procyclical? (ii) when is returns volatility countercyclical? etc.

6.5 Pricing kernels and Sharpe ratios

6.5.1 Market portfolios and pricing kernels

The market portfolio cannot be perfectly correlated with the stochastic discount factor, in general [cite some literature such as Cecchetti, Lam, and Mark (1994)]. Let $r_{i,t+1}^e = \tilde{R}_{i,t+1} - R_{t+1}$ be the excess return on a risky asset. We have:

$$0 = E_t(m_{t+1}r_{i,t+1}^e) = E_t(m_{t+1})E_t(r_{i,t+1}^e) + \rho_{i,t} \cdot \text{Std}_t(m_{t+1}) \cdot \text{Std}_t(r_{i,t+1}^e),$$

where $\text{Std}_t(u_{t+1})$ denotes the standard deviation of a variable u_{t+1} , conditionally upon the information available at time t , and $\rho_{i,t} \equiv \text{corr}_t(m_{t+1}, r_{i,t+1}^e)$, a conditional correlation. Hence, the Sharpe Ratio, $\mathcal{S} \equiv \frac{E_t(r_{i,t+1}^e)}{\text{Std}_t(r_{i,t+1}^e)}$, satisfies:

$$|\mathcal{S}| \leq \frac{\text{Std}_t(m_{t+1})}{E_t(m_{t+1})} = \text{Std}_t(m_{t+1}) \cdot R_{t+1}, \quad (6.19)$$

The highest possible Sharpe ratio is bounded. The equality holds for a hypothetical portfolio M , say, yielding excess returns perfectly conditionally negatively correlated with the stochastic discount factor, $\rho_{M,t} = -1$. We shall say of M that it is a β -CAPM generating portfolio. Is it also a market portfolio? After all, a feasible and attainable portfolio lying on the volatility bounds for the stochastic discount factor is clearly mean-variance efficient. The answer is subtle. As explained in the context of the static model of Chapter 1, the Sharpe ratio, \mathcal{S} , equals the slope of the Capital Market Line, and bears the interpretation of unit market risk-premium. If $\rho_{M,t} = -1$, then, by Eq. (6.19), the slope of the Capital Market Line reduces to $\frac{\text{Std}_t(m_{t+1})}{E_t(m_{t+1})}$. For example, with the Lucas model in Section 6.2,

$$\frac{\text{Std}_t(m_{t+1})}{E_t(m_{t+1})} = \sqrt{e^{\eta^2\sigma_D^2} - 1} \approx \eta\sigma_D.$$

In Section 6.2, we also explained that $(\mu_S - r)/\sigma_D = \eta\sigma_D$, which is only approximately true, according to the previous relation. Not even a simple model with a *single* tree, such as that in Section 6.2, would be capable of leading to a β -CAPM generating portfolio, or a market portfolio! Indeed, for this model, we have that $E(\tilde{R}) = e^{\mu_S}$, $R = e^{-\ln\beta + \eta(\mu_D - \frac{1}{2}\sigma_D^2) - \frac{1}{2}\eta^2\sigma_D^2}$, and $\text{var}(\tilde{R}) = e^{2\mu_S}(e^{\sigma_D^2} - 1)$, with a Sharpe ratio equal to:

$$\mathcal{S} = \frac{1 - e^{-\eta\sigma_D^2}}{\sqrt{e^{\sigma_D^2} - 1}}.$$

By simple computations, $\rho = -\frac{1 - e^{-\eta\sigma_D^2}}{\sqrt{e^{\eta^2\sigma_D^2} - 1}\sqrt{e^{\sigma_D^2} - 1}}$, which is not precisely “-1”—only approximately equal to -1, for low values of σ_D .

A further complication arises, a β -CAPM generating portfolio is not necessarily the tangency portfolio. We can show that there is another portfolio leading to the very same β -pricing relation predicted by the tangency portfolio. Such a portfolio is referred to as the *maximum correlation portfolio*, for reasons developed below. Let $\bar{R} = \frac{1}{E(m)}$. By the CCAPM in Chapter 2,

$$E(R^i) - \bar{R} = \frac{\beta_{R^i,m}}{\beta_{R_p,m}} (E(R_p) - \bar{R}),$$

where R_p is a portfolio return. Next, let $R_p = R^m \equiv \frac{m}{E(m^2)}$, which is clearly perfectly correlated with the stochastic discount factor. By results in Chapter 2,

$$E(R^i) - \bar{R} = \beta_{R^i, R^m} (E(R^m) - \bar{R}).$$

This is not yet the β -representation of the CAPM, because we have yet to show that there is a way to construct R^m as a portfolio return. In fact, there is a natural choice: pick $m = m^*$, where m^* is the minimum-variance kernel leading to the Hansen-Jagannathan bounds. Since m^* is linear in all asset returns, R^{m^*} can be thought of as a return that can be obtained by investing in all assets. Furthermore, in the appendix we show that R^{m^*} satisfies,

$$1 = E(m \cdot R^{m^*}).$$

Where is this portfolio located? The Appendix shows that there is no portfolio yielding the same expected return with lower variance, that is, R^{m^*} is mean-variance efficient), and that:

$$E(R^{m^*}) - 1 = \frac{r - Sh}{1 + Sh} = r - \frac{1 + r}{1 + Sh} Sh < r.$$

Mean-variance efficiency of R^{m^*} and the previous inequality imply that this portfolio lies in the *lower branch* of the mean-variance efficient portfolios. And this is so because this portfolio is *positively* correlated with the *true* pricing kernel. Naturally, the fact that this portfolio is β -CAPM generating doesn't necessarily imply that it is also perfectly correlated with the true stochastic discount factor. As shown in the appendix, R^{m^*} has only the maximum possible correlation with all possible m . Perfect correlation occurs exactly in correspondence of the stochastic discount factor $m = m^*$ (i.e. when the economy exhibits a stochastic discount factor exactly equal to m^*).

PROOF THAT R^{m^*} IS β -CAPM GENERATING. The relations, $1 = E(m^* R^i)$ and $1 = E(m^* R^{m^*})$, imply:

$$E(R^i) - R = -R \cdot \text{cov}(m^*, R^i), \quad E(R^{m^*}) - R = -R \cdot \text{cov}(m^*, R^{m^*}),$$

or,

$$\frac{E(R^i) - R}{E(R^{m^*}) - R} = \frac{\text{cov}(m^*, R^i)}{\text{cov}(m^*, R^{m^*})}.$$

By construction, R^{m^*} is perfectly correlated with m^* . Precisely, $R^{m^*} = m^* / E(m^{*2}) \equiv \gamma^{-1} m^*$, $\gamma \equiv E(m^{*2})$. Therefore,

$$\frac{\text{cov}(m^*, R^i)}{\text{cov}(m^*, R^{m^*})} = \frac{\text{cov}(\gamma R^{m^*}, R^i)}{\text{cov}(\gamma R^{m^*}, R^{m^*})} = \frac{\gamma \cdot \text{cov}(R^{m^*}, R^i)}{\gamma \cdot \text{var}(R^{m^*})} = \beta_{R^i, R^{m^*}}.$$

||

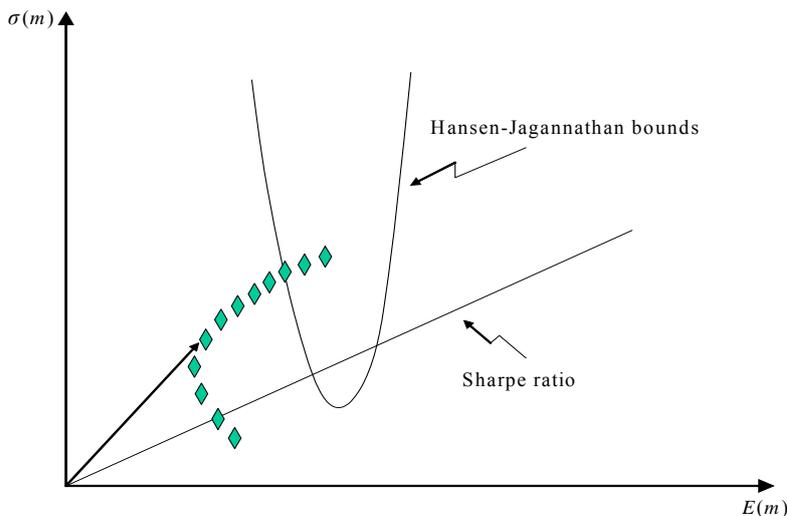
6.5.2 Pricing kernel bounds

Figure 6.2 depicts the typical situation the neoclassical asset pricing model has to face. Points \blacklozenge are those generated by the Lucas model for various values of η . The model has to be such that points \blacklozenge lie *above* the observed Sharpe ratio ($\sigma(m)/E(m) \geq$ greatest Sharpe ratio ever observed in the data—Sharpe ratio on the market portfolio) *and* inside the Hansen-Jagannathan bounds. Typically, we need high values of η to enter the Hansen-Jagannathan bounds.

There is an interesting connection between these facts and the classical mean-variance portfolio frontier described in Chapter 1. As shown in Figure 6.3, every asset or portfolio must lie inside the region bounded by two straight lines with slopes $\mp \sigma(m)/E(m)$. It must be so, as for any asset (or portfolio) priced by a stochastic discount factor m , we have that

$$|E(R^i) - R| \leq \frac{\sigma(m)}{E(m)} \cdot \sigma(R^i).$$

As seen in the previous section, the equality is only achieved by asset (or portfolio) returns that are perfectly correlated with m . A tangency portfolio such as T doesn't necessarily attain the volatility bounds for the stochastic discount factor. Moreover, the market portfolio has no reasons to lie on the volatility bound for the stochastic discount factor. As an example, for the simple Lucas model, the (only existing) asset has a Sharpe ratio, which doesn't lie on the volatility bounds for the stochastic discount factor. In a sense, the CCAPM does not need to imply the CAPM: there are necessarily no assets performing, at the same time, as market portfolios and β -CAPM generating, which are also priced consistently by the true stochastic discount factor. These conditions would only simultaneously hold if the candidate market portfolio were perfectly negatively correlated with the stochastic discount factor, which is a quite specific circumstance, the only circumstance where we can really say the CAPM is a particular case of the CCAPM. We still do not know conditions on general families of stochastic discount factors, which are consistent with the previous properties.



However, we know that there exists another portfolio, the maximum correlation portfolio, which is also β -CAPM generating. In other terms, if $\exists R_* : R_* = -\gamma m$, for some positive constant γ , then the β -CAPM representation holds, but this doesn't necessarily mean that R_* is also a market portfolio. More generally, if there is a return R_* that is β -CAPM generating, then,

$$\rho_{i,R_*} = \frac{\rho_{i,m}}{\rho_{R_*,m}}, \text{ all } i. \quad (6.20)$$

Therefore, we don't need an asset or portfolio return that is *perfectly* correlated with m to make the CCAPM collapse to the CAPM. All in all, the existence of an asset return that is

perfectly negatively correlated with the stochastic discount factor is a sufficient condition for the CCAPM to collapse to the CAPM, not a necessary condition. The proof of Eq. (6.20) is simple. By the CCAPM,

$$E(R^i) - R = -\rho_{i,m} \frac{\sigma(m)}{E(m)} \sigma(R^i), \quad \text{and} \quad E(R^*) - R = -\rho_{R^*,m} \frac{\sigma(m)}{E(m)} \sigma(R^*).$$

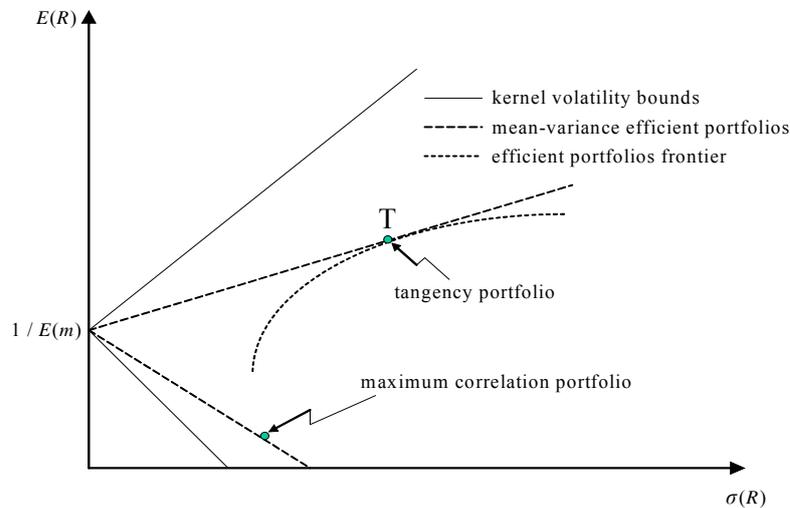
That is,

$$\frac{E(R^i) - R}{E(R^*) - R} = \frac{\rho_{i,m} \sigma(R^i)}{\rho_{R^*,m} \sigma(R^*)} \quad (6.21)$$

But if R^* is β -CAPM generating,

$$\frac{E(R^i) - R}{E(R^*) - R} = \frac{\text{cov}(R^i, R^*)}{\sigma(R^*)^2} = \rho_{i,R^*} \frac{\sigma(R^i)}{\sigma(R^*)}. \quad (6.22)$$

Comparing Eq. (6.21) with Eq. (6.22) produces Eq. (6.20).



A final thought. In many pieces of applied research, we often read that because we observe time-varying Sharpe ratios on (proxies of) the market portfolio, we should also model the market risk-premium $\pi_t \equiv \sqrt{\text{Var}_t(m_{t+1})} / E_t(m_{t+1})$ as time-varying. While Chapter 7 explains that the evidence for time-varying risk-premiums is overwhelming, a criticism to this motivation is that π_t is only an upper bound to the Sharpe ratio of the market portfolio. On a strictly theoretical point of view, then, a time-varying π_t is neither a necessary or a sufficient condition to have time-varying Sharpe ratios, as Figure 6.3 illustrates.

6.6 Conditioning bounds

The Hansen-Jagannathan bounds in Eq. (6.12) can be improved by using conditioning information, as originally shown by Gallant, Hansen and Tauchen (1990) and in Ferson and Siegel (2003). A difficulty with these bounds is that they may display a finite sample bias, in that they tend to overstate the true bounds and thus reject too often a given model. Finite sample corrections are considered by Ferson and Siegel (2003). [Discuss, analytically]

Alvarez and Jermann (2005). [Discuss, analytically]

6.7 The cross section of stock returns and volatilities

6.7.1 Returns

Consider the Security Market Line (SML) in Eq. (1.18) of Chapter 1,

$$b_i - r = \beta_i (\mu_M - r), \quad i = 1, \dots, m, \quad (6.23)$$

where $b_i - r$ is the average excess return on the i -th asset, β_i is its beta, and $\mu_M - r$ is the average excess return on the market. According to the one-factor CAPM model, each asset should display an average excess return lying precisely on the SML. Assets delivering average excess returns and betas above the SML, as the points A , B , C , and D in Figure 7.14 below, would be simply evidence that this single factor version of the CAPM does not work. Consider, for example, the asset leading to point A . A regression of the excess return of this asset onto the excess return on the market would produce a positive intercept, some $\alpha > 0$, such that its average excess return would equal $\alpha + \beta_i (\mu_M - r)$, thereby invalidating Eq. (6.23). There exist at least two pieces of evidence against the one-factor CAPM, which were systematically pointed out by Fama and French (1992, 1993):

- (i) *Size effect* (Banz, 1981): Average returns for “small firms,” or low capitalized firms (in terms of market equity, defined as stock price times outstanding shares) are too high given their beta.
- (ii) *Value effect* (Stattman, 1980; Rosenberg, Reid and Lanstein, 1985): Average returns on stocks of firms with high book-to-market (BM, henceforth) ratios, or “value stocks,” are too high given their beta. In general, average returns on value stocks are higher than those on “growth” stocks, i.e. those stocks with low BM ratios. As an example, the points D , C , B , and A in Figure 7.14 might typically refer to stocks with low-to-high BM ratios.

A third piece of evidence against the standard CAPM is the “momentum” effect:

- (iii) *Momentum effect* (Jegadeesh and Titman, 1993): Stocks with the highest returns in the previous twelve months will outperform in the next future.

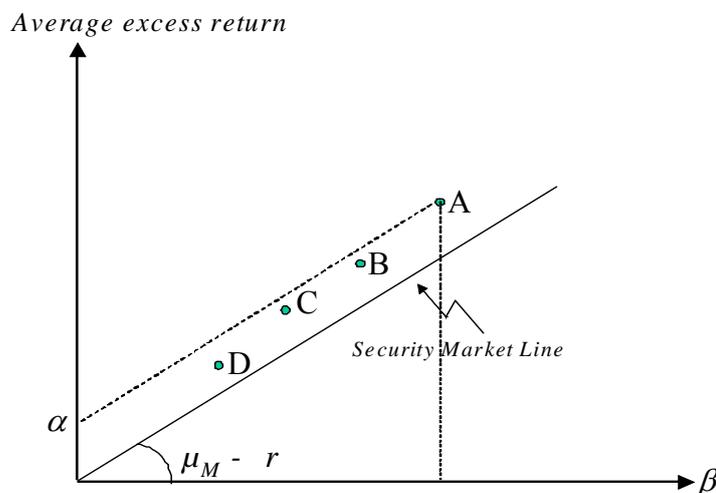


FIGURE 6.3.
217

The one-factor CAPM has no power in explaining the cross-section of asset returns, sorted by size, BM or momentum. Assets sorted in this way command a size premium, a value premium, and a momentum premium. For example, one can create portfolios sorted by size and BM, say 25 portfolios, out of a 5×5 matrix with dimensions given by size and BM. The puzzle, then, at least from the standard CAPM perspective, is that this model cannot explain the returns on these portfolios. Fama and French (1993) show that the returns on these portfolios can be very much better understood by means of a multifactor model, where both size and value premiums are explicitly taken into account. They consider three factors: (i) the excess return on the market; (ii) an “HML” factor, defined as the monthly difference between the returns on assets with high and low BM ratios (“high minus low”); an “SMB” factor, defined as the difference between the asset returns of firms with small and big size (“small minus big”). The HML and SMB factors are defined as the differences between the returns on the appropriate cells of a 2×3 matrix, obtained through percentiles of the distribution of asset returns over the previous year.

	Book-to-Market		
Size	L	M	H
S			
L			

The resulting model is the celebrated Fama-French three factor model. Carhart (1997) extends this model to a four-factor model with a momentum factor: the monthly difference between the returns on the high and low prior return portfolios.

6.7.2 Volatilities

6.8 Appendix

PROOF OF THE EQUATION, $\mathbf{1}_n = E[m_t^*(\bar{m}) \cdot (\mathbf{1}_n + R_t)]$. We have,

$$\begin{aligned}
E[m_t^*(\bar{m}) \cdot (\mathbf{1}_n + R_t)] &= E\left[\left(\bar{m} + (R_t - E(R_t))^\top \beta_{\bar{m}}\right) (\mathbf{1}_n + R_t)\right] \\
&= \bar{m}E(\mathbf{1}_n + R_t) + E\left[(R_t - E(R_t))^\top \beta_{\bar{m}} (\mathbf{1}_n + R_t)\right] \\
&= \bar{m}E(\mathbf{1}_n + R_t) + E\left[(\mathbf{1}_n + R_t) (R_t - E(R_t))^\top\right] \beta_{\bar{m}} \\
&= \bar{m}E(\mathbf{1}_n + R_t) + E\left[\left((\mathbf{1}_n + E(R_t)) + (R_t - E(R_t))\right) (R_t - E(R_t))^\top\right] \beta_{\bar{m}} \\
&= \bar{m}E(\mathbf{1}_n + R_t) + E\left[(R_t - E(R_t)) (R_t - E(R_t))^\top\right] \beta_{\bar{m}} \\
&= \bar{m}E(\mathbf{1}_n + R_t) + \Sigma \beta_{\bar{m}} \\
&= \bar{m}E(\mathbf{1}_n + R_t) + \mathbf{1}_n - \bar{m}E(\mathbf{1}_n + R_t),
\end{aligned}$$

where the last line follows by the definition of $\beta_{\bar{m}}$.

PROOF THAT R^{m^*} CAN BE GENERATED BY A FEASIBLE PORTFOLIO

PROOF OF THE EQUATION, $1 = E(m \cdot R^{m^*})$. We have,

$$E(m \cdot R^{m^*}) = \frac{E(m \cdot m^*)}{E[(m^*)^2]},$$

where

$$\begin{aligned}
E(m \cdot m^*) &= \bar{m}^2 + E\left[m (R_t - E(R_t))^\top \beta_{\bar{m}}\right] \\
&= \bar{m}^2 + E\left[m (1 + R_t)^\top\right] \beta_{\bar{m}} - E\left[m (1 + E(R_t))^\top\right] \beta_{\bar{m}} \\
&= \bar{m}^2 + \beta_{\bar{m}} - E(m) [1 + E(R_t)]^\top \beta_{\bar{m}} \\
&= \bar{m}^2 + \left[\mathbf{1}_n - \bar{m} (1 + E(R_t))^\top\right] \beta_{\bar{m}} \\
&= \bar{m}^2 + \left[\mathbf{1}_n - \bar{m} (1 + E(R_t))^\top\right] \Sigma^{-1} [\mathbf{1}_n - \bar{m} (\mathbf{1}_n + E(R_t))] \\
&= \bar{m}^2 + \text{var}(m^*),
\end{aligned}$$

where the last line is due to the definition of m^* .

PROOF THAT R^{m^*} IS MEAN-VARIANCE EFFICIENT. Let $p = (p_0, p_1, \dots, p_n)^\top$ the vector of $n + 1$ portfolio weights (here $p_i \equiv \pi^i / w$ is the portfolio weight of asset i , $i = 0, 1, \dots, n$). We have,

$$p^\top \mathbf{1}_{n+1} = 1.$$

The returns we consider are $\underline{r}_t = (\bar{m}^{-1} - 1, r_{1,t}, \dots, r_{n,t})^\top$. We denote our “benchmark” portfolio return as $r_{bt} = r^{m^*} - 1$. Next, we build up an arbitrary portfolio yielding the same expected return $E(r_{bt})$ and then we show that this has a variance greater than the variance of r_{bt} . Since this portfolio

is arbitrary, the proof will be complete. Let $r_{pt} = p^\top \underline{r}_t$ such that $E(r_{pt}) = E(r_{bt})$. We have:

$$\begin{aligned}
\text{cov}(r_{bt}, r_{pt} - r_{bt}) &= E[r_{bt} \cdot (r_{pt} - r_{bt})] \\
&= E[R_{bt} \cdot (R_{pt} - R_{bt})] \\
&= E(R_{bt} \cdot R_{pt}) - E(R_{bt}^2) \\
&= \frac{E[m^* (1 + p^\top \underline{r}_t)]}{E(m^{*2})} - \frac{E(m^{*2})}{[E(m^{*2})]^2} \\
&= \frac{1}{E(m^{*2})} [p^\top E(m^* (\mathbf{1}_{n+1} + \underline{r}_t)) - 1] \\
&= 0.
\end{aligned}$$

The first line follows by construction since $E(r_{pt}) = E(r_{bt})$. The last line follows because

$$p^\top E[m^* (\mathbf{1}_{n+1} + \underline{r}_t)] = p^\top \mathbf{1}_{n+1} = 1.$$

Given this, the claim follows directly from the fact that

$$\text{var}(R_{pt}) = \text{var}[R_{bt} + (R_{pt} - R_{bt})] = \text{var}(R_{bt}) + \text{var}(R_{pt} - R_{bt}) \geq \text{var}(R_{bt}).$$

PROOF OF THE EQUATION, $E(R^{m^*}) - 1 = r - \frac{1+r}{1+Sh}Sh$. We have,

$$E(R^{m^*}) - 1 = \frac{\bar{m}}{E[(m^*)^2]} - 1.$$

In terms of the notation introduced in Section 6.8, m^* is:

$$m^* = \bar{m} + (a\epsilon)^\top \beta_{\bar{m}}, \quad \beta_{\bar{m}} = \sigma^{-1} [\mathbf{1}_n - \bar{m} (\mathbf{1}_n + b)].$$

We have,

$$\begin{aligned}
E[(m^*)^2] &= [\bar{m} + (a\epsilon)^\top \beta_{\bar{m}}]^2 \\
&= \bar{m}^2 + E[(a\epsilon)^\top \beta_{\bar{m}}]^2 \\
&= \bar{m}^2 + E[(a\epsilon)^\top \beta_{\bar{m}} \cdot (a\epsilon)^\top \beta_{\bar{m}}] \\
&= \bar{m}^2 + E[(\beta_{\bar{m}}^\top a\epsilon)(\epsilon^\top a^\top \beta_{\bar{m}})] \\
&= \bar{m}^2 + \beta_{\bar{m}}^\top \cdot \sigma \cdot \beta_{\bar{m}} \\
&= \bar{m}^2 + [\mathbf{1}_n^\top - \bar{m}(\mathbf{1}_n^\top + b^\top)] \sigma^{-1} [\mathbf{1}_n - \bar{m} (\mathbf{1}_n + b)] \\
&= \bar{m}^2 + \mathbf{1}_n^\top \sigma^{-1} \mathbf{1}_n - \bar{m} (\mathbf{1}_n^\top \sigma^{-1} \mathbf{1}_n + \mathbf{1}_n^\top \sigma^{-1} b) \\
&\quad - \bar{m} [\mathbf{1}_n^\top \sigma^{-1} \mathbf{1}_n + b^\top \sigma^{-1} \mathbf{1}_n - \bar{m} (\mathbf{1}_n^\top \sigma^{-1} \mathbf{1}_n + b^\top \sigma^{-1} \mathbf{1}_n + \mathbf{1}_n^\top \sigma^{-1} b + b^\top \sigma^{-1} b)]
\end{aligned}$$

Again in terms of the notation of Section 6.8 ($\gamma \equiv \mathbf{1}_n^\top \sigma^{-1} \mathbf{1}_n$ and $\beta \equiv \mathbf{1}_n^\top \sigma^{-1} b$), this is:

$$E[(m^*)^2] = \gamma - 2\bar{m}(\gamma + \beta) + \bar{m}^2 (1 + \gamma + 2\beta + b^\top \sigma^{-1} b).$$

The expected return is thus,

$$E(R^{m^*}) - 1 = \frac{E(m^*)}{E[(m^*)^2]} - 1 = \frac{\bar{m} - \gamma + 2\bar{m}(\gamma + \beta) - \bar{m}^2 (1 + \gamma + 2\beta + b^\top \sigma^{-1} b)}{\gamma - 2\bar{m}(\gamma + \beta) + \bar{m}^2 (1 + \gamma + 2\beta + b^\top \sigma^{-1} b)}.$$

Now recall two definitions:

$$\bar{m} = \frac{1}{1+r}, \quad Sh = (b - \mathbf{1}_m r)^\top \sigma^{-1} (b - \mathbf{1}_m r) = b^\top \sigma^{-1} b - 2\beta r + \gamma r^2.$$

In terms of r and Sh , we have,

$$\begin{aligned} E(R^{m^*}) - 1 &= \frac{E(m^*)}{E[(m^*)^2]} - 1 \\ &= -\frac{\gamma(1+r)^2 - (1+r)(1+2\gamma+2\beta) + 1 + \gamma + 2\beta + b^\top \sigma^{-1} b}{\gamma(1+r)^2 - (1+r)(2\gamma+2\beta) + 1 + \gamma + 2\beta + b^\top \sigma^{-1} b} \\ &= \frac{r - Sh}{1 + Sh} \\ &= r - \frac{1+r}{1+Sh} Sh \\ &< r. \end{aligned}$$

This is positive if $r - Sh > 0$, i.e. if $b^\top \sigma^{-1} b - (2\beta + 1)r + \gamma r^2 < 0$, which is possible for sufficiently low (or sufficiently high) values of r .

PROOF THAT R^{m^*} IS THE m -MAXIMUM CORRELATION PORTFOLIO. We have to show that for any stochastic discount factor m , $|corr(m, R_{bt})| \geq |corr(m, R_{pt})|$. Define a ℓ -parametrized portfolio such that:

$$E[(1-\ell)R_o + \ell R_{pt}] = E(R_{bt}), \quad R_o \equiv \bar{m}^{-1}.$$

We have

$$\begin{aligned} corr(m, R_{pt}) &= corr[m, (1-\ell)R_o + \ell R_{pt}] \\ &= corr[m, R_{bt} + ((1-\ell)R_o + \ell R_{pt} - R_{bt})] \\ &= \frac{cov(m, R_{bt}) + cov(m, (1-\ell)R_o + \ell R_{pt} - R_{bt})}{\sigma(m) \cdot \sqrt{var((1-\ell)R_o + \ell R_{pt})}} \\ &= \frac{cov(m, R_{bt})}{\sigma(m) \cdot \sqrt{var((1-\ell)R_o + \ell R_{pt})}} \end{aligned}$$

The first line follows because $(1-\ell)R_o + \ell R_{pt}$ is a nonstochastic affine translation of R_{pt} . The last equality follows because

$$\begin{aligned} cov(m, (1-\ell)R_o + \ell R_{pt} - R_{bt}) &= E[m \cdot ((1-\ell)R_o + \ell R_{pt} - R_{bt})] \\ &= (1-\ell) \cdot \underbrace{E(mR_o)}_{=1} + \ell \cdot \underbrace{E(mR_{pt})}_{=1} - \underbrace{E(m \cdot R_{bt})}_{=1} \\ &= 0. \end{aligned}$$

where the first line follows because $E((1-\ell)R_o + \ell R_{pt}) = E(R_{bt})$. Therefore,

$$corr(m, R_{pt}) = \frac{cov(m, R_{bt})}{\sigma(m) \cdot \sqrt{var((1-\ell)R_o + \ell R_{pt})}} \leq \frac{cov(m, R_{bt})}{\sigma(m) \cdot \sqrt{var(R_{bt})}} = corr(m, R_{bt}),$$

where the inequality follows because R_{bt} is mean-variance efficient (i.e. \nexists feasible portfolios with the same expected return as R_{bt} and variance less than $var(R_{bt})$), and then $var((1-\ell)R_o + \ell R_{pt}) \geq var(R_{bt})$, all R_{pt} .

References

- Alvarez, F. and U.J. Jermann (2005): "Using Asset Prices to Measure the Persistence of the Marginal Utility of Wealth." *Econometrica* 73, 1977-2016.
- Banz, R.W. (1981): "The Relationship Between Return and Market Value of Common Stocks." *Journal of Financial Economics* 9, 3-18.
- Carhart, M. (1997): "On Persistence of Mutual Fund Performance." *Journal of Finance* 52, 57-82.
- Cecchetti, S., Lam, P-S. and N. C. Mark (1994): "Testing Volatility Restrictions on Intertemporal Rates of Substitution Implied by Euler Equations and Asset Returns." *Journal of Finance* 49, 123-152.
- Cochrane, J. (1996): "A Cross-Sectional Test of an Investment-Based Asset Pricing Model." *Journal of Political Economy* 104, 572-621.
- Epstein, L.G. and S.E. Zin (1989): "Substitution, Risk-Aversion and the Temporal Behavior of Consumption and Asset Returns: A Theoretical Framework." *Econometrica* 57, 937-969.
- Epstein, L.G. and S.E. Zin (1991): "Substitution, Risk-Aversion and the Temporal Behavior of Consumption and Asset Returns: An Empirical Analysis." *Journal of Political Economy* 99, 263-286.
- Fama, E. F. and K. R. French (1992): "The Cross-Section of Expected Stock Returns." *Journal of Finance* 47, 427-465.
- Fama, E. F. and K. R. French (1993): "Common Risk Factors in the Returns on Stocks and Bonds." *Journal of Financial Economics* 33, 3-56.
- Ferson, W. E. and A. F. Siegel (2003): "Stochastic Discount Factor Bounds with Conditioning Information." *Review of Financial Studies* 16, 567-595.
- Gallant, R. A., L. P. Hansen and G. Tauchen (1990): "Using the Conditional Moments of Asset Payoffs to Infer the Volatility of Intertemporal Marginal Rates of Substitution." *Journal of Econometrics* 45, 141-179.
- Gordon, M. (1962): *The Investment, Financing, and Valuation of the Corporation*. Homewood, IL: Irwin.
- Güvener, F. (2009): "A Parsimonious Macroeconomic Model for Asset Pricing." *Econometrica* 77, 1711-1740.
- Hansen, L. P. and R. Jagannathan (1991): "Implications of Security Market Data for Models of Dynamic Economies." *Journal of Political Economy* 99, 225-262.
- Jegadeesh, N. and S. Titman (1993): "Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency." *Journal of Finance* 48, 65-91.
- Mehra, R. and E. C. Prescott (1985): "The Equity Premium: A Puzzle." *Journal of Monetary Economics* 15, 145-161.

- Rosenberg, B. K. Reid and R. Lanstein (1985): "Persuasive Evidence of Market Inefficiency." *Journal of Portfolio Management* 11, 9-17.
- Stattman, D. (1980): "Book Values and Stock Returns." *The Chicago MBA: A Journal of Selected Papers* 4, 25-45.
- Weil, Ph. (1989): "The Equity Premium Puzzle and the Risk-Free Rate Puzzle." *Journal of Monetary Economics* 24, 401-421.

7

Aggregate fluctuations in equity markets

7.1 Introduction

This chapter documents empirical regularities of the aggregate stock market, those relating to the business cycle, and points to general issues about what we need to do with the neo-classical asset pricing model developed in Part I of these Lectures, so as to address these empirical puzzles. It is a natural development of the previous chapter, going beyond the fundamental equity premium and interest rate puzzles. Its scope is, indeed, a search for general classes of pricing kernels, which accommodate the business cycle properties of the aggregate stock market. We need to develop tools of analysis, which are somehow unusual in economics, aiming to reverse-engineer from any pricing kernel, the properties we need to ensure asset markets behave as in the data, thereby laying down foundations to many of the specific models discussed in the next chapter. Needless to mention, the pricing kernel properties that we look for are not necessarily those that ensure a resolution of the puzzles surveyed in the previous chapter. Once we ascertain a given class of kernels is likely to lead to the right asset price properties, further scrutiny is required, aimed to clarify whether this class leads to reasonable conclusions about, say, the equity premium—we may find a model to predict countercyclical volatility, as in the data, yet the level of this volatility can be one order of magnitude less than that in the data.

Section 7.2 provides a succinct overview of the empirical regularities of aggregate equity markets, which we want to match with our models. For example, we shall explain that price-dividend ratios and stock returns are procyclical, and that stock volatility and risk-premiums are both time-varying and countercyclical. Section 7.3 analyzes in deeper detail the empirical behavior of aggregate stock market volatility, and puts forward preliminary and certainly not exhaustive explanations for it. Section 7.4 is where we develop the framework of analysis aiming to assess whether the empirical behavior of price-dividend ratios, stock returns, risk-premiums and volatility can be rationalized within the neoclassical framework. Section 7.5 provides two examples of economies, which illustrate the predictions in Section 7.4: one economy, with habit formation, and a second, with uncertain fundamentals and a learning process about them. It also presents a class of analytically convenient models, which we shall use, at times, in the next chapter.

7.2 The empirical evidence: bird's eye view

Aggregate stock market fluctuations are intimately related to the business cycle. The evidence is striking and well-known (see, e.g., the survey in Campbell, 2003), although the emphasis in this section is to streamline how these fluctuations relate to general macroeconomic conditions. We use data sampled at a monthly frequency, covering the period from January 1948 through December 2002. We compute ex-post, yearly returns at month t as $\sum_{i=1}^{12} \tilde{R}_{t+1-i}$, where $\tilde{R}_t = \ln\left(\frac{S_t + D_t}{S_{t-1}}\right)$, S_t is the S&P Composite index as of month t , and D_t is the aggregate dividend, as calculated by Robert Shiller. Table 7.1 provides basic statistics for both row data such as P/D ratios, P/E ratios and ex-post returns, and stock volatility and expected returns. Stock volatility is computed as:

$$\text{Vol}_t \equiv \sqrt{6\pi} \cdot \bar{\sigma}_t, \quad \bar{\sigma}_t \equiv \frac{1}{12} \sum_{i=1}^{12} \left| \tilde{R}_{t+1-i} - R_{t+1-i} \right|, \quad (7.1)$$

where R_t is the risk-free rate, taken to be the one month bill return. The rationale behind this calculation is as follows. First, $\bar{\sigma}_t$ is an estimate of the average volatility occurring over the last 12 months. We annualize $\bar{\sigma}_t$ by multiplying it by $\sqrt{12}$. The term $\sqrt{6\pi}$ arises for the following reason. If we assume that a given return $R = \sigma u$, where σ is a positive constant and u is a standard unit normal, then $E(|R|) = \sigma \sqrt{\frac{2}{\pi}}$. The definition Vol_t in Eq. (8.12), then, follows by multiplying $\sqrt{12}\bar{\sigma}_t(\ell)$ by $\sqrt{\frac{\pi}{2}}$. This correction term, $\sqrt{\frac{\pi}{2}}$, has been suggested by Schwert (1989a) in a related context.

Expected returns are computed through the Fama and French (1989) predictive regressions of \tilde{R}_t on to default-premium, term-premium and the previously defined return volatility, Vol_t . With the exception of the P/D and P/E ratios, all figures are annualized percent.

We note the first main set of stylized facts:

FACT I. P/D, P/E ratios and ex-post returns are procyclical, although variations in the business cycle conditions do not seem to be the only driving force for them.

For example, Figure 7.1 reveals that price-dividend ratios decline during *all* of the economic slowdowns, as signaled by the recession indicator calculated by the National Bureau of Economic Research (NBER)—the NBER recessions. At the same time, during NBER expansions, price-dividend ratios seem to be driven by additional factors not necessarily related to the business cycle. For example, during the “roaring” 1960s, price-dividend ratios experienced two major drops with the same magnitude as the decline at the very beginning of the “chaotic” 1970s. Ex-post returns follow approximately the same pattern, although they are more volatile than price-dividend ratios (see Figure 7.2).

What about the first two conditional moments of asset returns?

FACT II. Stock volatility and expected returns are countercyclical. However, business cycle conditions do not seem to be the only forces explaining the swings of these variables.

Figures 7.3 through 7.5 are suggestive. For example, Figure 7.4 depicts the statistical relation between stock volatility and the industrial production growth rate over the last sixty years, which shows that stock volatility is largely countercyclical, being larger in bad times than in

good.¹ There are, of course, exceptions. For example, stock volatility rocketed to almost 23% during the 1987 crash—a crash occurring during one of the most enduring post-war expansions period. Countercyclical volatility is a stylized fact extensively discussed in Sections 7.3 and 7.4. In those sections, we shall learn that within the neoclassical modeling framework, this property does likely arise as soon as the volatility of the P/D ratios *changes* is countercyclical. Table 7.1 reveals, then, that the P/D ratios variations are more volatile in bad times than in good. Finally, and interestingly, Figure 7.4 suggests that stock volatility behaves asymmetrically over the business cycle, in that it increases more in bad times than it decreases in good. This asymmetric behavior of stock volatility echoes its high frequency behavior documented at least since Glosten, Jagannathan and Runkle (1993), whereby stock volatility increases more when returns are negative than it decreases when returns are positive.

A third set of stylized facts relates to the asymmetric behavior of the previous variables over the business cycle:

FACT III. P/D ratios and expected returns *changes* behave asymmetrically over the business cycle: the deepest variations in these variables occur during the contractionary phases of the business cycle.

During recessions, these variables move more than they do in good times. As an example, not only are expected returns countercyclical. On average, expected returns increase more during NBER recessions than they decrease during NBER expansions. Similarly, not only are P/D ratios procyclical. On average, P/D ratios increase less during NBER expansions than they decrease during NBER recessions. Moreover, this asymmetric behavior is, quantitatively, quite pronounced. Consider, for example, the changes in the P/D ratios: on average, their percentage (negative) changes during recessions is nearly twice as the percentage (positive) changes during expansions. Sections 7.3 and 7.4 aim to provide explanations of these facts within neoclassical models, and develop theoretical test conditions that the very same models would have to satisfy in order to be consistent with these facts.

	total		NBER expansions		NBER recessions	
	average	std dev	average	std dev	average	std dev
P/D ratio	31.99	15.88	33.21	15.79	26.20	14.89
P/E ratio	15.79	6.89	16.36	6.62	13.04	7.46
$\ln \frac{P/D_{t+1}}{P/D_t}$	2.01	12.13	3.95	10.81	-7.28	16.79
one year returns	8.59	15.86	12.41	13.04	-9.45	15.49
real risk-free rate	1.02	2.48	1.03	2.43	0.97	2.69
excess return volatility	14.55	4.68	14.05	4.47	16.91	4.91
expected returns	8.36	3.49	8.09	3.29	9.62	4.10

TABLE 7.1. Data are sampled monthly and cover the period from January 1948 through December 2002. With the exception of the P/D ratio levels, all figures are annualized percent.

¹The predictive regressions in Figures 7.4, 7.5 and 7.7 are obtained through least absolute deviations regressions, a technique known to be more robust to the presence of outliers than ordinary least squares (see Bloomfield and Steiger, 1983).

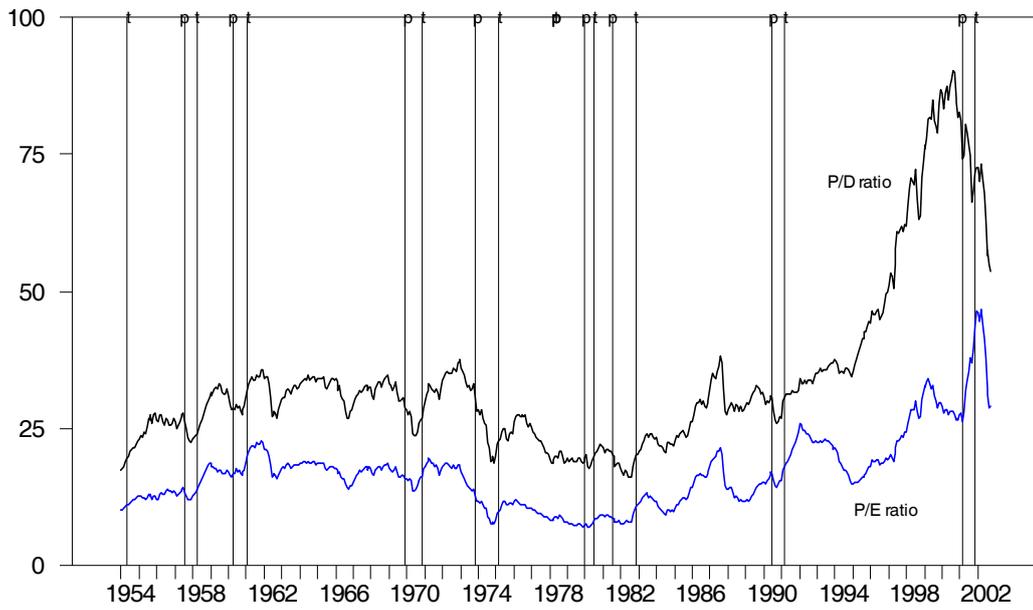


FIGURE 7.1. P/D and P/E ratios

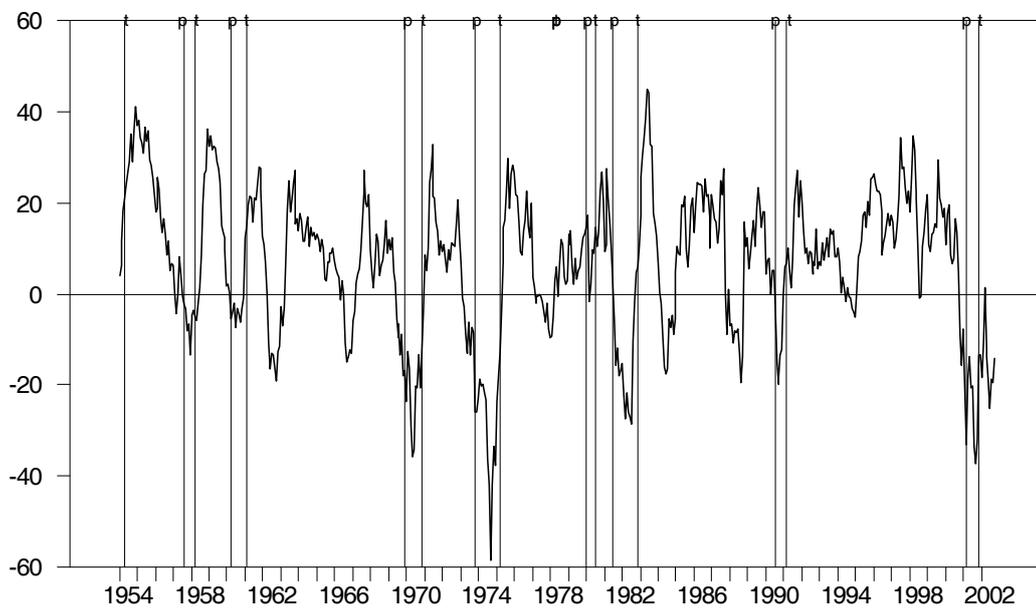


FIGURE 7.2. Monthly excess returns, in percentage, year-to-year.

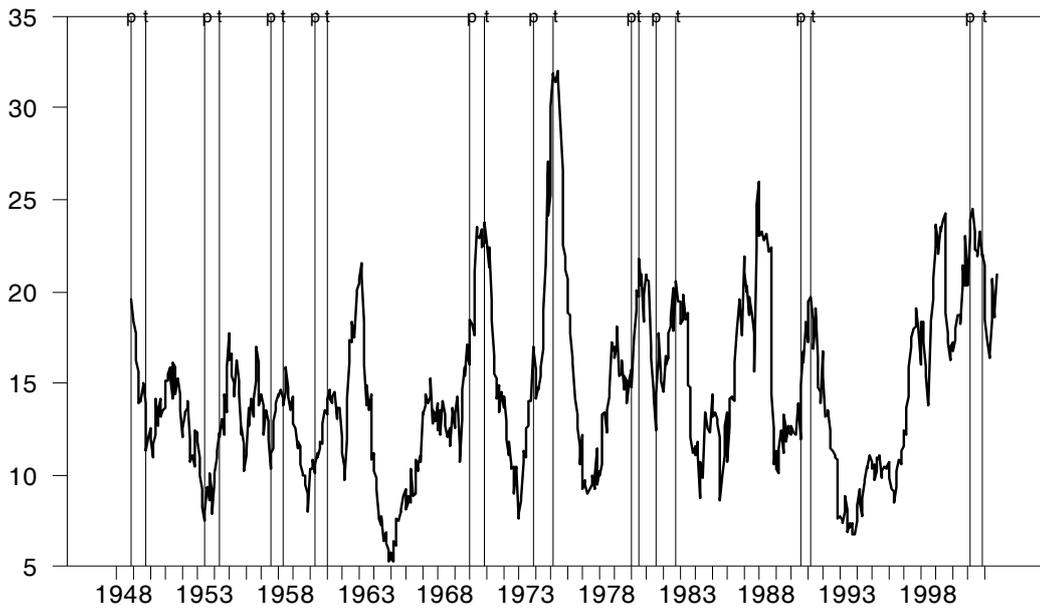


FIGURE 7.3. Aggregate stock market volatility.

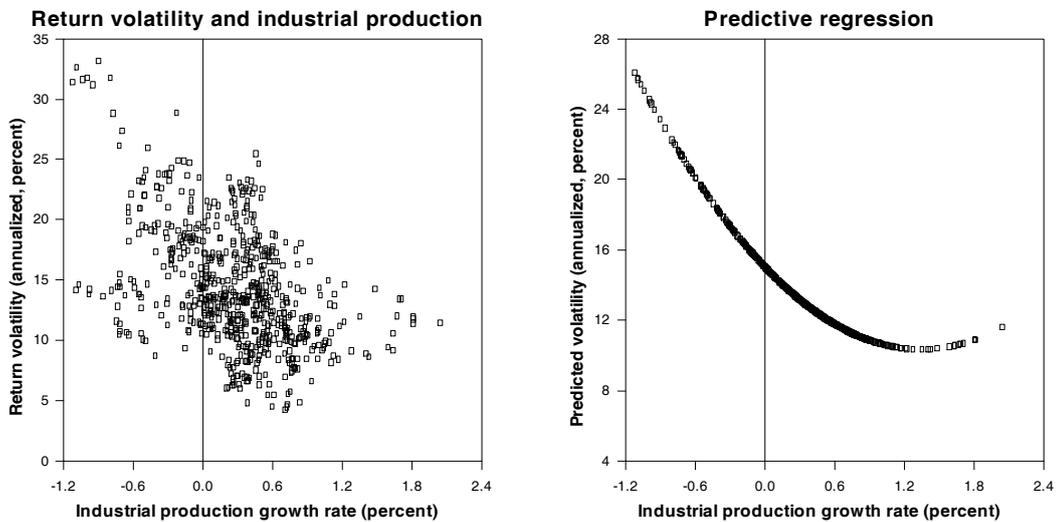


FIGURE 7.4. Stock volatility and business cycle conditions. The left panel plots stock volatility, Vol_t , against yearly (deseasoned) industrial production average growth rates, computed as $\text{IP}_t \equiv \frac{1}{12} \sum_{i=1}^{12} \text{Ind}_{t+1-i}$, where Ind_t is the real, seasonally adjusted industrial production growth as of month t . The right panel depicts the prediction of the static least absolute deviations regression: $\text{Vol}_t = 12.01 - 5.57 \cdot \text{IP}_t + 2.06 \cdot \text{IP}_t^2 + w_t$, where w_t is a residual term, and robust standard errors are in parenthesis. The data span the period from January 1948 to December 2002.

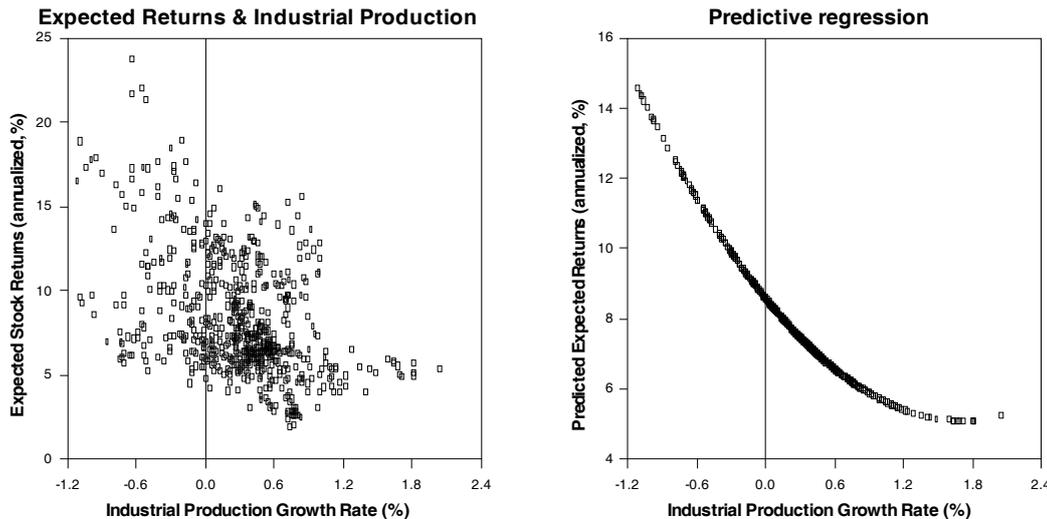


FIGURE 7.5. The left-hand side of this picture plots estimates of the expected returns (annualized, percent) ($\hat{\mathcal{E}}_t$ say) against yearly (deseasoned) industrial production average growth rates, computed as $IP_t \equiv \frac{1}{12} \sum_{i=1}^{12} \text{Ind}_{t+1-i}$, where Ind_t is the real, seasonally adjusted industrial production growth as of month t . Expected returns are estimated through the predictive regression of S&P returns on to default-premium, term-premium and return volatility, Vol_t . The right-hand side of this picture depicts the prediction of the static Least Absolute Deviations regression: $\hat{\mathcal{E}}_t = 8.56 - 4.05 \cdot IP_t + 1.18 \cdot IP_t^2 + w_t$, where w_t is a residual term, and robust standard errors are in parenthesis. Data are sampled monthly, and span the period from January 1948 to December 2002.

Fact I entails a quite intuitive consequence: price-dividend ratios might convey information relating to future returns. After all, expansions are followed by recessions. Therefore, in good times, the stock market predicts that in the future, returns will be negative. Define the excess return for the time period $[t, t+n]$ as $\tilde{R}_{t,t+n}^e \equiv \tilde{R}_{t,t+n} - R_{t,t+n}$, where $\tilde{R}_{t,t+n}$ is the asset return over $[t, t+n]$, and $R_{t,t+n}$ is the sum of the one-month Treasury bill rate, taken over $[t, t+n]$. Consider the following regressions,

$$\tilde{R}_{t,t+n}^e = a_n + b_n \times P/D_t + u_{n,t}, \quad n \geq 1, \quad (7.2)$$

where u_n is a residual term. Typically, then, the estimates of b_n are significantly negative, and the R^2 for these regressions increases with n . In turn, the previous regressions imply that $E[\tilde{R}_{t,t+n}^e | P/D_t] = a_n + b_n \times P/D_t$. They thus suggest that price-dividend ratios are driven by expected excess returns. In this restrictive sense, countercyclical expected returns (Fact II) and procyclical price-dividend ratios (Fact I) might be two sides of the same coin. To link these predictability results more closely to developments of the business cycle, consider the following regression, performed with monthly data from 1948:01 to 2002:12,

$$\tilde{R}_{t-12,t}^e = 14.64 - 9.09 \times IP_{t-12} - 14.27 \times \text{Infl}_{t-12} + u_t, \quad \text{with } R^2 = 11\%, \quad (7.3)$$

(1.04) (1.37) (2.67)

where robust standard errors are in parenthesis, u is a residual term, $\tilde{R}_{t-12,t}^e$ is the excess return from $t-12$ to t , IP_t is the average industrial production growth over the previous twelve months, as defined in Figure 7.4, and Infl_t is defined similarly as IP_t .

The negative signs of the coefficients in Eq. (7.3) are quite to be expected. Economic activity does display mean-reverting behavior, in that bad times are followed by good, in the sample size we consider. But good times are those where the stock market goes up. Therefore, a slowdown in economic activity is a predictor of high returns in the future. To illustrate with a simple example, consider a case where the aggregate stock market positively links to a single state variable tracking the business cycle conditions $x(t)$, say, such that the log of the aggregate equity index is $\ln S(t) = s_0 + s_x x(t)$, for two constant s_0 and s_x , and where $s_x > 0$. Assume, then, and critically, that $x(t)$ is mean-reverting, with unconditional expectation μ , speed of adjustment $\kappa > 0$, and some volatility coefficient $\sigma(x)$,

$$dx(t) = \kappa(\mu - x(t))dt + \sigma(x(t))dW(t),$$

where $W(t)$ is a Brownian motion. Then, it is straightforward to show that $E_{t-12} \left(\ln \frac{S(t)}{S(t-12)} \right) = a_0 - a_1 x(t-12)$, where E_t denotes the expectation taken conditionally upon the information set as of time t , and $a_0 \equiv \delta_x (1 - e^{-12\kappa}) \mu$ and $a_1 \equiv \delta_x (1 - e^{-12\kappa})$. That is, if $x(t)$ is mean-reverting, $\kappa > 0$, and the aggregate stock market is procyclical, $s_x > 0$, expected returns negatively link to past values of $x(t)$, i.e. $a_1 > 0$. This reasoning generalizes to a multivariate case, although the presence of feedbacks between macroeconomic variables might then dilute the contribution of each variable as a predictor of future expected returns.

Note that the nature of the regression results in Eq. (7.3) is the same as that in Eq. (7.2), for the simple reason that the price-dividend ratio is procyclical. Finally, note that at a contemporaneous level, excess returns are positively related to industrial production and negatively related to inflation,

$$\tilde{R}_{t-12,t}^e = \frac{10.47}{(1.07)} + \frac{7.27}{(1.19)} \times \text{IP}_t - \frac{16.33}{(2.91)} \times \text{Infl}_t + w_t, \quad \text{with } R^2 = 14\%, \quad (7.4)$$

with robust standard errors in parenthesis, and w is a residual term. Corradi, Distaso and Mele (2010) estimate a continuous time model where the aggregate stock price is driven by both industrial production and inflation, and one unobserved factor, and show that the links between returns and the two macroeconomic factors are similar to those summarized by the linear regression in Eq. (7.4).

Finally, an apparently puzzling feature is that price-dividend ratios do not predict future dividend growth. Let $g_t \equiv \ln(D_t/D_{t-1})$. In regressions taking the following format,

$$g_{t+n} = a_n + b_n \times P/D_t + u_{n,t}, \quad n \geq 1,$$

the predictive content of price-dividend ratios is poor, and estimates of b_n might often come with a wrong sign.

The previous regressions thus suggest that: (i) price-dividend ratios are driven by time-varying expected returns (i.e. by time-varying risk-premiums); and (ii) the role played by expected dividend growth is somewhat limited. As we shall see later in this chapter, this view can be challenged along several dimensions. First, it seems that expected *earning growth* does help predicting price-dividend ratios. Second, the fact expected dividend growth does not seem to affect price-dividend ratios can be a property to be expected in equilibrium.

Naturally, because expected returns and stock volatility are both strongly countercyclical, they then positively relate, at the business cycle frequency considered in this chapter, as illustrated by Figure 7.6 below.

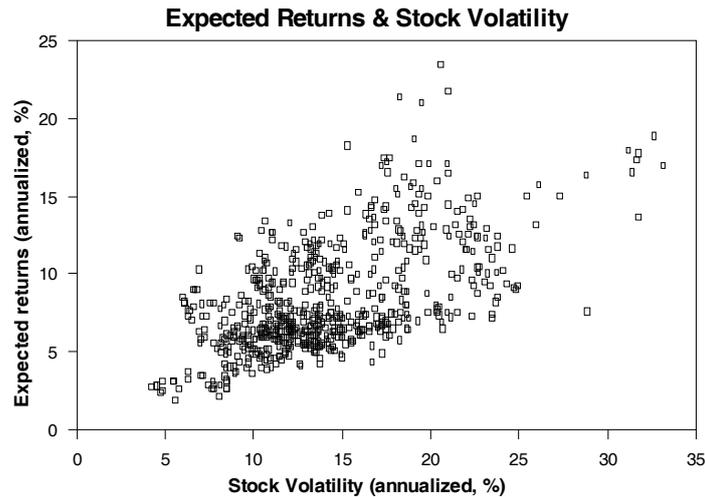


FIGURE 7.6.

7.3 Volatility: a business cycle perspective

A prominent feature of the U.S. stock market is the close connection between aggregate stock volatility and business cycle developments, as Figure 7.4 vividly illustrates. Understanding the origins and implications of these facts is extremely relevant to policy makers. Indeed, if stock market volatility is countercyclical, it must necessarily be encoding information about the development of the business cycle. Policy makers could then attempt at extracting the signals stock volatility brings about the development of the business cycle.

This section accomplishes three tasks. First, it delivers more details about stylized facts relating stock volatility, expected returns and P/D ratios over the business cycle (in Section 7.3.1). Second, it provides a few preliminary theoretical explanations of these facts (in Section 7.3.2). Third, it investigates whether stock volatility contains any useful information about the business cycle development (in Section 7.3.3). There are other exciting topics left over from this section. For example, we do not tackle statistical issues related to volatility measurement (see, e.g., Andersen, Bollerslev and Diebold, 2002, for a survey on the many available statistical techniques to estimate volatility). Nor do we consider the role of volatility in applied asset evaluation. Chapter 10, instead, provides details about how time-varying volatility affects derivative pricing. At a more fundamental level, the focus of this section is to explore the extent to which stock market volatility movements can be given a wider business cycle perspective, and to highlight some of the rational mechanisms underlying them.

7.3.1 Volatility cycles

Why is stock market volatility related to the business cycle? Financial economists seem to have overlooked this issue for decades. A notable exception is an early contribution by Schwert (1989a,b), who demonstrates how difficult it is to explain low frequency fluctuations in stock market volatility through low frequency variation in the volatility of other macroeconomic variables. A natural exercise at this juncture, is to look into the statistical properties of industrial production volatility and check whether this correlates with stock volatility. Accordingly, we compute industrial production volatility as, $\text{Vol}_{G,t} \equiv \frac{1}{\sqrt{12}} \sum_{i=1}^{12} |G_{t+1-i}|$, where G_t is the real,

seasonally adjusted industrial production growth rate as of month t , similarly as in Eq. (8.12). Figure 7.7 plots stock volatility against the volatility of industrial production growth, and does not reveal any statistically discernible pattern between these two variables. These results are in striking contrast with those available from Figure 7.4, where, instead, stock volatility exhibits a quite clear countercyclical behavior. More in detail, Table 7.1 reveals that stock market volatility is almost 30% higher during NBER recessions than during NBER expansions.

In fact, Schwert, also shows that stock volatility *is* countercyclical. The main focus of this section is to provide a few explanations for this seemingly puzzling evidence, in support of the view stock market volatility relates to the business cycle, although not precisely related to the *volatility* of other macroeconomic variables.

A seemingly separate, yet very well-known, stylized fact is that risk-premiums (i.e. the investors' expected return to invest in the stock market) are countercyclical (see, e.g., Fama and French, 1989, and Ferson and Harvey, 1991), as summarized by Fact II. Particularly important is also Fact III, that expected returns lower much less during expansions than they increase during recessions. Using post-war data, we find that compared to an average of 8.36%, the expected returns increase by nearly 19% during recessions and drop by a mere 3% during NBER expansions (see Table 7.1). A final stylized fact relates to the behavior of the price-dividend ratios over the business cycle. Table 7.1 reveals that not only are price-dividend ratios procyclical. Over the last fifty years at least, price-dividend ratios movements in the US have also been asymmetric over the business cycle: downward changes occurring during recessions have been more severe than upward movements occurring during expansions. Table 7.1 suggests that price-dividend ratios fluctuate nearly two times more in recessions than in expansions.

How can we rationalize these facts? A simple possibility is that the economy is frequently hit by shocks that display the same qualitative behavior of return volatility, expected returns and price-dividend ratios. However, the empirical evidence summarized in Figure 7.7 suggests this channel is unlikely. Another possibility is that the economy reacts to shocks, thanks to some mechanism endogenously related to the investors' maximizing behavior, which then activates the previous phenomena. The next section puts forward explanations for countercyclical stock volatility, which rely on such endogenous mechanisms. Section 7.3.3, instead, provides additional empirical results about cyclical properties of stock volatility. The motivation is simple: because stock volatility is countercyclical, it might contain useful information about ongoing business cycle developments. The section, then, aims to provide some answers to the following questions: (i) Do macroeconomic factors help explain the dynamics of stock market volatility? (ii) Conversely, what is the predictive content stock market volatility brings about the development of the business cycle? (iii) Finally, how does "risk-adjusted" volatility relate to the business cycle?

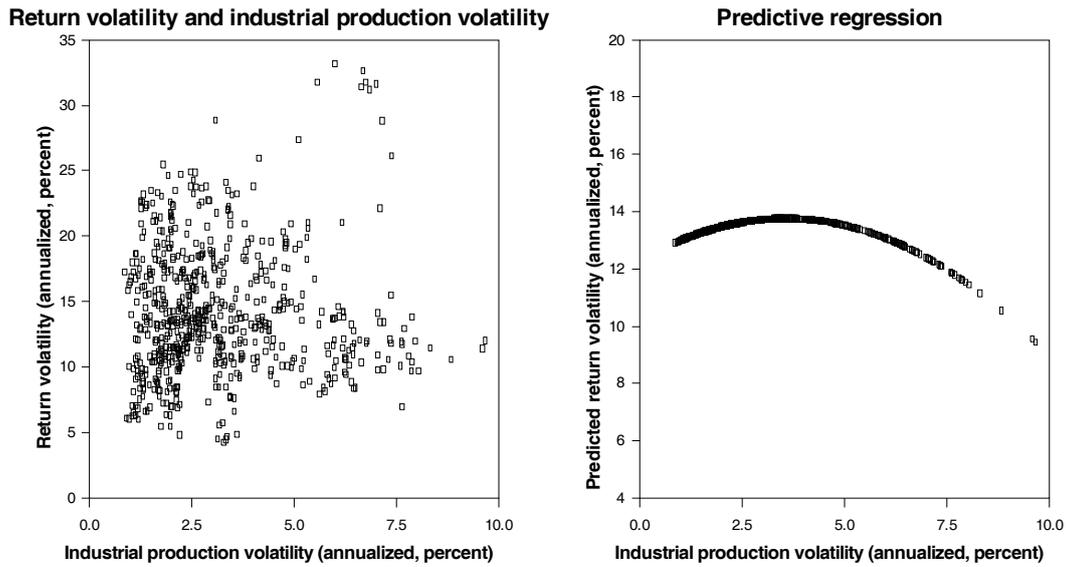


FIGURE 7.7. Return volatility and industrial production volatility. The left panel plots stock volatility, Vol_t , against industrial production volatility, $\text{Vol}_{G,t}$. The right panel of the picture depicts the prediction of the static least absolute deviations regression: $\text{Vol}_t = 12.28 - 0.83 \cdot \text{Vol}_{G,t} - 0.12 \cdot \text{Vol}_{G,t}^2 + w_t$, where w is a residual term, and standard errors are in parenthesis. The data span the period from January 1948 to December 2002.

7.3.2 Understanding the empirical evidence

This section aims to two tasks. First, it develops a simple example of an economy where countercyclical volatility arises in conjunction with the property that investors' required return are (i) countercyclical, and (ii) *asymmetrically* related business cycle development, an economy, that is, where risk-premiums increase more in bad times than they decrease in good, as suggested by the evidence in Table 7.1. Second, the section reviews additional plausible explanations for countercyclical volatility, where large price swings might relate to the investors' process of learning about the fundamentals of the economy. The aim of this section is to introduce to some of the main explanations of aggregate stock market fluctuations, which will be made deeper and deeper in the remaining parts of this and the following chapters.

7.3.2.1 Fluctuating compensation for risk

In frictionless markets, the price of a long-lived security and, hence, the aggregate stock market, is simply the risk-adjusted discounted expectation of the future dividends stream. Other things being equal, this price increases as the expected return from holding the asset and hence, the risk-premium, decreases. According to this mechanism, asset prices and price-dividend ratios are pro-cyclical because risk-adjusted discount rates are countercyclical.

Next, let us develop the intuition about why a countercyclical and asymmetric behavior of the risk-premium might lead to countercyclical volatility. Assume, first, that risk-premiums are countercyclical and that they decrease less in good times than they increase in bad times, consistently with the empirical evidence discussed in the previous section. Next, suppose the

economy enters a boom, in which case we expect risk-premiums to decrease and asset prices to increase, on average, as illustrated by Figure 7.8. The critical point is that during the boom, the economy is hit by shocks on the fundamentals, which makes risk-premiums and asset prices change. However, risk-premiums and, hence, asset prices, do not change as they would during a recession, since we are assuming that they behave asymmetrically over the business cycle. Then eventually, the boom ends and a recession begins. As the economy leads to a recession, the risk-premiums increase and asset prices decrease. Yet now, the shocks hitting the economy make risk-premiums and, hence, prices, increase more than they decreased during the boom. Once again, the reasons for this asymmetric behavior relate to our assumption that risk-premiums change asymmetrically over the business cycle.

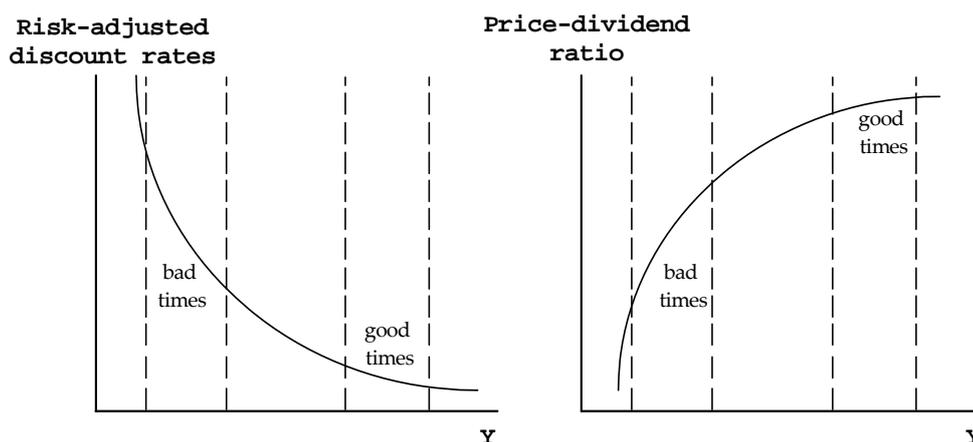


FIGURE 7.8. Countercyclical risk-premiums and stock volatility.

The empirical evidence in Table 7.1 is supportive of the channel described above: expected returns seem to move more during recessions than during expansions. Figure 7.9 connects such an asymmetric behavior of the expected returns with short-run macroeconomic fluctuations. It depicts how expected returns relate to the monthly Industrial production growth, according to whether the U.S. economy is in a booming or a recessionary phase.

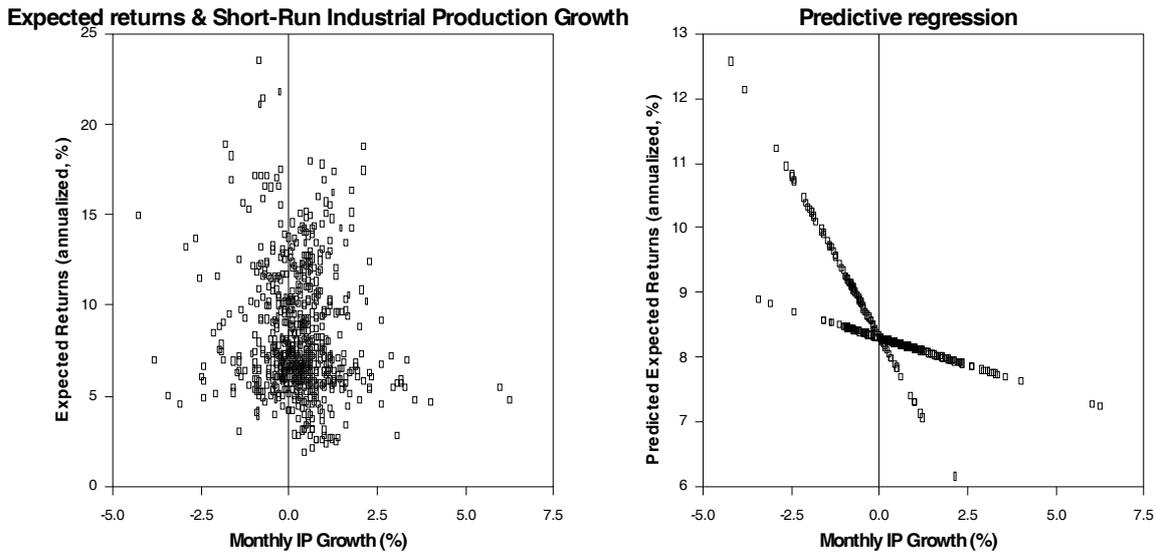


FIGURE 7.9. This picture is as Figure 7.5, except that it uses monthly IP growth. The predictive regression depicts the prediction of the Ordinary Least Squares: $\hat{\mathbb{E}} = 8.299 - (0.155) \mathbb{I}_{\text{recession}} \cdot 1.006 \cdot \text{Ind} - \mathbb{I}_{\text{expansion}} \cdot 0.169 \cdot \text{Ind} + w$, where $\mathbb{I}_{\text{recession}}$ (resp. $\mathbb{I}_{\text{expansion}}$) is the indicator function taking the value one if the economy is in a NBER-recession (resp. expansion) episode and zero otherwise, w is a residual term, and standard errors are in parenthesis.

To summarize, if risk-premiums are more volatile during recessions than booms, asset prices and, then, price-dividend ratios are more responsive to changes in economic conditions in bad times than in good, thereby leading to countercyclical volatility. These effects are precisely those we observe, as explained. The next section develops theoretical foundations for these facts, for a fairly general class of models with rational expectations, based on Mele (2007). A key result is that countercyclical volatility is likely to arise in many models, provided the previous asymmetry in discounting is sufficiently strong. More precisely, if the asymmetry in discounting is sufficiently strong, then, the price-dividend ratio is an increasing and *concave* function of some variables tracking the business cycle conditions. It is this concavity feature to make stock volatility increase on the downside. Under similar conditions, models with external habit formation predict countercyclical stock volatility along the same arguments (see, for example, Campbell and Cochrane, 1999; Menzly, Santos and Veronesi, 2004; Mele, 2007). Brunnermeier and Nagel (2007) find that US investors do not change the composition of their risky asset holdings in response to changes in wealth. The authors interpret their evidence against external habit formation. Naturally, time-varying risk-premiums do not exclusively arise through external habit formation. Barberis, Huang and Santos (2001) develop a theory distinct from habit formation, which leads to time-varying risk-premiums. The next chapter provides additional examples of economies where risk-premiums are time-varying are a result of alternative mechanisms, such as, say, market incompleteness.

These explanations of countercyclical volatility are elaborated further in the next section, within a fairly general continuous-time framework. While the tools of analysis of the next section are relatively unusual in economics, the intuition underlying the final results is that summarized by Figure 7.8. The scope of this section is to provide a quantitative illustration of these results, based upon a simple binomial tree model, which is solved in closed-form, and shown to predict a few of the stylized features of the aggregate market, surveyed in Section 7.2.

We consider an infinite horizon economy with a representative investor who in equilibrium consumes all the dividends promised by some asset. We assume that there exists a safe asset elastically supplied such that the safe interest rate is some constant $r > 0$. In the initial state, a dividend process takes a unit value (see Figure 7.10). In the second period, the dividend equals either $e^{-\delta}$ ($\delta > 0$) with probability p (the bad state) or e^{δ} with probability $1 - p$ (the good state). In the initial state, the investor's coefficient of constant relative risk-aversion (CRRA) is $\eta > 0$. In the good (resp., the bad) state, the investor's CRRA is η_G (resp., η_B) > 0 . In the third period, the investor receives the final payoffs depicted in Figure 7.10, where M_S is the price of a claim to all future dividends, discounted at a CRRA η_S , with $S \in \{G, B, GB\}$ and $\eta_{GB} = \eta$ (the “hybrid” state). This model is thus one with constant expected dividend growth, but random risk-aversion. Note that both random risk-aversion and dividend growth are acting as sources of “long-run risks”—once these risks are resolved, both risk-aversion and dividend growth remain fixed at their levels forever.

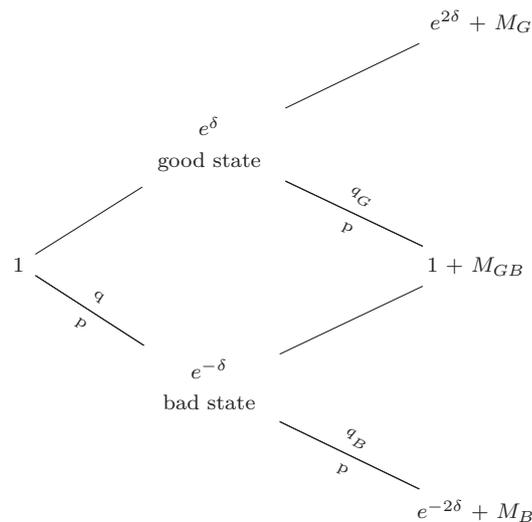


FIGURE 7.10. A tree model of random risk-aversion and countercyclical volatility. The dividend process takes a unit value at the initial node. With probability p , the dividend then decreases to $e^{-\delta}$ in the bad state. The corresponding risk-neutral probability is denoted as q . The risk-neutral probability of further dividends movements differs according to whether the economy is in the good or bad state (i.e. q_G or q_B). At the end of the tree, the investor receives the dividends plus the right to the stream of all future dividends. In the upper node, this right is worth M_G (the evaluation obtained through the risk-neutral probability q_G). In the central node it is worth M_{GB} (the evaluation obtained through the risk-neutral probability q). In the lower node it is worth M_B (the evaluation obtained through the risk-neutral probability q_B). The safe interest rate is taken to be constant.

The model is calibrated using the same U.S. data as in Table 7.1, and calibration results are in Table 7.2. Appendix A provides details about the solution and the calibration of the model. One important issue is that the calibration is made using data for aggregate dividend growth, which has a volatility around 6% annualized, almost the double as that on consumption growth. Consider, then, the simple calibration of the Lucas model in Section 6.2 of the previous chapter. As noted, this volatility would imply a relative risk-aversion of around $17 \approx \frac{0.06}{0.06^2}$, to match an equity premium of 6%. However, it was argued, the Lucas model of Chapter 6 cannot reproduce the high return volatility we observe in the data. In the simple model of the previous chapter, return volatility is simply dividend growth volatility and equals 6%, less than a half of the average volatility in the data, 14.55%. Nor would that model predict the countercyclical statistics in Table 7.1, as it predicts a constant price-dividend ratio.

	Data		
	expansions	average	recessions
P/D ratio	33.21	31.99	26.20
excess return volatility	14.05	14.55	16.91
	Model calibration		
	good state	average	bad state
P/D ratio	32.50	31.81	28.15
excess return volatility	7.29	8.20	13.03
risk-adjusted rate	8.95	9.07	9.71
expected returns	10.16	11.46	18.42
implied risk-aversion	13.69	13.89	14.96

TABLE 7.2. This table reports calibration results for the infinite horizon tree model in Figure 3. The expected returns and excess return volatility predicted by the model are computed using log-returns. The risk-adjusted rate is computed as $r + \hat{\sigma}_D \lambda_S$, where: r is the continuously compounded riskless rate; $\hat{\sigma}_D$ is the dividend volatility; λ_S is the Sharpe ratio on gross returns in state S , computed as $\lambda_S \equiv (q_S - p) \div \sqrt{p(1-p)}$ for $S = G$ (the good state) and $S = B$ (the bad state); p is the probability of the bad state; and q_S is the state dependent risk-adjusted probability of a bad state (for $S \in \{G, B\}$). Implied risk-aversion is the coefficient of relative risk aversion η_S in the good state ($S = G$) and in the bad state ($S = B$), implied by the calibrated model. The figures in the “average” column are the averages of the corresponding values in the good and bad states taken under the probability $p = 0.158$.

The model of this section can address these issues. First, it predicts return volatility to equal 8%, on average. Second, while the implied risk-aversion of the model is now around 11, its average expected excess returns are quite sustained. For example, fitting the Lucas model in the Section 6.2 of the previous chapter would lead to an implied risk-aversion of about $22 \approx \frac{0.08}{0.06^2}$, once we fix the expected excess returns to 8%. Third, the model reproduce swings of the stock volatility similar to those we observe in the data, with levels reaching 13% in the bad state of the world—although in this same state, the model might overstate expected returns by a few percentage points. Importantly, this calibration exercise illustrates in an exemplary manner

the asymmetric feature of expected returns and risk aversion. In this simple experiment, both expected returns and risk-aversion increase much more in bad times than they decrease in good.

7.3.2.2 Alternative channels

Rational explanations of stock market fluctuations must necessarily rely on some underlying state variable affecting the investors' decision environment. Two natural ways to accomplish this task are obtained through the introduction of (i) time-varying risk-premiums; and (ii) time-varying expected dividend growth. The previous tree model is one simple example addressing the first extension. More substantive examples of models predicting time-varying risk-premiums are the habit formation models mentioned in Section 7.3, and in Section 7.5 below.

Models addressing the second extension have also been produced. For example, Veronesi (1999, 2000) and Brennan and Xia (2001) have proposed models in which stock market volatility fluctuates as a result of a learning induced phenomenon. In these models, the growth rate of the economy is unknown and investors attempt to infer it from a variety of public signals. This inference process makes asset prices also depend on the investors' guesses about the dividends growth rate, and thus induces high return volatility. (In Veronesi, 1999, stock market volatility is also countercyclical.)

Finally, Bansal and Yaron (2004) formulate a model in which expected dividend growth is affected by some unobservable factor. This model, which will be discussed in detail in the next chapter, is also capable to generate countercyclical stock volatility. This property follows by the model's assumption that the volatilities of dividend growth and consumption are countercyclical. In contrast, in models with time-varying risk-premiums (such as the previous tree model), countercyclical stock market volatility emerges without the need to impose similar features on the fundamentals of the economy. Remarkably, in models with time-varying risk-premiums, countercyclical stock market volatility can be *endogenously* induced by rational fluctuations in the price-dividend ratio.

7.3.3 What to do with stock market volatility?

Both data and theory suggest that stock market volatility has a quite pronounced business cycle pattern. A natural purpose at this juncture is to exploit these patterns to perform some basic forecasting exercises. We consider three in-sample exercises. First, we forecast stock market volatility from past macroeconomic data (six month inflation, and six month industrial production growth). It is a natural exercise because stock volatility is an input to many decisions, ranging from portfolio decision to other decisions relating to risk-management. Second, we forecast industrial production growth from past stock market volatility, an exercise useful to policy makers, as further explained below. Third, we forecast the VIX index, an index of the risk-adjusted expectation of future volatility, from macroeconomic data, and attempt to measure the volatility risk-premium, defined as the amount of money a representative agent is willing to pay to be ensured against the event that future volatility will raise beyond his own expectations. Understanding the dynamics of the volatility risk-premium is, then, quite fundamental, as this is one of the most genuine measure of (lack of) risk appetite in capital markets. Finally, it is worth mentioning that the results of these forecasting exercises have to be interpreted with care. The next sections mention additional research that has elaborated on these exercises.

7.3.3.1 Macroeconomic constituents of stock market volatility

Table 7.3 reports the results for the first forecasting exercise. Volatility is positively related to past growth, a finding we can easily interpret. Bad times are followed by good times. Precisely, in our samples, high growth is inevitably followed by low growth. Since stock market volatility is countercyclical, high growth is followed by high stock market volatility. Stock market volatility is also related to past inflation, but in a more complex manner. Note that once we control for past values of volatility, the results remain significant. Figure 7.11 (top panel) depicts stock market volatility and its in-sample forecasts when the regression model is fed with *past* macroeconomic data only. This fit can even be improved through the joint use of both past volatility and macroeconomic factors such as industrial production and inflation. Nevertheless, it is remarkable that the fit from using past macro information is more than 60% better than just using past volatility, as witnessed by the R^2 s in Table 7.3. These results are somewhat in contrast with those Schwert (1989) reports. However, the results in this section hinge upon frequency scales lower than those in Schwert (1989).

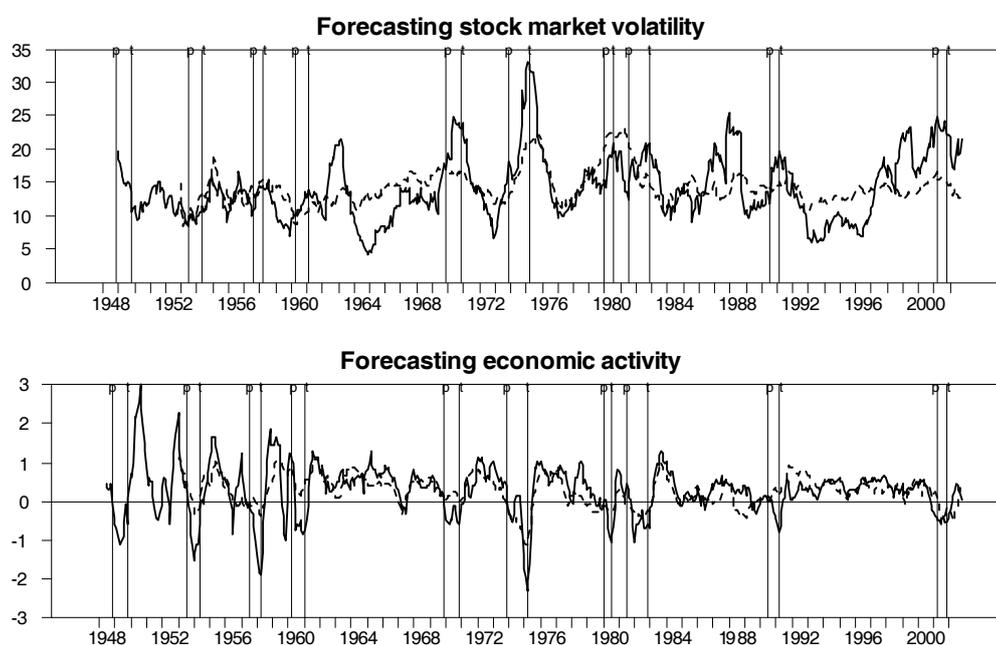


FIGURE 7.11. Forecasts. The top panel depicts stock market volatility (solid line) and stock market volatility forecasts obtained through the sole use of the macroeconomic indicators in Table 7.3 (dashed line). The bottom panel depicts 6-month moving average industrial production growth (solid line) and its forecasts based on the 6th regression in Table 7.4 (dashed line).

The previous findings, while certainly informal and preliminary, suggest that relating stock market volatility to macroeconomic factors might be a fertile avenue of research. The main question is how precisely stock market volatility relates to past macroeconomic factors. The regressions in Table 7.3 capture mere statistical relations between stock market volatility and macroeconomic factors. Yet, in the absence of arbitrage opportunities, stock market volatility is

certainly related to how the price responds to shocks in the fundamentals and, hence, macroeconomic conditions. Therefore, there should exist a no-arbitrage nexus between stock market volatility and macroeconomic factors. Corradi, Distaso and Mele (2010) pursue this topic in detail and build up a no-arbitrage model, which reproduces the previous predictability results. Christiansen, Schmeling and Schrimpf (2011) and Paye (2011) document that there is evidence of Granger causality from past values of several macroeconomic variables to stock volatility, in out of sample experiments, although we are still not able to exploit the relation linking these very same macroeconomic variables to stock volatility, for forecasting purposes. It is an important result, as it points to the possibility that in the future, alternative data sets might do a better job than the data set these authors are using.

The distinction between Granger causality and forecasting accuracy is subtle. A set of variables might well affect the probability distribution of stock volatility, which is the definition of Granger causality. At the same time, estimating, say, a linear regression linking past macroeconomic variables to stock volatility might not necessarily perform well. Intuitively, this relation can be subject to parameter estimation error, which increases the uncertainty surrounding the forecasts. Such an uncertainty might overwhelm the gain due to a bias reduction, due to a correctly specified model, without omitted variables (i.e. the macroeconomic variables), as illustrated more formally in Section 7.3.3.4. Statistical tests for Granger causality may rely on Clark and West (2007), and tests for forecasting accuracy may hinge upon Giacomini and White (2006).

	Past			Future	
Const.	6.92	7.76	2.48	Const.	8.28
Growth _{t-12}	–	0.29*	1.67	Growth _{t+12}	0.21*
Growth _{t-24}	–	0.74	1.09	Growth _{t+24}	1.62
Growth _{t-36}	–	2.17	2.44	Growth _{t+36}	–0.02*
Growth _{t-48}	–	1.77	1.91	Growth _{t+48}	0.12*
Infl _{t-12}	–	10.44	8.05	Infl _{t+12}	3.55
Infl _{t-24}	–	–5.96	–5.49	Infl _{t+24}	–0.81*
Infl _{t-36}	–	–1.42*	–0.97	Infl _{t+36}	–0.54*
Infl _{t-48}	–	3.73	3.31	Infl _{t+48}	4.33
Vol _{t-12}	0.43	–	0.37		
Vol _{t-24}	–0.17	–	–0.09		
Vol _{t-36}	0.02*	–	0.09		
Vol _{t-48}	0.12	–	0.09		
R ²	16.38	26.01	34.52	R ²	12.70

TABLE 7.3. Forecasting stock market volatility with economic activity. The first part of this table (“Past”) reports ordinary least square coefficient estimates in linear regression of volatility on to, *past* six month industrial production growth, *past* six month inflation, and *past* stock volatility. Growth_{t-12} is industrial production growth at time $t - 12$, etc. Time units are months. The second part of the table (“Future”) is similar, but it contains coefficient estimates in linear regressions of volatility on to *future* industrial production growth and *future* inflation. Starred figures are not statistically distinguishable from zero at the 95% level. R² is the percentage, adjusted R².

7.3.3.2 Macroeconomic implications of stock market volatility

Does stock market volatility also anticipate the business cycle? Fornari and Mele (2010) have tackled this issue, and concluded that stock volatility can help predict the business cycle. This issue is indeed quite a delicate one. Indeed, the fact stock volatility is countercyclical does not necessarily imply it anticipates real economic activity. And even if it anticipates it, there remains to know whether a sustained stock market volatility does really create the premises for future economic slowdowns. *Post hoc ergo propter hoc?* Does aggregate stock market volatility affect investment decisions in the real sector of the economy? Or, rather, does volatility help predict the business cycle? The policy implications of these issues are quite obvious. If volatility merely anticipates, without affecting, the business cycle, there is little policy makers can do about it, although of course its forecasting power is interesting per se. This theme is still unexplored.

Table 7.4 reports results obtained by regressing industrial production growth on to macroeconomic variables and return volatility (only R^2 s are reported). The volatility concept we use is purely related to volatility induced by price-dividend fluctuations (i.e. it is *not* related to dividend growth volatility). we find that the predictive power of traditional macroeconomic variables is considerably enhanced (almost doubled) with the inclusion of this new volatility concept and the price-dividend ratio. According to Figure 4 (bottom panel), stock market volatility does help predicting the business cycle. Fornari and Mele (2010) contain details on the forecasting performance of a new block, including stock market volatility and the slope of the yield curve. They show this block is quite successful and outperforms traditional models based on financial variables, both in sample and out of sample.

Predictors	R^2
(i) P/D Volatility	10.81
(ii) P/D ratio	15.57
(iii) P/D Volatility, P/D ratio	20.98
(iv) Growth, Inflation	21.20
(v) Growth, Inflation, P/D volatility	34.29
(vi) Growth, Inflation, P/D volatility, P/D ratio	41.76

TABLE 7.4. Forecasting economic activity with stock market volatility. This table reports the R^2 (adjusted, in percentage) from six linear regressions of 6 month moving average industrial production growth on to the listed set of predicting variables. Inflation is also 6 month moving average inflation. The regressor lags are 6 months, and 1, 2 and 3 years. P/D volatility is defined as a 12 month moving average of $abs(\ln(\frac{1+P/D_{t+1}}{P/D_t}))$, where $abs(\cdot)$ denotes the absolute value, and P/D is the price-dividend ratio.

7.3.3.3 Risk-adjusted volatility

Volatility trading

An important innovation for volatility trading was the introduction of the “variance swaps” during the beginning of the 2000s. Variance swaps are contracts allowing to trade future realized variance against a fixed swap rate. They allow to take pure views about volatility movements, without incurring into price-dependency issues arising from trading volatility through straddles, as we shall explain in detail in Chapter 10, which shall also explain the trading rationale underlying these contracts. All in all, the payoff guaranteed to the buyer of a swap equals the

difference between the realized volatility over the life of the contract and a fixed swap rate. Entering this contract at time of origination does not cost. Therefore, the fixed swap rate is equal to the expectation of the future realized volatility under the risk-neutral probability. In September 2003, the Chicago Board Options Exchange (CBOE) started to calculate the VIX index in a way that makes this index equal to such a risk-neutral expectation. The strength of this new index is that although it deals with risk-neutral expectations, it is nonparametric—it does not rely on any model of stochastic volatility. Precisely, it is based on a basket of all the available option prices, relying on the seminal work by Demeterfi et al. (1999), Bakshi and Madan (2000), Britten-Jones and Neuberger (2000), and Carr and Madan (2001).

Business cycle determinants of volatility trading

Figure 7.5 (top panel) depicts the VIX index, along with predictions obtained through a parametric model. The predicting model is based on the regression of the VIX index on the same macroeconomic variables considered in the previous sections: inflation and growth. Table 7.5 reports the estimation results, which reveal how important the contribution of macroeconomic factors is to explain the dynamics of the VIX.

	Past			Future	
Const.	2.60*	30.15	3.03*	Const.	25.53
Growth _{t-1}	–	–5.12	0.51*	Growth _{t+1}	25.53
Growth _{t-12}	–	–3.69*	–0.35*	Growth _{t+12}	–5.58
Growth _{t-24}	–	4.91	3.69	Growth _{t+24}	–8.34
Growth _{t-36}	–	11.19	4.33	Growth _{t+36}	–9.67
Infl _{t-1}	–	–26.96	–9.14*	Infl _{t+1}	1.04*
Infl _{t-12}	–	–22.62	–1.89*	Infl _{t+12}	–24.11
Infl _{t-24}	–	–1.59*	5.85*	Infl _{t+24}	9.32*
Infl _{t-36}	–	–6.02*	–2.56*	Infl _{t+36}	20.71
VIX _{t-1}	0.72	–	0.55		
VIX _{t-12}	0.18	–	0.14*		
VIX _{t-24}	–0.06*	–	–0.01*		
VIX _{t-36}	0.02*	–	0.12*		
R ²	66.87	54.12	71.03	R ²	55.04

TABLE 7.5. Forecasting the VIX index with economic activity. The first part of this table (“Past”) reports ordinary least square coefficient estimates in linear regression of the VIX index on to, *past* industrial production growth (as defined in Figure 1), *past* inflation (defined similarly as in Figure 1), and *past* volatility. Growth_{t-12} is industrial production growth at time $t - 12$, etc. Time units are months. The second part of the table (“Future”) is similar, but it contains coefficient estimates in linear regression of the VIX index on to *future* industrial production growth and *future* inflation. Starred figures are not statistically distinguishable from zero at the 95% level. R² is the percentage, adjusted R².

Figure 7.5 (bottom panel) depicts the volatility risk-premium, defined as the difference between the expectation of future volatility under the risk-neutral and the physical probability.

We estimated the risk-neutral expectation as the predicting part of the linear regression of the VIX index on the macroeconomic factors (inflation and growth only)—the dotted line in Figure 7.5, top panel. We estimated expected volatility as the predicting part of an AR(1) model fitted to the volatility depicted in Figure 4 (top panel). As we see, volatility risk-premiums are indeed strongly countercyclical. Once again, the results in these picture are suggestive, but they do represent mere statistical relations. The model considered by Corradi, Distaso and Mele (2010) has the strength to make these statistical relations emerge as a result of a fully articulated no-arbitrage model.

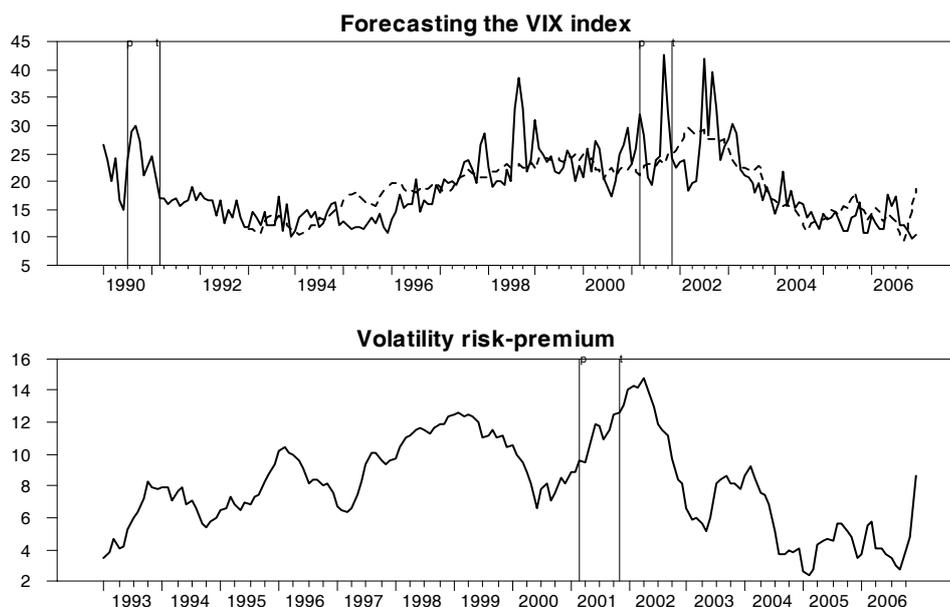


FIGURE 7.12. Forecasting the VIX Index, and the volatility risk-premium. The top panel depicts the VIX index (solid line) and the VIX forecasts obtained through the sole use of the macroeconomic indicators in Table 7.3 (dashed line). The bottom panel plots the volatility risk premium, defined as the difference between the one-month ahead volatility forecast calculated under the risk-neutral probability and the one-month ahead volatility forecast calculated under the physical probability.

7.3.3.4 Forecasting with the wrong model

The results in this section are in-sample, and it might turn out that real-time forecasts can be quite disappointing. One reason could be data-snooping: if we regress a variable of interest over thousands, there is a considerable chance that at least one out of these thousands nicely links to the endogenous variable, leading to a spectacular fit, in-sample, but only by chance, not because of a real economic link between this variable and the endogenous one. If this happens, then, naturally, the out-of-sample performance of the model can only be expected to disappoint. However, an opposite case may occur, where a link between two variables does really exist, but cannot be properly exploited in finite samples. Intuitively, we need to estimate this link using a finite sample, and the resulting finite-sample bias might turn to be substantial, leading to large forecasting errors. We develop an example to illustrate this point. Consider a data generating

process where a variable x_t Granger causes a second one, y_t , as follows:

$$y_t = c + \beta x_t + \epsilon_t, \quad \epsilon_t \sim \text{NID}(0, \omega_\epsilon), \quad x_t \sim \text{NID}(\mu_x, \omega_x), \quad (7.5)$$

for five constants c , β , ω_ϵ , μ_x and ω_x , the parameters of the model. We assume that μ_x and ω_x are known, and consider making predictions of the variable y_t through two models. The first model is misspecified, one where we simply neglect that x_t Granger causes y_t , i.e. $y_t = \psi + u_t$, for some constant ψ and some residual term u_t . We estimate the constant ψ of this misspecified model through ordinary least squares (OLS), obtaining:

$$\hat{\psi} - c = \beta \bar{x}_T + \bar{\epsilon}_T,$$

where \bar{x}_T and $\bar{\epsilon}_T$ denote the sample averages of x_t and ϵ_t , and T is the sample size. The prediction error generated by this model for time $T + 1$ is:

$$\eta_{1,T+1} \equiv y_{T+1} - \hat{\psi} = \beta (x_{T+1} - \bar{x}_T) + \epsilon_{T+1} - \bar{\epsilon}_T.$$

Note that although $\hat{\psi}$ is biased for c , even asymptotically, the predictor of this misspecified model is unbiased, as we have that $E(\eta_{1,T+1}) = 0$, where E denotes the unconditional expectation operator.

Next, consider using as a predictor, the predictive part of Eq. (7.5), obtained through the OLS estimators of c and β , say \hat{c} and $\hat{\beta}$. The resulting prediction error is,

$$\begin{aligned} \eta_{2,T+1} &\equiv y_{T+1} - \hat{c} - \hat{\beta} x_{T+1} \\ &= c - \hat{c} + (\beta - \hat{\beta}) x_{T+1} + \epsilon_{T+1} \\ &= (\beta - \hat{\beta}) (x_{T+1} - \bar{x}_T) + \epsilon_{T+1} - \bar{\epsilon}_T, \end{aligned} \quad (7.6)$$

where the second equality follows by (i) $\hat{c} = \bar{y}_T - \hat{\beta} \bar{x}_T$, with \bar{y}_T denoting the sample average of y_t , and (ii) Eq. (7.5), and:

$$\beta - \hat{\beta} = -\frac{\text{Cov}_T(x_t, \epsilon_t)}{V_T(x_t)},$$

and Cov_T and V_T stand for the sample covariance and variance of their arguments. The correctly specified model does, naturally, lead to an unbiased predictor, in that $E(\eta_{2,T+1}) = 0$, by the second line in Eq. (7.6). Therefore, the two models we consider—the misspecified and the correctly specified—both lead to unbiased predictors. However, the second predictor is plagued by parameter estimation error, and might actually lead to mean-squared prediction errors higher than those generated by the first predictor, especially when the the sample variance of $\beta - \hat{\beta}$ is large. In other words, for large samples, $\beta - \hat{\beta}$ is, of course, quite small, as $\hat{\beta}$ is consistent for β . In finite samples, however, this term can adversely affect the performance of the correctly specified model.

7.3.4 What did we learn?

Stock market volatility is higher in bad times than in good times. Explaining this basic fact is challenging. Indeed, economists know very well how to model risk-premiums and how these premiums should relate to the business cycle. We feel more embarrassed when we come to explain volatility. The ambition in this short essay is to explain that countercyclical volatility can be made consistent with the prediction of the neoclassical model of asset pricing - in

which asset prices are (risk-adjusted) expectations of future dividends. One condition activating countercyclical volatility is very simple: risk-premiums must swing sharply as the economy moves away from good states, just as the data seem to suggest.

The focus in this section relates to the *fluctuations* of aggregate stock volatility and risk premiums, not their *average levels*. Not surprisingly, the question whether these fluctuations can be consistent with rational evaluation is controversial, as for many topics at the intersection of financial economics and macroeconomics, as vividly illustrated in the old, early debate (see, e.g., Campbell, 2003; Mehra and Prescott, 2003). However, the simple model in this section suggests the neoclassical model might predict the swings that aggregate stock volatility experiences across states of nature. Do these theoretical insights have some additional empirical content? This section has discussed three empirical issues, which form the focus of ongoing research: (i) stock volatility links to the business cycle, in-sample, although it cannot necessarily be forecast through macroeconomic variables, out-of-sample; (ii) stock market volatility does contain relevant information related to business cycle developments; and (iii) the price of volatility does relate to the business cycle and, volatility risk-premiums are strongly countercyclical.

7.4 Rational market fluctuations

We explore mechanisms through which rational valuation could explain the countercyclical statistics described so far, by relying on a fairly general, yet parsimonious framework of analysis. The next section explains that the burden of the explanations should rely on the dynamic behavior of the aggregate price-dividend ratio. Section 7.4.2 develops the core tools of analysis, and provides broad examples where these tools will be applied in the remainder of this chapter.

7.4.1 The dynamics of asset returns

7.4.1.1 An asset return decomposition

Asset returns depend on developments of both payoffs *and* prices. Formally, consider the following identity, holding for the gross returns, $\tilde{R}_{t+1} \equiv \frac{S_{t+1} + D_{t+1}}{S_t}$,

$$\ln \tilde{R}_{t+1} \equiv g_{t+1} + \ln \left(\frac{p_{t+1} + 1}{p_t} \right), \quad (7.7)$$

where $g_t \equiv \ln \left(\frac{D_t}{D_{t-1}} \right)$, the dividend growth, and $p_t \equiv \frac{S_t}{D_t}$, the price-dividend ratio. The empirical evidence in Section 7.2 suggests that any model should make two minimal predictions about the aggregate price-dividend ratio. First, it needs to be volatile, and second, it needs to be more volatile in bad times than in good. Consider, for example, an economy where among other things, asset prices are driven by a state variable relating to the business cycle, as in the habit formation models of Section 7.5. A basic property we require from this particular model is that the price-dividend ratio be increasing and *concave* in the business cycle variable, as further explained in Section 7.3.2. Intuitively, this property ensures stock volatility increases on the downside—the very definition of countercyclical volatility.

7.4.1.2 Asymmetric behavior of the price-dividend ratio

How can we make sure the aggregate price-dividend ratio behaves asymmetrically over the business cycle? Assume the price-dividend ratio is driven by a vector of state variables y and,

everything that is needed, to have the following representation of the price-dividend ratio, as derived in Section 4.2.3 of Chapter 4:

$$p(y(t)) = \mathbb{E} \left[\int_t^\infty \frac{D_*(\tau)}{D(t)} \cdot e^{-\int_t^\tau \text{Disc}(y(u))du} \middle| y(t) \right], \quad \frac{D_*(\tau)}{D(t)} = e^{(g_0 - \frac{1}{2}\sigma_D^2)(\tau-t) + \sigma_D(\hat{W}(\tau) - \hat{W}(t))}, \quad (7.8)$$

where \mathbb{E} denotes the expectation under the risk-neutral probability, and $\text{Disc}(y(t)) = r(y(t)) + \sigma_D \lambda_{\text{CF}}(y(t))$ are what we termed *risk-adjusted discount discount rates*, an affine function of the short-term rate $r(y)$ and the cash-flow lambda $\lambda_{\text{CF}}(y)$.

Eq. (7.8) suggests the intuitive property that the sensitivity of the discount rates to $y(t)$ affects that of the price-dividend ratio and, then, return volatility. This section formalizes this intuition and shows that if the risk-adjusted discount rates increase in bad times sufficiently more than they decrease in bad, the price-dividend ratio is indeed concave in the variable y that correlates with the business cycle. Note that this search process makes precise the notion that we need additional state variables driving fluctuations in the price-dividend ratio. As explained in Chapter 6, multifactor model might not necessarily lead to the properties we observe in the data: we can easily imagine models where: (i) the variance of the pricing kernel can be arbitrarily increased, by adding more and more factors, and where (ii) price-dividend ratios are still constant. What we really need is a discipline on how to increase the dimension of a model, a task we pursue in detail in the remainder of this chapter. The tools the next section develops aim to uncover two types of properties, which can be streamlined into two categories: (i) “monotonicity” and (ii) “convexity”:

- (i) *Monotonicity.* Consider the price-dividend ratio in Eq. (7.8). By Itô’s lemma, stock volatility is $\sigma_D + \frac{p'(y)}{p(y)} \text{Vol}(y)$, where $\text{Vol}(y)$, the volatility of y , might help inflate stock volatility, should p be increasing in y . Such a monotonicity property is also important for a theoretical reason, as it ensures stock volatility is strictly positive, a crucial condition guaranteeing the agents’ budget constraints are well-defined.
- (ii.1) *Negative convexity.* If Y is a state variable related to the business cycle, and $\text{Vol}(y)$ is constant, stock volatility is countercyclical whenever p in Eq. (7.8) is *concave* in y , as in the simple reasoning underlying Figure 7.8.
- (ii.2) *Convexity.* Alternatively, suppose that expected dividend growth, g say, is stochastic, an assumption we explore in detail in Section 7.5. If p is increasing and *convex* in g , the price-dividend ratio would display “overreaction” to small changes in g in good times, i.e. when g is high. The empirical relevance of this point was first recognized by Barsky and De Long (1990, 1993), and formalized by Veronesi (1999) in a model with learning.

To clarify these properties, the next section revisits the option pricing literature about convexity properties of prices, in a context with risks, which are not necessarily traded.

7.4.2 Volatility, options and convexity

7.4.2.1 Proportional markets

Consider a two-period market, one for a right to receive a cash premium ψ at the second period. Assume that interest rates are zero, that the cash premium is a function of some random variable \tilde{y} , $\psi = \psi(\tilde{y})$, and denote the price of the of this right with $\bar{c} \equiv \mathbb{E}[\psi(\tilde{y})]$. What is the relation

between the volatility of \tilde{y} and \bar{c} ? Second-order stochastic dominance suggests \bar{c} is inversely related to *mean preserving* spreads in \tilde{y} , provided ψ is concave, as reviewed in the Appendix 4. Intuitively, a concave function “exaggerates” poor realizations of \tilde{y} and “dampens” the favorable ones.

Do these properties hold in a dynamic setting? Consider a continuous time extension of the previous risk-neutral environment. Assume the cash premium ψ is paid off at some future date T , and that $\tilde{y} = y(T)$, where Y is some underlying state process, with $y(0) = y$. If the yield curve is flat at zero, $c(y) \equiv \mathbb{E}[\psi(y(T))|y]$ is the price of the right. Clearly, the two pricing problems, $\mathbb{E}[\psi(y(T))|y]$ and $\mathbb{E}[\psi(\tilde{y})]$, are quite distinct. However, there are analogies. First, if y is a proportional process, i.e. one for which the risk-neutral distribution of $\frac{y(T)}{y}$ is independent of y , then,

$$c(y) = y \cdot \mathbb{E}[\psi(G(T))], \quad G(T) \equiv \frac{y(T)}{y}, \quad y > 0. \quad (7.9)$$

This simple formula suggests that standard stochastic dominance arguments still apply. That is, c decreases (resp. increases) after a mean-preserving spread in G whenever ψ is concave (resp. convex), consistently with the prediction of the Black and Scholes (1973) formula, as we further explain in Chapter 10—a point made by Jagannathan (1984, p. 429-430). In two independent papers, Bergman, Grundy and Wiener (1996) and El Karoui, Jeanblanc-Picqué and Shreve (1998) generalize these results to *any* diffusion process, not necessarily a proportional process. Bajeux-Besnainou and Rochet (1996, Section 5) and Romano and Touzi (1997) contain further extensions pertaining to stochastic volatility models.²

7.4.2.2 A canonical economy

The extensions mentioned in the previous subsections rely on the assumption Y is the price of a traded asset, which does not pay dividends. The assumption is crucial, for it makes the risk-neutralized drift of Y linear in y , such that the price of the claim in this context, c , inherits convexity properties from the final payoff only, ψ , as we shall clarify. This section studies markets with nontradable risks, where interesting nonlinearities arise. For example, Theorem 7.1 reveals that in the context of nontradable risks, the convexity of ψ is neither a necessary or a sufficient condition for the convexity of c . An important role in this context is that of the drift of the state variables underlying the markets.³ Moreover, “dynamic” stochastic dominance properties (those linking the volatility of the fundamentals to the price of contingent claims) are more intricate than in the classical second order stochastic dominance theory, as explained in the Appendix (see Theorem 7.A.4).

We substantiate these claims by hinging upon the following pricing problem, a benchmark for a variety of issues dealt with in this chapter.

²The proofs in these two articles are markedly distinct but they both rely on the convexity of the *price function*. We may consider an alternative proof, which directly hinges upon the convexity of the *payoff function*, and a result due to Hajek’s (1985). This result says that if ψ is increasing and convex, and x_1 and x_2 are two diffusion processes, both starting off from the same origin, with integrable drifts b_1 and b_2 and volatilities a_1 and a_2 , then, $E[\psi(x_1(\tau))] \leq E[\psi(x_2(\tau))]$, whenever $b_1(t) \leq b_2(t)$ and $a_1(t) \leq a_2(t)$ for all $t \in (0, \tau)$. This result allows for a more general approach than that in Bergman, Grundy and Wiener (1996) and El Karoui, Jeanblanc-Picqué and Shreve (1998), as it considers shifts in both b and a , which are particularly relevant thought-experiments in finance. Note that Hajek’s result generalizes the classic comparison theorem as given by Karatzas and Shreve (1991, p. 291-295), where ψ is an increasing function and $a_1 \equiv a_2$.

³Kijima (2002) produces a counterexample where convexity of option prices might break down even payoffs are convex in the underlying, and *traded*, assets. This counterexample relies on an extension of the Black-Scholes model where due to the presence of dividends, the drift of the underlying asset is concave in the asset price. Among other things, Theorem 7.1 unveils the origins of this counterexample.

CANONICAL PRICING PROBLEM. Let Y be the solution to:

$$dy(\tau) = b(y(\tau)) d\tau + a(y(\tau)) d\bar{W}(\tau), \quad (7.10)$$

where \bar{W} is a multidimensional \bar{P} -Brownian motion (for some \bar{P}), and b, a are some given functions. Let ψ and ρ be two twice continuously differentiable positive functions, and define

$$c(y, T) \equiv \bar{\mathbb{E}} \left[e^{-\int_0^T \rho(y(t)) dt} \cdot \psi(y(T)) \middle| y \right] \quad (7.11)$$

to be the price of an asset which promises to pay $\psi(y(T))$ at time T .

A simple market encompassed by our canonical pricing problem is one where Y is the price of a traded asset, and $\bar{P} = Q$, the risk-neutral probability, such that the drift in Eq. (7.10) is $b(y) = y\rho(y)$, by no-arbitrage. If Y is not a traded risk, $b(y) = b_0(y) - a(y)\lambda(y)$, where b_0 is the physical drift function of Y and λ is a risk-premium. Our canonical pricing problem now covers a number of interesting cases. For example, we may assume $\psi(y) = 1$, $\rho(y) = y$, and y is a short-term rate, in which case c is the price of a zero-coupon bond price.

Another example falling under the canonical pricing problem is one of a “scale-invariant economy” of the type of Eq. (7.9), in the following sense. Consider an endowment economy with a single risky asset, where a single consumption good, denoted with D , is solution to:

$$\begin{cases} \frac{dD(\tau)}{D(\tau)} = m(y(\tau)) d\tau + \sigma_0 dW_1(\tau) \\ dy(\tau) = \varphi(y(\tau)) d\tau + v_1(y(\tau)) dW_1(\tau) + v_2(y(\tau)) dW_2(\tau) \end{cases} \quad (7.12)$$

where W_1 and W_2 are standard Brownian motions, and y is some state variable, and m, φ, v_i are some functions: m is interpreted as the expected dividend growth, and φ, v_i are the drift of a state variable affecting this growth and, possibly, the pricing kernel, as we shall put forward below. Note that in this model, we take consumption volatility to be constant, σ_0 . It is easy to see that the distribution of $\frac{D(\tau)}{D}$ does not depend upon the initial value D . We also assume that the short-term interest rate, r , does not depend on D . By no arbitrage, and under regularity conditions, the price-dividend ratio for a *long-lived* security, p , satisfies:

$$p(y) = \int_0^\infty c(y, \tau) d\tau, \quad (7.13)$$

where

$$\begin{aligned} c(y, \tau) &\equiv \mathbb{E} \left[e^{-\int_0^\tau r(y(t)) dt} \cdot \frac{D(\tau)}{D} \middle| y \right] \\ &= \mathbb{E} \left[e^{\int_0^\tau m(y(t)) dt} \cdot e^{-\int_0^\tau \text{Disc}(y(u)) du} \left(e^{-\frac{1}{2}\sigma_0^2\tau + \sigma_0\hat{W}(\tau)} \right) \middle| y \right] \\ &= \bar{\mathbb{E}} \left[e^{\int_0^\tau m(y(t)) dt} \cdot e^{-\int_0^\tau \text{Disc}(y(u)) du} \middle| y \right], \end{aligned} \quad (7.14)$$

and \mathbb{E} is the expectation taken under the risk-neutral probability, under which D and y are diffusions with drifts $m(y) - \sigma_0\lambda_1(y)$ and $\varphi(y) - \sum_{i=1}^2 v_i(y)\lambda_i(y)$,⁴ and as usual, $\text{Disc}(y) =$

⁴See, for example, Huang and Pagès (1992, Theorem 3 p. 53) and Wang (1993, Lemma 1, p. 202), for regularity conditions underlying the Feynman-Kac theorem in infinite horizon settings; and Huang and Pagès (1992, Proposition 1, p. 41) for regularity conditions ensuring that the Girsanov’s theorem holds in infinite horizon settings.

$r(y) + \sigma_0 \lambda_{CF}(y)$, the *risk-adjusted discount discount rates*. Finally, the third line is obtained with a change of probability, and $\bar{\mathbb{E}}$ is the expectation taken under a conveniently changed probability \bar{Q} , defined by the Radon-Nikodym derivative,

$$\left. \frac{d\bar{Q}}{dQ} \right|_{F(\tau)} = e^{-\frac{1}{2}\sigma_0^2\tau + \sigma_0\hat{W}_1(\tau)}, \quad (7.15)$$

where $F(\tau)$ denotes the information set as of time τ generated by $\hat{W}_1(\tau)$, a Brownian motion under Q . By Girsanov theorem, we have that under \bar{Q} ,

$$\begin{aligned} dy(\tau) &= \bar{\varphi}(y(\tau)) d\tau + v_1(y(\tau)) d\bar{W}_1(\tau) + v_2(y(\tau)) d\bar{W}_2(\tau) \\ \text{where } \bar{\varphi}(y) &\equiv \varphi(y) - \sum_{i=1}^2 v_i(y) \lambda_i(y) + \sigma_0 v_1(y) \end{aligned} \quad (7.16)$$

and \bar{W}_i are Brownian motions under \bar{Q} . Mele (2005) considers a multivariate framework along these lines. Note the trick we have used to arrive to a relatively neat formula, by getting rid of term, $e^{-\frac{1}{2}\sigma_0^2\tau + \sigma_0\hat{W}_1(\tau)}$, arising because consumption and the state variable y are correlated. The density of y under \bar{Q} is right-shifted with respect to the same density under Q , due to the positive covariance between consumption growth and $dy(\tau)$.

The representation of the price-dividend ratio in Eqs. (7.13)-(7.14)-(7.16) relies on a framework slightly more general than that underlying Eq. (7.8), as it allows for stochastic growth. Our canonical pricing problem allows us to analyze properties of prices relating to long-lived assets, through those relating to c in Eq. (7.14), with y as in Eq. (7.16), once we set

$$\psi(y) \equiv 1; \quad \rho(y) \equiv \text{Disc}(y) - m(y); \quad \bar{\varphi}(y) = b(y). \quad (7.17)$$

The next theorem characterizes slope and convexity properties of the price c in the canonical pricing problem.

THEOREM 7.1. *We have:*

(i) *If $\psi' > 0$, then c is increasing whenever $\rho' \leq 0$. Furthermore, if $\psi' = 0$, then c is decreasing (resp. increasing) whenever $\rho' > 0$ (resp. < 0).*

(ii) *If $\psi'' \leq 0$ (resp. $\psi'' \geq 0$) and c is increasing (resp. decreasing), then c is concave (resp. convex) whenever $b'' < 2\rho'$ (resp. $b'' > 2\rho'$) and $\rho'' \geq 0$ (resp. $\rho'' \leq 0$). Finally, if $b'' = 2\rho'$, c is concave (resp. convex) whenever $\psi'' < 0$ (resp. > 0) and $\rho'' \geq 0$ (resp. ≤ 0).*

Theorem 7.1-(i) generalizes previous results about monotonicity of option prices, obtained by Bergman, Grundy and Wiener (1996). By the so-called “no-crossing property” of a diffusion, Y is not decreasing in its initial condition y . Therefore, c inherits the same monotonicity features of ψ if discounting does not operate adversely. This simple observation allows us to address monotonicity properties of long-lived asset prices, as we shall see in Section 7.5.

Theorem 7.1-(ii) generalizes a number of existing results on option price convexity. First, assume that ρ is constant and that Y is the price of a traded asset, such that $\rho' = b'' = 0$. The last part of Theorem 7.1-(ii) then says that the convexity of ψ propagates to the convexity of c . This result reproduces the findings in the literature surveyed earlier. Theorem 7.1-(ii) characterizes option price convexity within more general contingent claims models. As an example, suppose that $\psi'' = \rho' = 0$, and that Y is not a traded risk. Then, Theorem 7.1-(ii) suggests that c inherits the convexity properties of the drift of Y . As a final example, Theorem 7.1-(ii) extends a result in Mele (2003) relating to bond pricing: let $\psi(y) = 1$ and $\rho(y) = y$. Accordingly, c is

the price of a zero-coupon bond predicted by a short-term rate model, such as those we shall deal with in Chapter 11. By Theorem 7.1-(ii), then, c is convex whenever $b'' < 2$ (see Appendix 6 for further details and intuition on this bounding number).

Properties of asset prices with non-traded fundamentals, such as stock prices, heavily rely on both discounting and nonlinearities affecting the drift of the state variables. Section 7.5 shall systematically rely on the predictions of Theorem 7.1, to analyze the price of long-lived assets. In the next section, we illustrate the gist of the proofs underlying Theorem 7.1, by developing one example.

7.4.2.3 A digression on a “macro-asset” option

We discuss an example, that of a highly conceptual and abstract asset, a “macro-asset” option, and illustrate a few facts Theorem 7.1 can predict. Let $c(t)$ be the aggregate consumption process. The owner of the option has the right to receive a payoff $\psi(c(T))$, $\psi \in \mathcal{C}^2$, at some date T , where ψ is increasing and convex. We assume that $c(t)$ is solution to:

$$\frac{dc(t)}{c(t)} = g(t) dt,$$

where the consumption growth rate $g(t)$ satisfies

$$dg(t) = \varphi(g(t)) dt + v(g(t)) dW(t),$$

where φ and v are some well-behaved functions, and W is a standard Brownian motion. Let $p(c, g, t)$ be the rational price of the option when the state of the economy as of time $t \in [0, T]$ is $c(t) = c$ and $g(t) = g$. Let $p \in \mathcal{C}^{2,2,1}$. Assume that interest rates are constant and that all agents are risk-neutral.

By the usual connection between partial differential equations and conditional expectations, the price $p(c, g, t)$ is solution to the following partial differential equation:

$$0 = \frac{\partial p}{\partial t} + gcp_c + \frac{1}{2}v^2p_{gg} + \varphi p_g - rp, \quad \text{for all } c, g \text{ and } t \in [0, T], \quad (7.18)$$

with boundary condition $p(c, g, T) = \psi(c)$, all c, g , where subscripts denote partial differentiation. Monotonicity properties of the price function $p(c, g, t)$, with respect to both c and g , can be understood through two approaches. The first approach relies on the so-called *no-crossing property* of diffusion processes, and proceeds as follows. We have:

$$p(c, g, t) = e^{-r(T-t)} E \left[\psi \left(c(t) \cdot e^{\int_t^T g(u) du} \right) \middle| c(t) = c, g(t) = g \right]. \quad (7.19)$$

Since ψ is increasing, p is increasing in c as well. Furthermore, the no-crossing property of g implies that $g(u)$ is increasing in the initial condition $g(t)$. Therefore, $p(c, g)$ is also increasing in g .

To analyze convexity of p with respect to g , differentiate Eq. (7.18), and its boundary condition, with respect to c , and find that $w \equiv p_c$ is solution to:

$$0 = \frac{\partial w}{\partial t} + gcw_c + \frac{1}{2}v^2w_{gg} + \varphi w_g - (r - g)w, \quad \text{for all } c, g \text{ and } t \in [0, T],$$

with boundary condition $w(c, g, T) = \psi'(c)$, all c, g . The Feynman-Kac representation of the solution to the previous equation is:

$$p_c(c, g, t) = e^{-r(T-t)} E \left[e^{\int_t^T g(u) du} \cdot \psi'(c(T)) \middle| c(t) = c, g(t) = g \right], \quad (7.20)$$

which is positive, by the assumption that $\psi' > 0$, thereby confirming the monotonicity properties established previously through no-crossing arguments. So when is $p(c, g, t)$ convex in g ? By differentiating Eq. (7.18) with respect to g , one obtains that $u \equiv p_g$ is solution to:

$$0 = \frac{\partial u}{\partial t} + gc u_c + \frac{1}{2} v^2 u_{gg} + (\varphi + \frac{1}{2}(v^2)') u_g - (r - \varphi') u + c p_c, \quad \text{for all } c, g \text{ and } t \in [0, T], \quad (7.21)$$

with boundary condition $u(c, g, T) = 0$, all c, g . By the Feynman-Kac representation theorem,

$$p_g(c, g, t) = e^{-r(T-t)} E \left[\int_t^T e^{-r(u-t) + \int_t^u \varphi'(g(s)) ds} c(u) \cdot p_c(c(u), g(u), u) du \middle| c(t) = c, g(t) = g \right].$$

By Eq. (7.20), $p_c > 0$. Hence, p is increasing in g . We can now apply Theorem 7.1 and conclude that p is strictly convex in g whenever the drift function of g is weakly convex. Indeed, by differentiating Eq. (10.24) with respect to g , we obtain that $\omega \equiv p_{gg}$ is solution to:

$$\frac{\partial \omega}{\partial t} + gc \omega_c + \frac{1}{2} v^2 \omega_{gg} + (\varphi + (v^2)') \omega_g - (r - 2\varphi' - \frac{1}{2}(v^2)'') \omega + k, \quad \text{for all } c, g \text{ and } t \in [0, T], \quad (7.22)$$

where

$$k(c, g, t) \equiv 2c p_{cg}(c, g, t) + \varphi''(g) p_g(c, g, t), \quad (7.23)$$

and boundary condition $\omega(c, g, T) = 0$ all c, g . By Eq. (7.20), we have that $p_c(c, g, t) c = e^{-r(T-t)} E_t [c(T) \psi'(c(T))]$, which is increasing in g , by the assumption that ψ is increasing and convex, and the no-crossing property of a diffusion, by which $g(u)$ is increasing in the initial condition $g(t)$. Therefore, $p_{cg} > 0$. Furthermore, $p_g > 0$. Therefore, $k(c, g, t) > 0$ whenever $\varphi''(g) \geq 0$. By the Feynman-Kac theorem, then, p is convex in g whenever $\varphi''(g) \geq 0$.

The previous conclusions would hold even with a concave payoff function, say $\psi(c) = \ln c$. In this case, Eq. (7.20) implies that $p_c(c, g, t) = e^{-r(T-t)} \frac{1}{c}$, such that the function k in Eq. (7.23) collapses to, $k(c, g, t) = \varphi''(g) p_g(c, g, t)$. That is, the price function is convex (resp. concave) in g whenever φ is convex (concave) in g . Note, then, that the price is linear in g whenever $\varphi'' = 0$, as it can easily be verified by replacing $\psi(c) = \ln c$ into Eq. (7.19), leaving:

$$p(c, g, t) = e^{-r(T-t)} \ln c + e^{-r(T-t)} E \left[\int_t^T g(u) du \middle| g(t) = g \right].$$

7.5 Time-varying discount rates or uncertain growth?

Predictions about asset prices necessarily rely upon assumptions relating to both the pricing kernel (i.e. interest rates and risk-premiums) and the statistical distribution of dividend growth. As Figure 7.13 illustrates, we may seek for two basic types of predictions. A first type, shown by the two solid arrows, where we begin with a fully specified assumption for the dividends, e.g. the assumption dividend growth is independent and identically distributed, and then seek for properties of the pricing kernel consistent with a given set of dynamic properties of asset

prices (i.e. expected returns and return volatility). A second type of predictions, shown by the dashed arrows, relies, instead, on a search process where we ask which properties of dividends we expect to be consistent with a given set of properties of asset prices and a pricing kernel.

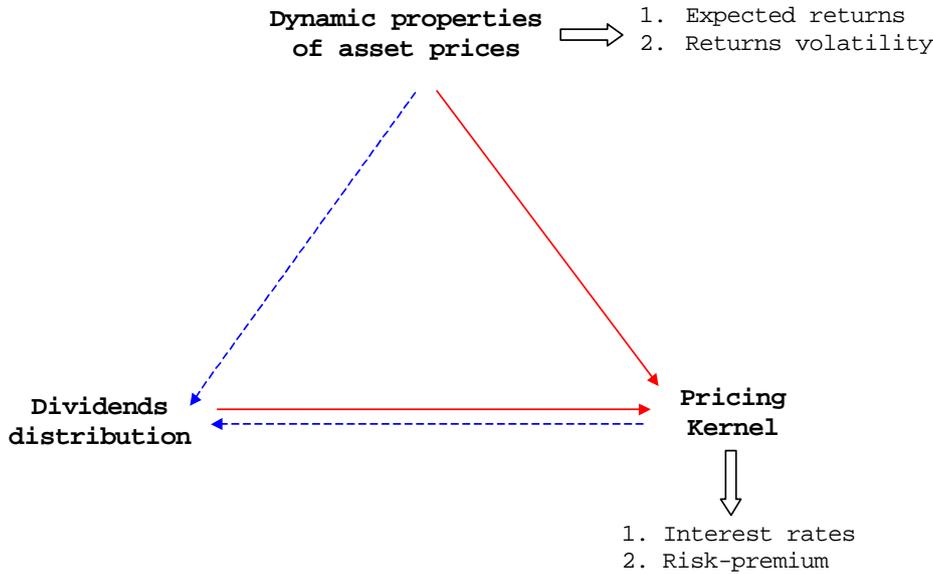


FIGURE 7.13.

This section hinges upon the methodology of the previous section, and implements both approaches. We consider two types of economies:

- (i) In the first economy, changes in the economic fundamentals determine cyclical variations in the discount rates (in Section 7.5.2)
- (ii) In the second economy, the economic fundamentals lead to time-varying expected dividend growth (in Section 7.5.3).

The next section provides preliminary results about pricing kernels, which we need to use while illustrating these two broad classes of economies. Finally, Section 7.5.4 is an introduction to a class of hopefully analytically convenient processes we can use to model long lived asset prices.

7.5.1 Markov pricing kernels

We derive interest rates and unit risk-premiums in a setting where the instantaneous utility function depends on additional state variables, on top of instantaneous consumption. Therefore, the following derivations extend the foundational issues on asset pricing with a representative agent, which we discussed in Section 4.5 of Chapter 4. Consider the stochastic discount factor introduced in Chapter 4, $m_t(\tau) \equiv \frac{\xi(\tau)}{\xi(t)}$, where the pricing kernel process, $\xi(\tau)$, is taken to satisfy:

$$\xi(\tau) \equiv \xi(D(\tau), y(\tau), \tau) = e^{-\int_0^\tau \delta(D(s), y(s)) ds} \Upsilon(D(\tau), y(\tau)), \quad \xi(0) = 1, \tag{7.24}$$

and $(D(\tau), y(\tau))$ are solutions to a slight generalization of Eqs. (7.12),

$$\begin{cases} dD(\tau) = m(D(\tau), y(\tau)) d\tau + \sigma_0(D(\tau)) dW_1(\tau) \\ dy(\tau) = \varphi(y(\tau)) d\tau + v_1(y(\tau)) dW_1(\tau) + v_2(y(\tau)) dW_2(\tau) \end{cases} \tag{7.25}$$

Naturally, the pricing kernel satisfies,

$$\frac{d\xi(\tau)}{\xi(\tau)} = -R(\tau) d\tau - \lambda_1(\tau) dW_1(\tau) - \lambda_2(\tau) dW_2(\tau), \quad (7.26)$$

where R is the short-term rate and $[\lambda_1, \lambda_2]$ is the vector of unit risk-premiums. We assume δ is bounded and positive, and that the function $\Upsilon(D, y)$ is twice continuously differentiable. By applying Itô's lemma to ξ in Eq. (7.24), and identifying the terms in Eq. (7.26), we find that interest rates and risk-premiums are both functions of the current values of (D, y) , and equal:

$$R(D, y) = \delta(D, y) - \frac{L\Upsilon(D, y)}{\Upsilon(D, y)},$$

$$\lambda_1(D, y) = -\sigma_0(D) \frac{\partial \ln \Upsilon(D, y)}{\partial D} - v_1(D, y) \frac{\partial \ln \Upsilon(D, y)}{\partial y}, \quad \lambda_2(D, y) = -v_2(D, y) \frac{\partial \ln \Upsilon(D, y)}{\partial y},$$

where L is the infinitesimal generator operator.

Consider, for example, an infinite horizon economy, where total consumption is solution to Eq. (7.12), with $v_2 \equiv 0$, and $v \equiv v_1$, and a single agent who solves the following program:

$$\max_{c(\tau)_{\tau \geq 0}} E \left[\int_0^{\infty} e^{-\delta\tau} u(c(\tau), x(\tau)) d\tau \right] \quad \text{s.t.} \quad V_0 = E \left[\int_0^{\infty} \xi(\tau) c(\tau) d\tau \right], \quad V_0 > 0,$$

where $\delta > 0$, the instantaneous utility u is continuous and three times continuously differentiable in its arguments, and $x \equiv y$, solution to

$$dx(\tau) = \varphi(D(\tau), x(\tau)) d\tau + v(D(\tau), x(\tau)) dW_1(\tau).$$

In equilibrium, $C = D$, where C is optimal consumption. In terms of ξ in Eq. (7.24), we have that $\delta(D, x) = \delta$, and $\Upsilon(D(\tau), x(\tau)) = \frac{u_1(D(\tau), x(\tau))}{u_1(D(0), x(0))}$, such that $\lambda_2 = 0$, and $\lambda \equiv \lambda_1$, and,

$$R(D, x) = \delta - \frac{u_{11}(D, x)}{u_1(D, x)} m(D, x) - \frac{1}{2} \sigma_0^2(D) \frac{u_{111}(D, x)}{u_1(D, x)}$$

$$- \frac{u_{12}(D, x)}{u_1(D, x)} \varphi(x) - \frac{1}{2} v^2(x) \frac{u_{122}(D, x)}{u_1(D, x)} - v(x) \sigma_0(D) \frac{u_{112}(D, x)}{u_1(D, x)} \quad (7.27)$$

$$\lambda(D, x) = -\frac{u_{11}(D, x)}{u_1(D, x)} \sigma_0(D) - \frac{u_{12}(D, x)}{u_1(D, x)} v(x). \quad (7.28)$$

7.5.2 External habit formation

We might think time-varying risk-premiums to be a plausibly natural engine of asset price fluctuations. Indeed, within the neoclassical asset pricing framework, the very properties of asset prices must necessarily inherit by those of the risk-premiums, when dividend growth is independent and identically distributed, as illustrated by Figure 7.13. Campbell and Cochrane (1999) model of external habit formation is certainly one of the most well-known attempts at explaining some of the empirical features outlined in Section 7.2, throughout the channel of time-varying risk-premiums. Consider an infinite horizon, complete markets economy, where a representative agent has undiscounted instantaneous utility:

$$u(c(\tau), x(\tau)) = \frac{(c(\tau) - x(\tau))^{1-\eta} - 1}{1-\eta}, \quad (7.29)$$

with c denoting consumption and x is a time-varying habit, or exogenous “subsistence level”. The properties of the habit process are defined in a residual way, by defining, first, those of the “surplus consumption ratio,” as put forward below. The total endowment process $D(\tau)$ satisfies,⁵

$$\frac{dD(\tau)}{D(\tau)} = g_0 d\tau + \sigma_0 dW(\tau). \quad (7.30)$$

A measure of distance between consumption and the level of habit is the “surplus consumption ratio,”

$$s(\tau) \equiv \frac{c(\tau) - x(\tau)}{c(\tau)}.$$

The curvature of the instantaneous utility is inversely related to s ,

$$-\frac{u_{cc}(c, x)c}{u_c(c, x)} = \eta \frac{c}{c-x} = \eta \frac{D}{D-x} \equiv \eta s^{-1}, \quad (7.31)$$

where subscripts denote partial derivatives, the second equality is the equilibrium condition, $c(\tau) = D(\tau)$, and the third is the definition of the surplus consumption ratio, in equilibrium. By assumption, $s(\tau)$ is solution to:

$$ds(\tau) = s(\tau) \left[(1 - \phi)(\bar{s}_l - \ln s(\tau)) + \frac{1}{2} \sigma_0^2 l(s(\tau))^2 \right] d\tau + \sigma_0 s(\tau) l(s(\tau)) dW(\tau), \quad (7.32)$$

where l is a positive function, defined below. This model of habit formation differs from previous formulations such as that of Ryder and Heal (1973), Sundaresan and Constantinides (1990), because of three fundamental reasons: (i) it is an “external” theory, in that the habit x is “aggregate,” not consumption chosen by the individual, similarly as with Abel’s (1990) “catching up with the Joneses” formulation, or Duesenberry’s (1949) relative income model; (ii) habit responds to consumption smoothly, not to each period past consumption, as in previous models of habit formation such as that of Ferson and Constantinides (1990); (iii) it guarantees marginal utility is always positive. The second property, (ii), produces slow mean reversions in the price-dividend ratio and long-horizon predictability, and large predictable movements in stock volatility, three empirical features reviewed in Section 7.2.

To derive the Sharpe ratio in this economy, we use Eq. (7.28) in Section 7.5.1,

$$\lambda(D, x) = \frac{\eta}{s} \left(\sigma_0 - \frac{1}{D} v(D, x) \right), \quad (7.33)$$

where v is the diffusion coefficient of the habit process, in equilibrium,

$$x(\tau) = D(\tau) (1 - s(\tau)),$$

and $s(\tau)$ is solution to Eq. (7.32). By Itô’s lemma, $v(D, x) = (1 - s - sl(s)) D \sigma_0$, which replaced into Eq. (7.33), leaves:

$$\lambda(s) = \eta \sigma_0 (1 + l(s)). \quad (7.34)$$

The real interest rate is, by Eq. (7.27),

$$R(s) = \delta + \eta \left(g_0 - \frac{1}{2} \sigma_0^2 \right) + \eta (1 - \phi) (\bar{s}_l - \ln s) - \frac{1}{2} \eta^2 \sigma_0^2 (1 + l(s))^2. \quad (7.35)$$

⁵Campbell and Cochrane (1999) consider a discrete-time model where log-consumption growth is Gaussian. Eq. (7.30) is, simply, the diffusion limit of their model.

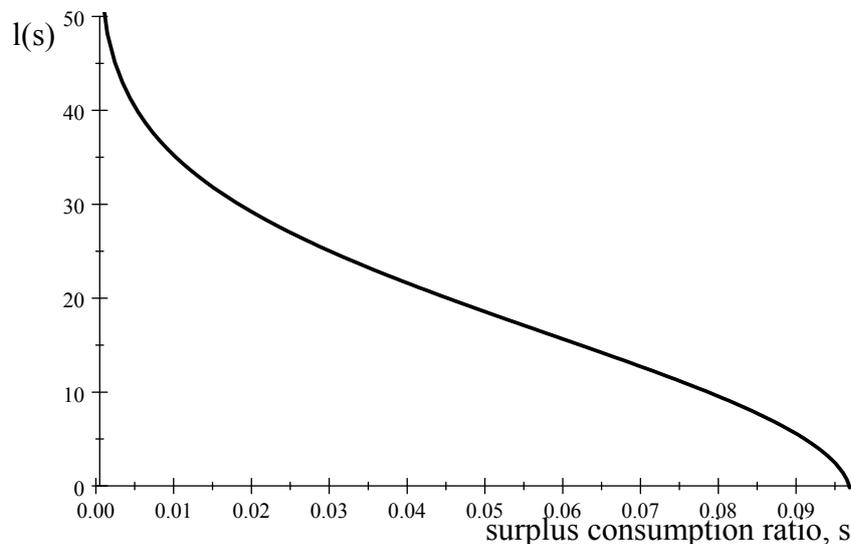
The third term reflects usual intertemporal substitution effects: bad times, when s is low, are those when agents expect the very same future surplus consumption ratio s will improve, due to mean reversion. Therefore, in bad times, agents expect their marginal utility to decrease in the future and to compensate for this fall, they will try to decrease future consumption, compared to today, by borrowing more, thereby pushing interest rates up. The last term is a precautionary savings term.

Campbell and Cochrane (1999) choose the function l so as to satisfy three conditions: (i) the short-term rate R is constant; and habit is predetermined both (ii) at the steady state, and (iii) near the steady state. The reason they choose R constant is motivated by the empirical features surveyed in Section 7.2, that real interest rates are really not volatile, compared to stock returns. Making habit predetermined at and near the steady state formalizes the idea that it takes time for consumption shocks to affect habit, at least at the steady state. The Appendix shows that under these conditions, the function l is:

$$l(s) = \bar{s}^{-1} \sqrt{1 + 2(\bar{s}_t - \ln s)} - 1, \quad (7.36)$$

where $\bar{s} = \sigma_0 \sqrt{\frac{\eta}{1-\phi}} = e^{\bar{s}_t}$. In turn, this function implies that the short-term rate in Eq. (7.35) is: $R = \delta + \eta \left(g_0 - \frac{1}{2} \sigma_0^2 \right) - \frac{1}{2} \eta (1 - \phi)$. The Appendix considers a slightly more general model, where the short-term rate is affine in $\ln s$.

The next picture depicts the function l in Eq. (7.36), computed through the parameter values utilized by Campbell and Cochrane, $\eta = 2$, $\sigma_0 = 0.0150$, $\phi = 0.870$. It is decreasing in s , and convex in s , over the empirically relevant range of variation of s .



Note that, then, these properties are inherited by the Sharpe ratio in Eq. (7.33). Quite simply, these fundamental properties of the Sharpe ratio arise simply because we want habit to be predetermined near the steady state, and the short-term rate constant or, at least, as shown in the Appendix, affine in $\ln s$.

The model makes a number of important predictions. Consider, first, the instantaneous utility in Eq. (7.29). By Eq. (7.31), $\text{CRRA} = \eta s^{-1}$. That is, risk aversion is countercyclical. Intuitively, during economic downturns, the surplus consumption ratio s decreases and agents become more risk-averse. As a result, prices decrease and expected returns increase. It is a very sensible

mechanism. Furthermore, the model generates realistic risk premiums. Intuitively, the stochastic discount factor is $e^{-\rho t} \left(\frac{s(t) D(t)}{s(0) D(0)} \right)^{-\eta}$, which due to $\eta > 0$, is quite countercyclical, due to the procyclicality of both s and D , and is more volatile than the standard stochastic discount factor, $e^{-\rho t} \left(\frac{D(t)}{D(0)} \right)^{-\eta}$. However, the economy is one with high risk-aversion, as on average, the calibrated model produces a value of ηs^{-1} around 40. Barberis, Huang and Santos (2001) have a similar mechanism, based on alternative preferences. **[Discuss]**

By Eq. (7.32), the log of s is a mean-reverting process. By taking logs, we are sure that s remains positive. Moreover, $\ln s$ is also conditionally heteroskedastic since its instantaneous volatility is $\sigma_0 l$. Because l is decreasing in s and s is clearly procyclical, the volatility of $\ln s$ is countercyclical. This feature is responsible of many interesting properties of the model, such as countercyclical returns volatility.

Finally, the Sharpe ratio λ in Eq. (7.34) is made up of two components. The first is $\eta \sigma_0$, which coincides with the Sharpe ratio predicted by the standard Gordon's (1962) model. The second is $\eta \sigma_0 l(s)$, and arises as a compensation related to the stochastic fluctuations of the habit, $x = D(1 - s)$. By the functional form of l Campbell and Cochrane assume, λ is therefore countercyclical. Combined with a high ϕ , this assumption leads to slowly varying, countercyclical expected returns. Finally, numerical simulations of the model leads the authors to conclude that the price-dividend ratio is concave in s . In the Appendix, we describe a simple algorithm that one may use to solve this and related models numerically, and in discrete time.

We now clarify the nature of the theoretical link between convexity of l and concavity of the price-dividend ratio in this and related models. We can use the canonical pricing problem of Section 7.4, and appeal to Theorem 7.1, to gain insights into this issue. What is the price-dividend ratio in this economy? Note that the short-term rate is constant in this model, as discussed. Yet for sake of generality, assume it is state-dependent (as, e.g., in the Appendix), although only a function of s . The price-dividend ratio is then as in Eqs. (7.13)-(7.14)-(7.16), with constant growth, viz $m(y) = g_0$:

$$p(s) = \int_0^\infty e^{g_0 \tau} \cdot \bar{\mathbb{E}} \left[e^{-\int_0^\tau \text{Disc}(s(u)) du} \middle| s \right] d\tau, \quad (7.37)$$

where $\text{Disc}(s) \equiv r(s) + \sigma_0 \lambda(s)$ are the risk-adjusted discount rates, and

$$ds(\tau) = \bar{\varphi}(s) d\tau + v(s(\tau)) d\bar{W}(\tau),$$

where

$$\begin{aligned} \bar{\varphi}(s) &= \varphi(s) - v(s) \lambda(s) + \sigma_0 v(s) \\ \varphi(s) &= s(1 - \phi) (\bar{s}_l - \ln s) + \frac{1}{2} \sigma_0^2 l^2(s), \quad v(s) = \sigma_0 s l(s) \end{aligned}$$

and $\bar{W}(\tau) = \hat{W}(\tau) - \sigma_0 \tau$ is a Brownian motion under the probability \bar{Q} , defined through the Radon-Nikodym derivative in Eq. (7.15), and, finally, $\hat{W}(\tau)$ is a Brownian motion under the risk-neutral probability Q .

The inner expectation in Eq. (7.37) can be analyzed through the canonical pricing problem of Section 7.4, such that, by Theorem 7.1, we can make the following conclusions:

- (i) *Suppose that the risk-adjusted discount rates are countercyclical, viz $\frac{d}{ds} \text{Disc}(s) \leq 0$. Then, the price-dividend ratio is procyclical, viz $\frac{d}{ds} p(s) > 0$.*
- (ii) *Suppose that the price-dividend ratio is procyclical. Then, the price-dividend ratio is also a concave function of s as soon as the risk-adjusted discount rates are convex in s , viz $\frac{d^2}{ds^2} \text{Disc}(s) > 0$, and $\frac{d^2}{ds^2} \bar{\varphi}(s) \leq 2 \frac{d}{ds} \text{Disc}(s)$.*

So we have found joint restrictions on the primitives such that the pricing function p is consistent with properties given in advance. What is the economic interpretation related to the convexity of risk-adjusted discount rates? If price-dividend ratios are concave in some state variable Y tracking the business cycle conditions, stock volatility increases on the downside, and is thus countercyclical, as illustrated by Figure 7.6. According to the previous predictions, price-dividend ratios are concave in Y whenever risk-adjusted discount rates are decreasing and *sufficiently* convex in Y . The economic significance of convexity in this context is that in good times, risk-adjusted discount rates are substantially constant. As a result, the evaluation of future dividends does not vary too much, and price-dividend ratios remain relatively constant. In bad times, however, risk-adjusted discount rates increase sharply, thus making price-dividend ratios more responsive to changes in the economic conditions.

One defect of this model is that the variables of interest are all driven by the same state variable, s . For this reason, the correlation between consumption growth and stock returns is, conditionally, one, whereas in the data, it is much less. Naturally, the correlation predicted by this model is less than one, unconditionally, but still too high, if compared with that in the data.

7.5.3 Large price swings as a learning induced phenomenon

The models of this section are those where expected consumption/dividend is unobserved, which leads to a natural question: how should we process available data to formulate guesses about the direction of expected consumption? These issues would naturally lead to models where on top of consumption, expected dividend is another state variable—hopefully likely to introduce interesting asset price dynamics. Note that although the focus of this section is about models with unobserved expected dividend growth, models where expected dividend growth is observed have always had an interest on their own (see for example, Campbell, 2003; Bansal and Yaron, 2004), and will be examined in Chapter 8.

7.5.3.1 Framework

Time variation in stock volatility may also arise as a result of the agents' learning process about the economic fundamentals. In models along these lines, public signals about the fundamentals hit the market, and agents make inference about them, thereby creating new state variables driving price fluctuations, which relate to the agents own guesses about the (unknown) state of the economic fundamentals. Timmermann (1993, 1996) provides models with exogenous discount rates, where learning effects increase stock volatility over and above that we might observe in a world without uncertainty, and learning effects, about the fundamentals. Brennan and Xia (2001) generalize these models to a stochastic general equilibrium. Veronesi (1999) provides a rational expectations model with learning about the fundamentals, with quite nonlinear learning effects. This section provides details about the mechanisms through which learning affects asset prices in general, and stock volatility in particular.

We shall assume that information about the fundamentals is incomplete, but symmetrically distributed among agents. The assumption of symmetric information might appear strong. It should not. The models of this section aim to capture the idea that markets function in a context of “incompressible” uncertainty, where agents are all unaware of the crucial aggregate, macroeconomic developments affecting asset prices. Chapter 9, instead, reviews models with both differential and asymmetric information, which are more useful whilst thinking about the functioning of markets for individual stocks, where it is, then, more plausible to have agents

with different information sets, who acquire information in dedicated information markets. Acquiring crucial information about, say, the direction of the business cycle, and having agents asymmetrically informed about it, thereby affecting asset prices, seems implausible—the cost of acquiring such information is incommensurable.

Note that the assumption of symmetric information simplifies the analysis, as the agents do not need to base their decisions upon the observation of the equilibrium price. For example, in a context with asymmetric information, agents can learn pieces of information other agents have, by “reading the equilibrium price,” because agents with superior information impinge part of their information on the asset price, through trading, as explained in Chapter 9. This complication does not arise in the context of this section: agents, then, need only to condition upon the realization of signals, which convey information about the fundamentals. There is no need for any agent to condition on prices, because prices merely convey the same information any such agent already has.

A final consideration pertains to the very nature of aggregate stock market fluctuations, which seems to be quite stable, historically, as reviewed in Section 7.2. It is an interesting aspect, because it is quite obvious that capital markets have undergone significant changes over time, which affected various aspects of their *microstructure*, such as the technology of transactions, the price discovery process, liquidity, and transaction volumes, to mention a few examples. How is it that the properties of the aggregate stock market reviewed in Section 7.2 do not appear to be affected by these changes? One possibility is, simply, that market microstructure is about the very high frequency behavior of markets, whereas the properties in Section 7.2 relate to slow, low frequency movements. The models in this section, and in this chapter, aim to rationalize some of these movements. Models addressing the previous market microstructure issues are reviewed in Chapter 9, as mentioned.

7.5.3.2 An introductory example of learning

Suppose consumption D is generated by $D = \theta + w$, where θ and w are independently distributed, with $p \equiv \Pr(\theta = A) = 1 - \Pr(\theta = -A)$, and $\Pr(w = A) = \Pr(w = -A) = \frac{1}{2}$. Suppose that the “state” θ is unobserved. How would we update our prior probability p of the “good” state upon the observation of D ? A simple application of the Bayes’ Theorem gives the posterior probabilities $\Pr(\theta = A | D_i)$ displayed in Table 7.3. Considered as a random variable defined over observable states D_i , the posterior probability $\Pr(\theta = A | D_i)$ has expectation $E[\Pr(\theta = A | D)] = p$ and variance $\text{var}[\Pr(\theta = A | D)] = \frac{1}{2}p(1 - p)$. Clearly, this variance is zero exactly where there is a degenerate prior on the state. More generally, it is a \cap -shaped function of the a priori probability p of the good state. Since the “filter,” $g \equiv E(\theta = A | D)$, is linear in $\Pr(\theta = A | D)$, the same qualitative conclusions are also valid for g .

		D_i (observable state)		
		$D_1 = 2A$	$D_2 = 0$	$D_3 = -2A$
$\Pr(D_i)$		$\frac{1}{2}p$	$\frac{1}{2}$	$\frac{1}{2}(1 - p)$
$\Pr(\theta = A D = D_i)$		1	p	0

TABLE 7.3. Randomization of the posterior probabilities $\Pr(\theta = A | D)$.

The probabilities in Table 7.3 follow by a simple application of Bayes' Theorem. Let $(E_i)_i$ be a partition of the state space Ω . (This partition can be finite or uncountable, i.e. the set of indexes i can be finite or uncountable—it does not really matter.) Then, by Bayes' Theorem,

$$\Pr(E_i|F) = \Pr(E_i) \cdot \frac{\Pr(F|E_i)}{\Pr(F)} = \Pr(E_i) \cdot \frac{\Pr(F|E_i)}{\sum_j \Pr(F|E_j) \Pr(E_j)}. \quad (7.38)$$

By applying Eq. (7.38) to our example, we have:

$$\Pr(\theta = A|D = D_1) = \Pr(\theta = A) \frac{\Pr(D = D_1|\theta = A)}{\Pr(D = D_1)} = p \frac{\Pr(D = D_1|\theta = A)}{\Pr(D = D_1)}.$$

But $\Pr(D = D_1|\theta = A) = \Pr(w = D_1 - A) = \Pr(w = A) = \frac{1}{2}$. Moreover, we have that $\Pr(D = D_1) = \frac{1}{2}p$. This leaves $\Pr(\theta = A|D = D_1) = 1$. It's trivial, but one proceeds similarly to compute the other probabilities.

This simple example conveys the main ideas underlying *Bayesian learning*. However, it leads to a nonlinear filter, g , which differs from those we usually encounter in the literature (see, e.g., Chapters 8 and 9 in Liptser and Shiryaev, 2001a), where the instantaneous variance of the posterior probability changes, $d\pi$ say, is, typically, proportional to $\pi^2(1-\pi)^2$, not to $\pi(1-\pi)$. This distinction is merely technical, and is due to the assumption w is a discrete random variable. Indeed, assume that w has some arbitrary, but continuous density ϕ , and zero mean and unit variance. Let $\pi(D) \equiv \Pr(\theta = A|D \in dD)$. By the Bayes rule in Eq. (7.38),

$$\pi(D) = \Pr(\theta = A) \cdot \frac{\Pr(D \in dD|\theta = A)}{\Pr(D \in dD|\theta = A) \Pr(\theta = A) + \Pr(D \in dD|\theta = -A) \Pr(\theta = -A)}.$$

But $\Pr(D \in dD|\theta = A) = \Pr(w = D - A) = \phi(D - A)$ and, similarly, $\Pr(D \in dD|\theta = -A) = \Pr(w = D + A) = \phi(D + A)$. Simple computations then leave,

$$\pi(D) - p = p(1-p) \frac{\phi(D - A) - \phi(D + A)}{p\phi(D - A) + (1-p)\phi(D + A)}. \quad (7.39)$$

That is, the variance of the “probability changes,” $\pi(D) - p$, is proportional to $p^2(1-p)^2$.

Next, in an attempt to add more structure to this model, assume that w is a Brownian motion, and set $A \equiv A d\tau$. Let $D_0 \equiv D(0) = 0$. In Appendix 9, we show that by applying Itô's lemma to $\pi(D)$, we obtain

$$d\pi(\tau) = 2A \cdot \pi(\tau)(1 - \pi(\tau))dW(\tau), \quad \pi(D_0) \equiv p, \quad (7.40)$$

where $dW(\tau) \equiv dD(\tau) - g(\tau)d\tau$ and $g(\tau) \equiv E(\theta|D(\tau)) = [A\pi(\tau) - A(1 - \pi(\tau))]$.

While our construction is highly heuristic, it also relies on rigorous foundations, such as those in Liptser and Shiryaev (2001a, theorem 8.1 p. 318; and example 1 p. 371). Quite critically, it can be shown (see, e.g., Liptser and Shiryaev (2001a, theorem 7.12 p. 273)) that W is a Brownian motion with respect to the agents' information set $\sigma(D(t), t \leq \tau)$. Therefore, the equilibrium in the original economy with incomplete information is isomorphic in its pricing implications to the equilibrium in a full information economy where the dividend process is solution to:

$$\begin{cases} dD(\tau) = (g(\tau) - \hat{\lambda}\sigma_0)d\tau + \sigma_0 d\hat{W}(\tau) \\ dg(\tau) = -\hat{\lambda}v(g(\tau))d\tau + v(g(\tau))d\hat{W}(\tau) \end{cases} \quad (7.41)$$

where \hat{W} is a Brownian motion under the risk-neutral probability, $v(g) \equiv (A - g)(g + A)/\sigma_0$, $\sigma_0 \equiv 1$, and where we are assuming that $\hat{\lambda}$, the risk-premium adjustment, is constant. In fact, Eqs. (7.41) hold for any $\sigma_0 > 0$, in that we have that $dW(\tau) = \sigma_0^{-1}(dz(\tau) - E(\theta|z(t)_{t \leq \tau})d\tau) = \sigma_0^{-1}(dz(\tau) - g(\tau)d\tau)$.

We can draw similar conclusions while dealing with *proportional* markets, i.e. markets where the dividend process, D , is solution to:

$$\begin{cases} \frac{dD(\tau)}{D(\tau)} = (g(\tau) - \sigma_0 \hat{\lambda})d\tau + \sigma_0 d\hat{W}(\tau) \\ dg(\tau) = -\hat{\lambda}v(g(\tau))d\tau + v(g(\tau))d\hat{W}(\tau) \end{cases} \quad (7.42)$$

where once again, $\hat{\lambda}$ denotes some risk-premium adjustment, which we assume constant. The instantaneous volatility of the expected dividend growth, g , is \cap -shaped in this example, too. In the presence of positive compensation for risk, $\hat{\lambda} > 0$, the risk-neutralized drift of g is, then, likely, a convex function of g . Our discussion of the canonical economy in Section 7.4 suggests that this convexity might have critical asset pricing implications—a convexity of the asset price with respect to expected growth, g . This convexity has a striking economic implication, meaning large price swings in good times, i.e. when agents living in a context with incomplete information interpret the signals they receives as ones leading to high growth—a convex price might lead to fluctuations we might interpret as arising out of a bubble even if no bubbles whatsoever are impinging on the markets. We now use the tools of Theorem 7.1, and analyze these convexity properties in two famous models of learning.

7.5.3.3 Convexity, and two models of learning

The model summarized by Eqs. (7.41) relates to that considered by Veronesi (1999), where an infinitely lived agent has constant *absolute* risk aversion equal to $\gamma > 0$, and observes realizations of D , generated by:

$$dD(\tau) = \theta d\tau + \sigma_0 dw_1(\tau), \quad (7.43)$$

where w_1 is a Brownian motion, and θ is the expected dividend change, supposed to follow a two-state $(\bar{\theta}, \underline{\theta})$ Markov chain. (See, also, David, 1997, for a related model.) The key issues in this economy are that the expected dividend change, θ , is unobserved, and that as a result of this, the agent attempts to learn about the state where he is living, following the same Bayesian rules described in the previous section. Using standard results generalizing those in the previous section, it is possible to show that the price in this economy is the same as the price in a full information economy where:

$$\begin{cases} dD(\tau) = g(\tau)d\tau + \sigma_0 dW(\tau) \\ dg(\tau) = k(\bar{g} - g(\tau))d\tau + v(g(\tau))dW(\tau) \end{cases} \quad (7.44)$$

where $v(g) = (\bar{\theta} - g)(g - \underline{\theta})/\sigma_0$, and k, \bar{g} are some positive constants. Note that while the diffusion terms in Eq. (7.41) and in Eq. (7.44) have the same functional form, the expected dividend change is a martingale in Eq. (7.41), and mean-reverting in Eq. (7.44), under the *physical* probability—i.e. when $\hat{\lambda} = 0$ in Eq. (7.41). Such a distinction arises because the model underlying Eq. (7.41) is one where θ is drawn at time 0, forever, whereas in Eq. (7.44), θ is a Markov chain.

Eqs. (7.44) amount to a special case of Eqs. (7.25). Therefore, by Eq. (7.28), the risk-premium is constant, and equal to $\lambda = \gamma\sigma_0$ such that the dynamics of dividends under the risk-neutral

probability are:

$$\begin{cases} dD(\tau) = (g(\tau) - \gamma\sigma_0^2) d\tau + \sigma_0 d\hat{W}(\tau) \\ dg(\tau) = (k(\bar{g} - g(\tau)) - \gamma\sigma_0 v(g(\tau))) d\tau + v(g(\tau)) d\hat{W}(\tau) \end{cases} \quad (7.45)$$

Veronesi (1999) also assumes the riskless asset is infinitely elastically supplied, and therefore that the interest rate r is a constant. Let us analyze the properties of the equilibrium price. It is easy to see that given Eq. (7.45), the asset price is:

$$S(D, g) = \mathbb{E} \left(\int_0^\infty e^{-r\tau} D(\tau) d\tau \middle| D, g \right) = \int_0^\infty C(D, g, \tau) d\tau,$$

where

$$C(D, g, \tau) = e^{-r\tau} (D - \gamma\sigma_0^2\tau) + G(g, \tau), \quad \text{and} \quad G(g, \tau) \equiv e^{-r\tau} \int_0^\tau \mathbb{E}(g(u) | g) du. \quad (7.46)$$

We apply Theorem 7.1 to study convexity properties of G , as the conditional expectation, $\mathbb{E}(g(u) | g)$, can be read as a special case of the canonical price in Eq. (7.11)—namely, for $\rho \equiv 1$ and $\psi(g) = g$. By Theorem 7.1-(ii), $\mathbb{E}(g(u) | g)$ is convex in g whenever the drift of g in Eq. (7.41) is convex. This condition always holds true, as $\gamma > 0$: the conditional expectation of the expected dividend change in Eq. (7.46), g , inherits the same second order properties (convexity) of this drift function.

The economic implications of this result are striking. In this economy, prices are convex in the expected dividend change. In good times, prices might react disproportionately with changes in the underlying fundamentals. The interpretation is equally interesting. Risk-aversion correction is basically zero over extreme situations—when expected dividend changes are at their boundaries. It is, however, the highest under relatively more normal circumstances. Formally, the risk-adjusted drift of g is $\hat{\varphi}(g) = \varphi(g) - \gamma\sigma_0 v(g)$, and is convex in g , because v is concave in g .

Would these properties persist, should we relax the assumption the riskless asset is infinitely elastically supplied? By Eq. (7.27), the short-term rate, when allowed to fluctuate is, $R(g) = \delta + \gamma g - \frac{1}{2}\gamma^2\sigma_0^2$, such that the asset price is, now,

$$S_r(D, g) = \mathbb{E} \left(\int_0^\infty e^{-\int_0^\tau R(g_s) ds} D(\tau) d\tau \middle| D, g \right) = \int_0^\infty C_r(D, g, \tau) d\tau,$$

where

$$C_r(D, g, \tau) = (D - \gamma\sigma_0^2\tau) B(g, \tau) + G_r(g, \tau) \\ B(g, \tau) \equiv \mathbb{E} \left(e^{-\int_0^\tau R(g_s) ds} \middle| g \right), \quad G_r(g, \tau) \equiv \mathbb{E} \left(e^{-\int_0^\tau R(g_s) ds} g(\tau) \middle| g \right)$$

By Theorem 7.1, the two functions, $B(g, \tau)$ and $G_r(g, \tau)$, are *affine* in g , if the risk-neutral drift of g in Eq. (7.45), $b(g) \equiv (k(\bar{g} - g) - \gamma\sigma_0 v(g))$, and the short-term rate $R(g)$, are such that $b''(g) = 2R'(g)$. A simple computation reveals that this is indeed the case: $b''(g) = -2\gamma$ and $2R'(g) = 2\gamma$. Endogeneizing the short-term rate destroys the convexity properties of the asset price. It is unpleasant arithmetics. We know the short-term rate should not fluctuate too much, compared to stocks. Assuming it is constant as Veronesi (1999) did, is reasonable, and leads to interesting properties, which would be destroyed under the counterfactual circumstance the short-term rate is allowed to considerably fluctuate.⁶

⁶Interestingly, this property of prices being linear in g once interest rates are endogeneous is not specific to cases where dividend is as in Eqs. (7.44) and agents have CARA. Veronesi (2000) and Mele (2005, Appendix B) show that convexity properties are lost even in proportional markets (i.e. markets where the distribution of dividend growth is independent of the initial level of dividends) and representative agents with CRRA.

Interestingly, this property that prices are linear in g , once interest rates are endogeneous, is not specific to cases where dividend is as in Eqs. (7.44) and agents have CARA. Veronesi (2000) shows that convexity properties are lost even in proportional markets (i.e. markets where the distribution of dividend growth is independent of the initial level of dividends) and representative agents with CRRA. In fact, Appendix 10 shows that there are no linear signal structures and representative CRRA economies with *complete securities markets* supporting the convexity property.

In all these examples, convexity properties of the asset price might arise, if any, as we are assuming agents learn about a *discrete* state space in a *continuous-time* economy. We now analyze another model, generalizing that in Eq. (7.42), and analyzed by Brennan and Xia (2001). In this model, the information structure is one where an infinitely lived agent with CRRA preferences observes D , solution to:

$$\frac{dD(\tau)}{D(\tau)} = \hat{g}(\tau) d\tau + \sigma_0 dw_1(\tau),$$

and the expected dividend growth, $\hat{g}(\tau)$, is unobserved. Rather than assuming $\hat{g}(\tau)$ to be on a countable number of states, Brennan and Xia postulate that it is an Ornstein-Uhlenbeck process:

$$d\hat{g}(\tau) = k(\bar{g} - \hat{g}(\tau))d\tau + \sigma_1 dw_1(\tau) + \sigma_2 dw_2(\tau),$$

where \bar{g} , σ_1 and σ_2 are positive constants. The agent implements a Bayesian learning procedure, similar to that described in the previous section. It can be shown that if our agent has a Gaussian prior on $\hat{g}(0)$ with variance γ_*^2 , as defined below, the asset price takes the form $S(D, g)$, where D and g are solution to Eqs. (7.25), with $m(D, g) = gD$, $\sigma_0(D) = \sigma_0 D$, $\varphi(g) = k(\bar{g} - g)$, $v_2 = 0$, and $v \equiv v_1 \equiv v_1(\gamma_*) = (\sigma_1 + \frac{1}{\sigma_0} \gamma_*)^2$, where γ_* is the positive solution to $v_1(\gamma) = \sigma_1^2 + \sigma_2^2 - 2k\gamma$.⁷ By Eq. (7.28), the risk-premium is constant, and equal to $\lambda = \eta\sigma_0$, where η is the CRRA coefficient, and by Eq. (7.27), the short-term rate is linear in g . Therefore, and by the same reasoning leading to Eq. (7.37), the price-dividend ratio is independent of D , and is given by:

$$p(g) = \int_0^\infty e^{-\sigma_0 \lambda \tau} \bar{\mathbb{E}} \left(e^{\int_0^\tau (g(u) - R(g(u))) du} \middle| g \right) d\tau, \quad (7.47)$$

where,

$$\begin{cases} \frac{dD(\tau)}{D(\tau)} = (g(\tau) - \sigma_0 \lambda) d\tau + \sigma_0 d\bar{W}(\tau) \\ dg(\tau) = (\varphi(g(\tau)) + \sigma_0 v(g(\tau))) d\tau + v(g(\tau)) d\bar{W}(\tau) \end{cases}$$

and $\bar{W}(\tau) = \hat{W}(\tau) - \sigma_0 \tau$ is a \bar{P} -Brownian motion, where the two functions, $\varphi(g)$ and $\sigma_0 v(g)$, are momentarily left unspecified, as we wish to provide more general results.

Under regularity conditions, monotonicity and convexity properties are inherited by the inner expectation in Eq. (7.47). Precisely, in the notation of the canonical pricing problem, we have that $\rho(g) \equiv -g + R(g) + \sigma_0 \lambda$ and $b(g) \equiv \varphi(g) + (\sigma_0 - \lambda) v(g)$. Therefore, by Theorem 7.1, we have:

⁷Brennan and Xia (2001) actually consider a slightly more general model, where consumption and dividends differ. They derive a model with a reduced-form identical to that in this example. In the calibrated model, Brennan and Xia found that the variance of the filtered \hat{g} is higher than the variance of the expected dividend growth in an economy with complete information. The results on γ_* in this example can be obtained through an application of theorem 12.1 in Liptser and Shiryaev (2001) (Vol. II, p. 22). They generalize results in Gennotte (1986) and are a special case of results in Detemple (1986). Both Gennotte and Detemple did not emphasize the impact of learning on the pricing function.

- (i) If $R'(g) < 1$, the price-dividend ratio is increasing in the dividend growth rate.
- (ii) Suppose that the price-dividend ratio is increasing in the dividend growth rate. Then, it is convex whenever $R''(g) > 0$, and $\frac{d^2}{dg^2} [\varphi(g) + (\sigma_0 - \lambda)v(g)] \geq -2 + 2R'(g)$.

Suppose, for instance, that the short-term rate is constant, because for example it is in infinite elastic supply. Then, the price-dividend ratio is increasing in the expected dividend growth, and it is convex in it,

$$\frac{d^2}{dg^2} (\varphi_0(g) + (\sigma_0 - \lambda)v(g)) \geq -2.$$

These conditions are satisfied by Brennan and Xia (2001). [...] Provide economic interpretation. [in progress] Moreover, explain that this is due to the fact the inner expectation in Eq. (7.47) is indeed one for an affine model to be introduced and explained in full detail in Chapter 12.

7.5.4 Linearity-generating processes

The focus of the previous sections is a search for pricing kernels that make asset pricing models qualitatively consistent with countercyclical statistics—a search relying on theoretical test conditions, those emanating from Theorem 7.1 in Section 7.4. We can actually use Theorem 7.1 for another, somehow surprising, purpose: a search for asset pricing models that have a closed-form solutions. The idea underlying this project relies on a simple remark. Theorem 7.1-(ii) provides conditions under which price-dividend ratios can be either concave or convex in the state variables driving them. Specifically, consider the representation of the price-dividend ratio in Eqs. (7.13)-(7.14)-(7.16), which fits the canonical pricing problem in Section 7.4, as observed, once we identify the primitives of the models with Eqs. (7.17), reported here for convenience,

$$\psi(y) \equiv 1; \quad \rho(y) \equiv \text{Disc}(y) - m(y); \quad b(y) = \bar{\varphi}(y),$$

where as usual, $\text{Disc}(y) \equiv R(y) + \sigma_0 \lambda_{\text{CF}}(y)$ and, by Eq. (7.16),

$$\bar{\varphi}(y) \equiv \varphi(y) - \sum_{i=1}^2 v_i(y) \lambda_i(y) + \sigma_0 v_1(y),$$

and $\varphi(y)$ denotes the drift of y under the physical probability. Then, given that $\psi' = \psi'' = 0$, we have that by Theorem 7.1-(i),

$$c \text{ is: } \begin{cases} \text{concave in } y & \text{if } (\text{Disc}'' - m'') \geq 0 \quad \text{and} \quad \psi'' < 0 \\ \text{convex in } y & \text{if } (\text{Disc}'' - m'') \leq 0 \quad \text{and} \quad \psi'' > 0 \end{cases} \quad (7.48)$$

regardless of whether c is increasing in y . It is quite unlikely a closed-form solution might exist for the price-dividend ratio when one of those conditions is satisfied. Yet how about assuming none of the two conditions for concavity or convexity hold true? In this case, obviously, price-dividend ratios are neither concave nor convex, so they must necessarily be affine in the state variables! Technically, then, we have that by the conditions in (7.48), the price-dividend ratios is affine in y if:

$$\bar{\varphi}'' = 2(\text{Disc}' - m') \quad \text{and} \quad \text{Disc}'' - m'' = 0. \quad (7.49)$$

Gabaix (2009) is the first paper to note that price-dividend ratios are affine in the state variables driving them, should the drift of these state variables be quadratic. His remarks are

consistent with Theorem 7.1, and the two conditions in (7.49). In fact, Gabaix develops a unified theory of “linearity-generating processes,” which generalizes the single state variable framework underlying Theorem 7.1 and the related conditions in (7.49).

To illustrate these facts, consider a model that fits the class of linearity-generating processes, one of external habit formation by Menzly, Santos and Veronesi (2004). In this model, a representative agent maximizes,

$$U = E \left(\int_0^\infty e^{-\delta t} \ln(c(t) - x(t)) dt \right), \quad (7.50)$$

where $x(t)$ is external habit. Relative risk-aversion equals the inverse of the surplus consumption ratio, $\frac{1}{s(t)}$, with $s(t) = \frac{c(t)-x(t)}{c(t)}$, which in equilibrium equals $\frac{D(t)-x(t)}{D(t)}$, where D is consumption endowment. Menzly, Santos and Veronesi assume that the surplus consumption ratio is a continuous-time autoregressive process, solution to,

$$d \left(\frac{1}{s(t)} \right) = \beta \left(\frac{1}{\bar{s}} - \frac{1}{s(t)} \right) dt - \alpha \left(\frac{1}{s(t)} - \frac{1}{v} \right) \sigma_0 dW(t), \quad (7.51)$$

for some constants β , \bar{s} , ω and v . It can be shown that if α is small enough, $s(t) \in (0, v)$. Ljungqvist and Uhlig (2000) utilize similar assumptions to model productivity shocks in their model of catching up with the Joneses. We can check that this model satisfies the two conditions in (7.49).

First, note that Eq. (7.51) implies that the surplus consumption ratio satisfies,

$$ds(t) = s(t) \left[\beta \left(1 - \frac{s(t)}{\bar{s}} \right) + \alpha^2 \sigma_0^2 \left(1 - \frac{1}{v} s(t) \right)^2 \right] dt + \alpha \sigma_0 s(t) \left(1 - \frac{s(t)}{v} \right) dW(t),$$

such that,

$$\bar{\varphi}(s) = s \left[\beta \left(1 - \frac{s}{\bar{s}} \right) + \alpha^2 \sigma_0^2 \left(1 - \frac{1}{v} s \right)^2 \right] + (\sigma_0 - \lambda_{\text{CF}}(s)) \alpha \sigma_0 s \left(1 - \frac{s}{v} \right).$$

Note that the drift function of the surplus ratio under \bar{Q} is quadratic, a property that would crucially lead to the first condition in (7.49) to be satisfied.

By results in Section 7.5.1, the market Sharpe ratio is $\lambda(D, x) = \frac{1}{s}(\sigma_0 - \frac{v(D, x)}{D})$, where $v(D, x)$ is the instantaneous volatility of the habit level, $x = D(1 - s)$ and, by Itô's lemma, equals $v(D, x) = D\sigma_0(1 - s) - D\text{Vol}(s)$, such that, by Itô's lemma again, $\text{Vol}(s) = \alpha\sigma_0s(1 - \frac{s}{v})$. Therefore, the market Sharpe ratio equals,

$$\lambda_{\text{CF}}(s) \equiv \lambda(D, x) = \sigma_0 \left(1 + \alpha \left(1 - \frac{s}{v} \right) \right).$$

Note, parenthetically, that the market Sharpe ratio is countercyclical. Finally, again by results in Section 7.5.1, we can also infer that:

$$R(s) = \delta + g_0 - \sigma_0^2 + \beta \left(1 - \frac{s}{\bar{s}} \right) - \sigma_0^2 \alpha \left(1 - \frac{s}{v} \right).$$

Therefore, we have that,

$$\bar{\varphi}''(s) = -2\frac{\beta}{\bar{s}}, \quad \text{Disc}'(s) = -\frac{\beta}{\bar{s}}, \quad m'(s) = m''(s) = 0.$$

As a result, the two conditions in (7.49) are satisfied, and the price-dividend ratio is affine in s . We can check that the price-dividend ratio is affine in s , through a direct computation. Denote the instantaneous utility with $u(c, x) \equiv \ln(c - x)$, with $c = D$. By the usual asset price representation, we have that the price-dividend ratio, $p(s_0)$, is,

$$\begin{aligned} p(s_0) &= E \left(\int_0^\infty e^{-\delta t} \frac{u_c(c(t), x(t)) c(t)}{u_c(c_0, x_0)} \frac{c(t)}{c_0} dt \right) \\ &= s_0 \cdot \int_0^\infty e^{-\delta t} E \left(\frac{1}{s(t)} \right) dt \\ &= s_0 \cdot \int_0^\infty e^{-\delta t} \left(\frac{1}{\bar{s}} + e^{-\beta t} \left(\frac{1}{s_0} - \frac{1}{\bar{s}} \right) \right) dt \\ &= \frac{1}{\delta + \beta} + \frac{\beta}{\delta \bar{s} (\delta + \beta)} s_0. \end{aligned}$$

Figure 7.14 depicts the price-dividend ratio as a function of the current surplus consumption ratio, using the following parameter values, $\rho = 0.04$, $\beta = 0.15$, and $\bar{s} = 0.03$.

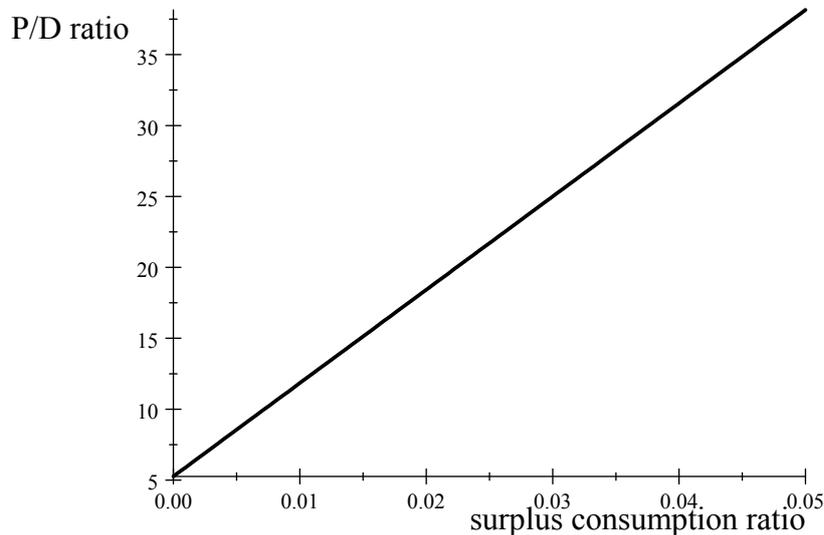


FIGURE 7.14. Price-dividend ratio for the aggregate consumption claim predicted by the Menzly, Santos and Veronesi (2004) model of external habit formation.

As an example of a simple case of a linearity-generating process, consider a model with a constant market Sharpe ratio λ and a constant short-term rate r , but with stochastic dividend growth,

$$\frac{dD(t)}{D(t)} = g(\tau) d\tau + \sigma_0 dW(\tau),$$

where the expected dividend growth, $y(\tau)$, is solution to,

$$dg(\tau) = -g^2(\tau) d\tau + vW(\tau),$$

such that,

$$m(g) = g, \quad \text{Disc}(g) = r + \sigma_0 \lambda, \quad \bar{\varphi}(g) = -g^2(\tau) + (\sigma_0 - v) \lambda.$$

It is straightforward to see that the two conditions in (7.49) hold true—the price-dividend ratio is affine in g . Indeed, assuming that the price-dividend ratio is independent of D , we have that it satisfies the following differential equation,

$$(-g^2 + \sigma_0 v) p' + \frac{1}{2} v^2 p'' + 1 = (r - g) p. \quad (7.52)$$

Let us conjecture, now, that the price-dividend ratio is affine in g , i.e. there are two constant α and β such that

$$p(g) = \alpha_0 + \alpha_1 g.$$

Replacing the previous expression into Eq. (7.52) allows us to pin down the two expressions for α_0 and α_1 such that the solution for the price-dividend ratio is,

$$p(g) = \frac{r + g}{r^2 - \sigma_0 v}.$$

7.6 Appendix 1: Calibration of the tree in Section 7.3

SOLUTION AND CALIBRATION OF THE MODEL. The initial step of the calibration reported in Table 2 involves estimating the two parameters p and δ of the dividend process. Let G be the dividend gross growth rate, computed at a yearly frequency. We calibrate p and δ by a perfect matching of the model's expected dividend growth, $\mu_D \equiv E(G) = pe^{-\delta} + (1-p)e^\delta$, and the model's dividend variance, $\sigma_D^2 \equiv \text{var}(G) = (e^\delta - e^{-\delta})^2 p(1-p)$, to their sample counterparts $\hat{\mu}_D = 1.0594$ and $\hat{\sigma}_D = 0.0602$ obtained on US aggregate dividend data. The result is $(p, \delta) = (0.158, 0.082)$. Given these calibrated values of (p, δ) , we fix $r = 1.0\%$, and proceed to calibrate the probabilities q, q_B and q_G .

To calibrate (q, q_B, q_G) , we need an explicit expression for all the payoffs at each node. By standard risk-neutral evaluation, we obtain a closed form solution for the price of the claim M_S , as follows. For each state $S \in \{G, B, GB\}$, M_S is solution to,

$$\frac{M_S}{D_S} = e^{-r} \mathbb{E}_S \left(\frac{D'_S}{D_S} + \frac{M'_S D'_S}{D'_S D_S} \right), \quad (7A.1)$$

where $\mathbb{E}_S(\cdot)$ is the expectation taken under the risk-neutral probability q_S in state S , $S \in \{G, B, GB\}$, and $q_{GB} = q$, $D_G = e^{2\delta}$, $D_B = e^{-2\delta}$, $D_{GB} = 1$, and D'_S and M'_S are the dividend and the price of the claim as of the next period. Since risk-aversion is constant from the third period on, the price-dividend ratio is constant as well, from the third period on, which implies that $\frac{M_S}{D_S} = \frac{M'_S}{D'_S}$. By using the equality $\frac{M_S}{D_S} = \frac{M'_S}{D'_S}$ in Eq. (7A.1), and solving for M_S , yields,

$$M_S = D_S \frac{q_S e^{-\delta} + (1 - q_S) e^\delta}{e^r - [q_S e^{-\delta} + (1 - q_S) e^\delta]}, \quad S \in \{G, B, GB\}. \quad (7A.2)$$

We calibrate $(q_G, q_B, q_{GB} = q)$ to make the “hybrid” price-dividend (P/D henceforth) ratio M_{GB} , the “good” P/D ratio $\frac{M_G}{e^{2\delta}}$ and the “bad” P/D ratio $\frac{M_B}{e^{-2\delta}}$ in Eq. (7A.2) perfectly match the average P/D ratio, the average P/D ratio during NBER expansion periods, and the average P/D ratio during NBER recession periods (i.e. 31.99, 33.21 and 26.20, from Table 7.1). Given $(p, \delta, r, q, q_S, q_G)$, we compute the P/D ratios in states G and B . For example, the price of the asset in state B is, $P_B = e^{-r} [q_B (e^{-2\delta} + M_B) + (1 - q_B) (1 + M_{GB})]$. Given P_B , we compute the log-return in the bad state as $\ln(\frac{\tilde{\Pi}}{P_B})$, where either $\tilde{\Pi} = e^{-2\delta} + M_B$ with probability p , or $\tilde{\Pi} = 1 + M_{GB}$ with probability $1 - p$. Then, we compute the return volatility in state B . The P/D ratios, the expected log-return and return volatility in state G are computed similarly. (Please notice that volatilities under p and under $\{q_S\}_{S \in \{G, B, GB\}}$ are not the same.)

Next, we recover the risk-aversion parameter η_S in the three states $S \in \{G, B, GB\}$ implied by the previously calibrated probabilities q, q_G and $q = q_{GB}$. As we shall show below, the relevant formula to use is,

$$\frac{q_S}{p} = \frac{e^{\eta_S \delta}}{pe^{\eta_S \delta} + (1-p)e^{-\eta_S \delta}}, \quad S \in \{G, B, GB\}. \quad (7A.3)$$

The values for the “implied” risk-aversion parameter in Table 7.2 are obtained by inverting Eq. (7A.3) for η_S , given the calibrated values of (p, δ, q_S, q_G) .

Finally, we compute the risk-adjusted discount rate as $r + \hat{\sigma}_D \lambda_S$, where λ_S is the Sharpe ratio, which we shall show below to equal,

$$\lambda_S = \frac{q_S - p}{\sqrt{p(1-p)}}, \quad S \in \{G, B, GB\}. \quad (7A.4)$$

PROOF OF EQ. (7A.3). We only provide the derivation of the risk-neutral probability q_B , since the proofs for the expressions of the risk-neutral probabilities q_G and $q = q_{GB}$ are nearly identical. In

equilibrium, the Euler equation for the stock price at the “bad” node is,

$$P_B = \beta E \left[\frac{u'_B(\tilde{D}_S)}{u'_B(e^{-\delta})} (\tilde{D}_S + M_S) \right] = \beta E \left[\tilde{G}_S^{-\eta_B} (\tilde{D}_S + M_S) \right], \quad S \in \{B, GB\}, \quad (7A.5)$$

where: (i) β is the discount rate; (ii) the utility function for consumption C is state dependent and equal to, $u_B(C) = C^{1-\eta_B}/(1-\eta_B)$; (iii) $E(\cdot)$ is the expectation taken under the probability p ; and (iv) the dividend \tilde{D}_S and the gross dividend growth rate \tilde{G}_S are either $\tilde{D}_B = e^{-2\delta}$ and $\tilde{G}_B = \frac{e^{-2\delta}}{e^{-\delta}} = e^{-\delta}$ with probability p , or $\tilde{D}_{GB} = 1$ and $\tilde{G}_{GB} = \frac{1}{e^{-\delta}} = e^{\delta}$ with probability $1-p$.

The model we set up assumes that the asset is elastically supplied or, equivalently, that there exists a storage technology with a fixed rate of return equal to $r = 1\%$. Let us derive the agent’s private evaluation of this asset. The Euler equation for the safe asset is,

$$e^{-r_B} = \beta E[\tilde{G}_S^{-\eta_B}] = \beta \sum_{S \in \{B, GB\}} p_S \tilde{G}_S^{-\eta_B}, \quad (7A.6)$$

where the safe interest rate, r_B , is state dependent, $p_B = p$ and $p_{GB} = 1-p$. Therefore,

$$q_B = \beta e^{r_B} p \tilde{G}_B^{-\eta_B}, \quad 1 - q_B = \beta e^{r_B} (1-p) \tilde{G}_{GB}^{-\eta_B} \quad (7A.7)$$

is a probability distribution. In fact, by plugging q_B and $1 - q_B$ into Eq. (7A.5), one sees that it is the risk-neutral probability distribution. To obtain Eq. (7A.3), note that by Eq. (7A.6), $\beta e^{r_B} = 1/E[\tilde{G}_S^{-\eta_B}]$, which replaced into Eq. (7A.7) yields,

$$\frac{q_B}{p} = \frac{\tilde{G}_B^{-\eta_B}}{E[\tilde{G}_S^{-\eta_B}]}.$$

Eq. (7A.3) follows by the definition of \tilde{G}_S given above.

PROOF OF EQ. (7A.4). Let e^μ the gross expected return of the risky asset. The asset return can take two values: e^{R_ℓ} with probability p , and e^{R_h} with probability $1-p$, and $R_h > R_\ell$. Therefore, for each state, we have that:

$$e^\mu = p e^{R_\ell} + (1-p) e^{R_h}, \quad e^r = q e^{R_\ell} + (1-q) e^{R_h}, \quad (7A.8)$$

where we have omitted the dependence on the state S to alleviate the presentation. The standard deviation of the asset return is $\text{Std}_R = (e^{R_h} - e^{R_\ell}) \sqrt{p(1-p)}$. The Sharpe ratio is defined as

$$\lambda = \frac{e^\mu - e^r}{\text{Std}_R}.$$

By subtracting the two equations in (7A.8),

$$q = p + \frac{e^\mu - e^r}{(e^{R_h} - e^{R_\ell}) \sqrt{p(1-p)}} \sqrt{p(1-p)} = p - \lambda \sqrt{p(1-p)},$$

from which Eq. (7A.4) follows immediately. Note, also, that in terms of this definition of the Sharpe ratio, the risk-neutral expectation of the dividend growth is, $\mathbb{E}(G) = E(G) - \lambda \sigma_D$.

7.7 Appendix 2: Asset prices in a multifactor model

Consider a reduced-form model, where an asset price, S_i say, is a twice-differentiable function of a number of factors, $S_i = S_i(y)$, $i = 1, \dots, m$, and $y = [y_1, \dots, y_d]^\top$ is a vector of factors. We assume that the i -th asset pays off an instantaneous dividend $D_i = D_i(y)$, and that y is a diffusion process:

$$dy(t) = \varphi(y(t)) dt + v(y(t)) dW(t),$$

where φ is d -valued, v is $d \times d$ valued, and W is a d -dimensional Brownian motion. We assume the number of assets does not exceed the number of factors, $m \leq d$, consistently with the framework in Chapter 4. By Itô's lemma:

$$\frac{dS_i}{S_i} = \frac{LS_i}{S_i} dt + \frac{\overbrace{\nabla S_i}^{1 \times d} \underbrace{v}_{d \times d}}{S_i} dW,$$

where LS_i is the infinitesimal operator. Let $r(t)$ be the instantaneous short-term rate. By no-arbitrage, and under regularity conditions, there exists a measurable d -vector process λ , the vector of unit prices of risk associated with the fluctuations of the factors, such that,

$$\begin{bmatrix} \frac{LS_1}{S_1} - r + \frac{D_1}{S_1} \\ \vdots \\ \frac{LS_m}{S_m} - r + \frac{D_m}{S_m} \end{bmatrix} = \underbrace{\sigma}_{m \times d} \underbrace{\lambda}_{d \times 1}, \quad \text{where } \sigma = \begin{bmatrix} \frac{\nabla S_1}{S_1} \\ \vdots \\ \frac{\nabla S_m}{S_m} \end{bmatrix} \cdot v. \quad (7A.9)$$

To simplify, we take this economy to be Markov, assuming that, $r(t) \equiv r(y(t))$ and $\lambda(t) \equiv \lambda(y(t))$.

Eqs. (7A.9) add up to a system of m uncoupled partial differential equations, and the solution is one no-arbitrage price system, assuming no bubbles.

7.8 Appendix 3: Arrow-Debreu PDEs

We develop an interesting connection. Note that by Eq. (7A.9), S is solution to,

$$LS + D = rS + (S_D \sigma_0 D + S_y v_1) \lambda_1 + S_y v_2 \lambda_2, \quad \forall (D, y) \in \mathbb{R} \times \mathbb{R}. \quad (7A.10)$$

Under regularity conditions, the Feynman-Kac representation of the solution to Eq. (7A.10) is,

$$S(D, y) = \int_0^\infty C(D, y, \tau),$$

where,

$$C(D, y, \tau) = \mathbb{E} \left[\exp \left(- \int_0^\tau r(y(t)) dt \right) \cdot D(\tau) \middle| D, y \right] = E[m(\tau) \cdot D(\tau) | D, y],$$

and m is the stochastic discount factor: $m(\tau) = \frac{\xi(\tau)}{\xi(0)}$, $\xi(0) = 1$.

Alternatively, we represent the price S under the physical probability. Given the previous assumptions, we have that ξ necessarily satisfies,

$$\frac{d\xi(\tau)}{\xi(\tau)} = -r(y(\tau)) d\tau - \lambda_1(y(\tau)) dW_1(\tau) + \lambda_2(y(\tau)) dW_2(\tau). \quad (7A.11)$$

Next, define the undiscounted ‘‘Arrow-Debreu adjusted’’ asset price process, defined as:

$$w(D, y) \equiv \Upsilon(D, y) \cdot S(D, y),$$

where $\Upsilon(D, y)$ is as in Eq. (7.24),

$$\Upsilon(D(\tau), y(\tau)) = \xi(\tau) e^{\int_0^\tau \delta(D(s), y(s)) ds}.$$

By results in Section 7.4.2, we know that the following price representation holds true:

$$S(\tau)\xi(\tau) = E \left[\int_\tau^\infty \xi(s) D(s) ds \right], \quad \tau \geq 0.$$

Under regularity conditions, the previous equation can then be understood as the unique Feynman-Kac stochastic representation of the solution to the following partial differential equation

$$Lw(D, y) + f(D, y) = \delta(D, y)w(D, y), \quad \forall (D, y) \in \mathbb{R} \times \mathbb{R},$$

where $f \equiv \Upsilon D$. Eq. (7A.10) then follows by the definition of $Lw(\tau) \equiv \frac{d}{ds} E[\Upsilon S] \big|_{s=\tau}$.

7.9 Appendix 4: The maximum principle

Consider the differential equation:

$$\frac{dx(\tau)}{d\tau} = \phi(\tau), \quad \tau \in (t, T), \quad (7A.12)$$

where ϕ satisfies regularity conditions so as to ensure x remains bounded on (t, T) . Assume that

$$x(T) = 0, \quad (7A.13)$$

and that

$$\text{sign}(\phi(\tau)) = \text{constant on } \tau \in (t, T). \quad (7A.14)$$

We wish to determine the sign of $x(t)$. Under the assumptions on $x(T)$ and ϕ in Eqs. (7A.13) and (7A.14), we have that:

$$\text{sign}(x(t)) = -\text{sign}(\phi). \quad (7A.15)$$

Figure 7.A.1 illustrates the intuitive reasons leading to Eq. (7A.15). For an analytical proof, note that,

$$0 = x(T) = x(t) + \int_t^T \phi(\tau) d\tau \iff x(t) = - \int_t^T \phi(\tau) d\tau,$$

which is Eq. (7A.15), under the assumption ϕ satisfies Eq. (7A.14).

Next, suppose that $x(\tau)$ still satisfies Eq. (7A.12), but that at the same time, is some function of a state variable $y(\tau)$, and time, viz

$$x(\tau) = f(y(\tau), \tau),$$

where the state variable satisfies:

$$\frac{dy(\tau)}{d\tau} = D(\tau), \quad \tau \in (t, T).$$

With enough regularity conditions on ϕ, f, D , we have that

$$\frac{dx}{d\tau} = \left(\frac{\partial}{\partial \tau} + L \right) f, \quad \tau \in (t, T), \quad (7A.16)$$

where $Lf = \frac{\partial f}{\partial y} \cdot D$. Therefore, comparing Eq. (7A.12) with Eq. (7A.16) leaves:

$$\left(\frac{\partial}{\partial \tau} + L \right) f = \phi, \quad \tau \in (t, T). \quad (7A.17)$$

Assume, then, that $x(T) = f(y(T), T) = 0$, such that the solution to Eq. (7A.17) is:

$$f(y(t), t) = - \int_t^T \phi(\tau) d\tau. \quad (7A.18)$$

We have, then, a conclusion similar to that in Eq. (7A.15). That is, suppose that $x(T) = f(y(T), T) = 0$, and that $\text{sign}(\phi(\tau)) = \text{constant on } \tau \in (t, T)$. By Eq. (7A.18):

$$\text{sign}(f(t)) = -\text{sign}(\phi).$$

These results can be extended to stochastic differential equations. Consider the more elaborate operator-theoretic format version of Eq. (7A.17), the one that arises in typical asset pricing models with Brownian motions:

$$0 = \left(\frac{\partial}{\partial \tau} + L - k \right) u + \zeta, \quad \tau \in (t, T). \quad (7A.19)$$

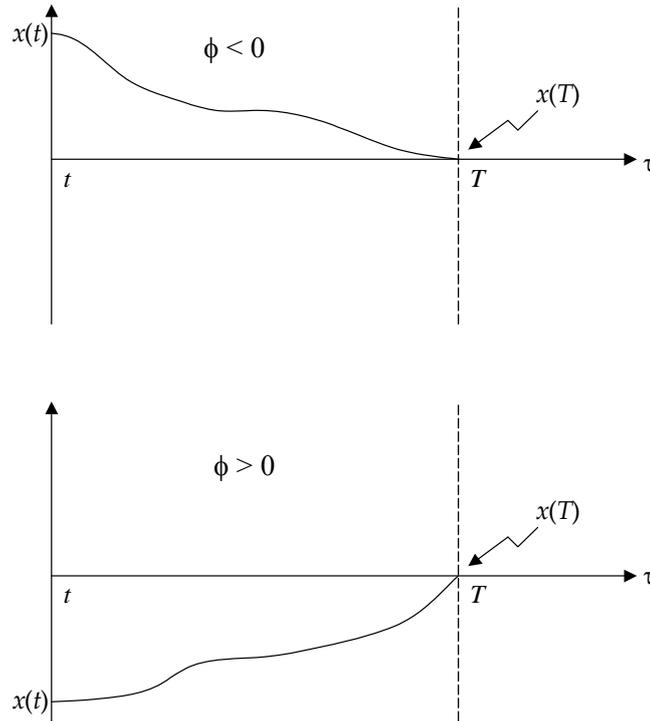


FIGURE 7A.1. Illustration of the maximum principle for ordinary differential equations

Let

$$y(\tau) \equiv e^{-\int_t^\tau k(u)du} u(\tau) + \int_t^\tau e^{-\int_t^u k(s)ds} \zeta(u) du.$$

We claim that if Eq. (7A.19) holds, then y is a martingale under some regularity conditions. Indeed,

$$\begin{aligned} dy(\tau) &= -k(\tau) e^{-\int_t^\tau k(u)du} u(\tau) d\tau + e^{-\int_t^\tau k(u)du} du(\tau) + e^{-\int_t^\tau k(u)du} \zeta(\tau) d\tau \\ &= -k(\tau) e^{-\int_t^\tau k(u)du} u(\tau) + e^{-\int_t^\tau k(u)du} \left[\left(\frac{\partial}{\partial \tau} + L \right) u(\tau) \right] d\tau + e^{-\int_t^\tau k(u)du} \zeta(\tau) d\tau \\ &+ \text{local martingale} \\ &= e^{-\int_t^\tau k(u)du} \left[-k(\tau) u(\tau) + \left(\frac{\partial}{\partial \tau} + L \right) u(\tau) + \zeta(\tau) \right] d\tau + \text{local martingale} \\ &= \text{local martingale,} \end{aligned}$$

where the last equality holds because $\left(\frac{\partial}{\partial \tau} + L - k \right) u + \zeta = 0$. If, finally, y is also a martingale, then,

$$y(t) = u(t) = E[y(T)] = E \left[e^{-\int_t^T k(u)du} u(T) \right] + E \left[\int_t^T e^{-\int_t^u k(s)ds} \zeta(u) du \right].$$

Starting from this relation, we can easily extend the previous proofs on differential equations to stochastic differential ones. Jumps can be dealt with in a similar fashion.

7.10 Appendix 5: Stochastic dominance

7.10.1 Classics

We begin with a review of Rothschild and Stiglitz (1970, 1971) theory definition of risk. Consider the following definition of stochastic dominance:

DEFINITION 7.A.1 (Second-order stochastic dominance). \tilde{x}_2 dominates \tilde{x}_1 if, for each utility function u satisfying $u' \geq 0$, we have also that $E[u(\tilde{x}_2)] \geq E[u(\tilde{x}_1)]$.

We have:

THEOREM 7.A.2. *The following statements are equivalent: (a) \tilde{x}_2 dominates \tilde{x}_1 , or $E[u(\tilde{x}_2)] \geq E[u(\tilde{x}_1)]$; (b) there exists a random variable $\eta > 0 : \tilde{x}_2 = \tilde{x}_1 + \eta$; (c) for every $x > 0$, we have that $F_1(x) \geq F_2(x)$.*

PROOF. We provide the proof when the support is compact, say $[a, b]$. First, we show that (b) \Rightarrow (c). We have: $\forall t_0 \in [a, b]$, $F_1(t_0) \equiv \Pr(\tilde{x}_1 \leq t_0) = \Pr(\tilde{x}_2 \leq t_0 + \eta) \geq \Pr(\tilde{x}_2 \leq t_0) \equiv F_2(t_0)$. Next, we show that (c) \Rightarrow (a). By integrating by parts,

$$E[u(x)] = \int_a^b u(x) dF(x) = u(b) - \int_a^b u'(x) F(x) dx,$$

where we have used the fact that: $F(a) = 0$ and $F(b) = 1$. Therefore,

$$E[u(\tilde{x}_2)] - E[u(\tilde{x}_1)] = \int_a^b u'(x) [F_1(x) - F_2(x)] dx.$$

Finally, it is easy to show that (a) \Rightarrow (b). ||

Next, we turn to the definition of “increasing risk.”

DEFINITION 7.A.3. \tilde{x}_1 is more risky than \tilde{x}_2 if, for each function u satisfying $u'' < 0$, we have also that $E[u(\tilde{x}_1)] \leq E[u(\tilde{x}_2)]$ for \tilde{x}_1 and \tilde{x}_2 having the same mean.

This definition of “increasing risk” does not rely on the sign of u' . Furthermore, if $\text{var}(\tilde{x}_1) > \text{var}(\tilde{x}_2)$, \tilde{x}_1 is not necessarily more risky than \tilde{x}_2 , according to the previous definition. The following is a standard counterexample. Let $\tilde{x}_2 = 1$ with probability (w.p.) 0.8, and 100 w.p. 0.2. Let $\tilde{x}_1 = 10$ w.p. 0.99, and 1090 w.p. 0.01. We have, $E(\tilde{x}_1) = E(\tilde{x}_2) = 20.8$, but $\text{var}(\tilde{x}_1) = 11762.204$ and $\text{var}(\tilde{x}_2) = 1647.368$. However, consider $u(x) = \ln x$. Then, $E(\log(\tilde{x}_1)) = 2.35 > E(\log(\tilde{x}_2)) = 0.92$. It is easily seen that in this particular example, the distribution function F_1 of \tilde{x}_1 “intersects” F_2 , which is in contradiction with the following theorem.

THEOREM 7.A.4. *The following statements are equivalent: (a) \tilde{x}_1 is more risky than \tilde{x}_2 ; (b) \tilde{x}_1 has more weight in the tails than \tilde{x}_2 , i.e. $\forall t, \int_{-\infty}^t [F_1(x) - F_2(x)] dx \geq 0$; (c) \tilde{x}_1 is a mean preserving spread of \tilde{x}_2 , i.e. there exists a random variable $\epsilon : \tilde{x}_1$ has the same distribution as $\tilde{x}_2 + \epsilon$, and $E(\epsilon | \tilde{x}_2 = x_2) = 0$.*

PROOF. Let us begin with (c) \Rightarrow (a). We have,

$$\begin{aligned} E[u(\tilde{x}_1)] &= E[u(\tilde{x}_2 + \epsilon)] \\ &= E[E(u(\tilde{x}_2 + \epsilon) | \tilde{x}_2 = x_2)] \\ &\leq E[u(E(\tilde{x}_2 + \epsilon | \tilde{x}_2 = x_2))] \\ &= E[u(E(\tilde{x}_2 | \tilde{x}_2 = x_2))] \\ &= E[u(\tilde{x}_2)]. \end{aligned}$$

As for (a) \Rightarrow (b), we have that:

$$\begin{aligned} E[u(\tilde{x}_1)] - E[u(\tilde{x}_2)] &= \int_a^b u(x) [f_1(x) - f_2(x)] dx \\ &= u(x) [F_1(x) - F_2(x)] \Big|_a^b - \int_a^b u'(x) [F_1(x) - F_2(x)] dx \\ &= - \int_a^b u'(x) [F_1(x) - F_2(x)] dx \\ &= - \left[u'(x) [\bar{F}_1(x) - \bar{F}_2(x)] \Big|_a^b - \int_a^b u''(x) [\bar{F}_1(x) - \bar{F}_2(x)] dx \right] \\ &= \int_a^b u''(x) [\bar{F}_1(x) - \bar{F}_2(x)] dx - u'(b) [\bar{F}_1(b) - \bar{F}_2(b)], \end{aligned}$$

where $\bar{F}_i(x) = \int_a^x F_i(u) du$. Now, \tilde{x}_1 is more risky than \tilde{x}_2 means that $E[u(\tilde{x}_1)] < E[u(\tilde{x}_2)]$ for $u'' < 0$. By the previous relation, $\bar{F}_1(x) > \bar{F}_2(x)$. Finally, Rothschild and Stiglitz (1970, p. 238) contain the proof that (b) \Rightarrow (c). ||

7.10.2 Dynamic

A quite old issue in financial economics is the relation between asset prices and the volatility of fundamentals (see, e.g., Malkiel, 1979; Pindyck, 1984; Poterba and Summers, 1985; Abel, 1988; Barsky, 1989), which the next theorem formalizes through a general statement.

THEOREM 7.A.5. (Dynamic Stochastic Dominance) *Consider two economies A and B with two fundamental volatilities a_A and a_B and let $\pi_i(x) \equiv a_i(x) \cdot \lambda^i(x)$ and $\rho_i(x)$ ($i = A, B$) the corresponding risk-premium and discount rate. If $a_A > a_B$, the price c^A in economy A is lower than the price c^B in economy B whenever for all $(x, \tau) \in \mathbb{R} \times [0, T]$,*

$$V(x, \tau) \equiv -[\rho_A(x) - \rho_B(x)] c^B(x, \tau) - [\pi_A(x) - \pi_B(x)] c_x^B(x, \tau) + \frac{1}{2} [a_A^2(x) - a_B^2(x)] c_{xx}^B(x, \tau) < 0. \quad (7A.20)$$

If X is the price of a traded asset, $\pi_A = \pi_B$. If in addition ρ is constant, c is decreasing (increasing) in volatility whenever it is concave (convex) in x . This phenomenon is tightly related to the ‘‘convexity effect’’ discussed in the main text. If X is not a traded risk, two additional effects are activated. The first one reflects a discounting adjustment, and is apparent through the first term in the definition of V . The second effect reflects risk-premiums adjustments and corresponds to the second term in the definition of V . Both signs at which these two terms show up in Eq. (7A.20) are intuitive.

Proof of Theorem 7.A.5. The function $c(x, T - s) \equiv \mathbb{E}[\exp(-\int_s^T \rho(x(t)) dt) \cdot \psi(x(T)) | x(s) = x]$ is solution to the following partial differential equation:

$$\begin{cases} 0 = -c_2(x, T - s) + L^* c(x, T - s) - \rho(x) c(x, T - s), & \forall (x, s) \in \mathbb{R} \times [0, T] \\ c(x, 0) = \psi(x), & \forall x \in \mathbb{R} \end{cases} \quad (7A.21)$$

where $L^*c(x, u) = \frac{1}{2}a(x)^2c_{xx}(x, u) + b(x)c_x(x, u)$ and subscripts denote partial derivatives. Clearly, c^A and c^B are both solutions to the partial differential equation (7A.21), but with different coefficients. Let $b_A(x) \equiv b_0(x) - \pi_A(x)$. The price difference $\Delta c(x, \tau) \equiv c^A(x, \tau) - c^B(x, \tau)$ is solution to the following partial differential equation: $\forall(x, s) \in \mathbb{R} \times [0, T)$,

$$0 = -\Delta c_2(x, T-s) + \frac{1}{2}\sigma^B(x)^2\Delta c_{xx}(x, T-s) + b_A(x)\Delta c_x(x, T-s) - \rho_A(x)\Delta c(x, T-s) + V(x, T-s),$$

with $\Delta c(x, 0) = 0$ for all $x \in \mathbb{R}$, and V is as in Eq. (7A.20) of the theorem. The result follows by the maximum principle for partial differential equations. ■

7.11 Appendix 6: Proof of Theorem 7.1

PROOF OF THEOREM 7.1. By differentiating twice the partial differential equation (7A.21) with respect to x , We find that $c^{(1)}(x, \tau) \equiv c_x(x, \tau)$ and $c^{(2)}(x, \tau) \equiv c_{xx}(x, \tau)$ are solutions to the following partial differential equations: $\forall(x, s) \in \mathbb{R}_{++} \times [0, T)$,

$$0 = -c_2^{(1)}(x, T-s) + \frac{1}{2}a(x)^2 c_{xx}^{(1)}(x, T-s) + [b(x) + \frac{1}{2}(a(x)^2)'] c_x^{(1)}(x, T-s) \\ - [\rho(x) - b'(x)] c^{(1)}(x, T-s) - \rho'(x) c(x, T-s),$$

with $c^{(1)}(x, 0) = \psi'(x) \forall x \in \mathbb{R}$, and $\forall(x, s) \in \mathbb{R} \times [0, T)$,

$$0 = -c_2^{(2)}(x, T-s) + \frac{1}{2}a(x)^2 c_{xx}^{(2)}(x, T-s) + [b(x) + (a(x)^2)'] c_x^{(2)}(x, T-s) \\ - \left[\rho(x) - 2b'(x) - \frac{1}{2}(a(x)^2)'' \right] c^{(2)}(x, T-s) \\ - [2\rho'(x) - b''(x)] c^{(1)}(x, T-s) - \rho''(x) c(x, T-s),$$

with $c^{(2)}(x, 0) = \psi''(x) \forall x \in \mathbb{R}$. By the maximum principle for partial differential equations, $c^{(1)}(x, T-s) > 0$ (resp. < 0) $\forall(x, s) \in \mathbb{R} \times [0, T)$ whenever $\psi'(x) > 0$ (resp. < 0) and $\rho'(x) < 0$ (resp. > 0) $\forall x \in \mathbb{R}$. This completes the proof of part (a) of the theorem. The proof of part (b) is obtained similarly. ■

7.12 Appendix 7: Dynamics of habit in Campbell and Cochrane (1999)

We derive Eq. (7.36), by making a slightly more general assumption that the short-term rate we wish to come up with, is affine in $\ln s$, meaning that, the last two terms in Eq. (7.35) sum up to,

$$\eta(1-\phi)(\bar{s}_l - \ln s) - \frac{1}{2}\eta^2\sigma_0^2(1+l(s))^2 = -\text{const.} + b(\bar{s}_l - \ln s), \quad (7A.22)$$

for some b , and where const. is to be determined. The working paper version of Campbell and Cochrane (1999) considers exactly this case.

Define the log of the surplus ratio as

$$s_l(\tau) \equiv \ln\left(1 - e^{x_l(\tau) - c_l(\tau)}\right), \quad (7A.23)$$

where $s_l \equiv \ln s$, $x_l \equiv \ln x$ and $c_l \equiv \ln c$, and consider its first-order Taylor's expansion around the steady state $\overline{x_l - c_l} \equiv E(x_l(\tau) - c_l(\tau))$,

$$s_l(\tau) \approx \bar{s}_l + \left(1 - \frac{1}{\bar{s}}\right)(x_l(\tau) - c_l(\tau) - \overline{x_l - c_l}), \quad (7A.24)$$

where $\bar{s}_l \equiv \ln\left(1 - e^{\overline{x_l - c_l}}\right)$ and $\bar{s} \equiv e^{\bar{s}_l}$. Consider the discrete-time version of Eq. (7.32),

$$s_{l,t+1} - \bar{s}_l \approx \phi(s_{l,t} - \bar{s}_l) + l(s_t) \left(c_{l,t+1} - c_{l,t} - \left(g_0 - \frac{1}{2}\sigma_0^2\right)\right), \quad (7A.25)$$

where $s_{l,t} \equiv s_l(\tau)$. Replacing Eq. (7A.24), evaluated at $\tau = t$ and $\tau = t + 1$, into both sides of the previous approximation, and rearranging terms, leaves:

$$x_{l,t+1} - c_{l,t+1} - \overline{x_l - c_l} = \phi(x_{l,t} - c_{l,t} - \overline{x_l - c_l}) + \frac{l(s_t)}{1 - \frac{1}{\bar{s}}} \left(c_{l,t+1} - c_{l,t} - \left(g_0 - \frac{1}{2}\sigma_0^2\right)\right). \quad (7A.26)$$

The function l in Eq. (7.36) is found by imposing the following three conditions, where the first is a slight generalization of that mentioned in the main text, and the remaining two are the last two conditions in the main text:

- First, the short-term rate in Eq. (7.35) is affine in $\ln s$, i.e. Eq. (7A.22) holds, such that:

$$l(s) = \sqrt{2\frac{1}{\eta^2\sigma_0^2}(\eta(1-\phi) - b)(\bar{s}_l - \ln s) + \frac{2}{\eta^2\sigma_0^2}\text{const.} - 1}. \quad (7A.27)$$

- Second, habit is predetermined at the steady state, meaning that $x_{l,t+1}$ does not change with $c_{l,t+1}$, which by Eq. (7A.26), it does not, when:

$$l(\bar{s}) = \frac{1}{\bar{s}} - 1. \quad (7A.28)$$

Evaluating Eq. (7A.27) at the steady state \bar{s} , and using the previous condition delivers, $\frac{2}{\eta^2\sigma_0^2}\text{const.} = \frac{1}{\bar{s}^2}$, such that Eq. (7A.27) is,

$$l(s) = \sqrt{2\frac{1}{\eta^2\sigma_0^2}(\eta(1-\phi) - b)(\bar{s}_l - \ln s) + \frac{1}{\bar{s}^2} - 1}. \quad (7A.29)$$

- Third, habit is predetermined near the steady state, meaning that,

$$\left. \frac{d}{ds_l} \left(\frac{dx_l}{dc_l} \right) \right|_{s=\bar{s}_l} = 0. \quad (7A.30)$$

We, then, need to find the dynamics of $x_{l,t+1}$, expressed as a function of $c_{l,t+1}$. By the definition of the log-surplus consumption ratio in Eq. (7A.23), we have that $x_{l,t+1} = \ln \left(1 - e^{\sigma(c_{l,t+1})} \right) + c_{l,t+1}$, where $\sigma(c_{l,t+1}) \equiv s_{l,t+1}$ and $s_{l,t+1}$ is as in Eq. (7A.25), such that, using Eq. (7A.25):

$$\frac{dx_{l,t+1}}{dc_{l,t+1}} = 1 - \frac{1}{e^{-\sigma(c_{l,t+1})} - 1} \sigma'(c_{l,t+1}) = 1 - \frac{l_o(s_{l,t})}{e^{-s_{l,t+1}} - 1},$$

where we have set $l_o(s_l) \equiv l(s)$. Therefore, Eq. (7A.30) is: $\left. \frac{d}{ds_l} \left(1 - \frac{l_o(s_l)}{e^{-s_l}} \right) \right|_{s=\bar{s}_l} = 0$, which leaves, after simple computation, and using Eq. (7A.28),

$$l'_o(\bar{s}_l) = -\frac{1}{\bar{s}_l}.$$

By taking the derivative in Eq. (7A.29), and replacing into the left hand side of the previous equation, and solving for \bar{s}_l , yields,

$$\bar{s} = \sigma_0 \sqrt{\frac{\eta^2}{\eta(1-\phi) - b}},$$

which is the expression of the main text, for $b = 0$. By replacing this expression of \bar{s} into Eq. (7A.29), leaves Eq. (7.36) in the main text.

Finally, note that, now, the expression of the short-term rate can be found after simple computations:

$$R(s) = \delta + \eta \left(g_0 - \frac{1}{2} \sigma_0^2 \right) - \frac{1}{2} (\eta(1-\phi) - b) + b(\bar{s} - \ln s).$$

7.13 Appendix 8: An algorithm to simulate discrete-time pricing models

Consider the pricing equation,

$$S = E [m \cdot (S' + D')], \quad m = \beta \frac{u_c(D', x')}{u_c(D, x)} = \beta \left(\frac{s'}{s} \right)^{-\eta} \left(\frac{D'}{D} \right)^{-\eta}.$$

The price-dividend ratio, $p \equiv S/D$ say, satisfies:

$$p = E \left[m \frac{D'}{D} (1 + p') \right], \quad \frac{D'}{D} = e^{g_0 + w}.$$

The previous equation is a functional equation in $p(s)$, say:

$$p(s) = E [g(s', s) (1 + p(s')) | s], \quad g(s', s) = \beta \left(\frac{s'}{s} \right)^{-\eta} \left(\frac{D'}{D} \right)^{1-\eta}.$$

A numerical solution can be implemented as follows. Create a grid and define $p_j = p(s_j)$, $j = 1, \dots, N$, for some N . We have,

$$\begin{bmatrix} p_1 \\ \vdots \\ p_N \end{bmatrix} = \begin{bmatrix} b_1 \\ \vdots \\ b_N \end{bmatrix} + \begin{bmatrix} a_{11} & \cdots & a_{1N} \\ \vdots & \ddots & \vdots \\ a_{1N} & \cdots & a_{NN} \end{bmatrix} \begin{bmatrix} p_1 \\ \vdots \\ p_N \end{bmatrix},$$

$$b_i = \sum_{j=1}^N a_{ji}, \quad a_{ji} = g_{ji} \cdot p_{ji}, \quad g_{ji} = g(s_j, s_i), \quad p_{ji} = \Pr(s_j | s_i) \cdot \Delta s,$$

where Δs is the integration step, $s_1 = s_{\min}$, $s_N = s_{\max}$, s_{\min} and s_{\max} are the boundaries in the approximation, and $\Pr(s_j | s_i)$ is the transition density from state i to state j - in this case, a Gaussian transition density. Let $p = [p_1 \cdots p_N]^\top$, $b = [b_1 \cdots b_N]^\top$, and let A be a matrix with elements a_{ji} . The solution is,

$$p = (I - A)^{-1} b. \tag{7A.31}$$

The model can be simulated in the following manner. Let \underline{s} and \bar{s} be the boundaries of the underlying state process. Fix $\Delta s = \frac{\bar{s} - \underline{s}}{N}$. Draw states. State s^* is drawn. Then,

1. If $\min \{s^* - \underline{s}, \bar{s} - s^*\} = s^* - \underline{s}$, let k be the smallest integer close to $\frac{s^* - \underline{s}}{\Delta s}$. Let $s_{\min} = s^* - k\Delta s$, and $s_{\max} = s_{\min} + N \cdot \Delta s$.
2. If $\min \{s^* - \underline{s}, \bar{s} - s^*\} = \bar{s} - s^*$, let k be the biggest integer close to $\frac{s^* - \underline{s}}{\Delta s}$. Let $s_{\max} = s^* + k\Delta s$, and $s_{\min} = s_{\max} - N \cdot \Delta s$.

The previous algorithm avoids interpolations, and ensures that during the simulations, p is computed in correspondence of exactly the state s^* that is drawn. Precisely, once s^* is drawn, we proceed to the following two steps: (i) create the corresponding grid $s_1 = s_{\min}$, $s_2 = s_{\min} + \Delta s, \dots, s_N = s_{\max}$ according to the previous rules; and (ii) compute the solution from Eq. (7A.31). In this way, one has $p(s^*)$ at hand—the simulated P/D ratio when state s^* is drawn.

7.14 Appendix 9: Heuristic details of learning in continuous time

We derive Eq. (7.40). We have,

$$dD = g d\tau + dW,$$

and, by Eq. (7.39),

$$\pi(D) = \frac{p\phi(D-A)}{p\phi(D-A) + (1-p)\phi(D+A)} = \frac{p}{p + (1-p)e^{-2AD}}$$

where the second equality follows by the Gaussian distribution assumption $\phi(x) \propto e^{-\frac{1}{2}x^2}$, and straight forward simplifications. By simple computations,

$$\frac{1 - \pi(D)}{\pi(D)} = \frac{(1-p)e^{-2AD}}{p}, \quad \pi'(D) = 2A\pi(D)^2 \frac{(1-p)e^{-2AD}}{p}, \quad \pi''(D) = 2A\pi'(D)[1 - 2\pi(D)]. \quad (7A.32)$$

By construction,

$$g = \pi(D)A + [1 - \pi(D)](-A) = A[2\pi(D) - 1].$$

Therefore, by Itô's lemma,

$$d\pi = \pi' dD + \frac{1}{2}\pi'' d\tau = \pi' dD + A\pi'(1 - 2\pi) d\tau = \pi'[g + A(1 - 2\pi)] d\tau + \pi' dW = \pi' dW.$$

By using the relations in (7A.32) once again,

$$d\pi = 2A\pi(1 - \pi) dW. \quad \blacksquare$$

7.15 Appendix 10: Linear regime-switching economies

We prove a claim made in Section 7.5.3. Consider a complete markets economy where dividends, consumption, and signals (D_t, c_t, a_t) satisfy:

$$\begin{pmatrix} dD_t/D_t \\ dc_t/c_t \\ da_t \end{pmatrix} = \begin{pmatrix} \theta \\ \bar{g} \\ \theta \end{pmatrix} dt + \begin{pmatrix} \bar{\sigma}_0 \\ \bar{\sigma} \\ \bar{\sigma}_a \end{pmatrix} dw_t, \quad \begin{aligned} \bar{\sigma}_0 &\equiv (\sigma_0 \ 0 \ 0) \\ \bar{\sigma} &\equiv (\sigma_1 \ \sigma_2 \ \sigma_3) \\ \bar{\sigma}_a &\equiv (\sigma_4 \ \sigma_5 \ \sigma_6) \end{aligned}$$

and $w = (w_1 \ w_2 \ w_3)^\top$ denotes a vector standard Brownian motion, with θ being a two-state $(\bar{\theta}, \underline{\theta})$ Markov chain, and \bar{g}, σ_i are constants. Let $\sigma_3\sigma_5 \neq \sigma_2\sigma_6$. Then, there are *no* CRRA representative agent equilibria in which price-dividend ratios are convex in expected dividend growth. To demonstrate this claim, we apply the filtering results of Liptser and Shiryaev (2001) (Vol. I), and find that the previous economy is isomorphic to one in which,

$$\begin{pmatrix} dD_t/D_t \\ dc_t/c_t \\ dy_t \end{pmatrix} = \begin{pmatrix} y_t \\ \bar{g} \\ k(y^* - y_t) \end{pmatrix} dt + \begin{pmatrix} \bar{\sigma}_0 \\ \bar{\sigma} \\ (\bar{\theta} - y_t)(y_t - \underline{\theta}) \cdot \bar{\sigma}_y \end{pmatrix} dW_t,$$

where $W = (W_1 \ W_2 \ W_3)^\top$ is a vector standard Brownian motion, $k, y^* > 0$, and $\bar{\sigma}_y$ is some vector satisfying $\bar{\sigma} \cdot \bar{\sigma}_y = 0$. Standard arguments lead that in equilibrium, the short-term rate, R , is constant and $\lambda_i \propto \sigma_i$ ($i = 1, 2, 3$). The claim follows by $\bar{\sigma} \perp \bar{\sigma}_y$ and Theorem 7.1.

7.16 Appendix 11: Bond price convexity revisited

Consider a short-term rate process $r(\tau)$, and let $u(r_0, T)$ be the price of a bond expiring at time T when the current short-term rate is r_0 :

$$u(r_0, T) = \mathbb{E} \left[\exp \left(- \int_0^T r(\tau) d\tau \right) \middle| r_0 \right].$$

As pointed out in Section 7.6, Theorem 7.1-(ii) implies that in scalar diffusion models of the short-term rate, such as those dealt with in Chapter 12, one has $u_{11}(r_0, T) < 0$ whenever $b'' < 2$, where b is the risk-neutralized drift of r . This result, obtained by Mele (2003), can be proved through the Feynman-Kac representation of u_{11} , and a similar proof can be used to show Theorem 7.1-(ii). This appendix provides a more intuitive derivation under a set of simplifying assumptions. By Eq. (6) p. 685 in Mele (2003),

$$u_{11}(r_0, T) = \mathbb{E} \left[\left(\left(\int_0^T \frac{\partial r}{\partial r_0}(\tau) d\tau \right)^2 - \int_0^T \frac{\partial^2 r}{\partial r_0^2}(\tau) d\tau \right) \exp \left(- \int_0^T r(\tau) d\tau \right) \right].$$

Hence $u_{11}(r_0, T) > 0$ whenever

$$\int_0^T \frac{\partial^2 r}{\partial r_0^2}(\tau) d\tau < \left(\int_0^T \frac{\partial r}{\partial r_0}(\tau) d\tau \right)^2. \quad (7A.33)$$

To keep the presentation simple, assume $r(\tau)$ is solution to:

$$dr(\tau) = b(r(\tau))dt + a_0 r(\tau) dW(\tau),$$

where a_0 is a constant. We have,

$$\frac{\partial r}{\partial r_0}(\tau) = \exp \left(\int_0^\tau b'(r(u)) du - \frac{1}{2} a_0^2 \tau + a_0 W(\tau) \right),$$

and

$$\frac{\partial^2 r}{\partial r_0^2}(\tau) = \frac{\partial r(\tau)}{\partial r_0} \left[\int_0^\tau b''(r(u)) \frac{\partial r(u)}{\partial r_0} du \right].$$

Therefore, if $b'' < 0$, then $\partial^2 r(\tau)/\partial r_0^2 < 0$, and by inequality (12.56), $u_{11} > 0$. Note that this result can considerably be improved. Suppose that $b'' < 2$, instead of $b'' < 0$. By the previous equality,

$$\frac{\partial^2 r}{\partial r_0^2}(\tau) < 2 \frac{\partial r(\tau)}{\partial r_0} \int_0^\tau \frac{\partial r(u)}{\partial r_0} du,$$

and consequently,

$$\int_0^T \frac{\partial^2 r}{\partial r_0^2}(\tau) d\tau < 2 \int_0^T \frac{\partial r(\tau)}{\partial r_0} \left(\int_0^\tau \frac{\partial r(u)}{\partial r_0} du \right) d\tau = \left(\int_0^T \frac{\partial r(u)}{\partial r_0} du \right)^2,$$

which is inequality (12.56).

References

- Abel, A.B. (1988): "Stock Prices under Time-Varying Dividend Risk: An Exact Solution in an Infinite-Horizon General Equilibrium Model." *Journal of Monetary Economics* 22, 375-393.
- Abel, A.B. (1990): "Asset Prices under Habit Formation and Catching Up with the Joneses." *American Economic Review Papers and Proceedings* 80, 38-42.
- Andersen, T. G., T. Bollerslev and F. X. Diebold (2002): "Parametric and Nonparametric Volatility Measurement." Forthcoming in Aït-Sahalia, Y. and L. P. Hansen (Eds.): *Handbook of Financial Econometrics*.
- Bakshi, G. and D. Madan (2000): "Spanning and Derivative Security Evaluation." *Journal of Financial Economics* 55, 205-238.
- Bajoux-Besnainou, I. and J.-C. Rochet (1996): "Dynamic Spanning: Are Options an Appropriate Instrument?" *Mathematical Finance* 6, 1-16.
- Bansal, R. and A. Yaron (2004): "Risks for the Long Run: A Potential Resolution of Asset Pricing Puzzles." *Journal of Finance* 59, 1481-1509.
- Barberis, N., M. Huang and T. Santos (2001): "Prospect Theory and Asset Prices." *Quarterly Journal of Economics* 116, 1-53.
- Barsky, R. B. (1989): "Why Don't the Prices of Stocks and Bonds Move Together?" *American Economic Review* 79, 1132-1145.
- Barsky, R. B. and J. B. De Long (1990): "Bull and Bear Markets in the Twentieth Century." *Journal of Economic History* 50, 265-281.
- Barsky, R. B. and J. B. De Long (1993): "Why Does the Stock Market Fluctuate?" *Quarterly Journal of Economics* 108, 291-311.
- Bergman, Y. Z., B. D. Grundy, and Z. Wiener (1996): "General Properties of Option Prices." *Journal of Finance* 51, 1573-1610.
- Black, F. and M. Scholes (1973): "The Pricing of Options and Corporate Liabilities." *Journal of Political Economy* 81, 637-659.
- Bloomfield, P. and Steiger, W. L. (1983): *Least Absolute Deviations*. Boston: Birkhäuser.
- Brennan, M. J. and Y. Xia (2001): "Stock Price Volatility and Equity Premium." *Journal of Monetary Economics* 47, 249-283.
- Britten-Jones, M. and A. Neuberger (2000): "Option Prices, Implied Price Processes and Stochastic Volatility." *Journal of Finance* 55, 839-866.
- Brunnermeier, M. K. and S. Nagel (2007): "Do Wealth Fluctuations Generate Time-Varying Risk Aversion? Micro-Evidence on Individuals' Asset Allocation." Forthcoming in *American Economic Review*.

- Campbell, J. Y. (2003): “Consumption-Based Asset Pricing.” In: Constantinides, G.M., M. Harris and R. M. Stulz (Editors): *Handbook of the Economics of Finance* (Volume 1B: Chapter 13), 803-887.
- Campbell, J. Y., and J. H. Cochrane (1999): “By Force of Habit: A Consumption-Based Explanation of Aggregate Stock Market Behavior.” *Journal of Political Economy* 107, 205-251.
- Carr, P. and D. Madan (2001): “Optimal Positioning in Derivative Securities.” *Quantitative Finance* 1, 19-37.
- Christiansen, C., M. Schmeling and A. Schrimpf (2011): “A Comprehensive Look at Financial Volatility Prediction by Economic Variables.” Working Paper, CREATES, Aarhus University.
- Clark, T.E. and K.D. West (2007): “Approximately Normal Tests for Equal Predictive Accuracy in Nested Models.” *Journal of Econometrics* 138, 291-311.
- Constantinides, G.M. (1990): “Habit Formation: A Resolution of the Equity Premium Puzzle.” *Journal of Political Economy* 98, 519-543.
- Corradi, V., W. Distaso and A. Mele (2010): “Macroeconomic Determinants of Stock Market Volatility and Volatility Risk-Premia.” Working paper University of Warwick, Imperial College, and London School of Economics.
- David, A. (1997): “Fluctuating Confidence in Stock Markets: Implications for Returns and Volatility.” *Journal of Financial and Quantitative Analysis* 32, 427-462.
- Demeterfi, K., E. Derman, M. Kamal and J. Zou (1999): “A Guide to Volatility and Variance Swaps.” *Journal of Derivatives* 6, 9-32.
- Detemple, J. B. (1986): “Asset Pricing in a Production Economy with Incomplete Information.” *Journal of Finance* 41, 383-391.
- Duesenberry, J.S. (1949): *Income, Saving, and the Theory of Consumer Behavior*. Cambridge, Mass.: Harvard University Press.
- El Karoui, N., M. Jeanblanc-Picqué and S. E. Shreve (1998): “Robustness of the Black and Scholes Formula.” *Mathematical Finance* 8, 93-126.
- Fama, E. F. and K. R. French (1989): “Business Conditions and Expected Returns on Stocks and Bonds.” *Journal of Financial Economics* 25, 23-49.
- Ferson, W. E. and C. R. Harvey (1991): “The Variation of Economic Risk Premiums.” *Journal of Political Economy* 99, 385-415.
- Fornari, F. and A. Mele (2010): “Financial Volatility and Real Economic Activity.” Working paper European Central Bank and London School of Economics.
- Gabaix, X. (2009): “Linearity-Generating Processes: A Modelling Tool Yielding Closed Forms for Asset Prices.” Working paper New York University.

- Genotte, G. (1986): “Optimal Portfolio Choice Under Incomplete Information.” *Journal of Finance* 41, 733-746.
- Giacomini, R. and H. White (2006): “Tests of Conditional Predictive Ability.” *Econometrica* 74, 1545-1578.
- Glosten, L., R. Jagannathan and D. Runkle (1993): “On the Relation between the Expected Value and the Volatility of the Nominal Excess Return on Stocks.” *Journal of Finance* 48, 1779-1801.
- Gordon, M. (1962): *The Investment, Financing, and Valuation of the Corporation*. Homewood, IL: Irwin.
- Hajek, B. (1985): “Mean Stochastic Comparison of Diffusions.” *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* 68, 315-329.
- Huang, C.-F. and Pagès, H. (1992): “Optimal Consumption and Portfolio Policies with an Infinite Horizon: Existence and Convergence.” *Annals of Applied Probability* 2, 36-64.
- Jagannathan, R. (1984): “Call Options and the Risk of Underlying Securities.” *Journal of Financial Economics* 13, 425-434.
- Karatzas, I. and S.E. Shreve (1991): *Brownian Motion and Stochastic Calculus*. Berlin: Springer Verlag.
- Kijima, M. (2002): “Monotonicity and Convexity of Option Prices Revisited.” *Mathematical Finance* 12, 411-426.
- Liptser, R. S. and A. N. Shiryaev (2001): *Statistics of Random Processes*. Berlin, Springer-Verlag. [2001a: Vol. I (*General Theory*). 2001b: Vol. II (*Applications*).]
- Ljungqvist, L. and H. Uhlig (2000): “Tax Policy and Aggregate Demand Management under Catching Up with the Joneses.” *American Economic Review* 90, 356-366.
- Malkiel, B. (1979): “The Capital Formation Problem in the United States.” *Journal of Finance* 34, 291-306.
- Mehra, R. and E.C. Prescott (2003): “The Equity Premium in Retrospect.” In Constantinides, G.M., M. Harris and R. M. Stulz (Editors): *Handbook of the Economics of Finance* (Volume 1B, chapter 14), 889-938.
- Mele, A. (2003): “Fundamental Properties of Bond Prices in Models of the Short-Term Rate.” *Review of Financial Studies* 16, 679-716.
- Mele, A. (2005): “Rational Stock Market Fluctuations.” WP FMG-LSE.
- Mele, A. (2007): “Asymmetric Stock Market Volatility and the Cyclical Behavior of Expected Returns.” *Journal of Financial Economics* 86, 446-478.
- Menzly, L., T. Santos and P. Veronesi (2004): “Understanding Predictability.” *Journal of Political Economy* 111, 1, 1-47.

- Pindyck, R. (1984): “Risk, Inflation and the Stock Market.” *American Economic Review* 74, 335-351.
- Paye, B. P. (2011): “Déjà Vol: Predictive Regressions for Aggregate Stock Market Volatility Using Macroeconomic Variables” Working Paper, Rice University.
- Poterba, J. and L. Summers (1985): “The Persistence of Volatility and Stock Market Fluctuations.” *American Economic Review* 75, 1142-1151.
- Romano, M. and N. Touzi (1997): “Contingent Claims and Market Completeness in a Stochastic Volatility Model.” *Mathematical Finance* 7, 399-412.
- Rothschild, M. and J. Stiglitz (1970): “Increasing Risk: I. A Definition.” *Journal of Economic Theory* 2, 225-243.
- Rothschild, M. and J. Stiglitz (1971): “Increasing Risk: II. Its Economic Consequences.” *Journal of Economic Theory* 5, 66-84.
- Ryder, H.E. and G.M. Heal (1973): “Optimal Growth with Intertemporally Dependent Preferences.” *Review of Economic Studies* 40, 1-33.
- Schwert, G. W. (1989a): “Why Does Stock Market Volatility Change Over Time?” *Journal of Finance* 44, 1115-1153.
- Schwert, G. W. (1989b): “Business Cycles, Financial Crises and Stock Volatility.” *Carnegie-Rochester Conference Series on Public Policy* 31, 83-125.
- Sundaresan, S.M. (1989): “Intertemporally Dependent Preferences and the Volatility of Consumption and Wealth.” *Review of Financial Studies* 2, 73-89.
- Timmermann, A. (1993): “How Learning in Financial Markets Generates Excess Volatility and Predictability in Stock Prices.” *Quarterly Journal of Economics* 108, 1135-1145.
- Timmermann, A. (1996): “Excess Volatility and Return Predictability of Stock Returns in Autoregressive Dividend Models with Learning.” *Review of Economic Studies* 63, 523-577.
- Veronesi, P. (1999): “Stock Market Overreaction to Bad News in Good Times: A Rational Expectations Equilibrium Model.” *Review of Financial Studies* 12, 975-1007.
- Veronesi, P. (2000): “How Does Information Quality Affect Stock Returns?” *Journal of Finance* 55, 807-837.
- Wang, S. (1993): “The Integrability Problem of Asset Prices.” *Journal of Economic Theory* 59, 199-213.

8

Tackling the puzzles

8.1 Introduction

This chapter discusses models that aim to address the empirical puzzles surveyed in the previous two chapters. The most prominent is the “equity premium puzzle”—the difficulty of the neoclassical model to predict an equity premium quantitatively consistent with the data. The difficulty lies in the circumstance that we would require a quite high level of risk-aversion, to reconcile models with data. Moreover, a high risk-aversion implies a low elasticity of intertemporal substitution and, hence, an implausibly high volatility of the interest rates—the “interest rate puzzle.”

In the early attempts to address the equity premium puzzle, the assumption of a representative agent with CRRA preferences was replaced with that of a representative agent with non-expected utility. In the non-expected utility framework, risk-aversion can be understood independently of the elasticity of intertemporal substitution. This approach, described in Section 8.2, does not necessarily deliver a satisfactory solution to the equity premium puzzle. We show that a possible resolution of the puzzle would require the existence of a number of state variables affecting the price-dividend ratio, for example, a long-run risk, such as the existence of a highly persistent consumption growth that turns even a small shock into an economic damage perduring for years and years. Section 8.3 explores an alternative channel, proposed to resolve the equity premium puzzle, based on a variant of habit formation. The external habit formation model reviewed in Section 7.5.2 of Chapter 7 relies on the existence of a representative agent with high risk-aversion. Alternatively, one may imagine economies with heterogeneous agents, each concerned with a reference consumption benchmark, and “catching up with the Joneses,” and each displaying a different local curvature of the utility function. This formulation has the potential to explain the equity premium subject to empirical nuances leading to a number of caveats.

Section 8.4 describes economies where agents are hit by risks that they cannot ensure against through security trading, such as idiosyncratic shocks relating to consumption shrinkages due to possible recessions. How come idiosyncratic shocks can have a price impact? The key assumption of the models we survey is that although agents might even be all equal, ex-ante, they might then be hit by shocks of different amplitude, i.e. idiosyncratic shocks. The perspective of a job

loss is one important instance of an idiosyncratic shock. Recessions do not necessarily affect all agents in the same way. Some agents might be hurt more than others. The likely occurrence of a job loss might induce agents to act prudently while investing in the stock market, and this agents' behavior leads to a potential resolution of the equity premium puzzle, theoretically at least. Despite the clarity of this explanation, the empirical implications of idiosyncratic risk are not necessarily exhaustive. Indeed, a natural hedge against idiosyncratic risk is self-insurance, i.e. the ability to save in good times to cope with adversities possibly occurring in bad times. Agents might actually eliminate a big portion of idiosyncratic risk by insuring themselves while having access to capital markets, thereby making idiosyncratic risk irrelevant, in practice, to the explanation of the equity premium. In order for idiosyncratic risk to really matter, we would need to observe a big and persistent idiosyncratic risk, or capital markets transactions to be so expensive, to prevent, or dissuade, agents from implementing self-insurance plans. But in reality, idiosyncratic risk is not as big and persistent, and market transaction costs are not as large, as the standard models with idiosyncratic require. Section 8.5 considers, then, economies with incomplete markets. These economies are unlikely to be consistent with the equity premium, when agents face the same degree of market incompleteness. Economies with agents confronting heterogeneous incompleteness have the potential to resolve the premium puzzle, at least in the case where there is only one agent who has access to the stock market. It is this agent's concern for the aggregate macroeconomic risk he bears, which leads him to require a premium consistent with data. It is an open question whether this quite encouraging model is resilient to a generalization to economies with many agents accessing the equity market.

Section 8.6 deals with issues arising within production-based economies. In these economies, consumption is endogenous, and an increase in the agents risk aversion might actually lead to a decreased consumption volatility. One additional important difficulty in these economies is that capital supply is infinitely elastic, such that the price of capital is quite smooth. To increase capital price volatility, we need hindrances in the capital formation process, such as the presence of adjustment costs, or an added volatility of the demand for capital, obtained for example when agents have habit formation over their consumption plans. Both rigidities in the capital formation process and volatility in the demand of capital are needed, in order to explain the equity premium. It is still unclear whether habit formation in production economies is exempt from the same criticism to which it is subject in endowment economies such as those of Section 8.3.

Section 8.7 presents a simple model, where we try to understand the extent to which equity volatility can be explained by firms leverage—a very old hypothesis. Section 8.8 discusses recent models capable to explain the cross-section of asset returns, and relying on multiple trees. Section 8.9 surveys predictions that the previous models make about the yield curve.

Section 8.10 concludes the chapter. It deals with an intriguing topic: what do financial economists and macroeconomists have really in common? Granted, in many of the models surveyed in this chapter, we try to understand asset prices in a context of the business cycle, but as we will see, our models necessitate a re-vamp of many assumptions underlying the neo-classical paradigm. Yet macroeconomists do not necessarily seem to acknowledge our asset pricing lessons. Are macroeconomists mistaken? Or is there a case for a modern version of a *dichotomy* between the real and the financial spheres of the economy? A simple model shows there is a potential for the hypothesis of a separation between finance and macroeconomics. At the same time, this potential is at the moment quite far from having revealed to lead to practical content.

8.2 Non-expected utility

The standard intertemporal additive separable utility function confounds intertemporal substitution effects and attitudes towards risk. This fact is problematic. Epstein and Zin (1989, 1991) and Weil (1989) consider a class of recursive, but not necessarily expected utility, preferences. In this section, we present some details of this approach, without insisting on the theoretic underpinnings, which the reader will find in Epstein and Zin (1989). We provide a basic definition and derivation of this class of preferences, and analyze its asset pricing implications.

8.2.1 Recursive formulation

Consider a decision-maker endowed with utility at time t , denoted as v_t , assumed to satisfy:

$$v_t = W(c_t, \hat{v}_{t+1}),$$

where W is the *aggregator*, and \hat{v}_{t+1} is the certainty-equivalent utility at $t + 1$, defined as $h(\hat{v}_{t+1}) = E_t[h(v_{t+1})]$, and, finally, h is a von Neumann-Morgenstern utility function. Note, obviously, that the certainty equivalent \hat{v}_{t+1} depends on the agent's risk-attitudes encoded into h . Inverting for the certainty equivalent, and replacing into W , leaves:

$$v_t = W(c_t, h^{-1}[E_t(h(v_{t+1}))]).$$

A common assumption made about W is:

$$W(c, \hat{v}) = (c^\rho + \beta \hat{v}^\rho)^{1/\rho} \quad \text{and} \quad h(\hat{v}) = \hat{v}^{1-\eta}, \quad (8.1)$$

for three positive constants ρ , η and β . In this formulation, risk-attitudes for static wealth gambles have still the classical CRRA flavor. Precisely, we say that η is the relative risk-aversion (RRA) for static wealth gambles, and $\psi \equiv (1 - \rho)^{-1}$ is the elasticity of intertemporal substitution (EIS). We have,

$$\hat{v}_{t+1} = h^{-1}[E_t(h(v_{t+1}))] = (E_t(v_{t+1}^{1-\eta}))^{\frac{1}{1-\eta}}.$$

Naturally, in the absence of uncertainty, $v_t^\rho = c_t^\rho + \beta v_{t+1}^\rho$ —another clear illustration that ψ is the EIS. The parametrization for the aggregator in Eq. (8.1) implies that:

$$v_t = \left(c_t^\rho + \beta (E_t(v_{t+1}^{1-\eta}))^{\frac{\rho}{1-\eta}} \right)^{\frac{1}{\rho}}. \quad (8.2)$$

It is straight forward to note that this utility function collapses to the standard intertemporal additively separable case, once the RRA equals the inverse of the EIS, i.e. when $\rho = 1 - \eta$, in which case $v_t^{1-\eta} = E(\sum_{n=0}^{\infty} \beta^n c_{t+n}^{1-\eta})$. Yet consider the general case, and note that, obviously, the function $V = v^{1-\eta}/(1-\eta)$ is ordinally equivalent to v_t in Eq. (8.2) and, satisfies,

$$V_t = \frac{1}{1-\eta} \left(c_t^\rho + \beta ((1-\eta) E_t(V_{t+1}))^{\frac{\rho}{1-\eta}} \right)^{\frac{1-\eta}{\rho}}. \quad (8.3)$$

We shall use Eq. (8.3), to derive asset pricing restriction emanating from portfolio choices of a representative agent, and to analyze the equilibrium.

8.2.2 Testable restrictions

Let us define cum-dividend wealth as $x_t \equiv \sum_{i=1}^m (P_{it} + D_{it}) \theta_{it}$. In the Appendix, we show that x_t evolves as follows:

$$x_{t+1} = (x_t - c_t) \omega_t^\top (\mathbf{1}_m + \mathbf{r}_{t+1}) \equiv (x_t - c_t) (1 + r_{M,t+1}), \quad (8.4)$$

where ω is the vector of proportions of wealth invested in the m assets, \mathbf{r}_{t+1} is the vector of asset returns, with any component i being equal to, $r_{it+1} \equiv \frac{P_{it+1} + D_{it+1} - P_{it}}{P_{it}}$, and $r_{M,t}$ is the return on the market portfolio, defined as,

$$r_{M,t+1} = \sum_{i=1}^m r_{it+1} \omega_{it}, \quad \omega_{it} \equiv \frac{P_{it} \theta_{it+1}}{\sum_i P_{it} \theta_{it+1}},$$

where P_{it} and D_{it} are the price and the dividend of asset i at time t .

Let us consider a Markov economy in which the underlying state is some process y . We consider stationary consumption and investment plans. Accordingly, let the stationary util be a function $V(x, y)$ when current wealth is x and the state is y . By Eq. (8.3),

$$V(x, y) = \frac{1}{1 - \eta} \max_{c, \omega} \left(c^\rho + \beta \left((1 - \eta) E(V(x', y')) \right)^{\frac{\rho}{1 - \eta}} \right)^{\frac{1 - \eta}{\rho}}. \quad (8.5)$$

In the Appendix, we show that the first order conditions for the representative agent lead to the following Euler equation,

$$E[m(x, y; x' y') (1 + r_i(y'))] = 1, \quad i = 1, \dots, m, \quad (8.6)$$

where the stochastic discount factor m is,

$$m(x, y; x' y') = \beta^\theta \left(\frac{c(x', y')}{c(x, y)} \right)^{-\frac{\theta}{\psi}} (1 + r_M(y'))^{\theta - 1}, \quad \theta \equiv \frac{1 - \eta}{1 - \frac{1}{\psi}}.$$

This stochastic discount factor displays the interesting property to be affected by the market portfolio return, r_M , at least as soon as $\eta \neq \frac{1}{\psi}$. In particular, when $\theta < 1$, the stochastic discount factor is countercyclical: it leads to larger cash-flows discounts when r_M is low than when r_M is high, which will make asset returns decrease when the market drops, and increase when the market grows. Potentially, the stochastic discount factor may inherit the excess volatility of market returns quite naturally. At the same time, these properties arise as a result of a fixed point problem: market returns affect the stochastic discount factor, which, then, affects market returns! It is not surprising, then, that except for isolated exceptions, asset prices predicted by these models are not known in closed-form. Moreover, these interesting properties need to be further qualified, as discussed in the next section.

8.2.3 Risk premiums and interest rates

So the Euler equation is,

$$E \left(\beta^\theta \left(\frac{c'}{c} \right)^{-\frac{\theta}{\psi}} (1 + r'_M)^{\theta - 1} (1 + r'_i) \right) = 1. \quad (8.7)$$

Eq. (8.7) obviously holds for the market portfolio and the risk-free asset. Therefore, by taking logs in Eq. (8.7) for $i = M$, and for the risk-free asset, $i = 0$ say, yields the following conditions:

$$0 = \ln E \left(\exp \left(-\delta\theta - \frac{\theta}{\psi} \ln \left(\frac{c'}{c} \right) + \theta R_M \right) \right), \quad R_M = \ln (1 + r'_M), \quad (8.8)$$

the constant $\delta \equiv -\ln \beta$, and,

$$-R_f = -\ln (1 + r_0) = \ln E \left(\exp \left(-\delta\theta - \frac{\theta}{\psi} \ln \left(\frac{c'}{c} \right) + (\theta - 1) R_M \right) \right). \quad (8.9)$$

Next, suppose that consumption growth, $\ln \left(\frac{c'}{c} \right)$, and the market portfolio return, R_M , are jointly normally distributed. In the appendix, we show that the expected excess return on the market portfolio is given by,

$$E(R_M) - R_f + \frac{1}{2} \sigma_{R_M}^2 = \frac{\theta}{\psi} \sigma_{R_M, c} + (1 - \theta) \sigma_{R_M}^2 \quad (8.10)$$

where $\sigma_{R_M}^2 = \text{var}(R_M)$ and $\sigma_{R_M, c} = \text{cov}(R_M, \ln(c'/c))$, and the term $\frac{1}{2} \sigma_{R_M}^2$ in the left hand side is a Jensen's inequality term. Note, Eq. (8.10) is a mixture of the Consumption CAPM (for the part $\frac{\theta}{\psi} \sigma_{R_M, c}$) and the CAPM (for the part $(1 - \theta) \sigma_{R_M}^2$).

The risk-free rate is given by,

$$R_f = \delta + \frac{1}{\psi} E \left(\ln \left(\frac{c'}{c} \right) \right) - \frac{1}{2} (1 - \theta) \sigma_{R_M}^2 - \frac{1}{2} \frac{\theta}{\psi^2} \sigma_c^2, \quad (8.11)$$

where $\sigma_c^2 = \text{var}(\ln(c'/c))$.

Eqs. (8.10) and (8.11) can be elaborated further. In equilibrium, the asset price and, hence, the return, is certainly related to consumption volatility. Precisely, let us assume that,

$$\sigma_{R_M}^2 = \sigma_c^2 + \sigma_*^2, \quad \sigma_{R_M, c} = \sigma_c^2, \quad (8.12)$$

where σ_*^2 is a positive constant that may arise when the asset return is driven by some additional state variable. (This is the case, for example, in the Bansal and Yaron (2004) model described below.) Under the assumption that the asset return volatility is as in Eq. (8.12), the equity premium in Eq. (8.10) is:

$$E(R_M) - R_f + \frac{1}{2} \sigma_{R_M}^2 = \eta \sigma_c^2 + (1 - \theta) \sigma_*^2 = \eta \sigma_c^2 + \frac{\eta - \frac{1}{\psi}}{1 - \frac{1}{\psi}} \sigma_*^2. \quad (8.13)$$

Disentangling risk-aversion from intertemporal substitution is not enough for the equity premium puzzle to be resolved. To raise the equity premium, we need that $\sigma_*^2 > 0$, meaning that additional state variables are needed, to drive variation of asset returns. At the same time, the volatility of these state variables has the power to affect asset returns only when risk-aversion is distinct from the inverse of the EIS. As an example, suppose that σ_*^2 does not depend on η and ψ and that $\psi > 1$. Then, the equity premium increases with σ_*^2 whenever $\eta > \psi^{-1}$. In other words, these state variables have the potential to affect the equity premium, once they enter the stochastic discount factor, as for example with the long-run risks models reviewed in Section 8.5.

Next, we derive the risk-free rate. Assume that $E[\ln(c'/c)] = g_0 - \frac{1}{2}\sigma_c^2$, where g_0 is the expected consumption growth, a constant. Furthermore, use the assumptions in Eq. (8.12) to obtain that the risk-free rate in Eq. (8.11) is,

$$R_f = \delta + \frac{1}{\psi}g_0 - \frac{1}{2}\eta \left(1 + \frac{1}{\psi}\right) \sigma_c^2 - \frac{1}{2} \frac{\eta - \frac{1}{\psi}}{1 - \frac{1}{\psi}} \sigma_*^2.$$

As we can see, we may increase the level of relative risk-aversion, η , without substantially affecting the level of the risk-free rate, R_f . This is because the effects of η on R_f are of a second-order importance (they multiply variances, which are orders of magnitude less than the expected consumption growth, g_0).

8.2.4 Campbell-Shiller approximation

Consider the definition of the return on the market portfolio,

$$R_{M,t+1} = \ln \left(\frac{P_{t+1} + C_{t+1}}{P_t} \right) = \ln \left(\frac{e^{z_{t+1}} + 1}{e^{z_t}} \right) + g_{t+1} \equiv f(z_{t+1}, z_t) + g_{t+1},$$

where P_t is the value of the market portfolio, $g_{t+1} = \ln \frac{C_{t+1}}{C_t}$ is the aggregate dividend growth, and $z_t = \ln \frac{P_t}{C_t}$ is the log of the aggregate price-dividend ratio. A first-order linear approximation of $f(z_{t+1}, z_t)$ around the ‘‘average’’ level of z leaves,

$$R_{M,t+1} \approx \kappa_0 + \kappa_1 z_{t+1} - z_t + g_{t+1}, \quad (8.14)$$

where $\kappa_0 = \ln \left(\frac{e^{\bar{z}} + 1}{e^{\bar{z}}} \right) + \frac{\bar{z}}{e^{\bar{z}} + 1}$, $\kappa_1 = \frac{e^{\bar{z}}}{e^{\bar{z}} + 1}$ and \bar{z} is the average level of the log price-dividend ratio, such that $\kappa_1 \approx 0.997$, using US data. The approximation in Eq. (8.14) appears for the first time in Campbell and Shiller (1988). We now use this approximation to illustrate how non-expected utility and long-run risks impart on the equity premium puzzle.

8.2.5 Risks for the long-run

Bansal and Yaron (2004) consider a model where *persistence* in the expected consumption growth can explain the equity premium puzzle. To capture the main points of this explanation, assume that consumption growth is solution to,

$$g_{t+1} = \ln \frac{C_{t+1}}{C_t} = g_0 - \frac{1}{2}\sigma_c^2 + x_t + \epsilon_{t+1}, \quad \epsilon_{t+1} \sim N(0, \sigma_c^2), \quad (8.15)$$

where x_t is a ‘‘small’’ persistent component in consumption growth, solution to,

$$x_{t+1} = \rho x_t + \eta_{t+1}, \quad \eta_{t+1} \sim N(0, \sigma_x^2). \quad (8.16)$$

To find an approximate solution to the log of the price-dividend ratio, replace the Campbell-Shiller approximation in Eq. (8.14) into the Euler equation (8.8) for the market portfolio,

$$0 = \ln E_t \left(\exp \left(-\delta\theta - \frac{\theta}{\psi} \ln \left(\frac{C_{t+1}}{C_t} \right) + \theta (\kappa_0 + \kappa_1 z_{t+1} - z_t + g_{t+1}) \right) \right). \quad (8.17)$$

Conjecture that the log of the price-dividend ratio takes the simple form, $z_t = a_0 + a_1 x_t$, where a_0 and a_1 are two coefficients to be determined. Substituting this guess into Eq. (8.17), and identifying terms, leaves:

$$z_t = a_0 + \frac{1 - \frac{1}{\psi}}{1 - \kappa_1 \rho} x_t, \quad (8.18)$$

for some constant a_0 given in the Appendix.

Note that in the presence of a very persistent growth process, $1 - \kappa_1 \rho \approx 1$, such that even small changes in the expected dividend growth, x_t , would lead to large swings in the price-dividend ratio. A model solved along these lines, where persistent processes would lead to highly volatile prices, was already available in the literature, at least since the discussion of Campbell, Lo and MacKinlay (1997, Chapter 7, p. 265) of a model with persistent expected returns:

“If this standard deviation is small [i.e. the variability of expected stock returns], it is tempting to conclude that changing expected returns have little influence on stock prices. [...] This conclusion is too hasty: [...] if expected returns vary in a persistent fashion, [prices] can be very variable even when the [expected returns are] not.”

The model discussed by Campbell, Lo and MacKinlay is one where expected returns are directly modeled as possibly persistent processes. Instead, the model of this section is one where expected growth is possibly persistent. We shall see soon, what the implications are, of such a broader perspective. A crucial point is, however, that high volatility of the price-dividend ratio does not necessarily lead to a resolution of the equity premium puzzle. For example, in the context of non-expected utility, Eq. (8.13) suggests that in equilibrium, relative risk-aversion and intertemporal elasticity of substitution should play together in the right direction, and the variance σ_* is equally important. For example, if $\eta = \psi^{-1}$, the price-dividend ratio would not even enter the Euler equation as we know. Therefore, we need to check how this high volatility of the price-dividend ratio translates into a high equity premium.

We use the expression of $R_{M,t+1}$ in Eq. (8.14), to compute σ_*^2 , volatility, risk-premium, etc. **[In progress]**

8.3 Heterogeneous agents and “catching up with the Joneses”

The attractive feature of the Campbell and Cochrane (1999) model of external habit formation, reviewed in Section 7.5.2 of Chapter 7, is to have the potential to generate the right properties of asset prices and volatilities, through the channel of a countercyclical price of risk. It does rely on a high risk-averse economy, though. Chan and Kogan (2002) show that a countercyclical price of risk might arise, without assuming the existence of a representative agent with a high risk-aversion. They consider an economy where heterogeneous agents have preferences displaying the “catching up with the Joneses” features introduced by Abel (1990, 1999). There is a continuum of agents, indexed by a parameter $\eta \in [1, \infty)$ in the instantaneous utility function,

$$u_\eta(c, x) = \frac{\left(\frac{c}{x}\right)^{1-\eta}}{1-\eta},$$

where c is consumption, and x is the “standard living of others,” to be defined below.

The total endowment in the economy, D , follows a geometric Brownian motion,

$$\frac{dD(\tau)}{D(\tau)} = g_0 d\tau + \sigma_0 dW(\tau). \quad (8.19)$$

By assumption, the standard of living of others, $x(\tau)$, is a weighted geometric average of the past realizations of the aggregate consumption D , viz

$$\ln x(\tau) = \ln x(0) e^{-\theta\tau} + \theta \int_0^\tau e^{-\theta(\tau-s)} \ln D(s) ds, \quad \text{with } \theta > 0.$$

Therefore, $x(\tau)$ satisfies,

$$dx(\tau) = \theta s(\tau) x(\tau) d\tau, \quad \text{where } s(\tau) \equiv \ln \left(\frac{D(\tau)}{x(\tau)} \right). \quad (8.20)$$

By Eqs. (8.19) and (8.20), $s(\tau)$ is solution to,

$$ds(\tau) = \left(g_0 - \frac{1}{2} \sigma_0^2 - \theta s(\tau) \right) d\tau + \sigma_0 dW(\tau).$$

This model is important. We already know that a realistically calibrated economy with habit formation and a representative agent relies on a high risk aversion. Moreover, an economy with “catching up with the Joneses” and a representative agent would also rely on a high risk-aversion, as argued below. Chan and Kogan (2002) show that their model, while capturing the spirit of habit formation through “catching up with the Joneses,” does not need to rely on a high risk-aversion, once we populate the economy with heterogenous agents, thereby allowing habit formation and “catching up with the Joneses” to perform a role in the explanation of the equity premium puzzle.

In this economy with complete markets, we can determine the asset price, and solve the model by relying on the centralization of competitive equilibrium through Pareto weightings, along lines similar to those in Theorem 2.7 of Chapter 2 in Part I. As explained in the Appendix, the equilibrium price process is the same as that in an economy with a representative agent endowed with the following instantaneous utility function,

$$u(D, x) \equiv \max_{c_\eta} \int_1^\infty u_\eta(c_\eta, x) f(\eta) d\eta \quad \text{s.t.} \quad \int_1^\infty c_\eta d\eta = D, \quad [\text{P1}]$$

where $f(\eta)^{-1}$ is the marginal utility of income of the agent η . The Appendix provides further details about the derivation of the value function of the program [P1], which is:

$$U(s) \equiv \int_1^\infty \frac{1}{1-\eta} f(\eta)^{\frac{1}{\eta}} V(s)^{\frac{\eta-1}{\eta}} d\eta, \quad (8.21)$$

where V is a Lagrange multiplier, a function of the state s , satisfying:

$$e^s = \int_1^\infty f(\eta)^{\frac{1}{\eta}} V(s)^{-\frac{1}{\eta}} d\eta. \quad (8.22)$$

Finally, the Appendix shows that the unit risk-premium predicted by this model is,

$$\lambda(s) = \sigma_0 \frac{\exp(s)}{\int_1^\infty \frac{1}{\eta} f(\eta)^{\frac{1}{\eta}} V(s)^{-\frac{1}{\eta}} d\eta}. \quad (8.23)$$

All in all, Eq. (8.22) determines the Lagrange multiplier, $V(s)$, which then feeds $\lambda(s)$ through Eq. (8.23). Empirically, the Pareto weighting function, $f(\eta)$, can be parametrized by a function, which can be calibrated to match selected characteristics of the asset returns and volatility. Note, finally, that this economy collapses to an otherwise identical homogeneous economy, once the social weighting function $f(\eta) = \delta(\eta - \bar{\eta})$, the Dirac's mass at $\bar{\eta}$. In this case, $\lambda(s) = \sigma_0 \bar{\eta}$, a constant. As anticipated, an economy with “catching up with the Joneses” and a representative agent, is unlikely to resolve the equity premium puzzle or address issues such as predictability of asset returns, because $\sigma_0 \bar{\eta}$ is both small and constant, a point made by Campbell (2003) under a different angle.

A technical, albeit crucial assumption of this model is that the standard of living X is a process with *bounded* variation, as Eq. (8.20) clearly shows. As a consequence, the standard living of others is not a risk agents require to be compensated for. The unit risk-premium in Eq. (8.23), then, is driven by s , through agents heterogeneity. By calibrating their model to US data, Chan and Kogan find that the risk-premium, $\lambda(s)$, is decreasing and convex in s .¹ The mechanism at the heart of this result relates to an endogenous redistribution of wealth. Note that the less risk-averse agents obviously invest a higher proportion of their wealth in risky assets, compared to the more risk-averse. In the poor states of the world, then, when stock prices decrease, the wealth of the less risk-averse agents lowers more than that of the more risk-averse. The result is a reduction in the fraction of wealth held by the less risk-averse individuals in the whole economy. Thus, in bad times, the contribution of these less risk-averse individuals to aggregate risk-aversion decreases and, hence, aggregate risk-aversion increases in the economy.

[Discuss the criticism of Xiouros and Zapatero (2010)]

8.4 Idiosyncratic risk

Aggregate risk is too small to justify the extent of the equity premium through a low level of risk-aversion. Do individuals bear some idiosyncratic risk, one that cannot be diversified away by trading in capital markets? And then, how can this risk affect asset evaluation? Shouldn't idiosyncratic risk be neutral to asset pricing? The answer to these questions is indeed subtle, and relies upon whether idiosyncratic risk affects agents' portfolio choices and, then, the stochastic discount factor.

Mankiw (1986) is the first contribution to point out asset pricing implications of idiosyncratic risks. In his model, aggregate shocks to consumption do not affect individuals in the same way, ex-post. Ex-ante, individuals know that the business cycle may adversely change—an *aggregate* shock—although they also anticipate that the very same same shock might be particularly severe to only a portion of them—an *idiosyncratic* shock. To illustrate, everyone faces a positive probability of experiencing a job loss during a recession, although then, only a part of the population will actually suffer from a job loss. These circumstances might significantly affect the agents' portfolio choice and, therefore, rational asset evaluation. Naturally, in the presence of contingent claims able to insure against these shocks, idiosyncratic risk would not matter. But the point is that in reality, these contingent claims do not exist yet, due perhaps to moral hazard or adverse selection reasons. This source of market incompleteness might then poten-

¹Their numerical results also revealed that in their model, the *log* of the price-dividend ratio is increasing and *concave* in s . Finally, their lemma 5 (p. 1281) establishes that in a homogeneous economy, the price-dividend ratio is increasing and *convex* in s .

tially explain the aggregate stock market behavior in a way that the model with a standard representative agent cannot.

Mankiw considers the pricing of a risky asset in a two-period model, with the first period budget constraint given by $p\theta + m = w$, and the second period consumption equal to:

$$c = \tilde{w} + R\theta + (w - p\theta)(1 + r) \equiv \tilde{w} + \tilde{X}\theta,$$

where m is the amount to invest in a money market account, r is the safe interest rate on the money market account, normalized to zero, w is the initial endowment, also normalized to zero, p is the price of the risky asset, R is the payoff promised by the risky asset and, finally, $\tilde{X} \equiv R - p$ is the asset “net payoff.” We may either endogenize the price p , given the payoff R , or, then, just the net payoff, \tilde{X} , as described next. The assets are in zero net supply, and because agents are ex-ante identical, we have that $\theta = 0$, such that in equilibrium, \tilde{w} equals per capita consumption, c .

There are two states of nature for the aggregate economy, which are equally likely. In the good state, the net asset payoff is $\tilde{X} = 1 + \pi$, and per capita consumption is $\tilde{w} = \mu$. In the bad state, where the net asset payoff is $\tilde{X} = -1$, per capita consumption is $\tilde{w} = (1 - \phi)\mu$. The payoff in the good state of the world, π , equals $2E(\tilde{X})$, and thus, it is a measure of the risk premium, which is determined in equilibrium. Table 8.1 summarizes payoffs, per capita consumption and individual consumption in this economy. Individual consumption is defined as the level of consumption pertaining to different individuals in different states of nature. In the good state of nature, everyone consumes μ . In the bad state of nature, a fraction $1 - \lambda$ of individuals are not hit by the aggregate shock, and consume μ . The fraction λ of individuals hit by the shock consume $\mu(1 - \frac{\phi}{\lambda})$. The ratio, $\frac{\phi}{\lambda}$, is the per capita fall in consumption for any individual hit by the shock in the bad state of nature. If $\lambda = 1$, the aggregate shock hits everyone. The highest concentration of the shock arises when $\lambda = \phi$, i.e. when the fall in consumption is borne by the lowest possible fraction of the population.

	Net Asset Payoff	Per capita consumption	Individual consumption
Bad state	-1	$(1 - \phi)\mu$	$\mu(1 - \frac{\phi}{\lambda})$ (consumed by λ)
Good state	$1 + \pi$	μ	μ (consumed by $1 - \lambda$)

TABLE 8.1. Aggregate fluctuations and idiosyncratic risk

The first order conditions for any individual are:

$$\begin{aligned} 0 &= E[\tilde{X}u'(\tilde{w} + \tilde{X}\theta)] \\ &= E[\tilde{X}u'(\tilde{w})] \\ &= -1 \cdot [u'(\mu(1 - \frac{\phi}{\lambda}))\lambda + u'(\mu)(1 - \lambda)] + (1 + \pi)u'(\mu) \end{aligned}$$

where the second equality follows by the equilibrium condition, $\theta = 0$, and u is an utility function satisfying standard regularity conditions. The premium, π , equals:

$$\pi = \lambda \frac{u'(\mu(1 - \frac{\phi}{\lambda})) + u'(\mu)}{u'(\mu)}.$$

Mankiw shows that for utility functions leading to prudent behavior, $u''' > 0$, the premium π is decreasing in λ : an increase in the concentration of aggregate shocks leads to higher premiums. Moreover, it is easy to see that π can be made arbitrarily large, for λ arbitrarily close to ϕ , once the utility function satisfies the Inada's condition, $\lim_{c \rightarrow 0} u'(c) = \infty$, as we have that $\lim_{\lambda \rightarrow \phi} \pi = \infty$. For example, in the log-utility case, we have that $\pi = \frac{\lambda\phi}{\lambda - \phi}$.

The critical assumption underlying Mankiw's model is that once agents are hit by an idiosyncratic shock, the game is over. What happens once we allow the agents to act in a multiperiod horizon? Intuitively, in a dynamic context, agents might implement self-insurance plans, by accumulating financial assets after good shocks and selling or short-selling after bad shocks have occurred. Telmer (1993) and Lucas (1994) show that if idiosyncratic shocks are not persistent, self-insurance is quite effective and asset prices behave substantially the same as they would do in a world without idiosyncratic risk. Therefore, to have asset prices significantly deviate from those arising within a complete market setting, one has to either (i) reduce the extent of risk-sharing, by assuming frictions such as transaction costs, short-selling constraints or in general severe forms of market incompleteness, or (ii) make idiosyncratic shocks persistent. With (i), we just merely eliminate the possibility that agents may implement self-insurance plans through capital market transactions. With (ii), we make idiosyncratic shocks so severe that no capital market transaction might allow agents to insure themselves and achieve portfolio solutions close to the complete market solution; intuitively, once any individual is hit by an idiosyncratic shock, he may short-sell financial assets in the short-run, although then, he cannot persistently do so, given his wealth constraints.

Heaton and Lucas (1996) calibrate a model with idiosyncratic shocks using PSID (Panel Study of Income Dynamics) and NIPA (National Income and Product Accounts) data. They find that idiosyncratic shocks are not quite persistent, and that a large amount of transaction costs is needed to generate sizeable levels of the equity premium. Constantinides and Duffie (1996), instead, take the issue of persistence in idiosyncratic risk to the extreme, and consider a model without any transaction costs, but with permanent idiosyncratic risk. They show that in fact, given an asset price process, it is always possible to find a cross-section of idiosyncratic risk processes compatible with the asset price given in advance. We now present this elegant model, which has a quite substantial theoretical importance per se, because of its feature to make so transparent how some state variables affecting consumer choices can be reverse-engineered from the observation of an asset price process.

Central to Constantinides and Duffie analysis is the assumption that each individual i has a consumption equal to $c_{i,t}$ at time t , given by:

$$c_{i,t} = \gamma_{i,t} c_t, \quad \gamma_{i,t} = \exp \left(\sum_{s=1}^t \left(\zeta_{i,s} y_s - \frac{1}{2} y_s^2 \right) \right),$$

where $\zeta_{i,t}$ are independent and standard normally distributed, and y_t is a sequence of random variables, interpreted as standard deviation of the cross-sectional distribution of the individual consumption growth shares, $\ln \frac{\gamma_{i,t}}{\gamma_{i,t-1}}$, i.e.,

$$\ln \left(\frac{c_{i,t+1}/c_{t+1}}{c_{i,t}/c_t} \right) \Big|_{F_t \cup \{y_{t+1}\}} \sim N \left(-\frac{1}{2} y_{t+1}^2, y_{t+1}^2 \right),$$

where F_t is the information set as of time t .

The meaning of the consumption share $\gamma_{i,t}$ is that of an idiosyncratic shock every agent i receives in his consumption share at time t . From the perspective of each agent, this shock

is uninsurable, in that it is unrelated to the asset returns. Moreover, by construction, the consumption share has a unit root, as $\ln \gamma_{i,t} - \ln \gamma_{i,t-1} = \zeta_{i,t} y_t - \frac{1}{2} y_t^2$: a change in y_t and/or a shock in $\zeta_{i,t}$ have a permanent effect on the future path of $\gamma_{i,t}$. All agents have a CRRA utility function. We want to make sure this setup is consistent with any given equilibrium asset price process, by requiring two conditions: (i) $\int c_{i,t} d\mu(i) = c_t$, i.e. $\int \gamma_{i,t} d\mu(i) = 1$, where $\mu(i)$ is the measure of agent i , a condition satisfied by the law of large numbers; (ii) the cross-sectional variances y_t^2 are reverse-engineered so as to be consistent with any stochastic discount factor and, hence, any asset price process given in advance. To achieve (ii), note that for any agent i , the value of an asset delivering a payoff equal to \tilde{X} at time $t+1$ is, by the law of iterated expectations,

$$\begin{aligned} & E \left[e^{-\rho} \left(\frac{c_{t+1}}{c_t} \right)^{-\eta} E \left(e^{-\eta(\zeta_{i,t+1} y_{t+1} - \frac{1}{2} y_{t+1}^2)} \middle| F_t \cup \{y_{t+1}\} \right) \cdot \tilde{X} \middle| F_t \right] \\ &= E \left[e^{-\rho} \left(\frac{c_{t+1}}{c_t} \right)^{-\eta} e^{\frac{1}{2}\eta(\eta+1)y_{t+1}^2} \cdot \tilde{X} \middle| F_t \right], \end{aligned}$$

where ρ is the discount rate and η is the CRRA coefficient. It is independent of any agent i , such that the stochastic discount factor is:

$$\frac{\xi_{t+1}}{\xi_t} \equiv m_{t+1} \equiv e^{-\rho} \left(\frac{c_{t+1}}{c_t} \right)^{-\eta} e^{\frac{1}{2}\eta(\eta+1)y_{t+1}^2}.$$

That is, given an aggregate consumption process, and an arbitrage free asset price process, there exists a cross-section of idiosyncratic risk processes that supports the given price process. As a trivial example, consider the standard Lucas stochastic discount factor, which obtains when $y_t \equiv 0$.

Which properties of the stochastic discount factor are we looking for? Naturally, we wish to make sure m_t is as countercyclical as ever, which might be the case should the dispersion of the cross-sectional distribution of the log-consumption growth, y_t^2 , be countercyclical. However, Lettau (2002) shows that empirically, such a dispersion seems to be not enough, even when multiplied by $\frac{1}{2}\eta(\eta+1)$, unless of course, we are willing to assume, again, a high level of risk-aversion. Note that Lettau analyzes a situation that favourably biases his final outcome towards not rejecting the null that idiosyncratic risk matters, as he assumes agents cannot insure themselves at all: once they are hit by an idiosyncratic shock, they just have to consume their income. Constantinides, Donaldson and Mehra (2002) consider an OLG to mitigate the issue of persistence in the idiosyncratic risk process.

8.5 Incomplete markets and heterogenous agents

[Begin this section, with a simple two period model, where agents face a system of incomplete markets, but share the same kind of incompleteness. Therefore, to make incomplete markets work, we need some heterogeneity. In the following model, one agent has access to the risk asset market, while a second agent does not. The equity premium is the expected excess return the first agent requires to enter the risk asset market, and can be quite large, even in the presence of small risk-aversion, because the agent is being willing to take on the entire aggregate macroeconomic risk.] [...]

Basak and Cuoco (1998) consider a model with two agents. One of these agents does not invest in the stock market, and has logarithmic instantaneous utility, $u_n(c) = \ln c$. From his perspective, markets are incomplete. The second agent, instead, invests in the stock market, and has instantaneous utility equal to $u_p(c) = (c^{1-\eta} - 1)/(1 - \eta)$. Both agents are infinitely lived. The competitive equilibrium of this economy cannot be Pareto efficient, and so aggregation results such as those underlying the economy in Section 8.2 cannot obtain. However, Basak and Cuoco show that aggregation still obtains in this economy, once we define social weights in a judicious way.

Let $\hat{c}_i(\tau)$ be the general equilibrium allocation of agent i , $i = p, n$. In equilibrium, $\hat{c}_p + \hat{c}_n = D$, where D is the instantaneous aggregate consumption, taken to be a geometric Brownian motion with parameters g_0 and σ_0 ,

$$\frac{dD(\tau)}{D(\tau)} = g_0 d\tau + \sigma_0 dW(\tau). \quad (8.24)$$

Define

$$s(\tau) \equiv \frac{\hat{c}_p(\tau)}{D(\tau)}, \quad (8.25)$$

which is the consumption share of the market participant.

The first order conditions pertaining to the two agents intertemporal consumption plans are:

$$u'_p(\hat{c}_p(\tau)) = w_p e^{\delta\tau} \xi(\tau), \quad \hat{c}_n(\tau)^{-1} = w_n e^{\delta\tau - \int_0^\tau R(s) ds}, \quad (8.26)$$

where w_p, w_n are two constants, ξ is the pricing kernel process, solution to,

$$\frac{d\xi(\tau)}{\xi(\tau)} = -R(\tau) dt - \lambda(\tau) \cdot dW(\tau), \quad (8.27)$$

and, finally, R is the short-term rate and λ is the unit risk-premium, which equal:

$$R(s) = \delta + \frac{\eta g_0}{s + \eta(1-s)} - \frac{1}{2} \frac{1}{s(s + \eta(1-s))} \sigma_0^2 \eta(1 + \eta), \quad (8.28)$$

and

$$\lambda(s) = \eta \sigma_0 s^{-1}. \quad (8.29)$$

The expressions for R and λ in Eqs. (8.28)-(8.29) are derived below. Appendix 2 provides a further derivation relying on the existence of a representative agent, as originally put forward by Basak and Cuoco (1998), and explained below.

In this economy, the marginal investor bears the entire macroeconomic risk. The risk premium he requires to invest in the aggregate stock market is large when his consumption share, $s(\tau)$, is small. With just a risk aversion of $\eta = 2$, and a consumption volatility of 1%, this model can explain the equity premium, as the plot of Eq. (8.29) in the Figure 8.1 illustrates. For example, Mankiw and Zeldes (1991) estimate that the share of aggregate consumption held by stock-holders is approximately 30%, which in terms of this model, would translate to an equity premium of more than 6.5%.

Guvenen (2009) makes an interesting extension of the Basak and Cuoco model. He consider two agents in which only the “rich” invests in the stock market, and is such that $EIS_{\text{rich}} > EIS_{\text{poor}}$. He shows that for the rich, a low EIS is needed to match the equity premium. However, US data show that the rich have a high EIS, which can not do the equity premium. (Guvenen considers an extension of the model where we can disentangle EIS and CRRA for the rich.)

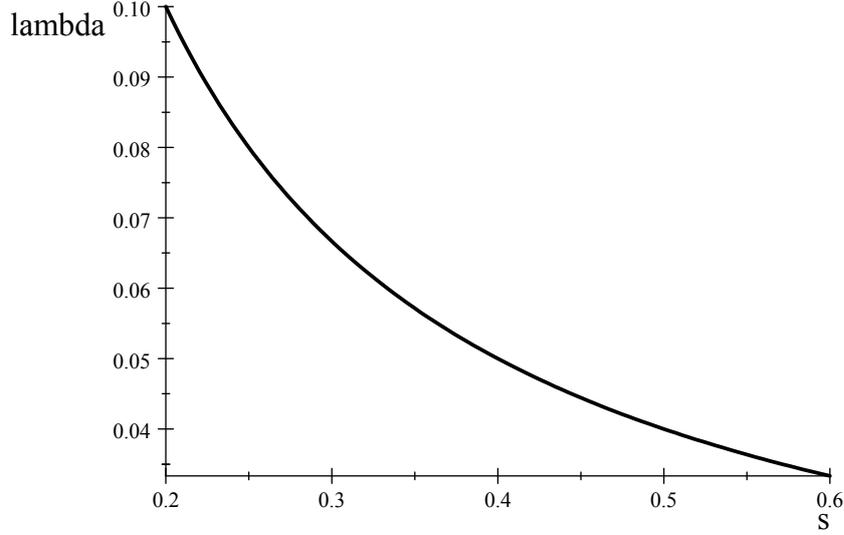


FIGURE 8.1. The equity premium in the Basak and Cuoco (1998) model, for $\eta = 2$ and $\sigma_0 = 1\%$.

To derive Eqs. (8.28)-(8.29), note that the consumption of the agent not participating in the stock market satisfies, by Eq. (8.26):

$$\frac{d\hat{c}_n(\tau)}{\hat{c}_n(\tau)} = (R(\tau) - \delta) d\tau. \quad (8.30)$$

Therefore, the consumption of the marginal investor, $\hat{c}_p = D - \hat{c}_n$, satisfies,

$$\begin{aligned} \frac{d\hat{c}_p(\tau)}{\hat{c}_p(\tau)} &= \frac{dD(\tau)}{D(\tau)} \frac{D(\tau)}{\hat{c}_p(\tau)} - \frac{d\hat{c}_n(\tau)}{\hat{c}_n(\tau)} \frac{\hat{c}_n(\tau)}{\hat{c}_p(\tau)} \\ &= \frac{1}{s(\tau)} (g_0 - (R(\tau) - \delta)(1 - s(\tau))) d\tau + \sigma_0 \frac{1}{s(\tau)} dW(\tau), \end{aligned} \quad (8.31)$$

where the second equality follows by the definition of s in Eq. (8.25), and the third by Eq. (8.24) and Eq. (8.30), and by rearranging terms. Moreover, by the first order conditions of the market participant in Eq. (8.26), and the CRRA assumption for u_p ,

$$\eta d \ln \hat{c}_p(\tau) = -\delta d\tau - d \ln \xi(\tau) = -\left(\delta - R(\tau) - \frac{1}{2} \lambda^2(\tau) \right) d\tau + \lambda(\tau) dW(\tau). \quad (8.32)$$

Using the relation, $d \ln c = \frac{dc}{c} - \frac{1}{2} \left(\frac{dc}{c} \right)^2$, then identifying terms in Eq. (8.31) and Eq. (8.32), delivers the two expressions for R and λ in Eqs. (8.28)-(8.29).

How do these results technically relate to aggregation? Basak and Cuoco define the utility of a representative agent, a social planner, as:

$$U(D, x) \equiv \max_{c_p + c_n = D} [u_p(c_p) + x \cdot u_n(c_n)], \quad (8.33)$$

where

$$x \equiv \frac{u'_p(\hat{c}_p)}{u'_n(\hat{c}_n)} = u'_p(\hat{c}_p)\hat{c}_n,$$

is a *stochastic social weight* and, once again, \hat{c}_p and \hat{c}_n are the private allocations, satisfying the first order conditions in Eqs. (8.26). By the definition of ξ , and Eqs. (8.26), $x(\tau)$ is solution to,

$$dx(\tau) = -x(\tau)\lambda(\tau)dW(\tau). \quad (8.34)$$

Then, the equilibrium in this economy is supported by a fictitious representative agent with utility $U(D, x)$. Intuitively, the social planner “allocations” satisfy, by construction,

$$\frac{u'_p(c_p^*(\tau))}{u'_n(c_n^*(\tau))} = \frac{u'_p(\hat{c}_p(\tau))}{u'_n(\hat{c}_n(\tau))} = x(\tau),$$

where the starred variables denote social planner’s “allocations.” In other words, Basak and Cuoco approach is to find a stochastic social weight process $x(\tau)$ such that the first order conditions of the representative agent leads to the market allocations. The utility in Eq. (8.33) can then be used to compute the short-term rate and risk premium, and lead precisely to Eqs. (8.28)-(8.29), as shown in Appendix 2.

8.6 Economies with production

Consider an economy with one representative firm producing one single good, as in Section 3.4.1.2 of Chapter 3, and paying off a dividend $D(K_t, I_t)$ in each period t , expressed as a function of capital K_t and investment I_t , with partial with respect to capital K_t equal to $D_K(K_t, I_t)$:

$$\begin{aligned} D(K_t, I_t) &\equiv \tilde{y}(K_t, N(K_t)) - w_t N(K_t) - p_t I_t - \phi\left(\frac{I_t}{K_t}\right) K_t \\ D_K(K_t, I_t) &\equiv \tilde{y}_K(K_t, N(K_t)) - \frac{\partial}{\partial K} \left(\phi\left(\frac{I_t}{K_t}\right) K_t \right) \end{aligned}$$

Remember, Tobin’s marginal q and average q are the same, by Theorem 3.2, meaning that the stock market value of the firm, $V(K_t)$, coincides with the value of installed capital, $V(K_t) = q_t K_{t+1}$, where q_t collapses to Tobin’s q, once we fix the price of uninstalled capital to one, $p_t \equiv 1$, which is the case as soon as the firm produces uninstalled capital, simply. A few calculations allow us to define equity returns in this economy. First, we note that:

$$\begin{aligned} V(K_t) &= q_t K_{t+1} \\ &= E_t [m_{t+1} (D_K(K_{t+1}, I_{t+1}) K_{t+1} + (1 - \delta) q_{t+1} K_{t+1})] \\ &= E_t [m_{t+1} (D_K(K_{t+1}, I_{t+1}) K_{t+1} + q_{t+1} (K_{t+2} - I_{t+1}))] \\ &= E_t [m_{t+1} (D_K(K_{t+1}, I_{t+1}) K_{t+1} - q_{t+1} I_{t+1} + V(K_{t+1}))] \\ &= E_t [m_{t+1} (D(K_{t+1}, I_{t+1}) + V(K_{t+1}))], \end{aligned}$$

where the second line follows by the q theory, as developed in Chapter 3, the third and fourth lines by the law of capital accumulation, and the expression for $V(K_{t+1})$, the fifth line by the

condition $q_{t+1} = -D_I(K_{t+1}, I_{t+1})$, and the homogeneity of the function D . Therefore, equity returns are:

$$\begin{aligned}\tilde{R}_{t+1} - 1 &\equiv \frac{D(K_{t+1}, I_{t+1}) + V(K_{t+1}) - V(K_t)}{V(K_t)} \\ &= \frac{D_K(K_{t+1}, I_{t+1})K_{t+1} - q_{t+1}I_{t+1} + V(K_{t+1}) - V(K_t)}{V(K_t)}.\end{aligned}$$

In the absence of adjustment costs, $\phi \equiv 0$, Tobin's q collapses to one, $q_t = 1$, such that the capital gains, $V(K_{t+1})/V(K_t) - 1 = -\delta + I_{t+1}/K_{t+1}$, bringing equity returns to:

$$\tilde{R}_{t+1} - 1 = \tilde{y}_K(K_t, N(K_t)) - \delta.$$

To match the volatility of equity returns, a model without adjustment costs would require a counterfactually large volatility of the marginal product of capital. Therefore, not only are adjustment costs needed to rationalize the existence of time-varying market-to-book ratios. Adjustment costs would have the potential to boost return volatility. But then, the equity premium puzzle can only be exacerbated in a setting without adjustment costs. Note, indeed, that by the usual representation of the equity premium in Section 6.5 of Chapter 6,

$$E_t(r_{t+1}^e) = -\text{corr}_t(m_{t+1}, r_{t+1}^e) \cdot \frac{\text{Std}_t(m_{t+1})}{E_t(m_{t+1})} \cdot \text{Std}_t(r_{t+1}^e),$$

where r^e denotes the equity return in excess of the risk-free rate. Unless the excess returns predicted by the model co-vary substantially, and negatively, with the stochastic discount factor, the equity premium can only be small, when $\text{Std}_t(r_{t+1}^e)$ is very small. One route to inflate the equity premium might seem to be one where risk-aversion is increased. However, in equilibrium, equity returns obviously relate to consumption, and in models with production, consumption smoothing may make the equity premium puzzle worsen: as originally pointed out by Rouwenhorst (1995), if consumption is endogenous, it becomes smoother as risk-aversion increases, thereby making $\text{Std}_t(r_{t+1}^e)$ smaller, in equilibrium.

The main issue with the neoclassical model is that capital supply is perfectly elastic, such that the price of capital and, hence, capital gains, are roughly constant, consistently with the previous arguments. As Jermann (1998) and Boldrin, Christiano and Fisher (2001) note, we need to introduce some sort of hindrance to the adjustment of capital supply to shocks. For example, Jermann (1998), assumes the presence of adjustment costs. Instead, Boldrin, Christiano and Fisher assume, among other things, that investment decisions can be thought to be determined prior to the realization of the shocks. Both Jermann (1998) and Boldrin, Christiano and Fisher (2001) consider economies with habit persistence anyway, which allows them to generate variability in the demand for capital and, hence, boost price volatility.

[In progress]

8.7 Leverage and volatility

Is firms' leverage responsible for a sustained stock volatility? Can leverage explain countercyclical stock volatility? We already know, from the previous chapter, that ex-post stock returns are high in good times, whence stock volatility is negatively related to ex-post returns. According to the *leverage effect* hypothesis, this negative relation arises because after a negative shock hits

a stock price, the debt/equity ratio increases and as a result, the firm becomes riskier, leading to an increase in the stock volatility. Empirically, it is often argued, the leverage effect is too weak. Most of the contributions to these issues are empirical (e.g., Black, 1976; Christie, 1982; Schwert, 1989a,b; Nelson, 1991). Naturally, another possibility is that stock volatility and returns are negatively related for reasons unrelated to the leverage effect. For example, stock volatility can be countercyclical because agents' preferences and beliefs, combined with macroeconomic conditions, lead precisely to this property, as in the models discussed in Chapter 7, and in the previous sections.

In this section, we explore an additional explanation, namely that countercyclical volatility arises as a result of a combined effect of the properties of the previous models, and leverage. A difficulty is that in many empirical studies, tests of the leverage effect hypothesis are performed without regard to a well specified economic model. Gallmeyer, Aydemir, Hollifield (2007) show that the reasoning underlying this hypothesis can be made rigorous. They formulate a general equilibrium model with levered firms, which they realistically calibrate, to disentangle leverage effects from "real" effects such as habit formation. They make use of a stochastic discount factor known to price assets fairly well, and conclude that leverage effects do indeed have little effects in general equilibrium. This section develops a variant of their model, which has the mere merit to admit a closed-form solution.

8.7.1 Model

8.7.1.1 Primitives

We consider an endowment economy, and denote endowment at time t with $\delta(t)$, assuming it is a Geometric Brownian motion with parameters g_0 and σ_0 . The reason we denote output with δ , rather than the usual D , is that we assume a representative firm issues debt, denoted with D , such that the value of the firm is, by Modigliani-Miller [to be discussed in Chapter 13], $V(t) \equiv E(t) + D(t)$, where $E(t)$ is the value of equity at time t . Let T denote debt maturity. The payoffs of the firm are such that $\delta(t) = \delta_E(t) + \delta_D(t)$, with obvious notation. We assume a representative agent has habit formation preferences, and to obtain closed-form solutions for the asset prices, we make reference to the Menzly, Santos and Veronesi (2004) economy in Section 7.5.4 of Chapter 7. We denote the equilibrium surplus consumption ratio with $s(t) = \frac{\delta(t) - x(t)}{\delta(t)}$, where, as explained extensively in Chapter 7, s is solution to,

$$d\left(\frac{1}{s(t)}\right) = \beta\left(\frac{1}{\bar{s}} - \frac{1}{s(t)}\right)dt - \alpha\left(\frac{1}{s(t)} - \frac{1}{v}\right)\sigma_0 dW(t),$$

and β , \bar{s} , α and v are parameters. In this economy, Sharpe ratios are countercyclical, being equal to $\lambda(s) = \sigma_0\left(1 + \alpha\left(1 - \frac{s}{v}\right)\right)$, as mentioned in Section 7.5.4. We assume debt services are $\delta_D = q\delta$, for some $q \in (0, 1)$, and set the benchmark for debt maturity to $T = 10$ years.

8.7.1.2 Model's predictions

We now show that a calibration of the model leads to the following results: (i) the price of debt is procyclical; (ii) return volatility is countercyclical; (iii) the leverage ratio is countercyclical; and finally, (iv) the contribution of leverage to equity returns volatility is quantitatively limited.

Equity volatility: a decomposition formula

Equity volatility is,

$$\text{Vol}\left(\frac{dE}{E}\right) = \sigma_V + (\sigma_V - \sigma_D)\frac{D}{E}.$$

From Chapter 7, we know that the price-dividend ratio for the aggregate consumption claim is $p(s) \equiv a + bs$, for two constants a and b , which we shall give below again. It is easy to show that the debt value is,

$$\frac{D(t)}{\delta(t)} = q(a_T + b_T s(t)),$$

where $a_T = \frac{1 - e^{-(\rho+\beta)T}}{\rho + \beta}$, $b_T = \frac{\beta(1 - e^{-\rho T}) + \rho(e^{-(\rho+\beta)T} - e^{-\rho T})}{\rho(\rho + \beta)\bar{s}}$, with $a = \lim_{T \rightarrow \infty} a_T$ and $b = \lim_{T \rightarrow \infty} b_T$. We are now in a position to derive an expression for equity volatility. We have,

$$\sigma_V(s(t)) = \sigma_0 + \frac{b}{a + bs(t)} \text{Vol}(ds(t)), \quad \sigma_D(s(t)) = \sigma_0 + \frac{b_T}{a_T + b_T s(t)} \text{Vol}(ds(t)),$$

where $\text{Vol}(ds(t)) = \alpha\sigma_0 s(t) \left(1 - \frac{1}{v} s(t)\right)$, such that

$$\begin{aligned} \underbrace{\text{Vol}\left(\frac{dE(t)}{E(t)}\right)}_{\approx 0.15} &= \underbrace{\sigma_0}_{=0.01} + \underbrace{\frac{b}{a + bs(t)}}_{= \text{endog. P/D fluct.} \approx 26.31} \cdot \underbrace{\text{Vol}(ds(t))}_{\approx 5 \cdot 10^{-3}} \\ &+ \underbrace{\left(\frac{b}{a + bs(t)} - \frac{b_T}{a_T + b_T s(t)}\right)}_{= \text{leverage multiplier} \approx 11.08} \cdot \underbrace{\text{Vol}(ds(t))}_{\approx 5 \cdot 10^{-3}} \cdot \underbrace{\frac{D(t)}{E(t)}}_{\approx 0.24}, \end{aligned} \quad (8.35)$$

where we have indicated the approximate average values taken by the variables of interest, and obtained by calibrating the model with the values in Table 7.5 in Section 7.5.4 of the previous chapter. Note, also, that the leverage ratio, $\frac{D(t)}{E(t)}$, is endogenous and equal to,

$$\frac{D(t)}{E(t)} = \frac{(a_T + b_T s(t))q}{a + bs(t) - (a_T + b_T s(t))q}.$$

In other words, we only “see” what happens to $\frac{D(t)}{E(t)}$ and $\text{Vol}\left(\frac{dE(t)}{E(t)}\right)$, as the surplus $s(t)$ changes. As the numerical values in Eq. (8.35) show, much of the action in this model derives from the large swings in the price-dividend ratio, $\frac{p'(s(t))}{p(s(t))} = \frac{b}{a + bs(t)}$.

These computations might suggest that debt maturity might lead to have obtain a greater leverage contribution to volatility. However, it is not the case, as we now show.

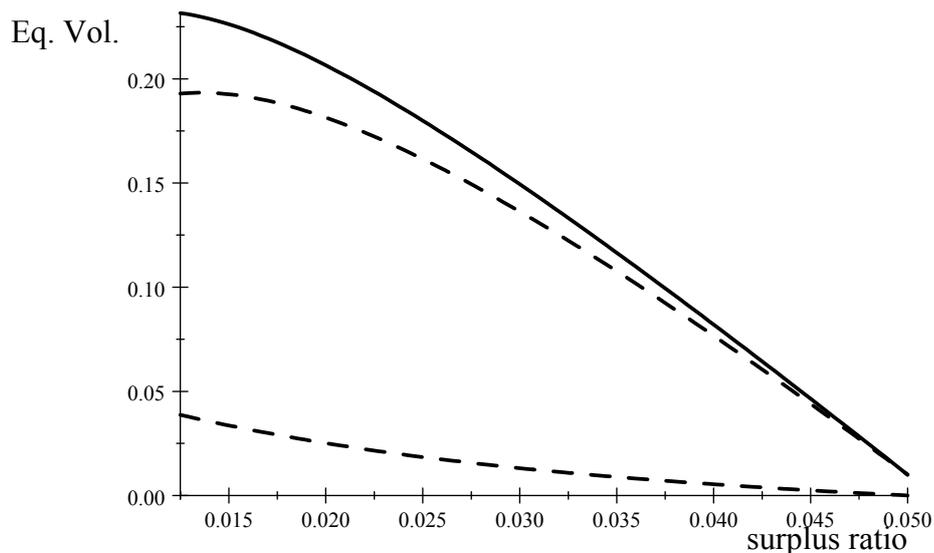


FIGURE 8.3. Equity volatility for $T = 10$. The solid line is total volatility. The top dashed is the contribution from “unlevered” volatility to total volatility, σ_V . The bottom dashed line is the contribution from “levered” volatility to total volatility, $(\sigma_V - \sigma_0) \frac{D}{E}$.

What is the statistical relation between the leverage ratio $\frac{D}{E}$ and return volatility that we should expect to find in the data?

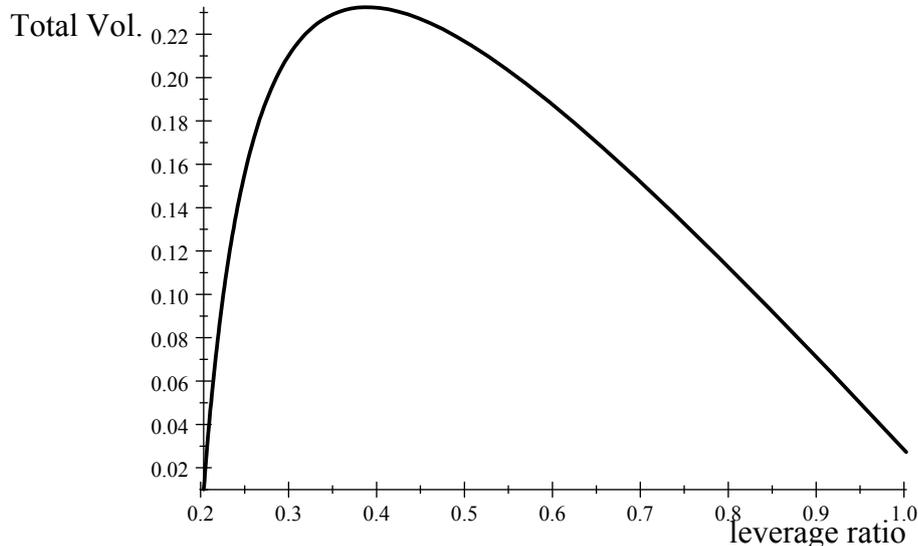


FIGURE 8.4. Leverage and equity volatility: a “naked” eye view.

Note, this is not a causal relation, both leverage and equity volatility are driven by the same state variable, the surplus consumption ratio.

The effects of debt maturity on the leverage effect are quite limited. Indeed, as debt maturity decreases, the leverage multiplier increases. However, the leverage ratio $\frac{D}{E}$ shrinks to zero as

maturity shrinks to zero. The overall effect is given by the third term on the right hand side of Eq. (8.35).

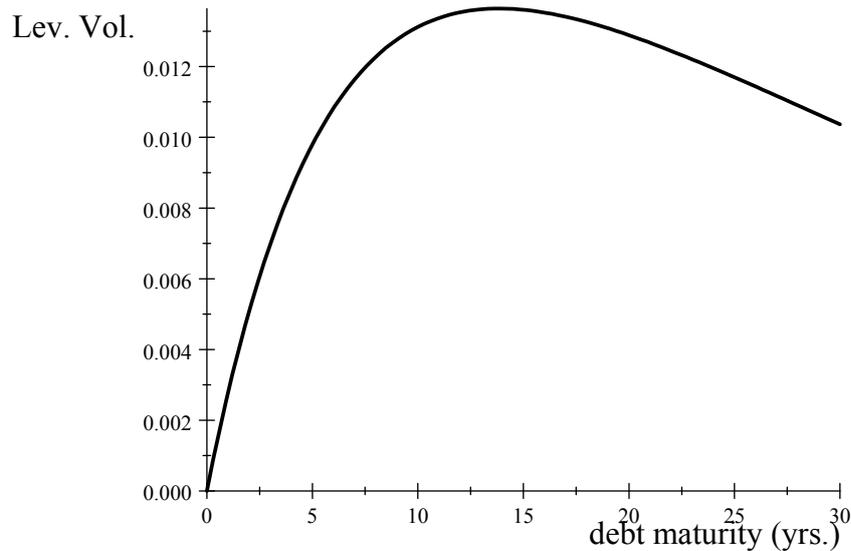


FIGURE 8.5. Leverage volatility at the steady state expectation, $s(t) = 0.03$.

Bankruptcy

The previous model had no role for bankruptcy, an obviously quite fundamental role, emphasized so many times in Chapter 13. Let us consider bankruptcy in a simple setting. Consider a two date economy, and suppose that the value of the firm in one year is, \tilde{V} , which equals $V_{\text{bad}} < \text{Nominal debt}$, with probability p , and $V_{\text{good}} > \text{Nominal debt}$, with probability $1 - p$. We assume risk-neutrality, and that there are no bankruptcy costs. Let $\tilde{R} = \frac{\tilde{S}_1 - S_0}{S_0}$ be the equity return, where \tilde{S}_1 is the equity value at the second period. Then, we have that $\text{vol}(\tilde{R}) = \sqrt{\frac{p}{1-p}}$. For example, if $p = 2\%$, then $\text{vol}(\tilde{R}) = 14\%$!

8.8 Multiple trees and the cross-section of asset returns

Menzly, Santos and Veronesi (2004), Cochrane, Longstaff and Santa-Clara (2008), Pavlova and Rigobon (2008), Martin (2011).

8.9 The term-structure of interest rates

What are the term-structure implications of the main paradigms considered so far? Consider the habit formation model introduced by Campbell and Cochrane (1999). While Campbell and Cochrane consider an economy where interest rates are constant, in their working paper they allow the short-term rate to be time-varying, and as explained in Appendix 5 of Chapter 7, set equal to:

$$R(s) = \delta + \eta \left(g_0 - \frac{1}{2} \sigma_0^2 \right) - \frac{1}{2} (\eta (1 - \phi) - b) + b (\bar{s} - \ln s), \quad (8.36)$$

where s is the surplus consumption ratio, b is a constant, and all the remaining parameters are as in Section 7.5.2 of Chapter 7. Wachter (2006) analyzes the term-structure implications of

this model in detail, both real and nominal, within an environment with time-varying expected inflation.

Note, the constant b does not depend on anything relating to the agents' preferences. Its mere role is to make interest rates time-varying. How to ensure that Eq. (8.36) is consistent with optimizing behavior? As explained in Chapter 7, the short-term rate depends on the sensitivity of habit to consumption shocks, a function of s , $l(s)$, through an effect due to precautionary savings: the higher this sensitivity, the higher the volatility of habit and, hence, the propensity to save, which drives interest rates down. This sensitivity $l(s)$ is “free,” in that it is not restricted by the theory—Campbell and Cochrane simply guide us with heuristic considerations leading to it. One of these considerations is that the short-term rate also relates to habit, due to intertemporal substitution effects, and negatively, due to mean-reversion. Campbell and Cochrane recipe is to choose $l(s)$ such that intertemporal substitution effects exactly offset precautionary savings, thereby making the short-term rate constant or, at most, affine in the log surplus consumption ratio, as in Eq. (8.36). Naturally, the sensitivity, $l(s)$, is a function of b , once this reverse engineering has unfold, as shown in Appendix 5 of Chapter 7.

The question arises as to which sign we should expect from the parameter b , empirically. Are real interest rates countercyclical? They *are*. It is somehow puzzling, from the perspective of the basic production economies analyzed in Chapter 3, where real interest rates are procyclical, being positively related to the marginal product of capital and, hence, to productivity shocks. However, economies with habit formation might be capable of generating countercyclical real rates, due to intertemporal substitution effects—It is the case, for example, for the models with frictions in the adjustment of capital supply to shocks of Boldrin, Christiano and Fisher (2001). In endowment economies and habit formation, countercyclical real rates are, then, quite likely to arise. Consider, for example, the Menzly, Santos and Veronesi (2004) model of external habit formation presented in Section 7.5.4. we remind that this model predicts that the short-term rate is:

$$R(s(t)) = \delta + g_0 - \sigma_0^2 + \beta \left(1 - \frac{s(t)}{\bar{s}}\right) - \sigma_0^2 \alpha \left(1 - \frac{s(t)}{v}\right). \quad (8.37)$$

The next picture depicts the short-term rate as a function of s , obtained using the parameter values in Table 8.2, which are similar to those used by Menzly, Santos and Veronesi.

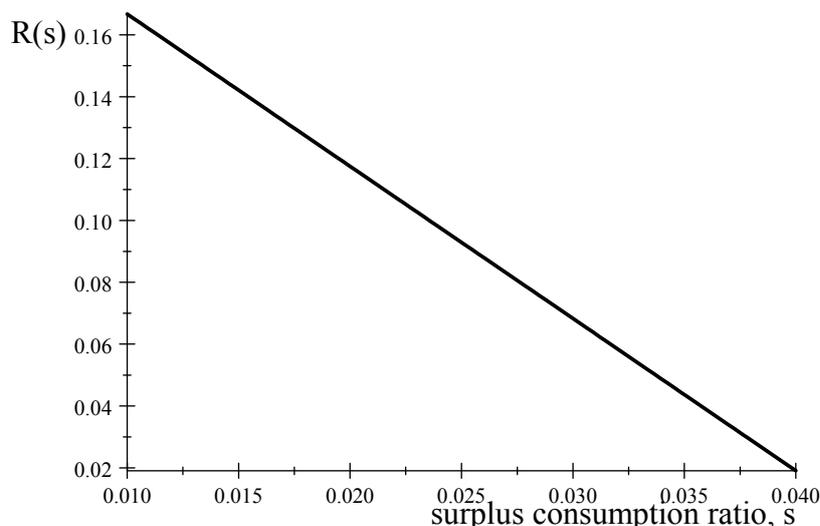


FIGURE 8.2. The short-term rate predicted by Menzly, Santos and Veronesi (2004) model of external habit formation, with parameter values as in Table 8.2.

g_0	σ_0	δ	β	\bar{s}	α	v	q
0.03	0.01	0.04	0.15	0.03	40	0.05	0.60

TABLE 8.2. Parameter values utilized for the Menzly, Santos and Veronesi (2004) model of external habit formation.

The fourth term of Eq. (8.37) reflects intertemporal substitution effects, and is the dominating term, leading to countercyclical interest rates, due to the mean reversion in the surplus consumption ratio, and similarly as in the Campbell and Cochrane model, as explained in Section 7.5.2 of Chapter 7. Finally, the catching-up model of Chan and Kogan (2002) reviewed in Section 8.3 leads to the same prediction: real interest rates are countercyclical.

[In progress]

8.10 Prices, quantities and the separation hypothesis

The compelling lesson from the current Part II of these lectures, is that to address the asset pricing puzzles the neoclassical model generates, we need to a substantial re-vamp of the standard paradigms underlying dynamic macroeconomic theory—namely, that underlying the basic version of the real business cycle theory reviewed in Chapter 3. For example, we need to consider adjustment costs, habit formation, or restricted stock market participation. How is it, then, that macroeconomists, in an attempt to explain quantity dynamics, would simply ignore the advances financial economists were introducing? Tallarini (2000) considers a different possibility, a real business cycle model where a representative agent has non-expected utility, as in Section 8.2,

$$U_t = \ln c_t + \theta \ln L_t + \frac{\beta}{\sigma} \ln E_t (e^{\sigma U_{t+1}}). \quad (8.38)$$

[Explain notation] Tallarini does not consider adjustment costs, and yet his model can explain the equity premium, through a simple increase in the risk aversion parameter σ —with intertemporal substitution kept constant, i.e. by keeping on assuming log consumption in the right hand side of Eq. (8.38). Interestingly, raising risk-aversion does not affect the quantity dynamics macroeconomists are interested in, only intertemporal substitution might affect it. Naturally, there are many other dimensions we should consider, to conclude on any model's prediction about asset prices. For example, Tallarini's assumption of no adjustment costs implies Tobin's q is one. Moreover, welfare calculations such as those in Lucas (19??) are likely to change, as Alvarez and Jermann (20??) demonstrate.

Finally, this model assumes there are no feedbacks from capital markets to real developments. Discuss the financial accelerator hypothesis through variants of Kiyotaki and Moore (1997) and Bernanke, Gertler and Gilchrist (1999), and argue the financial accelerator hypothesis might be quantitatively relevant, although at the same time, it might fail explain a number of empirical facts, and relies on quite abstract assumptions. [...] Chapter 13 discusses additional models of the credit crunch, where capital market turmoil can lead to adverse economic developments, which reinforce the capital market trends, over a spiral.

[In progress]

8.11 Appendix 1: Non-expected utility

8.11.1 Detailed derivation of optimality conditions and selected relations

DERIVATION OF EQ. (8.4). We have,

$$\begin{aligned}
 x_{t+1} &= \sum_i (P_{it+1} + D_{it+1}) \theta_{it+1} \\
 &= \sum_i (P_{it+1} + D_{it+1} - P_{it}) \theta_{it+1} + \sum_i P_{it} \theta_{it+1} \\
 &= \left(1 + \sum_i \frac{P_{it+1} + D_{it+1} - P_{it}}{P_{it}} \frac{P_{it} \theta_{it+1}}{\sum_i P_{it} \theta_{it+1}} \right) \sum_i P_{it} \theta_{it+1} \\
 &= \left(1 + \sum_i r_{it+1} \omega_{it} \right) (x_t - c_t)
 \end{aligned}$$

where the last line follows by the standard budget constraint $c_t + \sum P_{it} \theta_{it+1} = x_t$, the definition of r_{it+1} and the definition of ω_{it} given in the main text. ■

OPTIMALITY. Consider Eq. (8.5),

$$V(x, y) = \max_{c, \omega} \mathcal{W}(c, E(V(x', y'))) \equiv \frac{1}{1-\eta} \max_{c, \omega} \left[c^\rho + \beta \left((1-\eta) E(V(x', y')) \right)^{\frac{\rho}{1-\eta}} \right]^{\frac{1-\eta}{\rho}}.$$

The first order condition for c yields,

$$\mathcal{W}_1(c, E(V(x', y'))) = \mathcal{W}_2(c, E(V(x', y'))) \cdot E[V_1(x', y') (1 + r_M(y'))], \quad (8A.1)$$

where subscripts denote partial derivatives. Thus, optimal consumption is some function $c(x, y)$. Hence,

$$x' = (x - c(x, y)) (1 + r_M(y'))$$

We have,

$$V(x, y) = \mathcal{W}(c(x, y), E(V(x', y'))).$$

By differentiating the value function with respect to x ,

$$\begin{aligned}
 V_1(x, y) &= \mathcal{W}_1(c(x, y), E(V(x', y'))) c_1(x, y) \\
 &\quad + \mathcal{W}_2(c(x, y), E(V(x', y'))) E[V_1(x', y') (1 + r_M(y'))] (1 - c_1(x, y)),
 \end{aligned}$$

where subscripts denote partial derivatives. By replacing Eq. (8A.1) into the previous equation we get the Envelope Equation for this dynamic programming problem,

$$V_1(x, y) = \mathcal{W}_1(c(x, y), E(V(x', y'))). \quad (8A.2)$$

By replacing Eq. (8A.2) into Eq. (8A.1), and rearranging terms,

$$E \left[\frac{\mathcal{W}_2(c(x, y), \nu(x, y))}{\mathcal{W}_1(c(x, y), \nu(x, y))} \mathcal{W}_1(c(x', y'), \nu(x', y')) (1 + r_M(y')) \right] = 1, \quad \nu(x, y) \equiv E(V(x', y')).$$

Below, we show that by a similar argument the same Euler equation applies to any asset i ,

$$E \left[\frac{\mathcal{W}_2(c(x, y), \nu(x, y))}{\mathcal{W}_1(c(x, y), \nu(x, y))} \mathcal{W}_1(c(x', y'), \nu(x', y')) (1 + r_i(y')) \right] = 1, \quad i = 1, \dots, m. \quad (8A.3)$$

DERIVATION OF EQ. (8A.3). We have,

$$V(x, y) = \max_{c, \omega} \mathcal{W}(c, E(V(x', y'))) = \max_{\theta'} \mathcal{W}(x - \sum P_i \theta'_i, E(V(x', y'))); \quad x' = \sum (P'_i + D'_i) \theta'_i.$$

The set of first order conditions is,

$$\theta'_i : 0 = -\mathcal{W}_1(\cdot) P_i + \mathcal{W}_2(\cdot) E[V_1(x', y') (P'_i + D'_i)], \quad i = 1, \dots, m.$$

Optimal consumption is $c(x, y)$. Let $\nu(x, y) \equiv E(V(x', y'))$, as in the main text. By replacing Eq. (8A.2) into the previous equation,

$$E \left[\frac{\mathcal{W}_2(c(x, y), \nu(x, y))}{\mathcal{W}_1(c(x, y), \nu(x, y))} \mathcal{W}_1(c(x', y'), \nu(x', y')) \frac{P'_i + D'_i}{P_i} \right] = 1, \quad i = 1, \dots, m. \quad \blacksquare$$

DERIVATION OF EQ. (8.6). We need to compute explicitly the stochastic discount factor in Eq. (8A.3),

$$m(x, y; x' y') = \frac{\mathcal{W}_2(c(x, y), \nu(x, y))}{\mathcal{W}_1(c(x, y), \nu(x, y))} \mathcal{W}_1(c(x', y'), \nu(x', y')).$$

We have,

$$\mathcal{W}(c, \nu) = \frac{1}{1-\eta} \left[c^\rho + \beta ((1-\eta) \nu)^{\frac{\rho}{1-\eta}} \right]^{\frac{1-\eta}{\rho}}.$$

From this, it follows that,

$$\begin{aligned} \mathcal{W}_1(c, \nu) &= \left[c^\rho + \beta ((1-\eta) \nu)^{\frac{\rho}{1-\eta}} \right]^{\frac{1-\eta}{\rho}-1} c^{\rho-1} \\ \mathcal{W}_2(c, \nu) &= \left[c^\rho + \beta ((1-\eta) \nu)^{\frac{\rho}{1-\eta}} \right]^{\frac{1-\eta}{\rho}-1} \beta ((1-\eta) \nu)^{\frac{\rho}{1-\eta}-1}, \end{aligned}$$

and,

$$\mathcal{W}_1(c', \nu') = \left[c'^\rho + \beta ((1-\eta) \nu')^{\frac{\rho}{1-\eta}} \right]^{\frac{1-\eta}{\rho}-1} c'^{\rho-1} = \mathcal{W}(c', \nu')^{\frac{1-\eta-\rho}{1-\eta}} (1-\eta)^{\frac{1-\eta-\rho}{1-\eta}} c'^{\rho-1}, \quad (8A.4)$$

where $\nu' \equiv \nu(x', y')$. Therefore,

$$m(x, y; x' y') = \frac{\mathcal{W}_2(c, \nu)}{\mathcal{W}_1(c, \nu)} \mathcal{W}_1(c', \nu') = \beta \left(\frac{\nu}{\mathcal{W}(c', \nu')} \right)^{\frac{\rho}{1-\eta}-1} \left(\frac{c'}{c} \right)^{\rho-1}.$$

Along any optimal consumption path, $V(x, y) = \mathcal{W}(c(x, y), \nu(x, y))$. Therefore,

$$m(x, y; x' y') = \beta \left(\frac{E(V(x', y'))}{V(x', y')} \right)^{\frac{\rho}{1-\eta}-1} \left(\frac{c'}{c} \right)^{\rho-1}. \quad (8A.5)$$

We are left with evaluating the term $\frac{E(V(x', y'))}{V(x', y')}$. The conjecture to make is that $v(x, y) = b(y)^{1/(1-\eta)} x$, for some function b . From this, it follows that $V(x, y) = b(y) x^{1-\eta} / (1-\eta)$. We have,

$$\begin{aligned} V_1(x, y) &= \mathcal{W}_1(c(x, y), E(V(x', y'))) = \mathcal{W}(c, \nu)^{\frac{1-\eta-\rho}{1-\eta}} (1-\eta)^{\frac{1-\eta-\rho}{1-\eta}} c^{\rho-1} = V(x, y)^{\frac{1-\eta-\rho}{1-\eta}} (1-\eta)^{\frac{1-\eta-\rho}{1-\eta}} c^{\rho-1}. \end{aligned}$$

where the first equality follows by Eq. (8A.2), the second equality follows by Eq. (8A.4), and the last equality follows by optimality. By making use of the conjecture on V , and rearranging terms,

$$c(x, y) = a(y) x, \quad a(y) \equiv b(y)^{\frac{\rho}{(1-\eta)(\rho-1)}}. \quad (8A.6)$$

Hence, $V(x', y') = b(y') x'^{1-\eta} / (1-\eta)$, where

$$x' = (1 - a(y)) x (1 + r_M(y')), \quad (8A.7)$$

and

$$\frac{E(V(x', y'))}{V(x', y')} = \frac{E\left(\psi(y') (1 + r_M(y'))^{1-\eta}\right)}{\psi(y') (1 + r_M(y'))^{1-\eta}}. \quad (8A.8)$$

Along any optimal path, $V(x, y) = \mathcal{W}(c(x, y), E(V(x', y')))$. By plugging in \mathcal{W} (from Eq. 8.5)) and the conjecture for V ,

$$E\left[\psi(y') (1 + r_M(y'))^{1-\eta}\right] = \beta^{-\frac{1-\eta}{\rho}} \left(\frac{a(y)}{1-a(y)}\right)^{\frac{(1-\eta)(\rho-1)}{\rho}}. \quad (8A.9)$$

Moreover,

$$\psi(y') (1 + r_M(y'))^{1-\eta} = \left(a(y') (1 + r_M(y'))^{\frac{\rho}{\rho-1}}\right)^{\frac{(1-\eta)(\rho-1)}{\rho}}. \quad (8A.10)$$

By plugging Eqs. (8A.9)-(8A.10) into Eq. (8A.8),

$$\begin{aligned} \frac{E(V(x', y'))}{V(x', y')} &= \beta^{-\frac{1-\eta}{\rho}} \left(\frac{a(y)}{(1-a(y)) a(y') (1 + r_M(y'))^{\frac{\rho}{\rho-1}}}\right)^{\frac{(1-\eta)(\rho-1)}{\rho}} \\ &= \beta^{-\frac{1-\eta}{\rho}} \left(\left(\frac{c'}{c}\right)^{-1} \frac{x'}{(1-a(y)) x (1 + r_M(y'))^{\frac{\rho}{\rho-1}}}\right)^{\frac{(1-\eta)(\rho-1)}{\rho}} \\ &= \beta^{-\frac{1-\eta}{\rho}} \left(\left(\frac{c'}{c}\right)^{-1} \frac{1}{(1 + r_M(y'))^{\frac{1}{\rho-1}}}\right)^{\frac{(1-\eta)(\rho-1)}{\rho}} \end{aligned}$$

where the first equality follows by Eq. (8A.6), and the second equality follows by Eq. (8A.7). The result follows by replacing this into Eq. (8A.5). ■

PROOF OF EQS. (8.10) AND (8.11). By using the standard property that $\ln E(e^{\tilde{y}}) = E(\tilde{y}) + \frac{1}{2} \text{var}(\tilde{y})$, for \tilde{y} normally distributed, in Eq. (8.8), we obtain,

$$\begin{aligned} 0 &= \ln E\left(\exp\left(-\delta\theta - \frac{\theta}{\psi} \ln\left(\frac{c'}{c}\right) + \theta R_M\right)\right) \\ &= -\delta\theta - \frac{\theta}{\psi} E\left[\ln\left(\frac{c'}{c}\right)\right] + \theta E(R_M) + \frac{1}{2} \left(\left(\frac{\theta}{\psi}\right)^2 \sigma_c^2 + \theta^2 \sigma_{R_M}^2 - 2\frac{\theta^2}{\psi} \sigma_{R_M, c}\right). \end{aligned} \quad (8A.11)$$

We do the same in Eq. (8.9), and obtain,

$$R_f = \delta\theta + \frac{\theta}{\psi} E\left(\ln\left(\frac{c'}{c}\right)\right) - (\theta - 1) E(R_M) - \frac{1}{2} \left(\left(\frac{\theta}{\psi}\right)^2 \sigma_c^2 + (\theta - 1)^2 \sigma_{R_M}^2 - 2\frac{\theta(\theta - 1)}{\psi} \sigma_{R_M, c}\right). \quad (8A.12)$$

By replacing Eq. (8A.12) into Eq. (8A.11), we obtain Eq. (8.10) in the main text.

To obtain the risk-free rate R_f in Eq. (8.11), we replace the expression for $E(R_M)$ in Eq. (8.10) into Eq. (8A.12). ■

8.11.2 Details concerning models of long-run risks

PROOF OF EQ. (8.18). By substituting the guess $z_t = a_0 + a_1x_t$ into Eq. (8.17),

$$\begin{aligned} 0 &= (\kappa_0 - (1 - \kappa_1)a_0 - \delta)\theta + \ln E_t \left(\exp \left(-\frac{\theta}{\psi} \ln \left(\frac{C_{t+1}}{C_t} \right) + \theta\kappa_1 a_1 x_{t+1} - \theta a_1 x_t + \theta g_{t+1} \right) \right) \\ &= \theta \left(\left((\kappa_0 - (1 - \kappa_1)a_0 - \delta) + \left(1 - \frac{1}{\psi} \right) \left(g_0 - \frac{1}{2}\sigma_c^2 \right) \right) \right) + \ln E_t \left(\exp \left(\theta \left(1 - \frac{1}{\psi} \right) \epsilon_t + \theta\kappa_1 a_1 \eta_t \right) \right) \\ &+ \left((\kappa_1 \rho - 1)a_1 + \left(1 - \frac{1}{\psi} \right) \right) x_t \\ &\equiv \text{const}_1 + \text{const}_2 \cdot x_t, \end{aligned}$$

where the second equality follows by Eqs. (8.15) and (8.16). Note, then, that this equality can only hold if the two constants, const_1 and const_2 are both zero. Imposing $\text{const}_2 = 0$ yields,

$$a_1 = \frac{1 - \frac{1}{\psi}}{1 - \kappa_1 \rho},$$

as in Eq. (8.18) in the main text. Imposing $\text{const}_1 = 0$, and using the solution for a_1 , yields the solution for the constant a_0 . ■

8.11.3 Continuous time

Duffie and Epstein (1992a,b) extend the framework on non-expected utility to continuous time. Heuristically, the continuation utility is the continuous time limit of,

$$v_t = \left(c_t^\rho \Delta t + e^{-\delta \Delta t} \left(E(v_{t+\Delta t}^{1-\eta}) \right)^{\frac{\rho}{1-\eta}} \right)^{1/\rho}.$$

Continuation utility v_t solves the following stochastic differential equation,

$$dv_t = \left(-f(c_t, v_t) - \frac{1}{2}A(v_t) \|\sigma_{vt}\|^2 \right) dt + \sigma_{vt} dB_t, \quad \text{with } v_T = 0$$

Now, (f, A) is the aggregator, with A being a variance multiplier, placing a penalty proportional to utility volatility $\|\sigma_{vt}\|^2$. The aggregator (f, A) corresponds somehow to the aggregator (W, \hat{v}) of the discrete time case.

The solution to the previous “stochastic differential utility” is:

$$v_t = E \left[\int_t^T \left(f(c_s, v_s) + \frac{1}{2}A(v_s) \|\sigma_{vs}\|^2 \right) ds \right],$$

which collapses to the standard additive utility case once $f(c, v) = u(c) - \beta v$ and $A = 0$.

8.12 Appendix 2: Economies with heterogenous agents

ECONOMIES WITH A CONTINUUM OF AGENTS. We assume each agent faces a system of complete markets, in which case the equilibrium can be computed along the lines of Huang (1987), an extension of the classical approach described in Chapter 2 of Part I of these Lectures. We consider a continuum of agents indexed by an instantaneous utility function $u_a(c, x)$, where c is consumption, a is a parameter belonging to some set A , and x is some state variable affecting the utility function. For example, x is the “standard of living of others” in the Chan and Kogan (2002) model. Since agents access a complete market systems, the equilibrium allocation is Pareto efficient. By the second welfare theorem, then, we know that for each Pareto allocation $(c_a)_{a \in A}$, there exists a social weighting function f such that the Pareto allocation can be “implemented” by means of the following program,

$$\max_{c_a} E \left[\int_0^\infty e^{-\delta t} \left(\int_{a \in A} u_a(c_a(t), x(t)) f(a) da \right) dt \right], \quad \text{s.t.} \quad \int_{a \in A} c_a(t) da = D(t),$$

or, given that there is no intertemporal transfer of resources,

$$u(D, x) = \max_{c_a} \int_{a \in A} u_a(c_a, x) f(a) da, \quad \text{s.t.} \quad \int_{a \in A} c_a da = D, \quad [8A.P1]$$

where D is the aggregate endowment in the economy. Then, the equilibrium price system can be computed as the Arrow-Debreu state price density in an economy with a single agent endowed with the aggregate endowment D , instantaneous utility function $u(c, x)$, and where for $a \in A$, the social weighting function $f(a)$ equals the reciprocal of the marginal utility of income of the agent a .

The practical merit of this approach is that while the marginal utility of income is unobservable, the thusly constructed Arrow-Debreu state price density depends on the “infinite dimensional parameter”, f , which can be calibrated to reproduce the main quantitative features of consumption and asset price data. We now apply this approach and derive the equilibrium conditions in the Chan and Kogan (2002) model.

“CATCHING UP WITH THE JONESES” (CHAN AND KOGAN, 2002). In this model, markets are complete, and we have that $A = [1, \infty]$ and $u_\eta(c_\eta, x) = (c_\eta/x)^{1-\eta} / (1-\eta)$. The static optimization problem for the social planner in [8A.P1] can be written as,

$$u(D, x) = \max_{c_\eta} \int_1^\infty \frac{(c_\eta/x)^{1-\eta}}{1-\eta} f(\eta) d\eta, \quad \text{s.t.} \quad \int_1^\infty (c_\eta/x) d\eta = D/x. \quad (8A.13)$$

The first order conditions for this problem lead to,

$$\left(\frac{c_\eta}{x} \right)^{-\eta} f(\eta) = V(D/x), \quad (8A.14)$$

where V is a Lagrange multiplier, a function of the aggregate endowment D , normalized by x . It is determined by Eq. (8.22), which is obtained by replacing Eq. (8A.14) into the budget constraint of the social planner, the second of Eqs. (8A.13). The value function in Eq. (8.21) of the main text follows by replacing Eq. (8A.14) into the maximized value of the intertemporal utility, the first of Eqs. (8A.13). General equilibrium allocations, and prices, are obtained by setting $f(\eta)$ equal to the reciprocal of the marginal utility of income for agent η .

The expression for the unit risk-premium in Eq. (8.23) follows by results given in Section 7.5.1 of Chapter 7,

$$\lambda(s(D, x)) = - \left(\frac{\partial^2 U(s(D, x))}{\partial D^2} \bigg/ \frac{\partial U(s(D, x))}{\partial D} \right) \sigma_0 D, \quad s(D, x) \equiv \ln \frac{D}{x}, \quad (8A.15)$$

where $U(s)$ is the value function in Eq. (8.21). To evaluate the previous expression for λ , note, first, that:

$$\frac{\partial U(s(D, x))}{\partial D} = -\frac{V'(s(D, x))}{D} \int_1^\infty \frac{1}{\eta} f(\eta)^{\frac{1}{\eta}} V(s(D, x))^{-\frac{1}{\eta}} d\eta. \quad (8A.16)$$

Moreover, by differentiating Eq. (8.22) with respect to D , using Eq. (8A.16), and rearranging terms, leads to $\partial U(s(D, x))/\partial D = V(s(D, x))/x$. Differentiating this expression for $\partial U/\partial D$ with respect to D again, produces:

$$\frac{\partial^2 U(s(D, x))}{\partial D^2} = \frac{V'(s(D, x))}{x} \frac{1}{D}. \quad (8A.17)$$

Replacing Eqs. (8A.16)-(8A.17) into Eq. (8A.15) yields Eq. (8.23) in the main text. In this fictitious representative agent economy, the short-term rate is the expectation of the stochastic discount factor. It equals, again by results given in Section 7.5.1 of Chapter 7,

$$\begin{aligned} R(D, s, x) &= \rho - g_0 \frac{\partial^2 U(s(D, x))/\partial D^2}{\partial U(s(D, x))/\partial D} D - \theta \frac{\partial^2 U(s(D, x))/\partial D \partial x}{\partial U(s(D, x))/\partial D} s x - \frac{1}{2} \sigma_0^2 \frac{\partial^3 U(s(D, x))/\partial D^3}{\partial U(s(D, x))/\partial D} D^2 \\ &= \rho + \frac{g_0}{\sigma_0} \lambda(s(D, x)) - \theta \left(1 + \frac{V'(s(D, x))}{V(s(D, x))} \right) s(D, x) - \frac{1}{2} \sigma_0^2 \left(\frac{V''(s(D, x)) - V'(s(D, x))}{V(s(D, x))} \right). \end{aligned}$$

It is instructive to compare the first order conditions of the social planner in Eq. (8A.14) with those in the decentralized economy. Since markets are complete, we have that the first order conditions in the decentralized economy satisfy:

$$e^{-\delta t} (c_\eta(t)/x(t))^{-\eta} = \kappa(\eta) \xi(t) x(t), \quad (8A.18)$$

where $\kappa(\eta)$ is the marginal utility of income for agent η , and $\xi(t)$ is the usual pricing kernel.

By aggregating the market equilibrium allocations in Eq. (8A.18),

$$D(t) = \int_1^\infty c_\eta(t) d\eta = x(t) \int_1^\infty \left[e^{\delta t} \xi(t) x(t) \right]^{-\frac{1}{\eta}} \kappa(\eta)^{-\frac{1}{\eta}} d\eta.$$

By aggregating the social weighted allocations in Eq. (8A.14), with $f = \kappa^{-1}$,

$$D(t) = x(t) \int_1^\infty V(D(t)/x(t))^{-\frac{1}{\eta}} \kappa(\eta)^{-\frac{1}{\eta}} d\eta.$$

Hence, it must be that,

$$x(t)^{-1} V(D(t)/x(t)) = e^{\delta t} \xi(t). \quad (8A.19)$$

That is, if $f = \kappa^{-1}$, then, Eq. (8A.19) holds. The converse to this result is easy to obtain: eliminating $(c_\eta/x)^{-\eta}$ from Eq. (8A.14) and Eq. (8A.18) leaves:

$$\frac{e^{\delta t} x(t) \xi(t)}{V(D(t)/x(t))} = \frac{1}{f(\eta) \kappa(\eta)}, \quad (t, \eta) \in [0, \infty) \times [1, \infty).$$

Hence if Eq. (8A.19) holds, then, $f = \kappa^{-1}$.

To summarize, the equilibrium allocations and prices can be “centralized” through the social planner program in (8A.13), with $f = \kappa^{-1}$.

RESTRICTED STOCK MARKET PARTICIPATION (BASAK AND CUOCO, 1998). Given Eq. (8.34), and results given in Section 7.5.1 of Chapter 7, the unit risk-premium, $\lambda(D, x)$, solves a fixed point problem:

$$\lambda(D, x) = -\frac{U_{11}(D, x)}{U_1(D, x)} \sigma_0 D + \frac{U_{12}(D, x)x}{U_1(D, x)} \lambda(D, x).$$

That is,

$$\lambda(D, x) = -\frac{U_1(D, x)U_{11}(D, x)}{U_1(D, x) - U_{12}(D, x)x} \cdot \frac{\sigma_0 D}{U_1(D, x)}. \quad (8A.20)$$

We claim that:

$$U_1(D, x) = u'_p(c_p^*), \quad \text{and} \quad \frac{U_1(D, x)U_{11}(D, x)}{U_1(D, x) - U_{12}(D, x)x} = u''_p(c_p^*), \quad (8A.21)$$

where c_p^* and c_n^* are the social planner consumption allocations. By replacing Eqs. (8A.21) into Eq. (8A.20), and using the definition of u_p and s , leads to the expression for λ in Eq. (8.29). The expression for short-term rate R in Eq. (8.28) can be found similarly, and again, through the results given in Section 7.5.1 of Chapter 7.

We now show that Eq. (8A.21) hold true. Consider the Lagrangean for the maximization problem in Eq. (8.33),

$$L = u_p(c_p) + xu_n(c_n) + \nu(c - c_p - c_n),$$

where ν is a Lagrange multiplier, $x = \frac{u'_p(\hat{c}_p)}{u'_n(\hat{c}_n)}$, and \hat{c}_p and \hat{c}_n are the market consumption allocations. The first order conditions for the social planner lead to social allocation functions $c_p^* = c_p(D, x)$ and $c_n^* = c_n(D, x)$, and Lagrange multiplier $\nu(D, x)$, satisfying:

$$u'_p(c_p(D, x)) = \nu(D, x) = xu'_n(c_n(D, x)). \quad (8A.22)$$

Accordingly, the value of the problem in Eq. (8.33) is:

$$U(D, x) = u_p(c_p(D, x)) + xu_n(c_n(D, x)),$$

such that

$$\begin{aligned} U_1(D, x) &= u'_p(c_p(D, x)) \frac{\partial c_p(D, x)}{\partial D} + xu'_n(c_n(D, x)) \frac{\partial c_n(D, x)}{\partial D} \\ &= u'_p(c_p(D, x)) \left(\frac{\partial c_p(D, x)}{\partial D} + \frac{\partial c_n(D, x)}{\partial D} \right) \\ &= u'_p(c_p(D, x)), \end{aligned} \quad (8A.23)$$

where the second equality follows by the first order conditions in Eq. (8A.22), and the third equality holds by differentiating the equilibrium condition

$$D = c_p(D, x) + c_n(D, x), \quad (8A.24)$$

with respect to D .

Eq. (8A.23) establishes the first claim in Eqs. (8A.21). To prove the second claim, invert, first, the first order condition with respect to ν , obtaining, $c_p(D, x) = u'^{-1}_p[\nu(D, x)]$ and $c_n(D, x) = u'^{-1}_n[\nu(D, x)/x]$. Replace, then, these inverse functions into Eq. (8A.24),

$$D = u'^{-1}_p[\nu(D, x)] + u'^{-1}_n[\nu(D, x)/x], \quad (8A.25)$$

where, by Eq. (8A.24) and Eq. (8A.23),

$$\nu(D, x) = u'_p(c_p(D, x)) = U_1(D, x). \quad (8A.26)$$

Differentiating Eq. (8A.25) with respect to x and D , and using Eq. (8A.26), leaves:

$$0 = \frac{1}{u''_p(c_p(D, x))} U_{12}(D, x) + \frac{1}{u''_n(c_n(D, x))} \frac{U_{12}(D, x)x - U_1(D, x)}{x^2}, \quad (8A.27)$$

and

$$1 = \frac{1}{u_p''(c_p(D, x))} U_{11}(D, x) + \frac{1}{u_n''(c_n(D, x))} \frac{U_{11}(D, x)}{x}. \quad (8A.28)$$

Replacing Eq. (8A.28) into Eq. (8A.27) leaves:

$$0 = \frac{U_{12}(D, x)}{u_p''(c_p(D, x))} + \left(\frac{1}{U_{11}(D, x)} - \frac{1}{u_p''(c_p(D, x))} \right) \frac{U_{12}(D, x) x - U_1(D, x)}{x}.$$

The second relation in Eqs. (8A.21) follows by rearranging terms in the previous relation.

References

- Abel, A.B. (1990): “Asset Prices under Habit Formation and Catching Up with the Joneses.” *American Economic Review Papers and Proceedings* 80, 38-42.
- Abel, A.B. (1999): “Risk Premia and Term Premia in General Equilibrium.” *Journal of Monetary Economics* 43, 3-33.
- Alvarez, F. and U.J. Jermann (20??):
- Bansal, R. and A. Yaron (2004): “Risks for the Long Run: A Potential Resolution of Asset Pricing Puzzles.” *Journal of Finance* 59, 1481-1509.
- Basak, S. and D. Cuoco (1998): “An Equilibrium Model with Restricted Stock Market Participation.” *Review of Financial Studies* 11, 309-341.
- Bernanke, B. S., M. Gertler and S. Gilchrist (1999): “The Financial Accelerator in a Quantitative Business Cycle Framework.” In J. B. Taylor and M. Woodford (Eds.): *Handbook of Macroeconomics*, Vol. 1C, Chapter 21, 1341-1393.
- Black, F. (1976): “Studies of Stock Price Volatility Changes.” *Proceedings of the 1976 Meeting of the American Statistical Association*, 177-81.
- Boldrin, M., L. Christiano and J. Fisher (2001): “Habit Persistence, Asset Returns and the Business Cycle.” *American Economic Review* 91, 149-166.
- Campbell, J. Y., A. W. Lo and C. MacKinlay (1997): *The Econometrics of Financial Markets*. Princeton: Princeton University Press.
- Campbell, J. Y. (2003): “Consumption-Based Asset Pricing.” In Constantinides, G. M., M. Harris and R. M. Stulz (Editors): *Handbook of the Economics of Finance* (Volume 1B, chapter 13), North-Holland Elsevier, 803-887.
- Campbell, J. Y., and J. H. Cochrane (1999): “By Force of Habit: A Consumption-Based Explanation of Aggregate Stock Market Behavior.” *Journal of Political Economy* 107, 205-251.
- Campbell, J. and R. Shiller (1988): “The Dividend-Price Ratio and Expectations of Future Dividends and Discount Factors.” *Review of Financial Studies* 1, 195–228.
- Chan, Y.L. and L. Kogan (2002): “Catching Up with the Joneses: Heterogeneous Preferences and the Dynamics of Asset Prices.” *Journal of Political Economy* 110, 1255-1285.
- Christie, A.A. (1982): “The Stochastic Behavior of Common Stock Variances: Value, Leverage, and Interest Rate Effects.” *Journal of Financial Economics* 10, 407-432.
- Cochrane, J. H., F. A. Longstaff, and P. Santa-Clara (2008): “Two Trees.” *Review of Financial Studies* 21, 347-385.
- Constantinides, G.M. and D. Duffie (1996): “Asset Pricing with Heterogeneous Consumers.” *Journal of Political Economy* 104, 219-240.

- Constantinides, G.M., J.B. Donaldson and R. Mehra (2002): “Juniors Can’t Borrow: a New Perspective on the Equity Premium Puzzle.” *Quarterly Journal of Economics* 117, 269-296.
- Duffie, D. and L.G. Epstein (1992a): “Asset Pricing with Stochastic Differential Utility.” *Review of Financial Studies* 5, 411-436.
- Duffie, D. and L.G. Epstein (with C. Skiadas) (1992b): “Stochastic Differential Utility.” *Econometrica* 60, 353-394.
- Epstein, L.G. and S.E. Zin (1989): “Substitution, Risk-Aversion and the Temporal Behavior of Consumption and Asset Returns: A Theoretical Framework.” *Econometrica* 57, 937-969.
- Epstein, L.G. and S.E. Zin (1991): “Substitution, Risk-Aversion and the Temporal Behavior of Consumption and Asset Returns: An Empirical Analysis.” *Journal of Political Economy* 99, 263-286.
- Gallmeyer, M., Aydemir, A.C. and B. Hollifield (2007): “Financial Leverage and the Leverage Effect: A Market and a Firm Analysis.” working paper Carnegie Mellon.
- Guvenen, F. (2009): “A Parsimonious Macroeconomic Model for Asset Pricing.” *Econometrica* 77, 1711-1740.
- Heaton, J. and D.J. Lucas (1996): “Evaluating the Effects of Incomplete Markets on Risk Sharing and Asset Pricing.” *Journal of Political Economy* 104, 443-487.
- Huang, C.-f. (1987): “An Intertemporal General Equilibrium Asset Pricing Model: the Case of Diffusion Information.” *Econometrica* 55, 117-142.
- Jermann, U.J. (1998): “Asset Pricing in Production Economies.” *Journal of Monetary Economics* 41, 257-276.
- Kiyotaki, N. and J. Moore (1997): “Credit Cycles.” *Journal of Political Economy* 105, 211-248.
- Lettau, M. (2002): “Idiosyncratic Risk and Volatility Bounds, or, Can Models with Idiosyncratic Risk Solve the Equity Premium Puzzle?” *Review of Economics and Statistics* 84, 376-380.
- Lucas, R.E. (19??):
- Lucas, D.J. (1994): “Asset Pricing with Undiversifiable Income Risk and Short Sales Constraints: Deepening the Equity Premium Puzzle.” *Journal of Monetary Economics* 34, 325-341.
- Mankiw, N.G. (1986): “The Equity Premium and the Concentration of Aggregate Shocks.” *Journal of Financial Economics* 17, 211-219.
- Mankiw, N.G. and S.P. Zeldes (1991): “The Consumption of Stockholders and Non-Stockholders.” *Journal of Financial Economics* 29, 97-112.
- Martin, I. (2011): “The Lucas Orchard.” Working Paper Stanford University.

- Menzly, L., T. Santos and P. Veronesi (2004): “Understanding Predictability.” *Journal of Political Economy* 112, 1, 1-47.
- Nelson, D.B. (1991): “Conditional Heteroskedasticity in Asset Returns: A New Approach.” *Econometrica* 59, 347-370.
- Pavlova, A. and R. Rigobon (2008): “The Role of Portfolio Constraints in the International Propagation of Shocks.” *Review of Economic Studies* 75, 1215-1256.
- Rouwenhorst, G. K. (1995): “Asset Returns and Business Cycles.” In Cooley, T.F. (Ed.): *Frontiers of Business Cycle Research*, Princeton University Press, 294-330.
- Schwert, G.W. (1989a): “Why Does Stock Market Volatility Change Over Time?” *Journal of Finance* 44, 1115-1153.
- Schwert, G.W. (1989b): “Business Cycles, Financial Crises and Stock Volatility.” *Carnegie-Rochester Conference Series on Public Policy* 31, 83-125.
- Tallarini, T. (2000): “Risk-Sensitive Real Business Cycles.” *Journal of Monetary Economics* 45, 507-32.
- Telmer, C.I. (1993): “Asset-Pricing Puzzles and Incomplete Markets.” *Journal of Finance* 48, 1803-1832.
- Wachter, J.A. (2006): “A Consumption-Based Model of the Term Structure of Interest Rates.” *Journal of Financial Economics* 79, 365-399.
- Weil, Ph. (1989): “The Equity Premium Puzzle and the Risk-Free Rate Puzzle.” *Journal of Monetary Economics* 24, 401-421.
- Xiouros, C. and F. Zapatero (2010): “The Representative Agent of an Economy with External Habit Formation and Heterogeneous Risk Aversion.” *Review of Financial Studies* 23, 3017-3047.

9

Information and other market frictions

9.1 Introduction

In the models of the previous chapters, agents do not need to learn about the equilibrium price because information, whilst sometimes incomplete, is disseminated symmetrically across decision makers. This chapter considers settings where agents have access to differential or even asymmetric information about some attributes relating to economic developments. In these settings with imperfect information, there are gains to be made by learning about the equilibrium price, because the very same price, conveys some information every agents transmit while he trades, thereby making it public so to speak. Naturally, the price cannot convey all private information when information is costly, for otherwise there would not be any incentive to purchase information. To avoid this information paradox, prices need to convey “noisy,” or imperfect signals about the information private investors have. As Black (1986) discussed, noise makes markets function when information problems would otherwise lead them not to arise in the first place.

The assumption agents have imperfect information about the fundamentals of the economy was first used by Phelps (1970) and Lucas (1972), to explain the relation between monetary policy and the business cycle. This information-based approach to the business cycle, summarized in Lucas (1981), was, in fact, abandoned in favour of the real business cycle theory, reviewed in Chapter 3, partly because imperfect information cannot be considered as the sole engine of macroeconomic fluctuations. Instead, it is widely acknowledged that the merit of Lucas’ approach was the introduction of a systematic way of thinking about fluctuations, in a context with rational expectations. Moreover, his information approach has inspired work in financial economics, where imperfect information is likely to play a quite fundamental role. In Section 9.2, we provide a succinct account of the Lucas framework, and solve a model relying on a simplified version of Lucas (1973). We solve this model, following the perspective we think a finance theorist would typically have. It is quite useful to present this model, as this is very simple and at the same time, contributes to give us a big picture of where imperfect information can lead us, in general. Section 9.2 through 9.7 review the many models in financial economics that have been used to explain the price formation mechanism in contexts with imperfect information, be it asymmetric or differential, as we shall make precise below.

Sections 9.7 and 9.9 conclude this chapter, and present additional market frictions that are potentially apt to explain certain features in the asset price formation process.

9.2 Prelude: imperfect information in macroeconomics

There are n islands, where n goods are produced. Let y_i^s denote log-production supplied in the i -th island. (All prices and quantities are in logs, in this section.) It is assumed that this supply is set so as to equal the expected wedge of the price in the island, p_i , over the average price in the economy, p ,

$$y_i^s = E(p_i - p | p_i), \quad \text{where} \quad p = \frac{1}{n} \sum_{j=1}^n p_j.$$

The previous equation can be easily derived, once we assume p is common knowledge, as for example in the model of monopolistic competition of Blanchard and Kiyotaki (1987). If, instead, p is not common knowledge, it is more problematic to derive the *exact* functional form assumed for y_i^s , although this describes a quite plausible decision mechanism.

Information is disseminated differentially, not asymmetrically, in that producers in the i -th island do not know the price in the remaining islands, and guess economic developments in the other islands with the same precision. We assume and, later, verify, that *all* variables, exogenous and endogenous, are normally distributed. Under this presumption, we shall show, the price index p gathers all the available information in the economy efficiently, i.e. it is a sufficient statistics for all that information.

We have, by the Projection theorem,

$$y_i^s = E(p_i - p | p_i) \equiv \beta (p_i - E(p)),$$

where we have used the fact that information is symmetrically disseminated and, then, (i) the expectation $E(p_i) = E(p_j) = E(p)$ for every i and j , and (ii) both the numerator and denominator of the ratio, $\beta \equiv \frac{\text{cov}(p_i - p, p_i)}{\text{var}(p_i)}$, are the same across all islands. This coefficient will be determined below, as a result of the equilibrium.

Aggregating across all islands, yields the celebrated *Lucas supply* equation:

$$y^s \equiv \frac{1}{n} \sum_{j=1}^n y_j^s = \beta (p - E(p)). \quad (9.1)$$

Next, assume the demand for the good produced in the i -th island is given by:

$$y_i^d = m - p + u_i - \theta (p_i - p), \quad \text{where} \quad u_i \sim N(0, \sigma_u^2)$$

where money is

$$m = E(m) + \epsilon, \quad \text{where} \quad \epsilon \sim N(0, \sigma_\epsilon^2). \quad (9.2)$$

Finally, we assume that $E(u_i \epsilon) = 0$, and that u_i are a sectoral shocks, in that: $\sum_{j=1}^n u_j = 0$. The functional form assumed for the demand function, y_i^d , can be easily derived, assuming the goods in the islands are imperfect substitutes, as for example in Blanchard and Kiyotaki (1987).

In this context, the equilibrium price in the islands plays two roles. A first, standard role, is to clear the market in each island, being such that $y_i^s = y_i^d$, or:

$$\beta (p_i - E(p)) = m - p + u_i - \theta (p_i - p), \quad \text{for all } i. \quad (9.3)$$

Its second role is to convey information about the two shocks, the macroeconomic, monetary shock, ϵ , and the real shocks in all the islands, u_j , $j = 1, \dots, n$. Let us assume, then, that the only real shock that matters for the price in the i -th island is u_i . Below, we shall verify this conjecture holds, in equilibrium. Then, the price is a function $p_i = P(\epsilon, u_i)$, which we conjecture to be affine, in ϵ and u_i , viz

$$P(\epsilon, u_i) = a + b\epsilon + cu_i, \quad (9.4)$$

where the coefficients a , b and c have to be determined, in equilibrium. Under these conditions, the average price is a function $p = \bar{P}(\epsilon)$, equal to:

$$\bar{P}(\epsilon) = a + b\epsilon. \quad (9.5)$$

Let us replace Eqs. (9.4), (9.5) and (9.2) into Eq. (9.3). By rearranging terms, we obtain:

$$0 = (\beta b + b - 1)\epsilon + (\beta c + \theta c - 1)u_i + a - E(m).$$

This equation has to hold for all ϵ and u_i . Therefore,

$$a = E(m),$$

and the coefficients for ϵ and u_i must both equal zero, leading to the following expressions for b and c :

$$b = \frac{1}{1 + \beta}, \quad c = \frac{1}{\theta + \beta}. \quad (9.6)$$

We are left with determining β , which given Eqs. (9.4)-(9.5), and Eq. (9.6), is easily shown to equal:

$$\beta = \frac{\sigma_u^2}{\sigma_u^2 + \left(\frac{\theta + \beta}{1 + \beta}\right)^2 \sigma_\epsilon^2}. \quad (9.7)$$

The positive fixed point to this equation, which is easily shown to exist, delivers β , which can then be replaced back into Eqs. (9.6), to yield the solutions for b and c , which are both positive.

We can now figure out the implications of this equilibrium. By replacing Eqs. (9.4)-(9.5) into the Lucas supply equation (9.1), leaves:

$$y^s = \beta b \epsilon.$$

This is Lucas celebrated neutrality result. Anticipated monetary policy, $E(m)$, does not affect the equilibrium outcome, y^s . Instead, it is the monetary shock that affects y^s . Agents in any one island do not observe the price in the remaining islands and, hence, the aggregate price level, p . Therefore, they are unable to tell whether an increase in the price of the good they produce, p_i , is due to a real shock, u_i , or to a monetary shock, ϵ . In other words, they cannot disentangle a monetary shock from a real shock. If the agents were informed about the real shocks in the other islands, they would of course infer ϵ , and a monetary shock would not exert any effect on the equilibrium production. Formally, in equilibrium, the price difference, $p_i - p = cu_i$, which does not depend on ϵ , a standard ‘‘dichotomy’’ prediction reminiscent of classical theory. But $p_i - p$ is not observed, as p is not observed. Instead, the producers in the i -th island can only guess $p_i - E(p|p_i) = b\epsilon + cu_i$, which co-varies positively with the observed price, p_i , $cov(p_i - p, p_i) = c^2 \sigma_u^2$. This covariance is zero precisely when we remove the assumption of imperfect knowledge about the real shocks, so that $\sigma_u^2 = 0$, in which case $\beta = 0$. By contrast,

with imperfect knowledge, producers act so as to compensate for their partial lack of knowledge, and produce to the maximum extent they can justify, on the basis of the positive statistical co-movements, $cov(p_i - p, p_i) > 0$. Note, if $E(m) = m_{-1}$, i.e. money supply in the previous period, then from Eq. (9.5), the inflation rate, $p - p_{-1} = b\epsilon + (1 - b)\epsilon_{-1}$. Therefore, output and inflation are positively correlated, and generate a Phillips curve, which policy makers cannot exploit anyway, as anticipated monetary policy, $E(m)$, is rationally “factored out,” and does not affect output. This is the essence of the Lucas critique (Lucas, 1977).

In the next sections, we present a number of models that work due to a similar mechanism. Why should we ever purchase an asset from any one else, who is insisting in selling it to the market? Trading seems to be a difficult phenomenon to explain, in a world with imperfect information. Yet trading does occur, if imperfect information has the same nature as that of the Phelps-Lucas model. Agents might well be imperfectly informed about the nature of, say, unusually high market orders. For example, huge sell orders might arrive to the market, either because the asset is a lemon or because the agents selling it are hit by a liquidity shock. In the models of this section, an equilibrium with rational expectation exists, precisely because of this “noise”—liquidity, in this example. There is a chance the sell order arrives to the market, simply because the agents selling it are hit by a liquidity shock. Imperfectly informed agents, therefore, might be willing to buy, if it is in their interest to do so.

9.3 Grossman-Stiglitz paradox

Fama (1970) considers three forms of informational efficiency: (i) *Weak* efficiency, arising when asset prices convey all information relating to the past time-series of data; (ii) *Semi-strong* efficiency, relating to the situation where asset prices convey all public information, not only the past time series of data; (iii) *Strong* efficiency, whereby asset prices reflect private information. The third form of efficiency cannot exist, indefinitely, unless of course information is not costly. For it were costly, there would not incentive to purchase it in a world with informationally strong efficient markets. And if it were not costly, it would not, then, be private information. It is the Grossman-Stiglitz paradox

9.4 Noisy rational expectations equilibrium

9.4.1 Differential information

Hellwig (1980). Diamond and Verrecchia (1981).

9.4.2 Asymmetric information

Grossman and Stiglitz (1980).

9.4.3 Information acquisition

9.5 Strategic trading

Kyle (1985). Foster and Viswanathan (1996).

9.6 Dealers markets

Glosten and Milgrom (1985).

9.7 Noise traders

DeLong, Shleifer, Summers and Waldman (1990).

9.8 Demand-based derivative prices

9.8.1 Options

Gârleanu, Pedersen and Poteshman (2007).

9.8.2 Preferred habitat and the yield curve

Vayanos and Vila (2007), Greenwood and Vayanos (2008).

9.9 Over-the-counter markets

Duffie, Gârleanu and Pedersen (2005, 2007).

References

- Black, F. (1986): “Noise.” *Journal of Finance* 41, 529-543.
- Blanchard, O. and N. Kiyotaki (1987): “Monopolistic Competition and the Effects of Aggregate Demand.” *American Economic Review* 77, 647-666.
- Fama, E. (1970): “Efficient Capital Markets: A Review of Theory and Empirical Work.” *Journal of Finance* 25, 383-417.
- Lucas, R.E. (1972): “Expectations and the Neutrality of Money.” *Journal of Economic Theory* 4, 103-124.
- Lucas, R.E. (1973): “Some International Evidence on Output-Inflation Tradeoffs.” *American Economic Review* 63, 326-334.
- Lucas, R.E. (1977): “Econometric Policy Evaluation: A Critique.” *Carnegie-Rochester Conference Series on Public Policy* 1: 19-46.
- Lucas, R.E. (1981): *Studies in Business-Cycle Theory*. Boston, MIT Press.
- Phelps, E.S. (1970): “Introduction.” In: Phelps, E. S. (Editor): *Microeconomic Foundations of Employment and Inflation Theory*, New York: W. W. Norton.

Part III

Asset pricing and reality

10

Options and volatility

10.1 Introduction

This chapter is under construction. Will include exotics, evaluation through trees and calibration. Will cover some details on how to deal with market imperfections. This introduction will discuss markets for options, will give figures such as average trading volume re options and futures and volatility products on CBOE & CME.

10.2 Forwards

This section is a basic introduction to forwards and a slight generalization of them that allows for non-linear payoffs and the necessary changes of probability aimed to deal with these nonlinearities.

10.2.1 Pricing

Forwards can be synthesized, as follows. Let $P(t, T)$ be the price of a bond expiring at time T . Assuming the short-term rate r is constant, we have $P(t, T) = e^{-r(T-t)}$. At time t , borrow $P(t, T) F^*$ and buy a stock, with market price S_t , with $F^* : P(t, T) F^* - S(t) = 0$. Then, the payoff of this portfolio at time T is $S(T) - F^*$. But the portfolio is worthless at time t , so this trading is the same as a forward. Therefore, we have $F^* = F_T(t)$ (say), where:

$$F_T(t) = S(t) e^{r(T-t)}. \quad (10.1)$$

Forwards are insensitive to volatility, in general, although they might, under some circumstances clarified below.

10.2.2 Forwards as a means to borrow money

Forward contracts can be used to borrow money. We can do the following: (i) go long a forward, which at time T , delivers the payoff $-F + S(T)$; (ii) short-sell the underlying asset, which at time T , will give rise to a payoff of $-S(T)$. So, (i) and (ii) are such that now, we access to $S(t)$

dollars, due to (ii), and at time T , we need to pay $-F$, which is the sum of the two payoffs resulting from (i) and (ii). By Eq. (10.1), this is tantamount to borrowing money at the interest rate r .

10.2.3 A pricing formula

Consider, again, a contract similar to that in Section 10.2.1, where at time T , the payoff is given by $S(T) - K$, for some constant value of K . We know that the current value of this payoff is:

$$e^{-r(T-t)}\mathbb{E}_t(S(T) - K) = S(t) - e^{-r(T-t)}K, \quad (10.2)$$

which is zero, for $K = F^*$. We want to consider the situation where such a current value is not necessarily zero, and show that the previous expression is a special case of a quite important pricing formula. Consider a payoff at time T , equal to $S(T) - K$, provided the stock price at T is at least as large as some positive constant $\ell \geq 0$,

$$(S(T) - K)\mathbb{I}_{S(T) \geq \ell}.$$

For $\ell = 0$, this payoff is just that of a forward, and for $\ell = K$, the payoff is that of a European call. To price this payoff, we proceed as follows:

$$\begin{aligned} e^{-r(T-t)}\mathbb{E}_t[(S(T) - K)\mathbb{I}_{S(T) \geq \ell}] &= S(t)\mathbb{E}_t(\eta_t(T)\mathbb{I}_{S(T) \geq \ell}) - e^{-r(T-t)}K\mathbb{E}_t(\mathbb{I}_{S(T) \geq \ell}) \\ &= S(t)\hat{\mathbb{E}}_t(\mathbb{I}_{S(T) \geq \ell}) - e^{-r(T-t)}K\mathbb{E}_t(\mathbb{I}_{S(T) \geq \ell}) \\ &= S(t) \cdot \hat{Q}_t(S(T) \geq \ell) - e^{-r(T-t)}K \cdot Q(S(T) \geq \ell), \end{aligned} \quad (10.3)$$

where $\eta_t(T) \equiv \frac{e^{-r(T-t)}S(T)}{S(t)}$, Q_t is the risk-neutral probability given the information at time t , \hat{Q}_t is a new probability, with Radon-Nikodym derivative given by

$$\frac{d\hat{Q}_t}{dQ_t} = \eta_t(T), \quad (10.4)$$

and, finally, \mathbb{E}_t denotes the expectation under Q_t , and $\hat{\mathbb{E}}_t$ the expectation under \hat{Q}_t . Naturally, Eq. (10.3) collapses to Eq. (10.2), once $\ell = 0$, and to the celebrated Black and Scholes (1973) formula, once we take $S(t)$ to be a geometric Brownian Motion, and $\ell = K$, as explained in Section 10.4. It is a general formula, and a quite useful one, whilst dealing with difficult models where, for example, the volatility of the underlying asset return is not constant, as illustrated in Section 10.5.

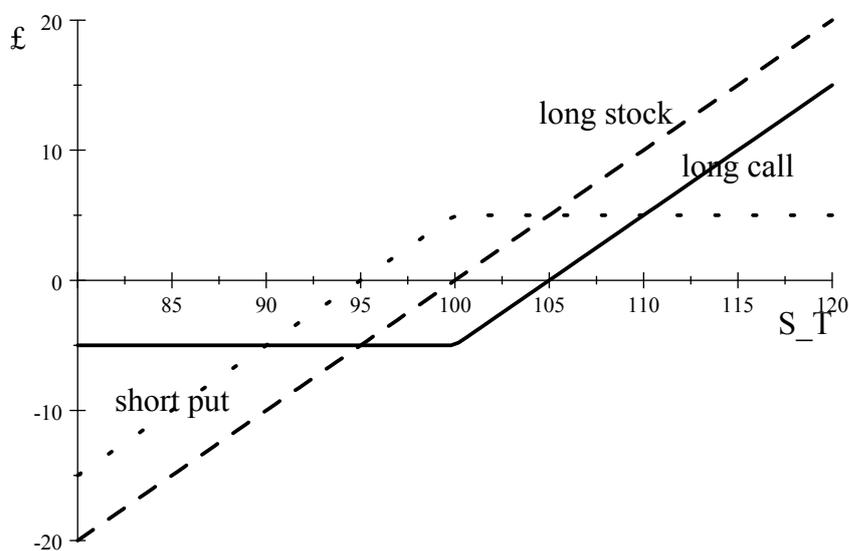
10.2.4 Forwards and volatility

10.3 Optionality and no-arb bounds

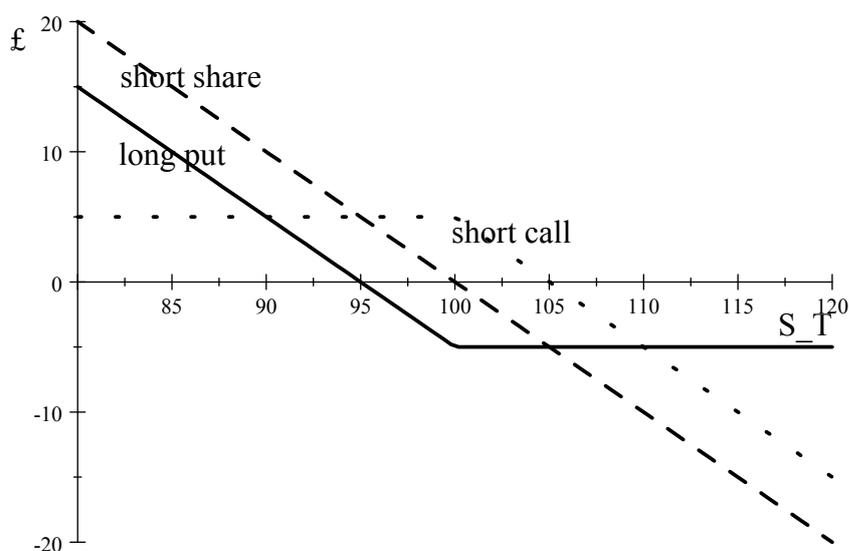
We analyze the properties of the simplest and extremely well-known, and yet fundamental, optionality a financial claim might display: the right, but not the obligation, to buy (in case of a call option) or sell (in case of a put option) a given asset at a given price, the strike, or exercise price, at some future date. While Chapter 4 illustrates foundational issues arising in the evaluation of these contracts, this chapter describes some of their properties, which might help explain their use in the market practice, and the reasons underlying some waves of financial innovation, such as that relating to volatility trading. This is the first section that illustrates a few fundamental properties of European options, with a few examples. American options, which can be exercised at any time before a reference maturity, are analyzed in Section 10.10.

10.3.1 Model-free properties

Let C and p the prices of the call and the put option, with S the price of the asset underlying these contracts, and with K and T the exercise prices and the expiration date. Let t be the current time. As we know, at the expiration, $C(T) = (S(T) - K)^+$ and $p(T) = (K - S(T))^+$. Figure 10.1 depicts the net profits generated by holding or short-selling one asset and one option written on the asset. We take the short-term rate $r = 0$ to simplify the presentation. The exposure to losses generated by going long an asset drops by trading the appropriate option, i.e by a long position in a call, as illustrated by the top panel, at least provided $C(t) < S(t)$, which is indeed a no-arbitrage condition, as shown below. It is this insurance feature that makes the option economically valuable. Likewise, the exposure to losses relating to a short position in the asset is mitigated by trading the appropriate option, i.e by a long position in the put option, as illustrated by the bottom panel, provided again that $p(t) < S(t)$, another no-arbitrage condition. Once again, this feature makes the option economically valuable.



Bullish view



Bearish view

FIGURE 10.1. Top panel: The solid line depicts the P&L of an at-the-money European call option when the interest rate is zero, $(S(T) - K)^+ - c(t)$, where $S(T)$ is the stock price at expiration, $K = 100$ is the strike price, and $c(t) = 5$ is the price of a call. The dashed line is the P&L from holding the stock, $S(T) - S(t)$, with $S(t) = 100$. The dotted line is the payoff arising from the sale of a put option $p(t) - (K - S(T))^+$, where by the put-call parity, $p(t) = c(t) - S(t) + K = 5$. The bottom panel depicts P&Ls arising from going long a put, $(K - S(T))^+ - p(t)$ (the solid line), shorting the stock, $S(T) - S(t)$ (the dashed line), and shorting a call $c(t) - (S(T) - K)^+$ (the dotted line).

The prices of the call and put options are related by the *put-call parity*. Let $P(t, T)$ be the time t price of a zero maturing at time $T > t$. Then, the prices of a put and a call option with the same exercise price K and the same expiration date T satisfy,

$$p(t) = C(t) - S(t) + KP(t, T). \tag{10.5}$$

To show Eq. (10.5), consider two portfolios: (a) long one call, short one underlying asset, and invest $KP(t, T)$; (b) long one put. The table below gives the value of the two portfolios at time t and at time T .

		Value at T	
		$S(T) \leq K$	$S(T) > K$
(a)	Value at t $C(t) - S(t) + KP(t, T)$	$-S(T) + K$	$S(T) - K - S(T) + K$
(b)	$p(t)$	$K - S(T)$	0

The two portfolios have the same value in each state of nature at time T . Therefore, their values at time t must be identical to rule out arbitrage. Mathematically, the put-call parity in Eq. (10.5), follows by taking conditional expectations of the identity: $e^{-r(T-t)}(S(T) - K)^+ \equiv e^{-r(T-t)}(K - S(T))^+ + e^{-r(T-t)}(S(T) - K)$.

By the put-call parity, the properties of European puts are easily found to follow from those of calls. Therefore, we only focus on calls, whenever possible. The price of a European call option satisfies the following bounds:

$$\max\{0, S(t) - KP(t, T)\} \leq C(S(t); K; T - t) \leq S(t). \tag{10.6}$$

Indeed, consider two portfolios: (a) long one call; (b) long the asset underlying the call and issue debt for an amount equal to $KP(t, T)$. The table below gives the value of the two portfolios at time t and at time T .

		Value at T	
		$S(T) \leq K$	$S(T) > K$
(a)	Value at t $C(t)$	0	$S(T) - K$
(b)	$S(t) - KP(t, T)$	$S(T) - K$	$S(T) - K$

The value of portfolio (a) dominates that of portfolio (b) at T , and the same must be true at time t . Moreover, the price C is positive because the payoff of the option is positive. Therefore, the first inequality in (10.6) is true. As for the second inequality, suppose the contrary, i.e. $C(t) > S(t)$. Then, at time t , we could sell one call and buy the underlying asset, thus making a profit equal to $C(t) - S(t)$. Come time T , the option will be exercised if $S(T) > K$, in which

case we shall sell the underlying assets and obtain K . If $S(T) < K$, the option will not be exercised, and we will still hold the asset or sell it and make a profit equal to $S(T)$. Eq. (10.6) implies the following asymptotic behavior of the call price: (i) $\lim_{S \rightarrow 0} C(S; K; T - t) \rightarrow 0$, (ii) $\lim_{K \rightarrow 0} C(S; K; T - t) \rightarrow S$, and (iii) $\lim_{T \rightarrow \infty} C(S; K; T - t) \rightarrow S$.

The top panel of Figure 10.2 illustrates the basic arbitrage bounds in (10.6), as well as the limiting behavior of the price for $S(t)$ small and for $S(t)$ large. First, the price $C(t)$ must be in the region within the AA and BB lines. Moreover, $C(t)$ is small when $S(t)$ is small, and large when $S(t)$ is large. However, $C(t)$ cannot lie outside the region within the AA line and BB lines, which implies that C gets large, by “sliding up” on the BB line.

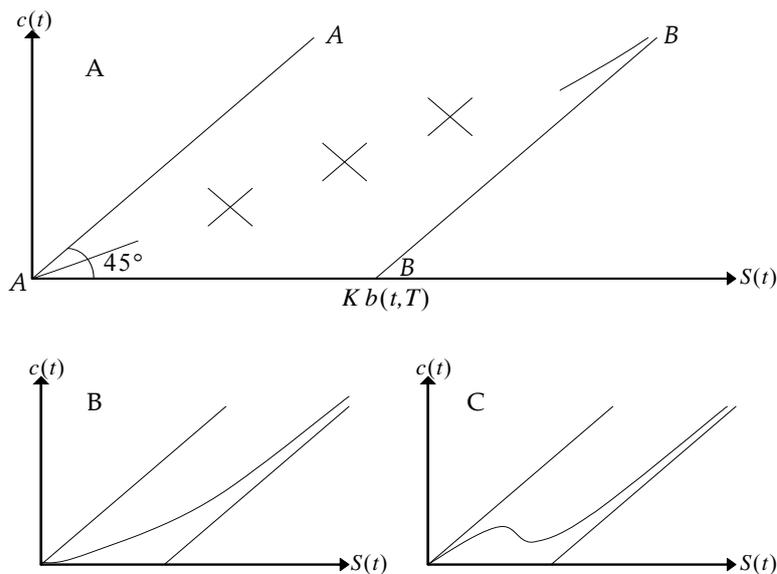


FIGURE 10.2.

How does the option price behave in the region within the AA and BB lines? We cannot tell. We may simply say that given the boundary behavior of $C(t)$, if $C(t)$ is convex in $S(t)$, it is also increasing in $S(t)$. Convexity of $C(t)$ is a reasonable property, which holds for basic diffusive models, as originally noted by Bergman, Grundy and Wiener (1996). In this case, $C(t)$ would behave as in the left-hand side of the bottom panel of Figure 10.2. This case seems to be relevant, empirically, and consistent with the predictions of the celebrated Black and Scholes (1973) formula, and some of its extensions. However, it is not a general property of option prices. Bergman, Grundy and Wiener (1996) provide several counter-examples where $C(t)$ can be decreasing over some range of $S(t)$, arising in models with jumps, or with stochastic volatility. Theoretically, we cannot rule out that the option price behaves as in the right-hand side of the bottom panel of Figure 10.2, as further developed in Section 10.5 [in progress].

The economic meaning of convexity is that the option is unlikely to be exercised when $S(t)$ is small. Therefore, changes in $S(t)$ have little effect on $C(t)$. However, the option is likely to be exercised when $S(t)$ is large. A percentage increase in S is then to be followed by an even higher percentage increase in C . In other terms, the elasticity of the option price with respect to the asset price is larger than one, $\epsilon \equiv \frac{dC}{dS} \cdot \frac{S}{C} > 1$. Mathematically, for an increasing and convex function, which is zero at the origin, it holds that the first order derivative is always higher than the secant. Therefore, option returns likely are more volatile than those on the underlying asset.

How does time-to-maturity affect the call price? Calls are known to be “wasting assets,” meaning that their value decreases over time, as illustrated by an hypothetical example in Figure 10.3, which plots the option price function relating to three maturity dates, $T_1 > T_2 > T_3$, as predicted by the Black & Scholes model.

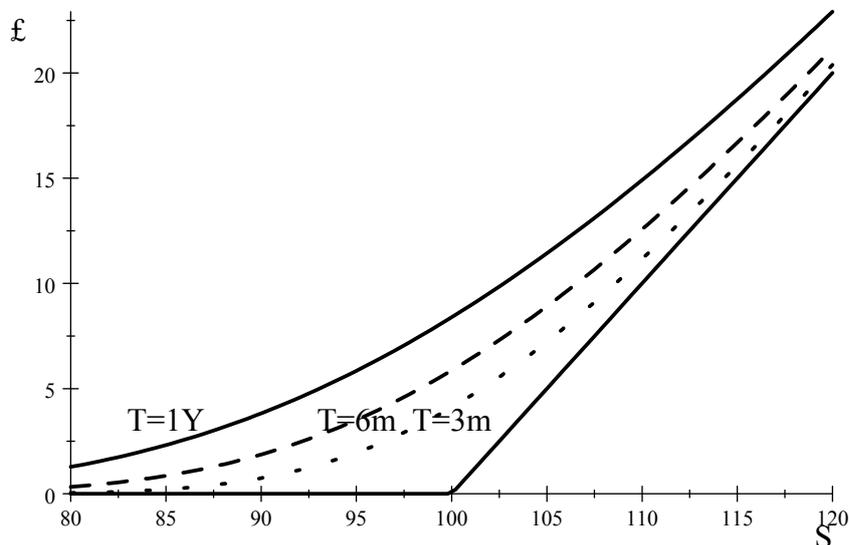


FIGURE 10.3. The value of a call option struckable at $K = 100$, as time to maturity shrinks, as predicted by the Black & Scholes model, with volatility parameter equal to 20% and short-term rate $r = 1\%$. The leftmost solid line is the price corresponding to time to maturity $T =$ one year, the dashed line is the price corresponding to time to maturity $T =$ six months, and the dotted line is the price corresponding to time to maturity $T =$ three months. The rightmost solid line is the no-arbitrage bound $(S(T) - K)^+$.

10.3.2 A case study: accumulators, decumulators

Options can be used to build up dedicated structured products, such as those relying on baskets of options. Consider, for example, an “accumulator.” Anecdotal evidence suggests that at times, accumulators might be quite popular amongst private bankers in Hong Kong, as for example during 2007, just before the market downturn. An accumulator is a portfolio which is long one call and short two or more puts. (A “decumulator” is short one accumulator.) The strike of the call is higher than the strike of the put. The rationale behind going long the call is to ensure profits are made once markets go up. Instead, puts are sold to finance the long position in the call, so as to make the value of the accumulator equal to zero at its inception. Consider, for example, Figure 10.4, where the strike of the call is $K_C = 100$. Note that since the value of the accumulator is zero at inception, the picture actually depicts net profits.

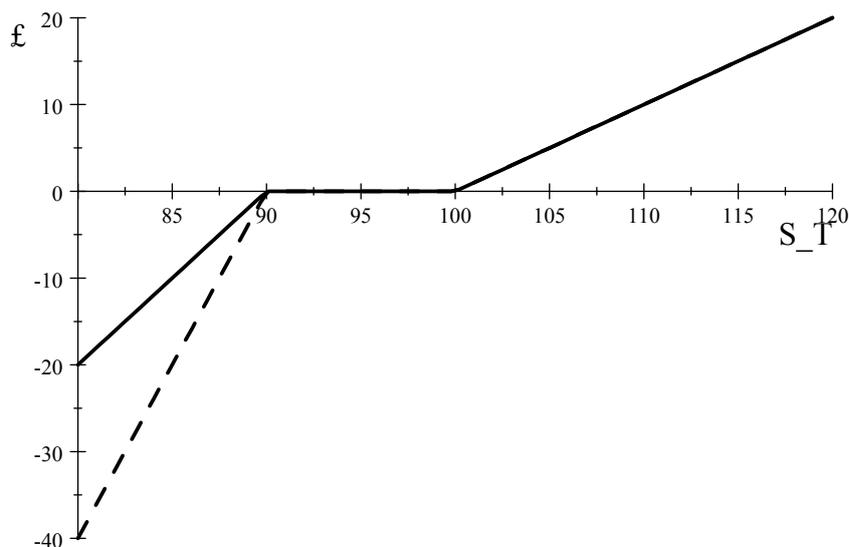


FIGURE 10.4. The solid line depicts the payoff guaranteed by an “accumulator,” a structured product that goes long one call option with strike price $K_C = 100$, and short n puts with strike price $K_p = 90$, and $n = 2$ (solid line), and $n = 4$ (dashed line).

If the current market level is $S = 102$, profits are likely to be made, at least provided the market does not fall below the strike at which the put is struck, which in this example is $K_p = 90$. However, accumulators are quite risky, in that during market downturns, they might lead to losses much more severe than the possible gains occurring in good times.

The losses that occur in bad times depend on how we choose the number of puts to sell, n , and their strike, K_p . As the previous picture reveals, losses widen as we increase the number of puts we go short. Therefore, we can decrease the probability of experiencing any losses, by just fixing $K_p = 90$, and decreasing n , although this might entail less resources left over to go long a call. A possibility might then be to go long a less expensive call, by adding a knock-out feature into the call contract (one that says that the option becomes worthless once the market reaches a certain level such as, say, 105 within the investment horizon), or by purchasing a call with a lower strike price. Alternatively, we might be willing to design a product with more risk, but also more upside, by choosing an appropriate strike for the puts. Obviously, puts become more valuable as we increase the strike price. Therefore, we could increase K_p , from 90 to 95 say, whilst keeping n constant. While selling put options with higher strikes increases the probability to have losses, it also allows us to purchase more expensive calls, those with lower strikes, which leads to a higher probability to achieve positive returns.

Naturally, the previous assumptions rely on the condition that the accumulator be self-financed, at its inception. This condition defines the type of options we can afford. Some call options can be too expensive and might require a large exposure relating to the short position in puts. For example, some calculations show that in a market with stochastic volatility such as that of Heston (1993b), we may need to sale short approximately six puts when the current index level is $S = 102$ and $K_p = 95$. Section 10.5.3 develops a case study where these risks are quantified under a variety of alternative assumptions about strikes and market volatility.

10.4 Evaluation and hedging

Investment banks sell options that they want to hedge against, to avoid the exposure to losses illustrated in Figure 10.1. Hedging is important when the only objective is to receive fees from the sale of derivatives. The portfolio that “mimics” the option price must display the properties discussed in Section 10.3.1. For example, we need to ensure that it behaves as the call price behaves in left-hand side of the bottom panel of Figure 10.2, which is the most relevant, empirically. We require this portfolio to exhibit a number of properties: (i) its value, V , should be increasing in S , which is ensured by including the asset underlying the option into the portfolio; (ii) the sensitivity of V with respect to S must be positive and bounded by one, $0 < \frac{dV}{dS} < 1$, which we can make, once the number of underlying assets is less than one; (iii) the elasticity of V with respect to S must be greater than one, $\frac{dV}{dS} \frac{S}{V} > 1$, a condition that could be met, by issuing debt. Mathematically, the value of the replicating portfolio should be $V = \theta S - D$, where θ denotes the number of the underlying assets, with $\theta \in (0, 1)$, and D is debt. In principle, this portfolio might lead these three properties to be satisfied.

In fact, hedging has a dynamic nature, because option prices obviously change over time. Therefore, we expect θ to be a function of the underlying asset price, S , and time to expiration of the option. The portfolio needs to satisfy additional properties: (iv) the number of the underlying assets must increase with S , and the value of the portfolio should be virtually insensitive to changes in S when S is low, and “slide up” through the BB line in Figure 10.3 when S is large. These conditions can be met by taking θ to be increasing in S , with $\lim_{S \rightarrow 0} \theta(S) \rightarrow 0$ and $\lim_{S \rightarrow \infty} \theta(S) \rightarrow 1$. Finally, the portfolio needs to be self-financed, as the long position in the option does not entail any additional inflows or outflows, until time to expiration: any additional purchases of the underlying asset has to be financed by issuing new debt, and any additional sale of the underlying asset must be used to shrink existing.

This section aims to make these arguments rigorous in contexts where the volatility of asset returns is stochastic. Section 10.4.1 is a quick overview of pricing issues, which are special cases of a more general framework developed in Chapter 4. Section 10.4.2 [In progress]

10.4.1 Spanning and cloning

A set of securities spans a set of payoffs, if any point in that set can be generated by a linear combination of the security prices. As explained in Chapter 4, the set of payoffs may include those promised by a contingent claim, for example, that promised by a European call, or final consumption, as in Harrison and Kreps (1979) and Duffie and Huang (1985). Chapter 4 relies on this “spanning” property and solves for consumption-portfolio choices through martingale methods. In this section, we show how spanning helps define replicating strategies with the purpose of pricing “redundant” assets. Consider the following model, where asset prices are assumed to be driven by a d -dimensional Brownian motion \mathbf{W} ,

$$d\mathbf{y}(t) = \varphi(\mathbf{y}(t)) dt + J(\mathbf{y}(t)) d\mathbf{W}(t), \quad (10.7)$$

for some vector and matrix valued functions φ and J that satisfy regularity conditions. The value of a portfolio, V , is $V(t) = \theta(t) \cdot S_+(t)$, where S_+ denotes the vector of the security prices and the money market account, and $\theta(t)$ is the vector including the units of all the assets and the money market account. A self-financed portfolio satisfies $dV(t) = \theta(t) \cdot dS_+(t)$, as we know from Section 4.3.1 of Chapter 4. Explicitly,

$$dV(t) = \left[\pi(t)^\top (\mu(t) - \mathbf{1}_m r(t)) + r(t) V(t) - C(t) \right] dt + \pi(t)^\top \sigma(t) dW(t), \quad (10.8)$$

where $\pi \equiv (\pi_1, \dots, \pi_m)^\top$, $\pi_i \equiv \theta_i S_i$, $\mu \equiv (\mu_1, \dots, \mu_m)^\top$, S_i is the price of the i -th asset, μ_i is its drift and σ is the volatility matrix of the price process. Chapter 4 utilizes the risk-neutral probability, \mathcal{Q} , to help characterize the span of the securities. This section characterizes spanning under the physical probability, P . In our context, asset prices are semimartingales under P , such as:

$$dA(t) = dF(t) + \tilde{\gamma}(t) dW(t), \quad (10.9)$$

where F is a process with finite variation, and $\tilde{\gamma}$ satisfies regularity conditions such as integrability. To replicate A , we first need to look for a portfolio π satisfying,

$$\tilde{\gamma}(t) = \pi^\top(t) \sigma(t). \quad (10.10)$$

Second, we equate the drift of V to the drift of F , obtaining,

$$\frac{dF(t)}{dt} = \pi(t)^\top (\mu(t) - \mathbf{1}_m r(t)) + r(t) V(t) = \pi(t)^\top (\mu(t) - \mathbf{1}_m r(t)) + r(t) A(t), \quad (10.11)$$

where the second equality holds because if drift and diffusion terms of F and V are identical, then $F(t) = V(t)$ —the *unique decomposition property* stated in Chapter 4. Clearly, if $m < d$, Eq. (10.10) has no solutions for π : the number of assets is so small that we cannot create a portfolio able to replicate all possible events in the future, a situation known as market incompleteness, as discussed in Chapter 4 (Definition 4.5).

Next, assume that the price of a call as of time t is $C(t, \mathbf{y}(t))$, for some function assumed to be differentiable in t , and twice differentiable in \mathbf{y} . By Itô's lemma,

$$dC = (LC) dt + (C_y \cdot J) dW,$$

where LC is the usual infinitesimal generator, C_y is the vector containing the partials of C with respect to each component of \mathbf{y} . The call price is a special case of Eq. (10.9). Therefore, we can replicate it with all the assets we have access to, once we take $C_y J = \pi^\top \sigma$.

10.4.2 Black & Scholes

Let $m = d = 1$, and assume the only state variable in Eq. (10.7) is the price of the asset underlying a call option, with $\varphi(s) = \mu s$, $J(s) = \sigma s$, where μ and σ^2 are constants—a geometric Brownian motion. Then, $C_y \cdot J = C_S \sigma S$, $\pi = C_S S$, or,

$$\theta(t) = \Delta_{\text{BS}}(t) \equiv \frac{\partial C_{\text{BS}}(S(t), T-t; K, \sigma)}{\partial S}, \quad (10.12)$$

where $C(\cdot) \equiv C_{\text{BS}}(\cdot)$ denotes the pricing formula as predicted by the Black and Scholes (1973) and Merton (1973), model, and by Eq. (10.11),

$$\frac{\partial C}{\partial t} + \frac{\partial C}{\partial S} \mu S + \frac{1}{2} \frac{\partial^2 C}{\partial S^2} \sigma^2 S^2 = \pi(\mu - r) + rC = \frac{\partial C}{\partial S} S(\mu - r) + rC, \quad (10.13)$$

subject to the boundary condition, $C(s, 0; K, \sigma) = (s - K)^+$. This derivation is similar to that in Chapter 4, although we now see it is a particular case of a more general multidimensional framework, which turns out to be useful whilst generalizing these arguments to a stochastic volatility setting, as we shall explain.

The solution to Eq. (10.13) is the celebrated Black and Scholes (1973) formula,

$$C_{\text{BS}}(S, T-t; K, \sigma) = S\Phi(d) - Ke^{-r(T-t)}\Phi(d - \sigma\sqrt{T-t}), \quad d = \frac{\ln(\frac{S}{K}) + (r + \frac{1}{2}\sigma^2)(T-t)}{\sigma\sqrt{T-t}}, \quad (10.14)$$

where Φ denotes the cumulative Normal distribution. Appendix 1 derives Eq. (10.13) hinging upon the original arguments developed by Black and Scholes (1973), and Merton (1973), where one considers a portfolio comprising the option, the underlying asset, and a portfolio strategy aiming to make the portfolio locally riskless.

Alternatively, we can use the general framework in Section 10.2.3 and arrive at Eq. (10.14). We simply set $\ell \equiv K$ in Eq. (10.3), and calculate the dynamics of the stock price under \hat{Q}_t and under Q_t , and determine $\hat{Q}(S(T) \geq K)$ and $Q(S(T) \geq K)$ in Eq. (10.3). Under Q , the stock price is the usual geometric Brownian motion with drift equal to rS , and then, $Q(S(T) \geq K) = \Phi(d - \sigma\sqrt{T-t})$, which explains the second term in Eq. (10.14). As for the first, term, we can show that the Radon-Nikodym derivative, $\frac{d\hat{Q}_t}{dQ_t} = \eta_t(T) \equiv \frac{e^{-r(T-t)}S(T)}{S(t)}$, is such that $\eta_\tau(T)$ is solution to:

$$\frac{d\eta_t(\tau)}{\eta_t(\tau)} = -(-\sigma)dW(\tau),$$

such that the stock price is solution to:

$$d \ln S(t) = \left(r + \frac{1}{2}\sigma^2\right)dt + \sigma d\hat{W}(t),$$

where \hat{W} is a Brownian motion under \hat{Q} . It easily follows that $\hat{Q}(S(T) \geq K) = \Phi(d)$, thereby completing the proof of the Black-Scholes formula.

A final comment. Eq. (10.14) holds even without requiring that a market exists for the option during the life of the option contract, or that the pricing function $C(t, S)$ is differentiable. That the option price is differentiable is a result, not an assumption. Let us define the function $C(t, S)$ that solves Eq. (10.13), with boundary condition $C(T, S) = (S - K)^+$. Note, we are not assuming this function is the option price. Rather, we shall show this is the option price. Consider a self-financed portfolio of bonds and stocks, with $\pi = C_S S$. Its value satisfies,

$$dV = [C_S S(\mu - r) + rV]dt + C_S \sigma S dW.$$

Moreover, by Itô's lemma, $C(t, s)$ is solution to

$$dC = \left(C_t + \mu S C_S + \frac{1}{2}\sigma^2 S^2 C_{SS}\right)dt + C_S \sigma S dW.$$

By subtracting the previous two equations, leaves:

$$dV - dC = \underbrace{[-C_t - rS C_S - \frac{1}{2}\sigma^2 S^2 C_{SS} + rV]}_{=-rC} dt = r(V - C) dt.$$

Hence, we have that $V(\tau) - C(\tau, S(\tau)) = [V(0) - C(0, S(0))]e^{r\tau}$, for all $\tau \in [0, T]$. Next, assume that $V(0) = C(0, S(0))$. Then, $V(\tau) = C(\tau, S(\tau))$ and $V(T) = C(T, S(T)) = (S(T) - K)^+$. That is, the portfolio $\pi = C_S S$ replicates the payoff underlying the option contract. Therefore, $V(0)$ is the value of the option at time zero, even when a market for the option does not exist over its life.

10.4.3 Surprising cancellations and “preference-free” formulae

Due to what Heston (1993a) (p. 933) quite aptly terms “a surprising cancellation,” the constant μ doesn’t show up in the final formula. Heston (1993a) shows that this property is not robust to modifications in the assumptions for the underlying asset price process. [In progress, Gamma processes, incomplete markets.]

However, even within a diffusion setting, the expected return on the option does of course depend on μ . By Eq. (10.13),

$$E\left(\frac{dC}{C}\right) = r + \underbrace{\frac{C_S}{C}\sigma S}_{\equiv \beta_S} \cdot \lambda, \quad (10.15)$$

where β_S is, simply, by Itô’s lemma, the instantaneous volatility of the option returns, and λ is the unit risk-premium related to the fluctuations of the asset price, $\lambda = (\mu - r) / \sigma$.

10.4.4 Future options and Black’s formula

Consider a *future option*, one that gives the buyer the right, not the obligation, to enter into a future contract for a specified price K , at time T , such that the payoff at time T is, $(F_S(T) - K)^+$, where $F_S(\tau)$ denotes the future price in Eq. (10.1), $F_S(\tau) = S(\tau) e^{r(S-\tau)}$. It is easy to see that $F_S(\tau)$ is martingale under the risk-neutral probability Q . For example, assuming that S is a Geometric Brownian motion volatility σ , we have that,

$$\frac{dF_S(\tau)}{F_S(\tau)} = \sigma d\tilde{W}(\tau), \quad \tau \in (t, S), \quad (10.16)$$

where \tilde{W} is a Brownian motion under Q . Therefore, the price of a future option as of time t is,

$$C_F(F_S(t), T - t; K) \equiv e^{-r(T-t)} \mathbb{E}_t(F_S(T) - K)^+ = e^{-r(T-t)} \left(F_S(t) \Phi(d) - K \Phi(d - \sigma\sqrt{T-t}) \right), \quad (10.17)$$

where

$$d = \frac{\ln\left(\frac{F_S(t)}{K}\right) + \frac{1}{2}\sigma^2(T-t)}{\sigma\sqrt{T-t}},$$

and the second equality of Eq. (10.17) follows by the Black & Scholes formula, Eq. (10.14). Eq. (10.17) is the celebrated Black’s (1976a) formula, which turns out to be very useful in the context of fixed income security pricing, as explained in Chapter 12. Appendix 2 provides an alternative derivation of Eq. (10.17), based on the pricing approach of Section 10.2, and the slightly more general assumption that the volatility of the future price in Eq. (10.16) is time-varying—but deterministic.

Chapter 12 explains that the property that future prices are martingales under the risk-neutral probability generalizes to one holding when interest rates are time-varying, under a certain probability called *forward probability* (see Chapter 12, Section 12.2).

10.4.5 Hedging

The “cloning” arguments suggest themselves as a mechanism to replicate a derivative by trading the asset underlying the derivative contract. Why do derivatives need to be replicated, in practice? Because most of them are dealt with by investment banks, which simply act as financial intermediaries, trading derivatives on behalf of third parties, being compensated through

fees. Suppose, for instance, an investment bank receives an order to sell a put. The bank would want to hedge against this put, by creating a replicating portfolio such that the value of this portfolio be the same as the final payoff the investment bank has to pay its buyer to honour its sale. So hedging is needed to replicate the final payoffs required to honour the contracts giving rise to these payoffs. Which hedge should we implement, in practice? Standard market practice is to use the “Black-Scholes delta” in Eq. (10.12). Calculating the derivative in Eq. (10.12) is easy. Note that the Black-Scholes formula is homogenous of degree one in S and K , that is, $C_{\text{BS}}(\lambda S, T - t; \lambda K, \sigma) = \lambda C_{\text{BS}}(S, T - t; K, \sigma)$ for any constant λ . Therefore, we have that by Euler’s theorem,

$$C_{\text{BS}}(S, T - t; K, \sigma) = \frac{C_{\text{BS}}(S, T - t; K, \sigma)}{\partial S} S + \frac{\partial C_{\text{BS}}(S, T - t; K, \sigma)}{\partial K} K, \quad (10.18)$$

and comparing this expression with the Black-Scholes formula in Eq. (10.14), produces,

$$\Delta^{\text{BS}}(t) = \Phi(d). \quad (10.19)$$

Naturally, investment banks, not to mention funds, can undertake speculative trading activities aimed to implement specific views, such as those described in Section 10.5.5 below, in which case hedging doesn’t necessarily need to be implemented. However, even in this case, hedging might be required to isolate the particular views a trading desk of the bank is taking. For example, Section 10.5.5 will explain that to express the view that equity volatility will raise, say, we cannot simply go long call options, because call prices are increasing both in volatility and the price underlying the option. A better solution is to go long an option, delta-hedged through Black-Scholes, as we shall explain.

10.4.6 Endogenous volatility

[In progress]

Hedges and crashes. Delta-hedging can lead to financial turmoil. The 1987 crash, and the conclusion of the Brady commission. The theory in the Brady’s report is that the market was initially hit by bad news, and fell, triggering further sell-offs originating from program trading. Uninformed investors, who did not fully understand the nature of these developments, exited the market, and the fall they created would trigger additional program trades until the market could not reach the usual equilibrium where it used to stay, switching to another equilibrium through a discrete change—an equity market crash of more than -22% in just one day, the largest of financial history.

The flash crash of May 6th, 2010 might be considered a modern prototype version of the 1987 crash. According to the explanation put forward in a joint report by the SEC and the CFTC (SEC-CFTC, 2010), a mutual fund sold a quite large number of E-mini futures on the S&P 500 for mere hedging motives, but high frequency trading firms would then initiate arbitrage adjustments, by going long them and simultaneously shorting the equities representing the SPY—the ETF that tracks the S&P 500 index. To unwind these positions would normally require traders with fundamental views, but the majority of the players over those few minutes of trading were high frequency trading firms, who would then transact with each other over a self-exciting process, where trading aggressiveness intensifies as volume increases, due to the nature of algorithmic trading. In the end, large volume and sell-offs reinforced each other, whereby increasing volumes triggered algorithms to sell into a falling market, leading to a crash of nearly -10% in the equity market over a few minutes.

[Give a short outline of this section]

10.4.6.1 Hedges

Gamma is always positive for long calls and puts, as these contracts have positive convexity, as illustrated by Figure 10.1. Naturally, short calls and puts have negative gamma. In order for the statement “when gamma is negative, delta hedging involves buying on the way up and selling on the way down” to be true, we also have to consider whether the delta is positive or not (that is, whether the derivative price is increasing or decreasing in the underlying asset price). So we have four instances of hedging portfolios:

- (i) Positive gamma: Buying on the way up and selling on the way down.
 - (i.1) Hedging portfolios with positive delta, as required, for example, to hedge against the *sale* of a call. Positive delta means that the hedging portfolio relies on *buying* the assets underlying the call. When the price of these assets are up, the delta is also up, which implies we need to keep on buying even more of the assets underlying the hedging portfolio. On the other hand, when prices are down, the delta is also down, which implies holding less of the assets underlying the hedging portfolio, thereby leading to sell some these assets precisely when the market is down.
 - (i.2) Hedging portfolios with negative delta, as required, for example, to hedge against the *sale* of a put. Negative delta means that the hedging portfolio relies on *selling* the assets underlying the put. In this case, delta is up when when prices are up. However, this now simply means that we need to sell less! For example delta might have been -2 before the market was up and now delta is -1 : that is, we need to buy back some of the assets underlying the hedging portfolio. When, instead, prices are down, delta is also down, which means we need to sell even more into a depressed market.
- (ii) Negative gamma: Buying on the way down and selling on the way up.
 - (ii.1) Hedging portfolios with positive delta, as required, for example, to hedge against having gone *long* a put. Positive delta means that the hedging portfolio relies on *buying* assets underlying the put. Negative gamma now means that as soon as the price of these asset goes up (resp. down), we need to buy less (resp. buy more), so we sell when prices go up and buy when prices go down.
 - (ii.2) Hedging portfolios with positive delta, as required, for example, to hedge against having gone *long* a call. We are now *selling* the assets underlying the call. Negative gamma, here, means that as the price of these assets goes up (resp. down), we need to sell more (resp. sell less), so once again, we sell when prices go up and buy when prices go down.

How to implement these hedging portfolios, in practice, is still an open question, as this issue is necessarily model-based. Section 10.5.4, for example, shows that delta hedging under the Black-Scholes assumptions would lead the bank to eliminate the risk of fluctuations in the underlying stock price. At the same time, however, hedging through Black-Scholes leads the derivatives book quite messy once the fundamental assumption underlying the Black-Scholes world does not hold, namely that volatility changes randomly. In this case, “hedging” would rather look like a volatility view. To appropriately hedge, one has to rely on more complicated hedging strategies. For example, to hedge against an option in a world of stochastic volatility, we would need to use a stock, a bond, and, another ... option!

10.4.6.2 Crashes

[In progress]

Use a simple model, by Grossman, to illustrate how volatility is pumped-up by automatic mechanisms. Then, discuss a streamlined version of Gennotte and Leland, with asymmetric information, to illustrate the 1987 crash.

10.4.7 Marking to market

Consider a derivative, which we go long at time $t = 0$, when it is worthless. As time unfolds, its value will change, which calls for marking to market it. Suppose the derivative pays off $\psi(S(T)) - K(0)$ at time T , where $S(T)$ is the price of some asset as of time T , and $K(0)$ is set so as to make the derivative worthless at time zero. Assuming that interest rates are constant, we have that $K(0) : e^{-rT} \mathbb{E}_0[\psi(S(T)) - K(0)] = 0$, where \mathbb{E}_0 is the expectation at time $t = 0$, taken under the risk-neutral probability. That is, $K(0) = \mathbb{E}_0[\psi(S(T))]$. The market value of the derivative at time t , say $\text{MtM}(t)$, is simply the present value of the expected payoff at T , under the risk-neutral probability, $e^{-r(T-t)} \mathbb{E}_t[\psi(S(T)) - K(0)]$, or

$$\text{MtM}(t) = e^{-r(T-t)} [K(t) - K(0)]. \quad (10.20)$$

For more elaborated payoffs, such as those depending on the realizations of the underlying risks over the life of the contract, the market to market updates might be more intricate than that in Eq. (10.20), as the case of the variance contracts illustrated in Section 10.7.3.

10.4.8 Properties of options in diffusive models

10.4.8.1 The comparative statics of dynamic models

We derive general properties of option prices arising in the context of diffusion processes, by hinging upon results and methods originally set forth by Bergman, Grundy and Wiener (1996), as well as the general framework in Chapter 7, as summarized by Proposition 7.1. We shall show that the price of European-style options inherits properties of the final payoff: monotonicity and convexity of the final payoff propagate through monotonicity and convexity of the option price. Convexity of the call price is, then, a condition we need to establish that the option price is increasing in the volatility of the underlying asset price.

We assume that the stock price is solution to the following stochastic differential equation:

$$\frac{dS(t)}{S(t)} = \mu(S(t)) dt + \sqrt{2v(S(t))} dW(t), \quad (10.21)$$

and let $C(S(t), t, T)$ be the price of a European-style option at time t , with a payoff function $\psi(S) > 0$ to be delivered at time T . We assume that $\psi(S)$ is differentiable, with $\psi'(S) > 0$. In the absence of arbitrage, C satisfies the following partial differential equation,

$$0 = C_t + C_S r S + C_{SS} S^2 v(S) - r C, \quad (10.22)$$

subject to the boundary condition, $C(S, T, T) = \psi(S)$. Let us differentiate Eq. (10.22) with respect to S . The result is that $H \equiv C_S$ satisfies another partial differential equation,

$$0 = H_t + H_S \left(r S + \frac{\partial}{\partial S} (S^2 v(S)) \right) + H_{SS} S^2 v(S), \quad (10.23)$$

subject to the boundary condition $H(S, \tau, T) = \psi'(S) > 0$. Therefore, we have that $H(S, \tau, T) = C_S(S, \tau, T) > 0$ for all realizations of the underlying asset price, by results reviewed in Proposition 1 and Appendix 1 of Chapter 7. That is, in a scalar diffusion setting, a European-style option price is increasing in the underlying asset price whenever the final payoff $\psi(S)$ is increasing.

Next, we consider the thought experiment to tilt the volatility of the underlying asset price. Consider two markets A and B with prices $(C^i, S^i)_{i=A,B}$, where the volatility of the asset price in market A is larger than that in market B, viz

$$\frac{dS^i(\tau)}{S^i(\tau)} = r d\tau + \sqrt{2v^i(S^i(\tau))} d\hat{W}(\tau), \quad i = A, B,$$

where \hat{W} is Brownian motion under the risk-neutral probability, and $v^A(s) > v^B(s)$ for all s . It is easy to see that the price difference in the two markets, $\Delta C \equiv C^A - C^B$, satisfies,

$$0 = \Delta C_t + \Delta C_S r S + \Delta C_{SS} S^2 v^A - r \Delta C + (v^A - v^B) C_{SS}^B S^2. \quad (10.24)$$

By the same results used to analyze Eq. (10.23), we now have that $\Delta C > 0$ whenever $C_{SS} > 0$. That is, if option prices are convex in the price of the underlying asset, they are increasing in the volatility of the asset price. As we say, volatility changes constitute mean-preserving spreads. [Develop this notion, and relate it to the Appendix of Chapter 7 and mention it doesn't work when it comes to analyze fixed income]. We wish to find conditions sufficient to guarantee that $C_{SS} > 0$. We differentiate Eq. (10.23) with respect to S , and $Z \equiv H_S = C_{SS}$ satisfies the following partial differential equation,

$$0 = Z_t + \left(rS + 2 \frac{\partial}{\partial S} (S^2 v(S)) \right) Z_S + Z_{SS} S^2 v(S) - \left(-r - \frac{\partial^2}{\partial S^2} (S^2 v(S)) \right) Z, \quad (10.25)$$

subject to the boundary condition $Z(S, \tau, T) = \psi''(S)$. By the usual reasoning, we have that if the final payoff is convex, $\psi''(S) > 0$, $H(S, \tau, T) = C_{SS}(S, \tau, T) > 0$, for all realizations of the underlying asset price. That is, in a scalar diffusion setting, a European-style option price is convex in the underlying asset price, provided the final payoff is convex. Therefore, in a scalar diffusion setting, a European-style option price is increasing in the volatility of the underlying asset price, provided the final payoff delivered by the option is convex.

10.4.8.2 Passage of time

Sometimes, we claim that options are “wasting” assets, in that their value decreases over time, due to a decrease in the value of the optionality. For call options, this is definitely true, at least within diffusive models. By the first equation in (10.22), we have:

$$C_t = -rC(\epsilon - 1) - C_{SS}v(S) < 0, \quad (10.26)$$

where $\epsilon \equiv \frac{S}{C} C_S$ is the elasticity of the option price with respect to the asset price, which for a call option, is larger than one, as noted in Section 10.3. However, for a put option, this elasticity is negative, and can make the right hand side of Eq. (10.26) change sign, especially

for far out-of-the-money options, as Figure 10.5 reveals.

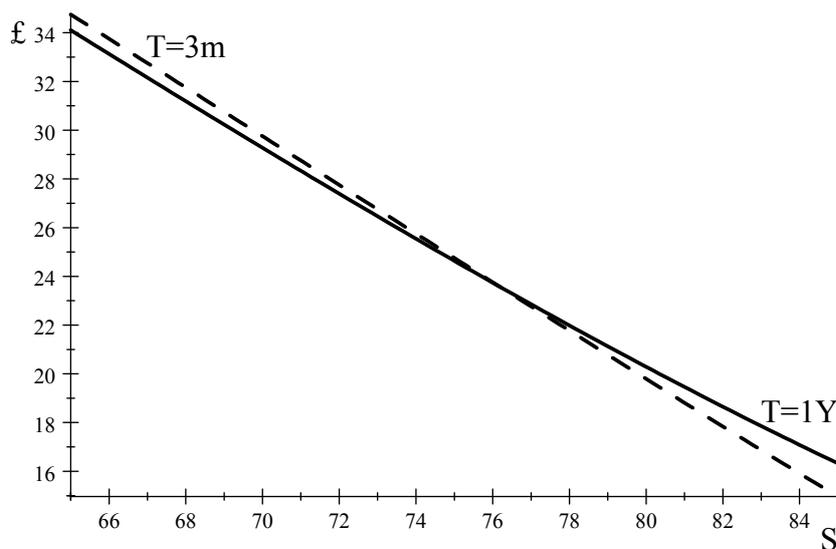


FIGURE 10.5. The value of a put option with strike $K = 100$, as time to maturity shrinks, as predicted by the Black & Scholes model, with volatility parameter equal to 20% and short-term rate $r = 1\%$. The solid line is the price corresponding to time to maturity $T =$ one year, and the dashed line is the price corresponding to time to maturity $T =$ three months.

10.4.8.3 Counterexamples

[In progress]

10.4.8.4 Recovering risk-neutral probabilities

Consider the price of a European call,

$$C(S(t), t, T; K) = P(t, T) \int_K^\infty (x - K) q(x | S(t)) dx,$$

where Q is the risk-neutral probability and $q(x^+ | x) dx \equiv dQ(x^+ | x)$. Assuming that $\lim_{x \rightarrow \infty} xq(x | S) = 0$, and differentiating with respect to K leaves:

$$e^{r(T-t)} \frac{\partial C(S(t), t, T; K)}{\partial K} = - \int_K^\infty q(x | S(t)) dx.$$

We can check this relation holds true in the Black-Scholes model, in Eq. (10.18). Let us differentiate again,

$$e^{r(T-t)} \frac{\partial^2 C(S(t), t, T; K)}{\partial K^2} = q(K | S(t)). \quad (10.27)$$

Eq. (10.27) allows us to recover the risk-neutral density using option prices. The Arrow-Debreu state density, $\mathcal{D}_{AD}(S^+ = u | S(t))$, is given by,

$$\mathcal{D}_{AD}(S^+ = u | S(t)) = e^{r(T-t)} q(S^+ | S(t)) \Big|_{S^+=u} = e^{2r(T-t)} \frac{\partial^2 C(S(t), t, T; K)}{\partial K^2} \Big|_{K=u}.$$

These results are quite useful in applied work. They also help deal with the pricing of volatility contracts reviewed in Section 10.6, as explained in Appendix 4.

10.5 Stochastic volatility

10.5.1 Statistical models of changing volatility

10.5.1.1 ARCH and random variance models

Asset returns exhibit both temporal dependence in their second order moments and heavy-peaked and tailed distributions, as reviewed in Chapter 7. While this feature was known since Mandelbrot (1963) and Fama (1965), its study was formalized only through the introduction of the so called Auto Regressive Conditionally Heteroscedastic (ARCH) model of Engle (1982) and Bollerslev (1986). An ARCH model works as follows. Let $\{y_t\}_{t=1}^N$ be a record of observations on some asset returns, $y_t = \ln S_t/S_{t-1}$, where S_t is the asset price. The empirical evidence suggests that a good model is:

$$y_t = a + \epsilon_t, \quad \epsilon_t | F_{t-1} \sim N(0, \sigma_t^2), \quad \sigma_t^2 = w + \alpha \epsilon_{t-1}^2 + \beta \sigma_{t-1}^2, \quad (10.28)$$

where a , w , α and β are parameters and F_t denotes the information set as of time t . This model is known as the GARCH(1,1) model (Generalized ARCH). It was introduced by Bollerslev (1986), and collapses to the ARCH(1) model introduced by Engle (1982) once we set $\beta = 0$. ARCH models have played a prominent role in the analysis of many aspects of financial econometrics, such as the term structure of interest rates, the pricing of options, or the presence of time varying risk premiums in the foreign exchange market, as summarized by the classic survey of Bollerslev, Engle and Nelson (1994).

The quintessence of ARCH models is to make volatility dependent on the variability of past observations. An alternative formulation, initiated by Taylor (1986), makes volatility driven by some unobserved components. This formulation gives rise to the *stochastic volatility* model. Consider, for example, the following stochastic volatility model,

$$\begin{aligned} y_t &= a + \epsilon_t, & \epsilon_t | F_{t-1} &\sim N(0, \sigma_t^2); \\ \ln \sigma_t^2 &= w + \alpha \ln \epsilon_{t-1}^2 + \beta \ln \sigma_{t-1}^2 + \eta_t; & \eta_t | F_{t-1} &\sim N(0, \sigma_\eta^2) \end{aligned}$$

where a , w , α , β and σ_η^2 are parameters. The main difference between this model and the GARCH(1,1) model in Eq. (10.28) is that the volatility as of time t , σ_t^2 , is not predetermined by the past forecast error, ϵ_{t-1} . Rather, this volatility depends on the realization of the stochastic volatility shock η_t at time t . This makes the stochastic volatility model considerably richer than a simple ARCH model. As for the ARCH models, SV models have also been intensively used, especially following the progress accomplished in the corresponding estimation techniques. The seminal contributions related to the estimation of this kind of models are mentioned in Mele and Fornari (2000). Early contributions that relate changes in volatility of asset returns to economic intuition include Clark (1973) and Tauchen and Pitts (1983), who assume that a stochastic process of information arrival generates a random number of intraday changes of the asset price.

10.5.1.2 ARCH and diffusive models

Under regularity conditions, ARCH models and stochastic volatility models behave essentially the same as the sampling frequency gets sufficiently high. Precisely, Nelson (1990) shows that ARCH models converge in distribution to the solution of the stochastic differential equations, in the sense that the finite-dimensional distributions of the volatility process generated by ARCH models converge towards the finite-dimensional distributions of some diffusion process, as the

sampling frequency goes to infinity. Mele and Fornari (2000) (Chapter 2) contain a review of results relating to this type of convergence, and Corradi (2000) develops a critique related to the conditions underlying these convergence results. To illustrate, heuristically, consider the following model,

$$\begin{cases} d \ln S(\tau) &= (\dots) d\tau + \sigma(\tau) dW(\tau) \\ d\sigma^2(\tau) &= (\omega - \varphi\sigma^2(\tau)) d\tau + \psi\sigma^2(\tau) dW_\sigma(\tau) \end{cases} \quad (10.29)$$

where $dW(\tau)$ and $dW_\sigma(\tau)$ are correlated, with correlation ρ , and ω , φ , and ψ are some constants. Consider, further, the ARCH model:

$$\begin{cases} y_{\Delta,n+1} &= \epsilon_{\Delta,n+1}, \quad \epsilon_{\Delta,n} = \sigma_{\Delta,n} \cdot u_{\Delta,n}, \quad u_{\Delta,n} \sim \text{NID}(0, 1) \\ \sigma_{\Delta,n+1}^2 &= w_\Delta + \alpha_\Delta (|\epsilon_{\Delta,n}| - \gamma\epsilon_{\Delta,n})^2 + \beta_\Delta \sigma_{\Delta,n}^2 \end{cases} \quad (10.30)$$

where $y_{\Delta,n+1} = \ln(S_{\Delta,n+1}/S_{\Delta,n}) - E(\ln(S_{\Delta,n+1}/S_{\Delta,n}))$, n and Δ refer to the indexing of observed data and the sampling frequency (weekly, say), and w_Δ , α_Δ , β_Δ are positive parameters, possibly depending on the sampling frequency, and $\gamma \in (-1, 1)$. The parameter γ allows to capture the Black-Christie-Nelson leverage effect (Black, 1976b; Christie, 1982; Nelson, 1991) discussed in Chapter 8. Note that the second of Eqs. (10.30) can be written as:

$$\begin{aligned} \sigma_{\Delta,n+1}^2 - \sigma_{\Delta,n}^2 &= [\Delta t^{-1} w_\Delta - \Delta t^{-1} (1 - \alpha_\Delta E(|u_{\Delta,n}| - \gamma u_{\Delta,n})^2 - \beta_\Delta) \sigma_{\Delta,n}^2] \Delta t \\ &\quad + \Delta t^{-\frac{1}{2}} \alpha_\Delta \cdot \sigma_{\Delta,n}^2 \sqrt{\Delta t} \cdot \Delta W_{\sigma,n}, \end{aligned} \quad (10.31)$$

and $\Delta W_{\sigma,n} \equiv (|u_{\Delta,n}| - \gamma u_{\Delta,n})^2 - E(|u_{\Delta,n}| - \gamma u_{\Delta,n})^2$. The first two terms define the drift term for the variance process, and the last term is the diffusive component. Suppose that $\lim_{\Delta t \downarrow 0} \Delta t^{-1} w_\Delta = \omega$, $\lim_{\Delta t \downarrow 0} \Delta t^{-1} (1 - \alpha_\Delta E(|u_{\Delta,n}| - \gamma u_{\Delta,n})^2 - \beta_\Delta) = \varphi$, and, finally, $\lim_{\Delta t \downarrow 0} \Delta t^{-1/2} \sqrt{\varpi} \alpha_\Delta = \psi < \infty$, where $\varpi \equiv \text{var}(\Delta W_{2,n})$. Then, under regularity conditions, the sample paths of S and σ^2 in Eqs. (10.30) converge to those of S and σ^2 in Eqs. (10.29), with a well-defined correlation coefficient ρ (see Fornari and Mele, 2006).¹

10.5.2 Implied volatility, smiles and skews

Parallel to time-series research into asset volatilities reviewed in the previous section, research on option prices over the 1980s considerably challenged the assumption of a constant volatility in the Black & Scholes and Merton model. As we know, the Black & Scholes model relies on the assumption that price of the asset underlying the option is a geometric Brownian motion with constant volatility,

$$\frac{dS(t)}{S(t)} = \mu dt + \sigma dW(t),$$

where W is a Brownian motion, and μ , σ are constants. As we also know, σ is the only parameter to enter the option pricing formula. A crucial point is that not only is the assumption of a constant σ inconsistent with the time-series evidence reviewed in the previous section. It is also inconsistent with empirical evidence on the *cross-section* of option prices. Let $C_t^\$(K, T)$ denote the time t market price of a call option struckable at K at time T , and consider the call option

¹For example, if $\gamma = 0$, the random component of the diffusive term in Eq. (10.31) collapses to $\Delta W_{\sigma,n} = u_{\Delta,n}^2 - E(u_{\Delta,n}^2)$, and the moment condition for the diffusive component is $\psi = \lim_{\Delta t \downarrow 0} \Delta t^{-1/2} \sqrt{2} \alpha_\Delta$. Intuitively, in this case ΔW_σ is an IID sequence of centered chi-square variates with one degree of freedom (and variance $\varpi = 2$), and stands for the discrete version of the Brownian motion increments dW_σ in the second of Eqs. (10.29).

price predicted by the Black-Scholes formula, $C_{\text{BS}}(S(t), T-t; K, \sigma)$ in Eq. (10.14). Define the Black-Scholes *implied volatility* as the value of σ that equates the Black-Scholes formula to the market price of the option, IV say,

$$\text{IV} : C_t^{\$}(K, T) = C_{\text{BS}}(S(t), T-t; K, \text{IV}). \quad (10.32)$$

By results given in Section 10.4.7, we know the Black-Scholes option price is strictly increasing in σ . Therefore, this definition of implied volatility makes sense, in that there exists a unique value for IV such that Eq. (10.32) holds true. In fact, the market practice is to quote options in terms of implied volatilities, not prices. Moreover, this same implied volatility relates to both the call and the put option prices, for the following reason. Consider the put-call parity in Eq. (10.5),

$$P_t(K, T) = C_t(K, T) - S(t) + Ke^{-r(T-t)}.$$

Naturally, this same equation must necessarily hold for the Black-Scholes model for each σ , i.e. $P_{\text{BS}}(S(t), T-t; K, \sigma) = C_{\text{BS}}(S(t), T-t; K, \sigma) - S(t) + Ke^{-r(T-t)}$. Subtracting this equation from the previous one, we see that, the implied volatilities for a call and for a put options are the same.

If the Black & Scholes model holds, implied volatilities would be the same for each K . Yet empirically, and at least since 1987, the cross section of implied volatilities exhibit striking characteristics, when gauged against the “moneyness of the option” defined as,

$$mo \equiv \frac{S(t)e^{r(T-t)}}{K}. \quad (10.33)$$

Prior to 1987, the pattern of implied volatilities was unclear or U-shaped in $\frac{1}{mo}$ at best—a “smile.” After the 1987 crash, the smile pattern turned into a “smirk,” also referred to as “volatility skew.” Figure 10.6 illustrates the prediction about smile and smirks emanating from a model where volatility is random, discussed at length over the next sections. But, is random volatility the only explanation to smirks? In fact, additional explanations might relate to the very fact that options (be they call or puts) that are deep-in-the-money and options (be they call or puts) that are deep-out-of-the money are relatively less liquid and, therefore, command a liquidity risk-premium. Since the Black-Scholes option price is increasing in volatility, the implied volatility is U-shaped in $\frac{1}{mo}$.

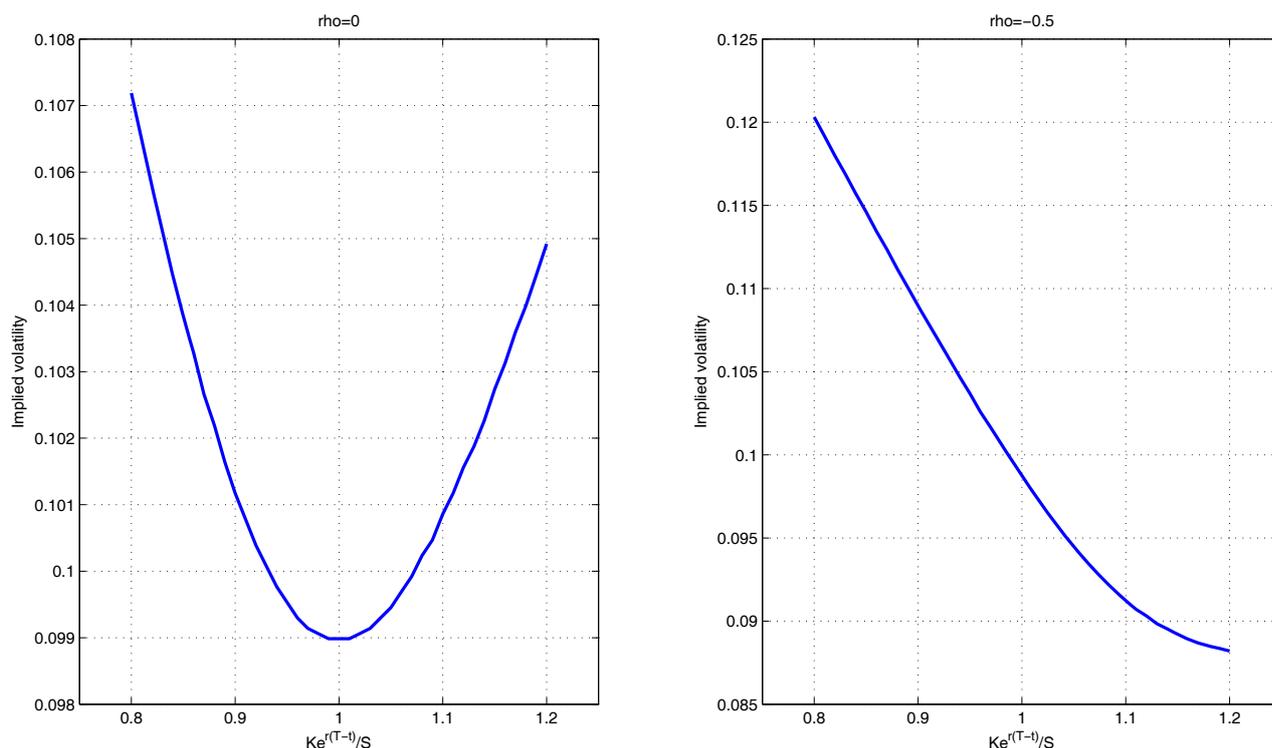


FIGURE 10.6. Smile and smirk predicted by the Heston model in Eq. (10.46), with parameters fixed at $\kappa = 2$, $\omega = 0.01$, $\xi = 0.1$ and, for the left-hand panel (the smile) $\rho = 0$, and the right-hand panel (the smirk) $\rho = -0.5$. The initial values of the asset price and volatility are $S = 100$, and $\sqrt{v(t)} = \sqrt{\omega}$, and the short-term rate $r = 0$, and the maturity of the option is six months.

Explanations of the skews relating to statistical properties of the underlying assets are easy to follow. The Black & Scholes model relies on the assumption asset returns are log-normally distributed. But this assumption may not be correct, as the market might be pricing through alternative distributions. These alternative distributions might put more weight on the tails, as a result of the market fears about the occurrence of extreme outcomes. For example, the market might fear the stock price will decrease under a certain level, say \underline{K} , more than the Black & Scholes model would predict. As a result, the market density should have a left tail thicker than that of the log-normal, for values of $S < \underline{K}$. This possibility is illustrated by the left panel of Figure 10.7, which depicts the risk-neutral distributions of both the Black & Scholes model, and one model with random volatility—a model that does generate thick tails, as discussed below. A market density with a left tail thicker than that of Black & Scholes implies that the probability deep-out-of-the-money *puts* (i.e., those with low strike prices) will be exercised is higher under the market density than under the log-normal. In other words, the volatility needed to price deep-out-of-the-money *puts* is larger than that needed to price at-the-money calls and puts.

At the other extreme, the market might assign a higher likelihood that the stock price will be above some \bar{K} than predicted by the Black & Scholes, which would translate into a market

density with a right tail thicker than the log-normal, for values of $S > \bar{K}$. This characteristic implies a larger probability that deep-out-of-the-money *calls* (i.e., those with high strike prices) will be exercised, compared to the log-normal. As a result, the implied volatility needed to price deep-out-of-the-money *calls* is larger than that needed to price at-the-money calls and puts, as illustrated by the left panel of Figure 10.7. As explained, the second effect has disappeared since the 1987 crash, leaving the “smirk” of the right panel of Figure 10.7.

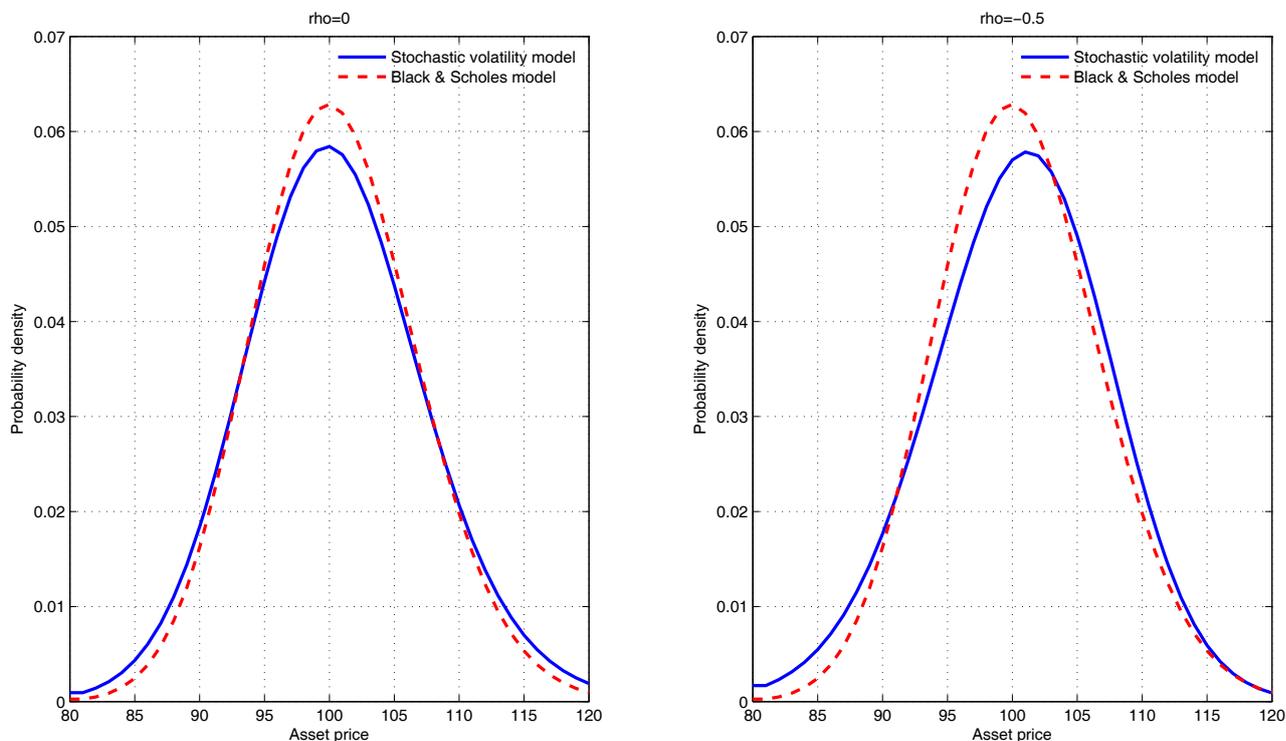


FIGURE 10.7. Risk-neutral densities predicted by the Black & Scholes model (dashed line) and the Heston model in Eq. (10.46). The Black & Scholes volatility parameter is $\sigma = 9\%$, and Heston’s parameters are fixed at $\kappa = 2$, $\omega = 0.01$, $\xi = 0.1$ and, for the left-hand panel $\rho = 0$, and the right-hand panel $\rho = -0.5$. The initial values of the asset price and volatility are $S = 100$, and $\sqrt{v(t)} = \sqrt{\omega}$, the short-term rate $r = 0$, and maturity is six months.

The results depicted in Figures 10.6 and 10.7 relate to the celebrated Heston’s (1993) model where volatility is random, and is elaborated in more detail in Section 10.5.4 (see Eq. (10.46)).² Although these figures and the previous explanations are suggestive, smiles have been rationalized by the presence of stochastic volatility during the early 1990s (Ball and Roma, 1994; Renault and Touzi, 1996). To illustrate, consider the continuous time model,

$$\begin{cases} \frac{dS(t)}{S(t)} = \mu dt + \sigma(t) dW(t) \\ d\sigma^2(t) = b(S(t), \sigma(t))dt + a(S(t), \sigma(t))dW_\sigma(t) \end{cases} \quad (10.34)$$

²The densities in Figure 10.6 are those of $1 - \Phi(d - \sigma\sqrt{T-t})$ in Eq. (10.14) for Black & Scholes, and of $1 - P_2(\ln S, \sigma^2, t)$ in Eq. (10.50) for Heston—both densities are with respect to K .

where W_σ is Brownian motion correlated with W , with instantaneous correlation equal to ρ , and b and a are some functions satisfying regularity conditions. The drift function, b , is needed to generate mean-reverting behavior in stochastic volatility, $\sigma^2(t)$, a characteristic we require in exactly the same spirit of what we need to assume from interest rates, as further elaborated in Chapter 12.

The option price is given by, $C(S(t), \sigma^2(t), T-t) = e^{-r(T-t)} \mathbb{E}_Q [(S(T) - K)^+ | S(t), \sigma^2(t)]$, where $\mathbb{E}_Q[\cdot]$ is the expectation taken under some risk-neutral probability Q . Then, the implied volatility from this model,

$$\text{IV} : C(S(t), \sigma^2(t), T-t) = C_{\text{BS}}(S(t), T-t; K, \text{IV}),$$

is U-shaped with respect to $\frac{1}{mo}$ whenever the correlation $\rho = 0$. Indeed, let $X = \ln(mo)$, where mo denotes the option moneyness, as defined in Eq. (10.33). In Appendix 3, we show that it holds, approximately, that in a model with stochastic volatility and zero correlation, IV is, approximately, a quadratic function of X , $\text{IV}(X, T-t)$ say, with a minimum occurring at $X = 0$, just as in the left panel of Figure 10.6, viz

$$\text{IV}(X, T-t) \approx \mu_V(t) + \frac{1}{2} \left(\frac{X^2 - \left(\frac{1}{2}\mu_V^2(t)(T-t)\right)^2}{\mu_V^3(t)(T-t)} \right) \cdot \text{var}_t(\sqrt{\tilde{V}_{t,T}}), \quad (10.35)$$

where $\mu_V(t) = \mathbb{E}_t(\sqrt{\tilde{V}_{t,T}})$.

Intuitively, we expect these interesting properties to hold because of a compelling lesson we learnt from both the early literature on random variance and the ARCH literature: random changes in volatility lead to a return distribution with tails thicker than the normal—one with kurtosis larger than three (Mandelbrot, 1963; Fama, 1965; Nelson, 1990; Mele and Fornari, 2000), as illustrated numerically by the Heston's model in Figure 10.7. For example, we know from Nelson (1990), that even if unexpected returns are conditionally normally distributed, they are approximately, and unconditionally, Student's t , once we assume their variance follows a GARCH(1,1) process. Mathematically, denote the unexpected returns as of time t with ϵ_t , and suppose that $\epsilon_t = z_t \sigma_t$, where $z_t \sim \text{NID}(0, 1)$ and σ_t , the conditional volatility of ϵ_t , is some random process. Then we have, by Jensen's inequality, that $E(\epsilon_t^4) = E(z_t^4) E(\sigma_t^4) \geq E(z_t^4) [E(\sigma_t^2)]^2 = E(z_t^4) [E(\epsilon_t^2)]^2$, which is an equality when σ_t is not random. It follows that the kurtosis, $\text{Kurt} \equiv \frac{E(\epsilon_t^4)}{[E(\epsilon_t^2)]^2} \geq E(z_t^4) = 3$. That is, random volatility makes the unconditional return density leptokurtotic even when the conditional is normal. Although these calculations relate to unconditional densities, similar conclusions would apply to conditional: random volatility makes a T -day conditional density leptokurtotic even when the one-day conditional is normal. As a result of this leptokurtoticity, the probability out-of-the money options is exercised is larger than that implied by the log-normal distribution—the smile effect. As for the smirk effect, we need $\rho < 0$, as shown by Figure 10.7. Intuitively, when $\rho < 0$, the left tail of the return distribution is thicker than the right, thereby making out-of-the money puts most valuable.

The model in Eqs. (10.34) has been extended to one with jumps, where the variance process follows a mean-reverting process such as:

$$d\sigma^2(t) = \kappa(\omega - \sigma^2(t)) dt + \psi\sigma(t) dW_\sigma(t) + \mathcal{S} \cdot dJ(t),$$

where $J(t)$ is a Poisson process with intensity v (see Section 4.7 in Chapter 4), $\mathcal{S} > 0$ is the size of the jump, which we suppose to be constant for illustration purposes only, and, finally,

κ , ω and ψ are constants. In this model, the presence of positive jumps, $v > 0$, makes the left tail of the return distribution thicker, when $\rho < 0$. Therefore, we need a high κ to avoid a too thicker distribution. With $v = 0$, instead, a thicker distribution can only be obtained through lower values of κ .

Naturally, the effects illustrated in this section mostly refer to explanations related to stochastic volatility although in general, they might arise for other reasons leading to leptokurticity, such feedback effects. [In progress, mention the literature on feedback effects of the 1990s.]

10.5.3 Option pricing with stochastic volatility

Models with stochastic volatility lead to *market incompleteness*. Market incompleteness, as we defined it in Chapter 4, is the situation where we cannot hedge against future contingencies by trading all of the available securities. In our context, market incompleteness relates to the fact that we cannot hedge, or replicate, the payoff of a derivative written on an asset, because the number of assets available for trading (one) is less than the sources of risk, the two Brownian motions in Eq. (10.34). Naturally, markets can be “completed” by the presence of the option. However, in this case the option price is not preference free, as we shall show. Intuitively, the option price in a stochastic volatility world, $C(S(\tau), \sigma^2(\tau), T - \tau)$, is driven by both the asset price, $S(\tau)$, and the stochastic variance, $\sigma^2(\tau)$, where $\sigma^2(\tau)$ is driven by the Brownian motion $W_\sigma(\tau)$. The value of a (“would be replicating”) portfolio that only includes the underlying asset would only be driven by the Brownian $W(\tau)$, without factoring in all the random fluctuations affecting volatility, $\sigma^2(\tau)$. But then, this portfolio cannot replicate the option price $C(S(\tau), \sigma^2(\tau), T - \tau)$, as this is also driven by $\sigma^2(\tau)$, and then by $W_\sigma(\tau)$. It is the reason why we cannot come up with a preference-free price for the option, as explained in a general context in Chapter 4.³

To summarize, stochastic volatility entails two inextricable consequences: (i) There are an infinity of option prices consistent with absence of arbitrage, which correspond to the many risk-neutral probabilities consistent with the model: there are many risk-adjustments that we can make to the drift term of the variance process in Eqs. (10.34), as we shall see below; (ii) there cannot be perfect hedging strategies only relying on the underlying asset. As regards point (ii), we might, alternatively, either (a) use a strategy, which albeit not self-financed, would still allow for a perfect replication of the claim, or (b) a self-financed strategy that would apply to some misspecified model. In case (a), the strategy leads to a hedging cost process. In case (b), the strategy leads to a tracking error process, although there might be situations where the claim can be “super-replicated,” as explained below.

10.5.3.1 Replication

Let us suppose that the price of the underlying asset is solution to Eqs. (10.34). The rational pricing function of a European-style option is $C^T(\tau) = C(S(\tau), \sigma^2(\tau), T - \tau)$. Suppose two such options are traded, with prices $C^{T_1}(\tau)$ and $C^{T_2}(\tau)$, where we take $T_1 < T_2$. We cannot replicate $C^{T_1}(\tau)$ by trading the underlying asset and the money market account. Indeed, let V be the value of a self-financed strategy including the asset price and the money market account,

³Stochastic volatility is not a source of market incompleteness per se. Mele (1998) (p. 88) considers a “circular” market with m asset prices, where (i) asset price no. i exhibits stochastic volatility, and (ii) this stochastic volatility is driven by the Brownian motion driving the $(i - 1)$ -th asset price. Therefore, in this market, each asset price is solution to Eqs. (10.34) and yet markets are complete.

which obviously satisfy:

$$dV = (\pi(\mu - r) + rV) dt + \pi\sigma dW, \quad (10.36)$$

where π is the value of the investment in the asset underlying the option. Instead, the price of the first option satisfies:

$$dC^{T_1} = \mathcal{L}(C^{T_1}) dt + C_S^{T_1} \sigma S dW + C_{\sigma^2}^{T_1} a dW_\sigma, \quad (10.37)$$

where \mathcal{L} is the infinitesimal generator, $\mathcal{L}C^{T_1} \equiv \frac{\partial C^{T_1}}{\partial t} + \mu S C_S^{T_1} + b C_{\sigma^2}^{T_1} + \frac{1}{2} \sigma^2 S^2 C_{SS}^{T_1} + \frac{1}{2} a^2 C_{\sigma^2 \sigma^2}^{T_1} + \rho a \sigma S C_{S \sigma^2}^{T_1}$. We see that we cannot match the diffusion coefficients of the option price in Eq. (10.37) through that in Eq. (10.36).

Instead, we might replicate the price of the first option, C_{T_1} , through a self-financed portfolio strategy comprising trading in (i) the asset underlying the option, (ii) the option expiring at T_2 , and (iii) the money market account. The value V of this strategy satisfies:

$$dV = \left(\pi_1 (\mu - r) + \pi_2 \left(\frac{\mathcal{L}(C^{T_2})}{C^{T_2}} - r \right) + rV \right) dt + \pi_1 \sigma dW + \pi_2 \left(\frac{C_S^{T_2}}{C^{T_2}} \sigma S dW + \frac{C_{\sigma^2}^{T_2}}{C^{T_2}} a dW_\sigma \right), \quad (10.38)$$

where π_1 is the value of the investment in the asset price, and π_2 is the value of the investment in the second option. We match the diffusion coefficients of Eq. (10.37) and Eq. (10.38), and obtain the following solutions for π_1 and π_2 :

$$\pi_1 = \left(\frac{C_S^{T_1}}{C^{T_1}} - \frac{C_{\sigma^2}^{T_1}}{C_{\sigma^2}^{T_1}} \right) C_S^{T_2} S, \quad \pi_2 = \frac{C_{\sigma^2}^{T_1}}{C_{\sigma^2}^{T_2}}. \quad (10.39)$$

Replacing these expressions into the drift of Eq. (10.36), and equating the drift of Eq. (10.36) to that of Eq. (10.37), leaves:

$$\frac{\mathcal{L}(C^{T_1}) - rC^{T_1} - C_S^{T_1}(\mu - r)S}{C_{\sigma^2}^{T_1}} = \frac{\mathcal{L}(C^{T_2}) - rC^{T_2} - C_S^{T_2}(\mu - r)S}{C_{\sigma^2}^{T_2}}. \quad (10.40)$$

These two ratios agree. They must then be equal to some process $a \cdot \Lambda^\sigma$, say, which is independent of the maturity of the option. Therefore, we obtain that,

$$\frac{\partial C}{\partial t} + rSC_S + (b - a\Lambda^\sigma)C_{\sigma^2} + \frac{1}{2}\sigma^2 S^2 C_{SS} + \frac{1}{2}a^2 C_{\sigma^2 \sigma^2} + \rho a \sigma S C_{S \sigma^2} = rC. \quad (10.41)$$

The interpretation of Λ^σ is that of the unit risk-premium required to face the risk of stochastic fluctuations in volatility. The problem, absence of arbitrage does not suffice to recover a unique Λ^σ . By the Feynman-Kac stochastic representation of the solution to a partial differential equation, there are many solutions to Eq. (10.41),

$$C(S(t), \sigma^2(t), T - t) = e^{-r(T-t)} \mathbb{E}_{Q^\Lambda} \left((S(T) - K)^+ \mid S(t), \sigma^2(t) \right), \quad (10.42)$$

where Q^Λ is a risk-neutral probability, induced by the many Λ that are consistent with absence of arbitrage.

Indeed, this derivation suggests two possible uses of the portfolio strategies of Eqs. (10.39). The first, obvious, is hedging. We can always hedge the first option with π_1 and π_2 in Eqs. (10.39). The second is more subtle. If we really think our evaluation model for the first option

is better than the market, we can always synthesize the first option with the portfolio in Eqs. (10.39), and replicate the payoff of the first option at expiration, $\psi(S_{T_1})$.

Eqs. (10.40) and (10.41) can be interpreted as APT relations. Indeed, let us define the unit risk-premium related to the fluctuations of the asset price, $\lambda = (\mu - r)/\sigma$. Then, Eq. (10.40) or Eq. (10.41) imply that,

$$\frac{\mathcal{L}(C)}{C} = E\left(\frac{dC}{C}\right) = r + \underbrace{\frac{C_S}{C}\sigma S}_{\equiv \beta_S} \cdot \lambda + \underbrace{\frac{C_{\sigma^2}}{C}a}_{\equiv \beta_{\sigma^2}} \cdot \Lambda^\sigma,$$

where β_S is the beta related to the volatility of the option price induced by fluctuations in the stock price, S , and β_{σ^2} is the beta related to the volatility of the option price induced by fluctuations in the return volatility. It is a generalization of the APT relation in Eq. (10.15) that holds for the Black & Scholes model.

10.5.3.2 Market completeness

Derivatives can complete markets. Show this is the case when $C_\sigma > 0$. Show conditions under which this is true. It's a generalization to the one-dimensional case analyzed in Section 10.4.7. Review the literature. [In progress]

10.5.3.3 Pricing formulae

Hull and White (1987), Scott (1987) and Wiggins (1987) develop the first option pricing models with stochastic volatility. Heston (1999b) provides an analytical solution assuming an affine model for the variance process, for otherwise, we need to solve through numerical methods relying on Montecarlo simulation or the numerical solution to partial differential equations.

Hull & White

Hull and White (1987) derive a first pricing formula based on a continuous-time model where asset returns and volatility are uncorrelated,

$$\begin{cases} \frac{dS(t)}{S(t)} &= rdt + \sigma(t) dW(t) \\ \frac{d\sigma^2(t)}{\sigma^2(t)} &= \mu_v d\tau + \xi dW_\sigma(t) \end{cases} \quad (10.43)$$

where W and W_σ are uncorrelated Brownian motions, defined under the risk-neutral probability. They show that the option price takes the following form:

$$C(S(t), \sigma^2(t), T-t) = \mathbb{E}_t^{\tilde{V}}[C_{BS}(S(t), T-t; K, \sqrt{\tilde{V}_{t,T}})], \quad (10.44)$$

where $C_{BS}(S(t), T-t; K, \sqrt{\tilde{V}_{t,T}})$ denotes the usual Black-Scholes formula in Eq. (10.14), evaluated at the average variance, defined as

$$\tilde{V}_{t,T} = \frac{1}{T-t} \int_t^T \sigma^2(\tau) d\tau,$$

and $\mathbb{E}_t^{\tilde{V}}$ denotes the conditional risk-neutral expectation taken with respect to laws generating $\tilde{V}_{t,T}$. According to Eq. (10.44), the option price is simply the Black & Scholes formula averaged over all possible “values” taken by future average variance, $\tilde{V}_{t,T}$. Accordingly, the authors

provide a Taylor's expansion around the conditional expectation of $\tilde{V}_{t,T}$, $m_V(t) = \mathbb{E}_t(\tilde{V}_{t,T})$,

$$\begin{aligned} C(S(t), \sigma^2(t), T-t) &= C_{\text{BS}}(S(t), T-t; K, \sqrt{m_V}) \\ &+ \frac{1}{2} \frac{\partial^2 C_{\text{BS}}(S(t), T-t; K, \sqrt{V})}{\partial V^2} \Bigg|_{V=m_V(t)} \cdot \text{var}_t(\tilde{V}_{t,T}) \\ &+ \frac{1}{6} \frac{\partial^3 C_{\text{BS}}(S(t), T-t; K, \sqrt{V})}{\partial V^3} \Bigg|_{V=m_V(t)} \cdot \text{skew}_t(\tilde{V}_{t,T}) + \dots \end{aligned}$$

In fact, this formula is quite general, in that it goes through in contexts more general than that in Eq. (10.43), such as that of Eqs. (10.34), provided the two Brownian motions W and W_σ are uncorrelated, as formally shown in Appendix 3. In Appendix 3, we also review the arguments leading to the following generalization of the Hull & White equation (10.44) to the case where the two Brownian motions W and W_σ are correlated, with a stochastic correlation equal to $\rho(\tau)$, say. Romano and Touzi (1997) show that in this case, and denoting the Brownian motion $W(\tau) = \rho(\tau)W_\sigma(\tau) + \sqrt{1-\rho^2(\tau)}dZ(\tau)$, from some standard Brownian motion Z , the Hull & White equation generalizes to,

$$\begin{aligned} C(S(t), \sigma^2(t), \rho(\tau), T-t) &= \mathbb{E}_{\tilde{V}}[C_{\text{BS}}(S(t), L_{t,T}, T-t; K, \sqrt{\tilde{V}_{t,T}^\rho})] \\ L_{t,T} &= e^{-\frac{1}{2} \int_t^T \rho^2(\tau) \sigma^2(\tau) d\tau + \int_t^T \sigma(\tau) \rho(\tau) dW_\sigma(\tau)}, \quad \tilde{V}_{t,T}^\rho = \frac{1}{T-t} \int_t^T \sigma^2(\tau) (1-\rho^2(\tau)) d\tau \end{aligned} \quad (10.45)$$

The assumption underlying Eq. (10.45) is that the correlation process $\rho(\tau)$, and the coefficients b and a in the second of Eqs. (10.34), do not depend on the asset price.

Heston

Heston's (1993b) develops an analytical solution to a model with stochastic volatility, relying on the following dynamics of the stock price:

$$\begin{cases} d \ln S(t) &= \left(r - \frac{1}{2} \sigma^2(t) \right) dt + \sigma(t) dW(t) \\ d\sigma^2(t) &= \kappa(\omega - \sigma^2(t)) dt + \xi \sigma(t) dW_\sigma(t) \end{cases} \quad (10.46)$$

where W and W_σ are two Brownian motions that have a constant correlation, ρ . The return variance is a square-root process. We supply hints about the derivation of Heston's formula, by developing a general formula that follows the same line of reasoning leading to Eq. (10.3) in Section 10.2, as follows:

$$\begin{aligned} &e^{-r(T-t)} \mathbb{E}_t[(S(T) - K)^+] \\ &= e^{-r(T-t)} \int_0^\infty \int_0^\infty (S(T) - K)^+ q_t(S(T), \sigma^2(T)) dS(T) d\sigma^2(T) \\ &= e^{-r(T-t)} \int_0^\infty \mathbb{I}_{S(T) \geq K} S(T) q_t^m(S(T)) dS(T) - e^{-r(T-t)} K \int_0^\infty \mathbb{I}_{S(T) \geq K} q_t^m(S(T)) dS(T) \\ &= S(t) \int_0^\infty \mathbb{I}_{S(T) \geq K} \hat{q}_t^m(S(T)) dS(T) - e^{-r(T-t)} K \int_0^\infty \mathbb{I}_{S(T) \geq K} q_t^m(S(T)) dS(T) \\ &= S(t) \cdot \hat{Q}_t(S(T) \geq K) - e^{-r(T-t)} K \cdot Q_t(S(T) \geq K), \end{aligned} \quad (10.47)$$

where $q_t(S(T), \sigma^2(T))$ is the risk-neutral joint density of the stock price and variance at T , $q_t^m(S(T))$ is the risk-neutral marginal density of the stock price at T , and finally, $\hat{q}_t^m(S(T))$ is a new marginal density of the stock price at T , with Radon-Nikodym derivative with respect to $q_t^m(S(T))$ given by the expression in Eq. (10.4):

$$\eta_t(T) = \frac{\hat{q}_t^m(S(T))}{q_t^m(S(T))} = \frac{e^{-r(T-t)} S(T)}{S(t)},$$

and finally, $\hat{Q}_t(S(T) \geq K)$ and $Q_t(S(T) \geq K)$ are two probabilities with densities \hat{q}_t^m and q_t^m , respectively. All these densities and probabilities are conditional upon the information at time t .

It is easy to see that the state process, $\eta_\tau(T)$, is solution to:

$$\frac{d\eta_t(\tau)}{\eta_t(\tau)} = -(-\sigma(\tau)) dW(\tau),$$

such that the stock price is solution to:

$$\begin{cases} d \ln S(t) &= (r + \frac{1}{2}\sigma^2(t)) dt + \sigma(t) d\hat{W}(t) \\ d\sigma^2(t) &= \kappa(\omega - \sigma^2(t)) dt + \xi\sigma(t) dW_\sigma(t) \end{cases} \quad (10.48)$$

under \hat{q}_t^m .

Let $x \equiv \ln S$. In the Black-Scholes case, $\sigma^2(t)$ is a constant, and the two probabilities, $\hat{Q}_t(x(T) \geq \ln K)$ and $Q_t(x(T) \geq \ln K)$, can be expressed in closed-form, using Eq. (10.48) and Eq. (10.46), respectively, leading to the celebrated formula in Eq. (10.14), as explained in Section 10.4.2.

In the Heston's model, the two probabilities, $P_1(x(t), \sigma^2(t), t) \equiv \hat{Q}_t(x(T) \geq \ln K)$ and $P_2(x(t), \sigma^2(t), t) \equiv Q_t(x(T) \geq \ln K)$, are solutions to:

$$\hat{L}P_1(x, \sigma^2, t) = 0, \quad LP_2(x, \sigma^2, t) = 0, \quad (10.49)$$

with the same boundary condition $P_j(x, \sigma^2, T) = \mathbb{1}_{x \geq \ln K}$, $j = 1, 2$, and where \hat{L} and L are the infinitesimal generators associated to Eq. (10.48) and Eq. (10.46). While the solution to these probabilities is unknown in closed-form, their characteristic functions are exponential affine in x and σ^2 . Precisely, define the two characteristic functions:

$$f_1(x, \sigma^2, t; \phi) = \hat{\mathbb{E}}_t(e^{-i\phi x(T)}), \quad f_2(x, \sigma^2, t; \phi) = \mathbb{E}_t(e^{-i\phi x(T)}), \quad i = \sqrt{-1},$$

where $\hat{\mathbb{E}}_t$ denotes the expectation taken with respect to \hat{q}_t^m , and \mathbb{E}_t denotes the conditional expectation taken against q_t^m .

The two functions f_j satisfy the same partial differential equations (10.49), but they can be solved in closed-form, because their boundary conditions are simply $f_j(x, \sigma^2, T) = e^{-i\phi x}$. Indeed, a fundamental definition is that a model is affine if its characteristic function is exponential-affine in its state variables. Affine models were already used to analyze the term structure of interest rates, since at least Vasicek (1977) and Cox, Ingersoll and Ross (1985), as we shall discuss in Chapter 12. Heston's model is the option pricing counterpart to these models of the yield curve.

The solution to the two characteristic functions is given by:

$$f_j(x, \sigma^2, t; \phi) = e^{C_j(T-t; \phi) + D_j(T-t; \phi)\sigma^2 + i\phi x},$$

where

$$\begin{aligned}
 C_j(T-t; \phi) &= r\phi i(T-t) + \frac{\kappa\omega}{\xi^2} \left[(b_j - \rho\xi\phi i + d_j)(T-t) - 2 \ln \left(\frac{1 - g_j e^{d_j(T-t)}}{1 - g_j} \right) \right] \\
 D_j(T-t; \phi) &= \frac{b_j - \rho\xi\phi i + d_j}{\xi^2} \left(\frac{1 - e^{d_j(T-t)}}{1 - g_j e^{d_j(T-t)}} \right) \\
 g_j &= \frac{b_j - \rho\xi\phi i + d_j}{b_j - \rho\xi\phi i - d_j}, \quad d_j = \sqrt{(b_j - \rho\xi\phi i)^2 - \xi^2 (2u_j\phi i - \phi^2)} \\
 b_1 &= \kappa - \xi\rho, \quad b_2 = \kappa, \quad u_1 = \frac{1}{2}, \quad u_2 = -\frac{1}{2}
 \end{aligned}$$

such that,

$$P_j(x(t), \sigma^2(t), t) = \frac{1}{2} + \frac{1}{\pi} \int_0^\infty \operatorname{Re} \left[\frac{e^{-i\phi \ln K} f_j(x(t), \sigma^2(t), t; \phi)}{i\phi} \right] d\phi. \quad (10.50)$$

[Write a small technical Appendix on inversions of characteristic functions] Replacing these two probabilities into Eq. (10.47), yields the celebrated Heston's formula.

10.5.3.4 Case study: assessing accumulators in markets with stochastic volatility

Is stochastic volatility important, in practice? Consider the “accumulators” described in Section 10.3.2, which are portfolios short n puts having the same strike K_p , and long one call with strike K_C , such that the initial cost of the portfolio is equal to zero. Let $K_C = 100$, and assume all options expire in six months. We consider two markets. The first market is one where asset prices are generated by the Black & Scholes model. In the second market, asset prices follow the Heston's model in Eq. (10.46). We assume that the parameter values are the same as those in Figure 10.7. We also assume that random fluctuations in volatility are not priced, such that the parameters κ and ω under the physical probability are the same as those under the risk-neutral. We address a number of issues aiming at exploring how stochastic volatility affects risk and return of these accumulators.

First, assume the current level of the index is $S = 98$. How many puts with strike $K_p = 90$ should we sell, to finance the long position in the call option, in both the Black & Scholes and Heston's markets? Assuming that the current volatility in Heston's model equals that related to the steady state variance, $\sigma(\tau) = \sqrt{\omega} = \sqrt{0.01}$, we find that in the Heston's market, $n = 3.9399$, whereas in Black & Scholes, $n = 6.5362$.

Next, we assume the current level of the index is $S = 102$ and re-do the previous calculations assuming, now, that the strike of the put is $K_p = 95$. We find that in Heston's market, $n = 5.9338$, whereas in Black & Scholes, $n = 8.8767$. The interpretation is that by increasing the strike from $K_p = 90$ to $K_p = 95$, we move towards more expensive puts. At the same time, by assuming that $S = 102$ (instead of $S = 98$), we are also considering a better market, which has an opposite effect on the put price than that arising following an increase in K_p . Moreover, a better market obviously makes call options more expensive. These effects lead to a higher number of puts to sell, in order to finance a long position in the call.

We can assess how the portfolio changes, once we consider a value of the index still equal to $S = 98$, but two levels of volatility: one higher than that corresponding to the steady state, $\sigma(t) = \sqrt{0.02}$, and another, lower, $\sigma(t) = \sqrt{0.008}$. In the high volatility case, $\sigma(t) = \sqrt{0.02}$, we find that $n = 2.9252$, and in the low volatility case, we find that $\sigma(t) = \sqrt{0.008}$, $n = 4.3249$.

An increase in volatility makes both puts and calls increase in value, although the price of the two puts increases more than that of the call.

Finally, assume that the rate of appreciation μ of the asset price is constant under the physical probability, and equal to $\mu = 3\%$. We can use Montecarlo simulations to calculate the average and standard deviation of the Profits & Losses generated by the accumulator. We summarize our findings in the table below.

P&Ls for a six month accumulator		
	average	std dev
B&S	1.3716	6.5114
Heston	1.2925	7.0352
Heston (high vol)	1.3571	8.9136
Heston (low vol)	1.2700	6.5906

Figure 10.8 depicts the relative frequency of the P&Ls corresponding to all cases we consider.

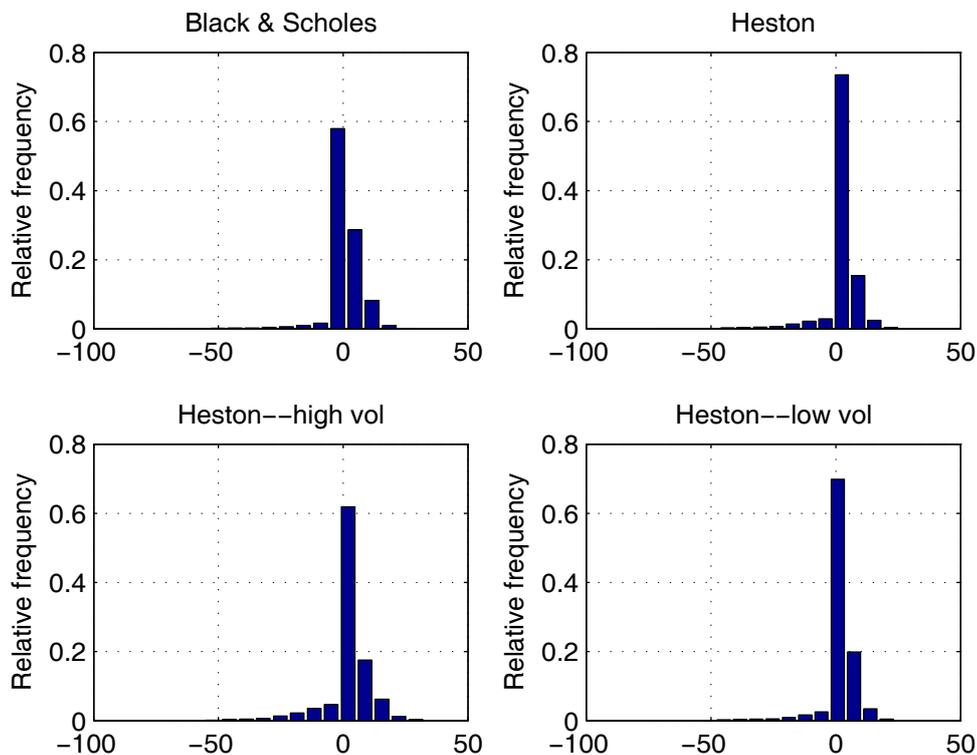


FIGURE 10.8. Frequency of Profit and Losses regarding a six month “accumulator” under different market assumptions: Black & Scholes (NW quadrant), Heston’s with current volatility fixed at its long-term value (NE), Heston’s with current volatility lower than its long-term value (SE), and Heston’s with current volatility higher than its long-term value (SW).

Note that although the average P&Ls are positive in all cases, the frequency distributions of the P&Ls exhibit quite high standard deviations, with long left-tails—downside risk is quite substantial, consistently with the payoff structure that we depicted in Figure 10.4. In the Heston’s market, average profits are higher than in Black & Scholes because accumulators necessitate less puts. Within the Heston’s market, profits lower when we move to both a low volatility and high volatility scenarios. In the low volatility scenario, average profits lower because the number of puts in the accumulator needs to increase. In the high volatility scenario, profits lower because whilst the number of puts in the accumulator decreases, a more volatile market also makes the accumulator more likely to generate adverse outcomes, thereby leveling down the expected profits.

10.6 Trading volatility with options

10.6.1 Payoffs

Positioning in volatility is a trading strategy relying on expectations about future volatility movements. It is a non-directional strategy, as it does not rely on the direction of the markets, only on their changes. It is market practice to distinguish between two types of volatility trading:

- (i) *Vega trading*, or *volatility surface* trading. It refers to a trade aiming to profit from a view that the term-structure of implied volatilities will change—for example, from the expectation of a flattening or a steepening term-structure of implied volatilities. It requires positioning into multiple types of options according to the nature of the expectation. For example, a “bull flattener” relies on the expectation that long-term implied volatilities will decrease faster than short-term implied volatilities, leading the term-structure of implied volatilities to flatten, which could be implemented through a portfolio which is: (i.1) long short-term options, and (i.2) short the long-term ones. (This portfolio would need to be delta-hedged, for reasons explained below).
- (ii) *Gamma trading*, which aims to generate profits from a realized volatility exceeding the current implied volatility. It relies therefore on directional views of ongoing volatility developments and for this reason has, obviously, an horizon much shorter than that of vega trading.

Option-based strategies might allow us to have views about these volatility developments, and include trading “straddles,” “strangles,” “butterflies,” “calendars,” and even delta-hedged option positions, as we shall explain below. They consist of portfolios comprising options and assets underlying these options, and aim to make P&Ls consistent with views about volatility developments.

A natural question arises. We know option prices are, generally, increasing in volatility. So why do we need to create portfolios of options and underlyings, in order to trade volatility? The reason is that option prices are increasing in both volatility and the asset price. For example, in a stochastic volatility setting, the option price is $C(S_t, \sigma_t^2, t)$, and if the volatility σ_t increases, the option price $C(S_t, \sigma_t^2, t)$ increases as well, in general. However, it might be possible that the increase in volatility occurs exactly when the asset price decreases. Incidentally, this circumstance is quite likely to occur, given the empirical evidence about the negative correlation between σ_t and S_t reviewed in Section 10.5. The implication would be that the increase in C determined by an increase in σ_t might be offset by the fall in C following the drop in S_t . To

isolate movements in the asset price volatility, we need to consider portfolios reverse-engineered so as to be insensitive to changes in the underlying asset price. [Mention here and in the next section Goldman Sachs approach to VIX]

To mitigate the effects of the movements in the underlying price, we may consider Black-Scholes hedges, such that the long position in the call option is offset by the short-position in the “Black-Scholes replicating” portfolio—which, by construction, only neutralizes movements in S_t , not σ_t . An alternative is a portfolio comprising options with final payoffs driven by the stock price, and negatively correlated, such as a European put and call options. For example, a *straddle* is a portfolio of one call option and one put option that have the same strike price and the same maturity. (A *strangle* is the same as a straddle, with the difference that the strike of the call differs from that of the put.)

Figure 10.9 depicts an example of payoffs arising from going long a straddle. The left panel shows the final payoff, equal to $(K - S_T)^+ + (S_T - K)^+$, for $K = 100$, as well as the value of this payoff at t , assuming a Black and Scholes market, with a risk-free rate $r = 1\%$, instantaneous volatility $\sigma = 20\%$, and under two assumptions about the maturity of the straddle, three months and one month. The right panel shows, instead, the P&L of this straddle, defined as $(K - S_T)^+ + (S_T - K)^+ - e^{r(T_i-t)} \text{Cost}_{i,t}$, where $\text{Cost}_{i,t} = p_{\text{BS}}(S_t, T_i - t; K, \sigma) + C_{\text{BS}}(S_t, T_i - t; K, \sigma)$, $C_{\text{BS}}(\cdot)$ is the Black-Scholes formula in Eq. (10.14), $p_{\text{BS}}(\cdot)$ is the corresponding put price, and $T_i - t$ is the maturity of the straddle, with $T_1 - t = \frac{1}{12}$ and $T_2 - t = \frac{3}{12}$. We assume that the index level at t is $S_t = 100$, such that the straddles are approximately at the money—the strike leading to at-the-money straddles is: $K_i = S_t e^{r(T_i-t)}$, but for comparison reasons, we keep on setting $K = 100$ for both straddles.

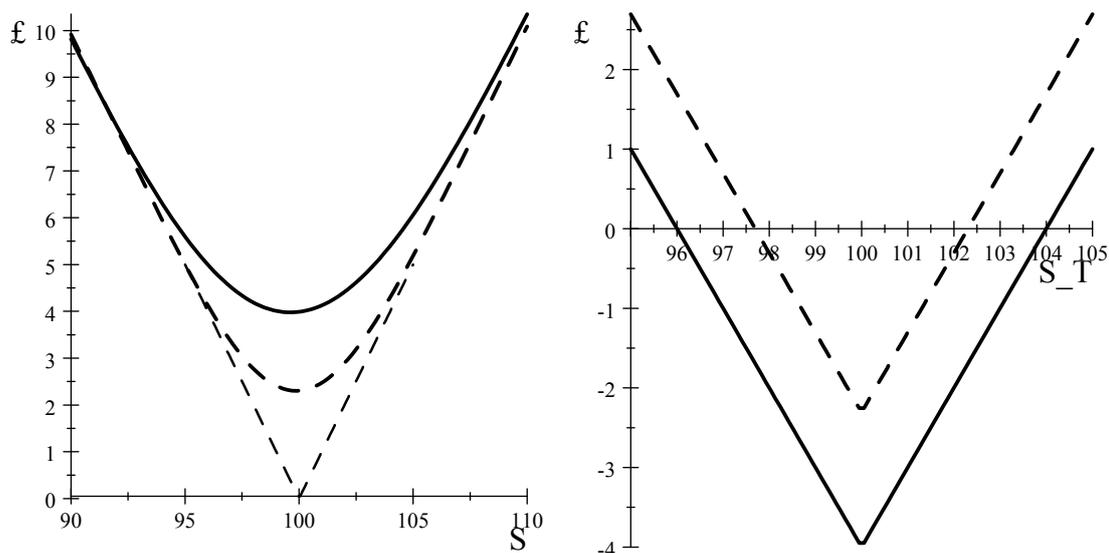


FIGURE 10.9. The left panel depicts the payoff of a “straddle” with strike price $K = 100$ (thin dashed line), as well as the value of a straddle with maturity $T_2 - t = \frac{3}{12}$ (solid line) and $T_1 - t = \frac{1}{12}$ (dashed line). The right panel depicts the P&Ls of roughly at-the-money straddles bought at time t , defined as $(K - S_T)^+ + (S_T - K)^+ - e^{r(T_i-t)} \text{Cost}_{i,t}$, where $\text{Cost}_{i,t} = p_{\text{BS}}(S_t, T_i - t; K, \sigma) + C_{\text{BS}}(S_t, T_i - t; K, \sigma)$, with $T_2 - t = \frac{3}{12}$ months (solid line), and $T_1 - t = \frac{1}{12}$ (dashed line), $r = 1\%$, $\sigma = 10\%$, and the index level $S_t = 100$.

The logic behind a straddle is that a call and a put have deltas that roughly compensate with each other, thereby allowing this portfolio to change primarily because of volatility movements. Figure 10.9 illustrates that a straddle helps express views about volatility, in that it pays off whenever the stock price moves sufficiently away from the initial level of the index, $S_t = 100$. Note that there is a small technical complication, due to the circumstance that the delta of the straddle is not precisely always zero, especially when the index level drifts away from moneyness. By Eq. (10.19), and the put-call parity in Eq. (10.5), it is: $2\Phi(d) - 1$. It is depicted in Figure 10.10. The reason the delta of the straddle deviates from zero while the index is away from its initial level is that the straddle becomes a short or a long position as soon as the index moves. When the index is up, the delta of the call is higher than the delta of the put, in absolute value, and when the index is down, the opposite happens.

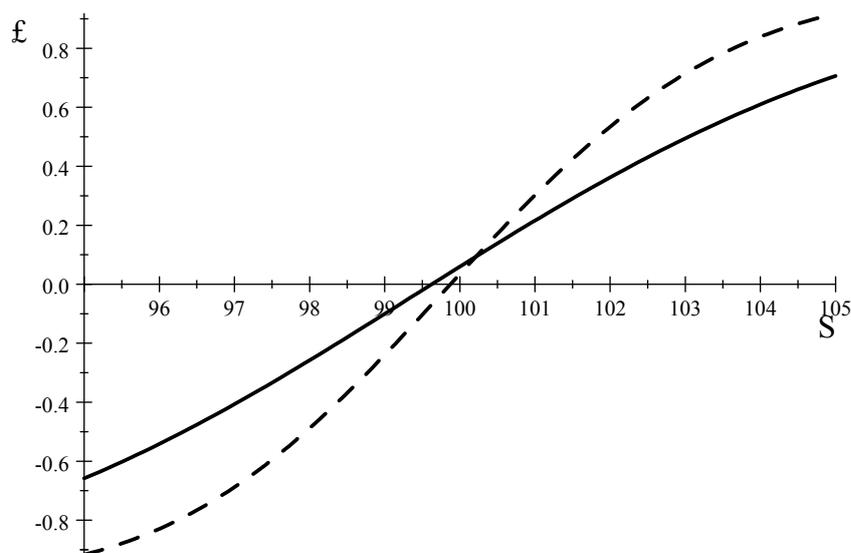


FIGURE 10.10. The delta of a straddle, $2\Phi(d) - 1$, where $\Phi(d)$ is the Black and Scholes delta in Eq. (10.19), the strike $K = 100$, for maturity $T_2 - t = \frac{3}{12}$ (solid line) and $T_1 - t = \frac{1}{12}$ (dashed line), and $r = 1\%$ and $\sigma = 10\%$.

Naturally, the strategy to short a straddle lead to opposite payoffs than those depicted in Figure 10.9—shorting a straddle relies on the expectation that markets are going to be stable. Straddles bear some inglorious history. In 1995, the 233-year old Barings Bank collapsed, because of the famous short-straddle one of its traders, Nick Leeson, was implementing on the Nikkei Index. A short-straddle is, of course, a view volatility will not raise. However, in January 1995, a violent earthquake made the Nikkei index crash by almost 7% in a week. The straddle was “naked,” i.e. delta-hedged, at most, and led to losses Leeson was not only unable to absorb, but also to amplify, given he was insisting on having views the Index would stabilize. The Index did not.

Potential losses arising from a short position in straddles can be reduced, by going long one additional portfolio comprising: (i) an out-of-the money put, which pays exactly when the underlying goes down, and (ii) an out-of-the money call, which pays when the underlying goes up. Combining this portfolio with a short-straddle leads to what is known as *butterfly spread*. Figure 10.11 depicts payoffs and P&L relating to a butterfly, where the straddle has strike $K = 100$ and maturity one month (as one of the straddles in Figure 10.9), and the strikes of the

out-of-the-money call and put are $K_C = 102$ and $K_p = 98$. The right panel shows the P&L of the butterfly, defined as $-[(K - S_T)^+ + (S_T - K)^+] + (K_p - S_T)^+ + (S_T - K_C)^+ - e^{r(T-t)}\text{Cost}_t$, where $T - t = \frac{1}{12}$, and Cost_t is the value of the butterfly at time t in the same Black & Scholes market considered in Figures 10.9-10.10.

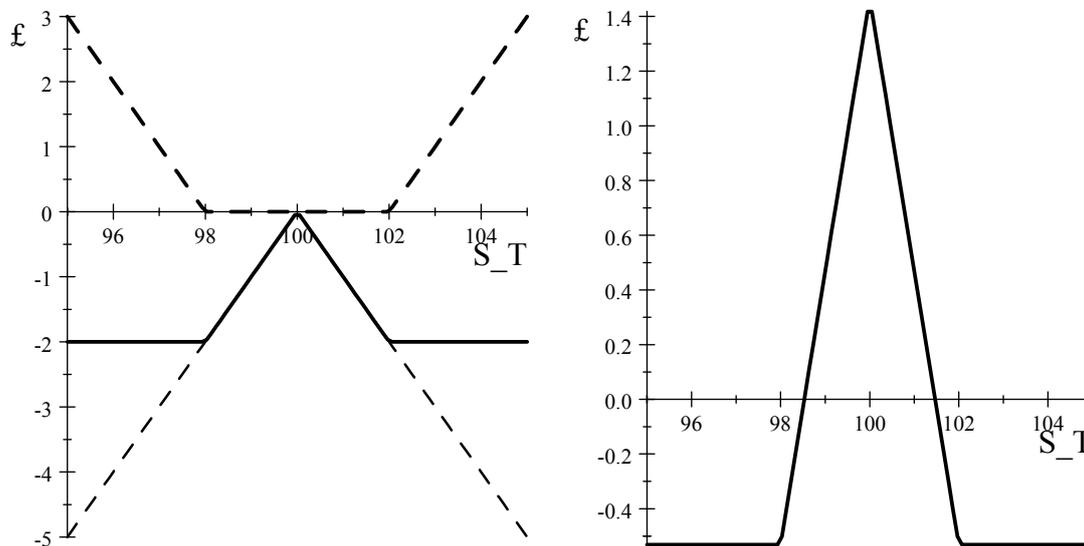


FIGURE 10.11. The left panel depicts the payoff of a “butterfly” with maturity equal to one month (solid line), which is (i) short one straddle (the thin dashed line) with strike $K = 100$, and (ii) long one out-of-the-money put, with strike $K_p = 98$, and one out-of-the-money call, with strike $K_C = 102$ (the dashed line). The right panel depicts the P&L of the butterfly, $-[(K - S_T)^+ + (S_T - K)^+] + (K_p - S_T)^+ + (S_T - K_C)^+ - e^{r(T-t)}\text{Cost}_t$, where $T - t = \frac{1}{12}$, and Cost_t is the value of the butterfly at time t , obtained in a Black & Scholes market with $r = 1\%$ and $\sigma = 10\%$ and the index level $S_t = 100$.

Alternatives to straddles are *calendar spreads*, which are portfolios long one call with maturity $T_1 - t$ and short one call with maturity $T_2 - t$, where $T_1 < T_2$, and where the two calls have the same strike price. If the underlying asset price does not move too much, the calendar spread value drops, because the price decay due to the passage of time (see Section 10.4.6.2) is more severe for the call with lower time to maturity. Naturally, if the price of the underlying increases, the value of the calendar spread increases as well, due to the positions in the two call options. Therefore, going long a calendar is consistent with the view that market volatility is about to increase. Figure 10.12 depicts the payoff and the P&Ls of a calendar, with strike $K = 100$, and maturities $T_1 - t = \frac{1}{12}$, and $T_2 - t = \frac{3}{12}$, assuming Black & Scholes market considered in Figures 10.9-10.11, with an index level $S_t = 100$. We calculate the payoff one month after the initial positioning. This payoff is, therefore, $(S_T - K)^+ - C_{BS}(S_T, \frac{2}{12}; K, \sigma)$, whereas the P&L is given by $(S_T - K)^+ - C_{BS}(S_T, \frac{2}{12}; K, \sigma) - e^{r\frac{1}{12}}(C_{BS}(S_t, T_1 - t; K, \sigma) - C_{BS}(S_t, T_2 - t; K, \sigma))$.

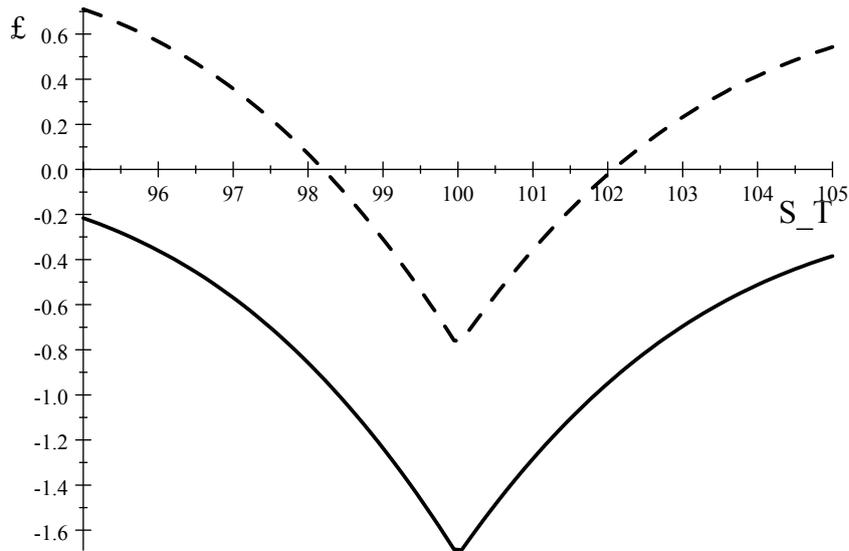


FIGURE 10.12. Payoff and P&L of a “calendar,” a portfolio which is (i) long a call option with strike price $K = 100$ and time to maturity $T_1 - t = \frac{1}{12}$, and (ii) short a call option with strike price $K = 100$ and time to maturity $T_2 - t = \frac{3}{12}$. The solid line plots the payoff after one month from inception, defined as $(S_T - K)^+ - C_{BS}(S_T, \frac{2}{12}; K, \sigma)$, whereas the dashed line is the payoff inclusive of the initial value of the position, $\pi(S_T) - e^{r\frac{1}{12}}(C_{BS}(S_t, T_1 - t; K, \sigma) - C_{BS}(S_t, T_2 - t; K, \sigma))$, with the index level $S_t = 100$.

10.6.2 P&Ls of Δ -hedged strategies

Straddles, or Black-Scholes hedged strategies, are not necessarily the best way to take views about volatility developments. To understand volatility trading through option-based strategies, and their related shortcomings, consider the simplest strategy, where one buys an option and hedges it through the Black and Scholes formula.⁴ Suppose to live in a world with stochastic volatility, where the asset price moves as in Eqs. (10.34). Assume that at time t , we go long a call option with market price equal to $C(S_t, \sigma_t^2, t)$. Let us build up a self-financed portfolio with value V_t ,

$$V_t = a_t S_t + b_t B_t, \tag{10.51}$$

where $B_t = e^{rt}$ is the money market account,

$$V_0 = C(S_0, \sigma_0^2, 0), \quad a_t = \Delta_{BS}(S_t, t; IV_0), \tag{10.52}$$

and IV_0 is the Black-Scholes implied volatility as of time $t = 0$, i.e. the time at which we are to take a view on future volatility.

Consider, first, the following heuristic arguments. Assume a Black & Scholes market, where the short-term rate, r , is zero and μ is also zero. Although volatility is constant in this market, on average, there might be periods where realized volatility, $\left(\frac{\Delta S_t}{S_t}\right)^2$, is higher than IV_0^2 . What is the P&L of a call option delta-hedged through Black & Scholes? Note that a call option delta-hedged with Black & Scholes is simply a portfolio with valued equal to $\Pi_t = C_t - \Delta_{BS} S_t - b_t B_t$,

⁴The following arguments also apply to the hypothetical situation where an investment bank, say, purchases an option for a mere market making scope, and then tries to hedge against it through Black-Scholes. It is, however, an unrealistic situation, as investment banks hedge through books, not through the single units adding up to the books.

such that, approximately,

$$\begin{aligned}\Delta\Pi_t &= \underbrace{\Theta_t\Delta t + \frac{1}{2}\Gamma_t(\Delta S_t)^2 + \Delta_{\text{BS}}\Delta S_t - \Delta_{\text{BS}}\Delta S_t}_{=\Delta C_t} \\ &= \left(-\frac{1}{2}\Gamma_t S_t^2 \text{IV}_0^2\right)\Delta t + \frac{1}{2}\Gamma_t(\Delta S_t)^2 = \frac{1}{2}\Gamma_t S_t^2 \left[\left(\frac{\Delta S_t}{S_t}\right)^2 - \text{IV}_0^2\Delta t\right],\end{aligned}\quad (10.53)$$

where $\Theta = \frac{\partial C}{\partial t}$, $\Delta_{\text{BS}} = \frac{\partial C}{\partial S}$, the Delta, $\Gamma = \frac{\partial^2 C}{\partial S^2}$, the Gamma, and the second equality follows by the Black-Scholes pricing equation (10.13). Aggregating until the maturity of the option, delivers the P&L at T :

$$\text{P\&L}_T \equiv \sum_{t=1}^T \Delta\Pi_t = \frac{1}{2} \sum_{t=1}^T \Gamma_t S_t^2 \left[\left(\frac{\Delta S_t}{S_t}\right)^2 - \text{IV}_0^2\Delta t \right]. \quad (10.54)$$

Note that the Black & Scholes delta is needed to compensate for those portions of the call price movements arising due to the asset price movements—the term $\Delta_{\text{BS}}\Delta S_t$ in the brackets of Eq. (10.53) contributes positively to the P&L, when $\Delta S_t > 0$, and negatively otherwise. So the hedging is natural: the call price can, say, go up because volatility goes up or because the underlying price goes up. To isolate pure views about volatility, we need to hedge against movements in the underlying price. As is clear, hedging through the Black-Scholes delta helps neutralize this effect. Hedging is actually effective in the short-term, as the period-by-period profits $\Delta\Pi_t$ in Eq. (10.53), only depend on how far realized volatility is from the initial Black-Scholes implied volatility. However, hedging might lead to a P&L that is inconsistent with views, because the difference between realized volatility and Black-Scholes implied volatility at time t , $\left(\frac{\Delta S_t}{S_t}\right)^2 - \text{IV}_0^2\Delta t$, is weighted with the Dollar Gamma, $\Gamma_t S_t^2$, which is positive as the Black-Scholes price is convex in S . As a result, we might end up with a situation where the P\&L_T might be negative even if the terms $\left(\frac{\Delta S_t}{S_t}\right)^2 - \text{IV}_0^2\Delta t$ are positive for most of the time, a feature known as “price-dependency.” We now illustrate these facts through a continuous-time model with random volatility.

So consider a general situation where volatility is not constant, such that the model is misspecified. El Karoui, Jeanblanc-Picqué and Shreve (1998) make the following observation. Consider the value of the self-financed portfolio in Eq (10.51). Because this portfolio is self-financed,

$$\begin{aligned}dV_t &= a_t dS_t + r b_t B_t dt \\ &= r V_t dt + a_t (dS_t - r S_t dt) \\ &= [r V_t + a_t (\mu - r) S_t] dt + a_t \sigma_t S_t dW_t.\end{aligned}$$

Moreover, by Itô’s lemma,

$$dC_{\text{BS}}(S_t, T - t; K, \text{IV}_0) = \left(\frac{\partial C_{\text{BS}}}{\partial t} + \mu S_t \frac{\partial C_{\text{BS}}}{\partial S} + \frac{1}{2} \sigma_t^2 S_t^2 \frac{\partial^2 C_{\text{BS}}}{\partial S^2} \right) dt + \sigma_t S_t \frac{\partial C_{\text{BS}}}{\partial S} dW_t,$$

where,

$$\begin{aligned} & \frac{\partial C_{BS}}{\partial t} + \mu S_t \frac{\partial C_{BS}}{\partial S} + \frac{1}{2} \sigma_t^2 S_t^2 \frac{\partial^2 C_{BS}}{\partial S^2} \\ &= \underbrace{\frac{\partial C_{BS}}{\partial t} + r S_t \frac{\partial C_{BS}}{\partial S} + IV_0^2 S_t^2 \frac{\partial^2 C_{BS}}{\partial S^2}}_{\equiv r C_{BS}} + (\mu - r) S_t \frac{\partial C_{BS}}{\partial S} + \frac{1}{2} (\sigma_t^2 - IV_0^2) S_t^2 \frac{\partial^2 C_{BS}}{\partial S^2} \\ &= r C_{BS} + (\mu - r) S_t \frac{\partial C_{BS}}{\partial S} + \frac{1}{2} (\sigma_t^2 - IV_0^2) S_t^2 \frac{\partial^2 C_{BS}}{\partial S^2}. \end{aligned}$$

Therefore, the tracking error, or P&L_t, defined as the difference between the Black-Scholes price and the portfolio value,

$$P\&L_t \equiv C_{BS}(S_t, T - t; K, IV_0) - V_t,$$

satisfies,

$$dP\&L_t = \left(r P\&L_t + \frac{1}{2} (\sigma_t^2 - IV_0^2) S_t^2 \frac{\partial^2 C_{BS}}{\partial S^2} \right) dt.$$

At maturity T :

$$\begin{aligned} P\&L_T &\equiv C_{BS}(S_T, T - T; K, IV_0) - V_T \\ &= \max\{S_T - K, 0\} - V_T \\ &= \frac{1}{2} e^{rT} \int_0^T e^{-rt} (\sigma_t^2 - IV_0^2) \frac{\partial^2 C_{BS}}{\partial S^2} S_t^2 dt. \end{aligned} \tag{10.55}$$

This expression is the continuous-time counterpart to Eq. (10.54). Moreover, it can be shown that a delta-hedged straddle strategy leads to twice the expression in Eq. (10.55), with the second partial of the straddle replacing the Black-Scholes Gamma. Because the Black-Scholes price is convex, Eq. (10.55) tells us that even if we do not exactly know the law of movement of volatility, but still hold the view it will be persistently higher than the initial Black-Scholes implied volatility, we can obtain positive profits through (i) a long position in a call, and (ii) a short position in the Black-Scholes replicating portfolio. Naturally, it isn't an arbitrage opportunity. The critical assumption is that volatility will increase.

Eq. (10.55) is problematic. Even if the volatility σ_t is higher than IV_0 for most of the time, the final P&L may not necessarily lead to a profit. The reason is that each volatility view, $\sigma_t^2 - IV_0^2$, is weighted by the Dollar Gamma, $S_t^2 \frac{\partial^2 C_{BS}}{\partial S^2}$. It may be that “bad” realization of the volatility views, i.e. $\sigma_t^2 < IV_0^2$, occur precisely when the Dollar Gamma is large—the “price-dependency” issue raised whilst discussing Eq. (10.54). Moreover, the strategy is costly, as it relies on Δ -hedging. The volatility contracts of Section 10.6 overcome these difficulties.

10.7 Local volatility

10.7.1 Issues

Stochastic volatility models might provide interesting explanations, such as the smile effect, as discussed in Section 10.5.2. However, these models cannot allow for a *perfect* fit of the smile. Towards the end of 1980s and the beginning of the 1990s, a modeling approach emerged to cope with issues relating to a perfect fit of the yield curve. As reviewed in Chapters 11 and 12, this modeling approach arose as a response to the need to make the pricing of interest rate

derivatives rely on models where the underlying assets in the books of the banks, bonds say, are priced without any error. In 1993 and 1994, Derman & Kani, Dupire and Rubinstein [cite exact references] come up with a technology that could be applied to options on tradable assets.

Why is it important to exactly fit the structure of already existing plain vanilla options? Banks trade both plain vanilla and less liquid, or “exotic” derivatives. Suppose we wish to price exotic derivatives. We want to make sure the model we use to price the illiquid option must predict that the plain vanilla option prices are identical to those we are trading. How can we trust a model that is not even able to pin down all outstanding contracts? A model like this could give rise to arbitrage opportunities to unscrupulous users.

10.7.2 The perfect fit

Let us outline the steps we need to pricing new derivatives while avoiding any pricing errors for the existing ones:

- (i) We take as given the prices of a set of actively traded European options. Let K and T be strikes and time-to-maturity of these liquid options. We aim to match the model to the data:

$$C_{\S}(K, T) = C(K, T), \quad K, T \text{ varying}, \quad (10.56)$$

where $C_{\S}(K, T)$ are market data, and $C(K, T)$ are the model’s prediction.

Is it mathematically possible to consider a diffusive model for the stock price, such that the initial collection of European option prices, $C_{\S}(K, T)$, is predicted without errors by the resulting model, as in Eq. (10.56)?

- (ii) The answer is in the affirmative. Consider a diffusion process for the stock price:

$$\frac{dS_t}{S_t} = rdt + \sigma(S_t, t) d\hat{W}_t,$$

where \hat{W} is a Brownian motion under the risk-neutral probability. The only function to “calibrate” to make Eq. (10.56) hold is the volatility function, $\sigma(S_t, t)$.

- (iii) The Appendix shows that Eq. (10.56) holds when $\sigma(S_t, t) = \sigma_{\text{loc}}(S_t, t)$, where:

$$\sigma_{\text{loc}}(K, T) = \sqrt{2 \frac{\frac{\partial C(K, T)}{\partial T} + rK \frac{\partial C(K, T)}{\partial K}}{K^2 \frac{\partial^2 C(K, T)}{\partial K^2}}}. \quad (10.57)$$

The function $\sigma_{\text{loc}}(S, t)$ is referred to as *local volatility*. Its square is the local variance, defined as the conditional expectation under Q of the instantaneous variance given the market level S at some future date τ ,

$$\sigma_{\text{loc}}^2(S, \tau) = \mathbb{E}_t[\sigma^2(S_\tau, \tau) | S_\tau = S] \quad (10.58)$$

where $\mathbb{E}[\cdot | \cdot]$ is the conditional expectation taken under the risk-neutral probability. All in all, local volatility is a volatility function such that the theoretical price equals the market price of all available options.

- (iv) Finally, we can price the illiquid options through numerical methods, for example through simulations. In the simulations, we use

$$\frac{dS_t}{S_t} = rdt + \sigma_{\text{loc}}(S_t, t) d\hat{W}_t.$$

Empirically, the local volatility surface, $\sigma_{\text{loc}}(S, t)$ is typically decreasing in S for fixed t , a phenomenon known as the Black-Christie-Nelson leverage effect discussed in Chapter 8 and Section 10.5.2. This fact might lead to assume from the outset that $\sigma(x, t) = x^\alpha f(t)$, for some function f and some constant $\alpha < 0$, as simplification leading to the so-called CEV (Constant Elasticity of Variance) model. Practitioners are increasingly relying on the so-called SABR model, which combines “local vols” with “stoch vol,” as follows:

$$\begin{aligned} \frac{dS_t}{S_t} &= rdt + \sigma(S_t, t) \cdot v_t \cdot d\hat{W}_t \\ dv_t &= \phi(v_t)dt + \psi(v_t)d\hat{W}_t^v \end{aligned} \quad (10.59)$$

where \hat{W}^v is another Brownian motion, and ϕ, ψ are some functions. [Provide references.] The appendix shows that in this specific case, the initial structure of European options prices is pinned down by:

$$\tilde{\sigma}_{\text{loc}}(K, T) = \frac{\sigma_{\text{loc}}(K, T)}{\sqrt{\mathbb{E}(v_T^2 | S_T)}}, \quad (10.60)$$

where $\sigma_{\text{loc}}(K, T)$ is the same as in Eq. (10.57). For this model, we simulate

$$\begin{cases} \frac{dS_t}{S_t} = rdt + \tilde{\sigma}_{\text{loc}}(S_t, t) \cdot v_t \cdot d\hat{W}_t \\ dv_t = \phi(v_t)dt + \psi(v_t)d\hat{W}_t^v \end{cases}$$

A note on recalibration. Clearly, local surfaces are obviously functions of the initial state where the calibration starts off. The calibration has to be re-performed all the time to reflect new information.

10.7.3 Relations with implied volatility

10.7.3.1 Implied volatility as expected local volatility

Section 10.5.5 provides the expression of the P&L relating to a long position in a call option, delta-hedged with Black and Scholes using an implied volatility fixed at an initial level IV_0 ,

$$\text{P\&L}_T \equiv C_{\text{BS}}(S_T, T; K, T, IV_0) - V_T = \frac{1}{2} e^{rT} \int_0^T e^{-rt} (\sigma_t^2 - IV_0^2) S_t^2 \frac{\partial^2 C_{\text{BS}}}{\partial S^2} dt. \quad (10.61)$$

Naturally, we have that $e^{-rT} \mathbb{E}(\text{P\&L}_T) = 0 \Leftrightarrow C_{\text{BS}}(S_0, 0; K, T, IV_0) = V_0 = C(S_0, \sigma_0^2, 0)$, the true market price, consistently with Eq. (10.52), such that, setting the expression of the last equality of Eq. (10.61), and solving for IV_0 , delivers:

$$IV_0^2 = \frac{\mathbb{E} \left[\int_0^T e^{-rt} S_t^2 \frac{\partial^2 C_{\text{BS}}}{\partial S^2} \cdot \sigma_t^2 dt \right]}{\mathbb{E} \left[\int_0^T e^{-rt} S_t^2 \frac{\partial^2 C_{\text{BS}}}{\partial S^2} dt \right]}.$$

Alternatively, we may consider another hedging positioning, suggested by Gatheral (2006, Chapter 3), where the delta-hedging is made through some fictitious time-varying instantaneous, but deterministic, volatility, equal to $\bar{\sigma}_t$, say, where

$$\bar{\sigma}_t^2 = \frac{1}{T-t} \int_t^T \nu_u du, \quad \bar{\sigma}_0^2 = \text{IV}_0^2, \quad (10.62)$$

for some deterministic ν_t . In this case, the P&L would be similar to that in Eq. (10.61), with

$$\text{P\&L}_T = \frac{1}{2} e^{rT} \int_0^T e^{-rt} (\sigma_t^2 - \nu_t) S_t^2 \frac{\partial^2 C_{\text{BS}}}{\partial S^2} dt.$$

Imposing the zero profit condition under Q , leaves:

$$\nu_t = \frac{\mathbb{E} \left(S_t^2 \frac{\partial^2 C_{\text{BS}}}{\partial S^2} \sigma_t^2 \right)}{\mathbb{E} \left(S_t^2 \frac{\partial^2 C_{\text{BS}}}{\partial S^2} \right)} = \mathbb{E}_{Q^\Gamma} (\sigma_t^2), \quad (10.63)$$

where \mathbb{E}_{Q^Γ} is the expectation taken under the probability Q^Γ , defined as,

$$\frac{dQ^\Gamma}{dQ} = \frac{S_t^2 \frac{\partial^2 C_{\text{BS}}}{\partial S^2}}{\mathbb{E} \left(S_t^2 \frac{\partial^2 C_{\text{BS}}}{\partial S^2} \right)}.$$

We term Q^Γ “Dollar-Gamma” probability. By Eqs. (10.62) and (10.63),

$$\text{IV}_0^2 = \frac{1}{T} \int_0^T \mathbb{E}_{Q^\Gamma} (\sigma_t^2) dt. \quad (10.64)$$

So implied vols *are* expectations of future realized vols, but only under the Dollar-Gamma probability.

We can elaborate on Eq. (10.64). We have:

$$\begin{aligned} \mathbb{E}_{Q^\Gamma} (\sigma_t^2) &= \mathbb{E} \left[\mathbb{E} \left(\sigma_t^2 \frac{dQ^\Gamma}{dQ} \middle| S_t \right) \right] \\ &= \mathbb{E} \left[\frac{dQ^\Gamma}{dQ} \mathbb{E} (\sigma_t^2 | S_t) \right] \\ &= \mathbb{E} \left[\frac{dQ^\Gamma}{dQ} \sigma_{\text{loc}}^2 (S_t, t) \right] \\ &= \mathbb{E}_{Q^\Gamma} [\sigma_{\text{loc}}^2 (S_t, t)] \\ &= \int \sigma_{\text{loc}}^2 (S_t, t) \frac{S_t^2 \frac{\partial^2 C_{\text{BS}}}{\partial S^2} q(S_t | S_0)}{\mathbb{E} \left(S_t^2 \frac{\partial^2 C_{\text{BS}}}{\partial S^2} \right)} dS_t \\ &\approx \sigma_{\text{loc}}^2 (\tilde{S}_t, t), \end{aligned} \quad (10.65)$$

where the first equality follows by the law of iterated expectations, $q(S_t | S_0)$ denotes the conditional density of S_t given S_0 , and $\sigma_{\text{loc}}^2(S_t, t)$ is the local variance, as defined in Section 10.6.2. Finally, \tilde{S}_t is a deterministic, “most likely path” of S_t , after Gatheral (2006, Chapter 6), a

sort of certainty equivalent for the local variance, for a fixed t . We also know that at $t = T$, $\partial^2 C_{BS}/\partial S^2$ is Dirac's delta centered at K , such that we may safely condition $\tilde{S}_T = K$ and, then, view \tilde{S}_t as a bridge starting from S_0 and ending at K . As a simple example, $\tilde{S}_t \approx S_0 (K/S_0)^{t/T}$. As a second example, $\mathbb{E}(S_t | S_T = K)$, which we may approximate assuming S_t is a Geometric Brownian motion with parameters r and σ , in which case $\tilde{S}_t \approx S_0 e^{rt} \left(\frac{K}{S_0 e^{rT}}\right)^{t/T} e^{\frac{1}{2}\sigma^2 t}$. Gatheral argues, with a numerical example, that these approximations are quite reasonable, at least for options with time to maturity less than a year.

Using the approximation in Eq. (10.65) delivers:

$$IV_0^2 = \frac{1}{T} \int_0^T \sigma_{\text{loc}}^2(\tilde{S}_t, t) dt. \quad (10.66)$$

Surfaces depend on the initial state, as mentioned in Section 10.6.2. “Sticky smiles” might be defined as those where the skew does not depend on the initial state, roughly. Suppose a very simple example, where $IV(t, T, K; S_t) = a - b(K + S_t) = a - b\left(\frac{K}{S_t} + 1\right) S_t$. As S_t falls, the skew goes up, consistent with the leverage effect. This skew does not depend on the initial price S_0 . We can generate this skew, by assuming the local variance does not depend on time, $\sigma_{\text{loc}}^2(K, t) = a - b(K + S_0)$. Indeed, in light of Eq. (10.66), we would then have that $IV^2(0, T, K; S_0) = IV_0^2 = a - b(K + S_0)$ and then, $IV(t, T, K; S_t) = a - b(K + S_t)$ for each t .

10.7.3.2 Local volatility as a function of implied volatility

[In progress]

10.8 The price of volatility

How much volatility do we expect to prevail in the future, after controlling for risk? An informal answer to this question has long been the volatility implied by at-the-money options. In fact, it is not. Expected volatility, adjusted for risk, is a weighted average of implied volatilities of a continuum of options, as explained below. It is not mere academic purism. Knowing expected volatility under the risk-neutral probability allows to trade assets with payoffs linked to future realized volatility, known as variance swaps. In fact, in September 2003, the Chicago Board Options Exchange (CBOE) changed its volatility index VIX to approximate the variance swap rate of the S&P 500 index return (for 30 days), as in Eq. (10.69) below. In March 2004, the CBOE launched the CBOE Future Exchange for trading futures on the new VIX. Options on VIX are also available for trading.

There are a number of compelling reasons explaining the interest investors may have in these contracts. One is, undeniably, related to the possibility to take views about developments in stock market volatility, without incurring into the price-dependency issues pointed out in Section 10.5.4. Passive funds managers might also find these contracts useful, as in times of high volatility, tracking errors widen and, then, index tracking performance deteriorates. Hedge funds might find this type of contracts attractive as well, as they invest in “relative value” strategies, attempting to profit from temporary price discrepancies. In times of high volatility, price discrepancies typically widen, and volatility contracts help these institutions hedge against these events.

10.8.1 Evaluation

Let us consider the following price process S_t under the risk-neutral probability:

$$\frac{dS_t}{S_t} = rdt + \sigma_t d\hat{W}_t,$$

where σ_t is \mathcal{F}_t -adapted, i.e. \mathcal{F}_t can be larger than that generated by the stock price, $\mathcal{F}_t^S \equiv \sigma(S_\tau : \tau \leq t)$. By Eqs. (10.57) and (10.58), and the result on risk-neutral densities summarized by Eq. (10.27), we have that:

$$e^{-r(T-t)} \mathbb{E}_t(\sigma_T^2) = 2 \int_0^\infty \frac{\frac{\partial C_t(K,T)}{\partial T} + rK \frac{\partial C_t(K,T)}{\partial K}}{K^2} dK, \quad (10.67)$$

where $C_t(K, T)$ is the price as of time t of a call option expiring at T and struck at K .

Next, define the realized “integrated” variance within the time interval $[T_1, T_2]$, with $T_1 > t$:

$$\text{var}(T_1, T_2) \equiv \int_{T_1}^{T_2} \sigma_u^2 du.$$

Finally, we compute the risk-neutral expectation of this realized variance, under the assumption $r = 0$, By Eq. (10.67),

$$\mathbb{E}_t[\text{var}(T_1, T_2)] = 2 \int_0^\infty \frac{C_t(K, T_2) - C_t(K, T_1)}{K^2} dK. \quad (10.68)$$

Eq. (10.68) can be generalized to the case where $r > 0$. In the Appendix, we show that for $T_1 = t$, $T_2 \equiv T$,

$$\mathbb{E}_t[\text{var}(t, T)] = 2e^{r(T-t)} \left[\int_0^{F(t)} \frac{P_t(K, T)}{K^2} dK + \int_{F(t)}^\infty \frac{C_t(K, T)}{K^2} dK \right], \quad (10.69)$$

where $F(t)$ is the forward price: $F(t) = e^{r(T-t)} S(t)$, and $P_t(K, T)$ is the price as of time t of a put option expiring at T and struck at K . A proof of Eq. (10.69) is in the Appendix. Intuitively, one use the following representation of the so-called log-contract payoff:

$$-\ln \frac{F_T}{F_t} = -\frac{1}{F_t} (F_T - F_t) + \left(\int_0^{F_t} (K - S_T)^+ \frac{1}{K^2} dK + \int_{F_t}^\infty (S_T - K)^+ \frac{1}{K^2} dK \right). \quad (10.70)$$

By Itô’s lemma, the risk-neutral expectation of the term on the left-hand side of Eq. (10.70) is one-half the risk-neutral expectation of $\text{var}(t, T)$, normalized by $e^{-r(T-t)}$. Instead, the expectation of the right-hand side of Eq. (10.70), normalized by $e^{-r(T-t)}$ is half the right-hand side of Eq. (10.69).

The new VIX index calculated by CBOE is an approximation to the square root of $\mathbb{E}_t[\text{var}(t, T)]$ in Eq. (10.69), re-normalized by time-to-maturity:

$$\text{VIX}(t, T) \equiv \sqrt{\frac{1}{T-t} \mathbb{E}_t[\text{var}(t, T)]},$$

obtained using a finite number of out-of-the-money options. The VIX index can be used to price and trade *variance swaps*, which are contracts that have zero value at the inception date, t . At maturity T , the buyer of the swap receives,

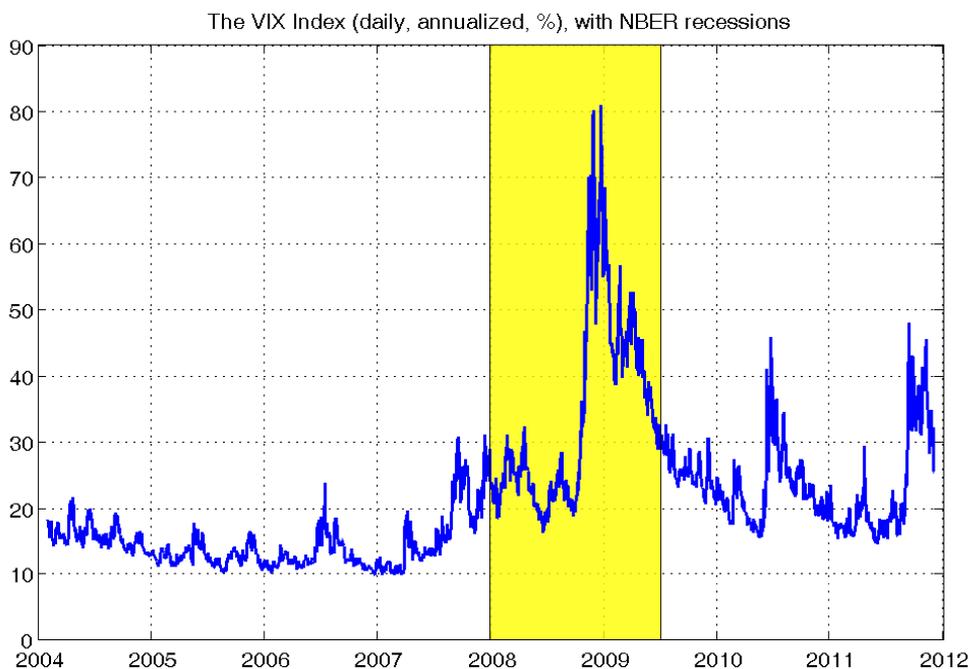
$$\pi_T^{\text{var}} = (\text{var}(t, T) - \mathbb{P}_{\text{var}}(t, T)) \times \text{Notional}, \quad (10.71)$$

where Notional is the notional value of the contract, and $\mathbb{P}_{\text{var}}(t, T)$ is the swap rate agreed at t , and paid off at time T .⁵ Therefore, this contract is a forward, not a swap really. If r is deterministic,

$$\mathbb{P}_{\text{var}}(t, T) = \mathbb{E}[\text{var}(t, T)],$$

where $\mathbb{E}[\text{var}(t, T)]$ is given by Eq. (10.69). Therefore, (10.69) is used to evaluate these variance swaps. Finally, it is worth mentioning that the previous contracts rely on some notions of *realized* volatility as a continuous record of returns is obviously unavailable. Sometimes it is said that variance swaps are profitable to protection sellers, because “The derivative house has the statistical edge,” meaning that realized variance from t to T , say, is general lower than future expected variance under the risk-neutral probability, reflecting variance risk-premiums.

The following picture depicts the development of the new VIX index since its inception.



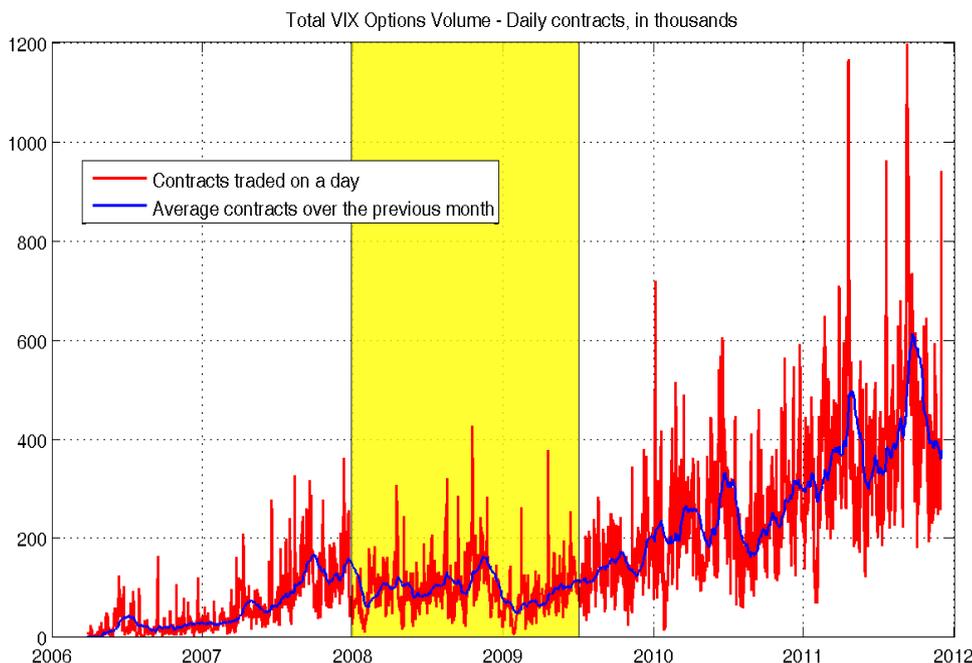
The trading of derivatives based on the new VIX index has increased at a very fast pace, because it aims to replace expensive straddles and makes books less messy with outcomes consistent with views—eliminates price dependency. It is not a mere theoretical curiosity. The following table depicts transaction data for options on the VIX index as compared to other options cleared by CBOE. (Note that the notionals relating to VIX options and futures are not the same, being \$1000 for VIX futures and \$100 for VIX options, as of August 2011.)

⁵A market practice has long been to define the variance notional in such a way that $\text{Notional} = \text{Vega Notional} / 2\sqrt{\text{var-p}}$, where Vega Notional is the notional expressed in volatility percentage points. Suppose, for example, that realized volatility is 1 “vega” (i.e., one volatility point) above the square root of the variance swap rate, $\text{var}(t, T) = (\sqrt{\text{var-p}(t, T)} + 1)^2$, such that $\pi_T^{\text{var}} = (1 + \frac{1}{2\sqrt{\text{var-p}(t, T)}}) \times \text{Vega Notional} \approx \text{Vega Notional}$. The Vega Notional is, then, approximately, the notional for each vega realized volatility exceeds the square root of the variance swap rate.

CBOE trading volume (contracts)—AVERAGE PER DAY, AUGUST 2011

Total trading volume	~ 12,000,000	
CBOE Index Options	~ 2,000,000	(i)
S&P 500 Options	~ 1,300,000	(ii)
CBOE VIX Options	~ 582,000	~ $\frac{1}{6}$ of (i)+(ii) ~ $\frac{1}{2}$ of (ii)
CBOE VIX Futures	~ 79,000	

The following picture provides information about the actual developments of volume on VIX options.



Section 10.6.3 explains how the skew relates to local volatility, but how is the expected variance in Eq. (10.69) related to the skew? Demeterfi, Derman, Kamal and Zou (1999) show that if the implied volatility varies linearly with the strike,

$$IV = IV_{\text{atm}} - b \frac{K - S}{S},$$

for some constant b , then,

$$\frac{1}{T - t} \mathbb{E} [\text{var} (t, T)] \approx IV_{\text{atm}}^2 \cdot (1 + 3 (T - t) b^2).$$

That is, the existence of a skew, $b \neq 0$, increases the value of the fair variance above the at-the-money implied volatility.

10.8.2 Forward volatility trading

Let us consider the following example of structured volatility trading. Suppose we hold the view that market volatility will rise in one year time, to an extent that is inefficiently priced in by the term structure of the currently traded variance swaps. Precisely, our view is that the spot price of the variance swap in one year will exceed the “implied forward variance swap price,” i.e.

$$\mathbb{P}_{\text{var}}(1, 2) > \mathbb{P}_{\text{var}}(0, 2) - \mathbb{P}_{\text{var}}(0, 1). \quad (10.72)$$

To implement a trade consistent with this view, we may proceed as follows:

- (i) long a two year variance swap, struck at $\mathbb{P}_{\text{var}}(0, 2)$, with notional one
 - (ii) short a one year variance swap, struck at $\mathbb{P}_{\text{var}}(0, 1)$, with notional e^{-r}
- (10.73)

Obviously, this strategy does not cost, at time zero.

The strategy in 10.73 generates profits whenever Eq. (10.72) holds true. Indeed, suppose Eq. (10.72) holds true at time 1. Then, come time 1, we can short another one year variance swap, struck at $\mathbb{P}_{\text{var}}(1, 2)$. Intuitively, we do so because “we bought it cheap,” according to Eq. (10.72). Shorting this variance swap at time 1 generates the following payoff at time 2:

$$\pi_1(2) \equiv \mathbb{P}_{\text{var}}(1, 2) - \text{var}(1, 2). \quad (10.74)$$

Moreover, the two year variance swap we went long at time zero (component (i) of 10.73) gives rise to the following payoff at time 2:

$$\pi_2(2) \equiv \text{var}(0, 2) - \mathbb{P}_{\text{var}}(0, 2). \quad (10.75)$$

Adding Eq. (10.74) and Eq. (10.75), and using the relation, $\text{var}(0, 2) = \text{var}(0, 1) + \text{var}(1, 2)$, leads to:

$$\pi(2) \equiv \pi_1(2) + \pi_2(2) = \mathbb{P}_{\text{var}}(1, 2) + \text{var}(0, 1) - \mathbb{P}_{\text{var}}(0, 2).$$

Finally, the one year variance swap with notional e^{-r} we shorted at time zero (component (ii) of 10.73) leads to the following payoff at time 1:

$$\pi(1) \equiv (\mathbb{P}_{\text{var}}(0, 1) - \text{var}(0, 1)) e^{-r}. \quad (10.76)$$

Investing $\pi(1)$ for a further year at the safe interest rate delivers $\pi(1) e^r$ at time 2, such that the total profits at time 2 are:

$$\pi_{\text{tot}} \equiv \pi(2) + \pi(1) e^r = \mathbb{P}_{\text{var}}(1, 2) - \mathbb{P}_{\text{var}}(0, 2) + \mathbb{P}_{\text{var}}(0, 1) > 0, \quad (10.77)$$

where the inequality follows by Eq. (10.72).

10.8.3 Marking to market

Suppose a variance contract expiring at time T is issued at time t , when it is costless. How is this contract worth at time $\tau \in (t, T)$? Let us take the time τ risk-neutral discounted expectation of π_T^{var} in Eq. (10.71),

$$\begin{aligned} \frac{\mathbb{E}_\tau(\pi_T^{\text{var}})}{\text{Notional}} &= e^{-r(T-\tau)} \mathbb{E}_\tau(\text{var}(t, \tau) + \text{var}(\tau, T) - \mathbb{P}_{\text{var}}(t, T)) \\ &= e^{-r(T-\tau)} (\text{var}(t, \tau) + \mathbb{P}_{\text{var}}(\tau, T) - \mathbb{P}_{\text{var}}(t, T)). \end{aligned} \quad (10.78)$$

where \mathbb{E}_τ denotes the risk-neutral expectation conditional upon the information available at time τ .

Marking to market suggests an alternative way to implement the forward volatility trading of the previous section. Suppose, then, again, to have the view that markets for volatility will be such that Eq. (10.72) holds true at time 1, and, accordingly, consider the strategy in 10.73. If Eq. (10.72) holds true at time 1, we may close the position (i) in 10.73 at time 1. By Eq. (10.78), the market value of the two year variance swap we went long at time 0 is,

$$\hat{\pi}(1) \equiv (\text{var}(0, 1) + \mathbb{P}_{\text{var}}(1, 2) - \mathbb{P}_{\text{var}}(0, 2)) e^{-r}. \quad (10.79)$$

At time 1, we obtain $\hat{\pi}(1) + \pi(1)$, which we can invest at the safe interest rate for one additional period, delivering the profit π_{tot} in Eq. (10.77), for time 2.

10.8.4 Stochastic interest rates

When interest rates are stochastic, but still independent of volatility, the expressions given for the contract and indexes do not hold anymore, and there are a number of qualifications, which we make in Remark A.1 of Appendix 5. Moreover, the forward volatility trading strategy in 10.73 should be modified. For example, we might use the following strategy:

- (i) long a two year variance swap, struck at $\mathbb{P}_{\text{var}}(0, 2)$, with notional one
- (ii) short a one year variance swap, struck at $\mathbb{P}_{\text{var}}(0, 1)$, with notional $\frac{P(0,2)}{P(0,1)}$

If come time 1, Eq. (10.72) holds true, we may liquidate (i), thereby accessing the payoff relating to (ii), for a total payoff equal to:

$$\begin{aligned} & (\text{var}(0, 1) + \mathbb{P}_{\text{var}}(1, 2) - \mathbb{P}_{\text{var}}(0, 2)) P(1, 2) + (\mathbb{P}_{\text{var}}(0, 1) - \text{var}(0, 1)) \frac{P(0, 2)}{P(0, 1)} \\ &= (\mathbb{P}_{\text{var}}(1, 2) - \mathbb{P}_{\text{var}}(0, 2) + \mathbb{P}_{\text{var}}(0, 1)) P(1, 2) \\ & \quad + (\text{var}(0, 1) - \mathbb{P}_{\text{var}}(0, 1)) \left(P(1, 2) - \frac{P(0, 2)}{P(0, 1)} \right), \end{aligned}$$

where the first term on the left hand side arises by the liquidation of (i) and by Eq. (10.79), and the second term on the left hand side arises by (ii). By Eq. (10.72), the first term on the right hand side is positive. If the short-term interest rate was deterministic, $P(1, 2) = \frac{P(0,2)}{P(0,1)}$, and the second term on the right hand side would be zero. When interest rates are stochastic, the second term can take on any sign although then, its absolute value should be quite low, compared to the first term on the right hand side.

10.8.5 Hedging

A financial institution might be merely interested in intermediating the contract, which then needs to be hedged against. Suppose, for example, that the financial institution sells protection at time t , thereby promising to pay the realized integrated variance $\text{var}(t, T)$ at time T . We want to replicate this integrated variance. By Itô's lemma:

$$\text{var}(t, T) = 2 \int_t^T \frac{1}{S_u} dS_u - 2 \ln \left(\frac{S_T}{S_t} \right) = 2 \left(\int_t^T \frac{1}{S_u} dS_u - r(T-t) \right) - 2 \ln \frac{F_T}{F_t}. \quad (10.80)$$

The first term can be replicated by continuously rebalancing a stock position, which is always long $\theta_t = \frac{2}{S_t}$ shares of the stock, adjusted for the time value of money. Precisely, consider a self-financed portfolio (θ_τ, ψ_τ) , such that its value satisfies:

$$V_\tau = \theta_\tau S_\tau + \psi_\tau M_\tau,$$

where M_τ denotes the money market account. We choose:

$$\hat{\theta}_\tau = \frac{1}{S_\tau} \frac{M_\tau}{M_T}, \quad \hat{\psi}_\tau = \left[\int_t^\tau \frac{1}{S_u} dS_u - 1 - r(\tau - t) \right] \frac{1}{M_T}. \quad (10.81)$$

It is easy to see that

$$\hat{V}_\tau = \left[\int_t^\tau \frac{1}{S_u} dS_u - r(\tau - t) \right] \frac{M_\tau}{M_T}, \quad (10.82)$$

such that: (i) $\hat{V}_t = 0$, and (ii) $\hat{V}_T = \int_t^T \frac{1}{S_u} dS_u - r(T - t)$. In Appendix 5, we show that $(\hat{\theta}_\tau, \hat{\psi}_\tau)$ is self-financed. The bottom line is that we can hedge the first term in Eq. (10.80) through a self-financed portfolio that costs nothing at time t . This portfolio is simply $(2\hat{\theta}_\tau, 2\hat{\psi}_\tau)$.

To replicate the second term in Eq. (10.80), the payoff of the so-called log-contract, note that we simply have to make reference to twice Eq. (10.70). Therefore, the log-contract can be replicated by shorting $2/F_t$ units of forwards, which are of course costless at time t , and going long a continuum of out-of-the-money options with weights $2/K^2$, which cost

$$2 \int_0^{F_t} \frac{P_t(K, T)}{K^2} dK + 2 \int_{F_t}^\infty \frac{C_t(K, T)}{K^2} dK = e^{-r(T-t)} \mathbb{E}[\text{var}(t, T)]$$

where the equality follows by Eq. (10.69). We borrow $e^{-r(T-t)} \mathbb{E}[\text{var}(t, T)]$ to purchase these options, and once this is done, we are guaranteed $\text{var}(t, T)$ is replicated at time T , as we now have replicated both the first term and the second term in Eq. (10.80). Finally, come time T , we pay back the loan, worth $\mathbb{E}[\text{var}(t, T)]$, and receive a payoff equal to $\text{var}(t, T) - \mathbb{E}[\text{var}(t, T)]$, due to the sale of insurance. Since $\text{var}(t, T)$ is replicated, no additional funds are needed at time T .

10.9 Skewness contracts

We might be interested in left tail risks. Model-free skewness contracts and indexes can be designed and built up to cope with this risk. (These indexes are maintained by the CBOE since January 2011.) Let's take a general perspective. In Appendix 6, we show that for any twice differentiable function f , we have that

$$e^{-r(T-t)} \mathbb{E}[f(F_T) - f(F_t)] = \int_0^{F_t} f''(K) \cdot \text{Put}_t(K, T) dK + \int_{F_t}^\infty f''(K) \cdot \text{Call}_t(K, T) dK, \quad (10.83)$$

where $F_t = e^{r(T-t)} S_t$, the forward rate, and the remaining usual notation. Eq. (10.83) shows that any Markov payoff can be spanned through a set of European options. For example, if $f(x) = \ln x$, we can price a log-contract, which leads to the new VIX index, as explained in the previous section. We are now interested in skewness contracts.

Consider the following payoff, $f_v(F_T) \equiv \left(\ln \frac{F_T}{F_t}\right)^2$. Note that by construction, $f_v(F_t) = 0$, such that Eq. (10.83) allows us to price this payoff as follow,

$$\mathbb{E}[f_v(F_T)] = e^{r(T-t)} \left(\int_0^{F_t} \omega_{v,t}(K) \cdot \text{Put}_t(K, T) dK + \int_{F_t}^{\infty} \omega_{v,t}(K) \cdot \text{Call}_t(K, T) dK \right)$$

$$\omega_{v,t}(K) \equiv f_v''(F_T) = \frac{2}{K^2} \left(1 - \ln \frac{K}{F_t} \right)$$

Which volatility contract does the payoff $f_v(F_T)$ relate to? It's a contract relating to the second moment of the cumulative return $\ln \frac{F_T}{F_t}$, rather than the realized volatility of the previous section, defined as the sum of the instantaneous return variances, $\int_t^T \left(\frac{dF_u}{F_u}\right)^2$. Precisely, note that by Itô's lemma,

$$\mathbb{E}[f_v(F_T)] = \mathbb{E} \left[2 \int_t^T \ln \left(\frac{F_\tau}{F_t} \right) \frac{dF_u}{F_u} + \int_t^T \left(1 - \ln \frac{F_\tau}{F_t} \right) \left(\frac{dF_u}{F_u} \right)^2 \right]. \quad (10.84)$$

This volatility contract is a bit unusual, as the standard notion of volatility we typically price is $\int_t^T \left(\frac{dF_u}{F_u}\right)^2$, rather than the expression on the right hand side of Eq. (10.84).

The current literature and practice on skewness contracts have a similar “cumulative return” flavor. Consider the following payoff, introduced by Bakshi, Kapadia and Madan (2003),

$$f_{\text{sk}}(F_T) \equiv \left[\ln \frac{F_T}{F_t} - \mathbb{E} \left(\ln \frac{F_T}{F_t} \right) \right]^3, \quad f_{\text{sk}}(F_t) = 0.$$

The payoff $f_{\text{sk}}(F_T)$ refers to the third moment of the cumulative return over a certain investment horizon. Instead, the notion of a “realized skewness” would rely on the third moments of the instantaneous returns, averaged over the given investment horizon. Pricing results relating to realized skewness are not available at the time of writing. Let us keep on relying on the payoff $f_{\text{sk}}(F_T)$, and consider the definition of skewness, adjusted for risk,

$$\text{Skew} \equiv \frac{\mathbb{E}[f_{\text{sk}}(F_T)]}{\mathbb{E}[f_{\text{vv}}(F_T)]^{\frac{3}{2}}},$$

where,

$$f_{\text{vv}}(F_T) \equiv \left[\ln \frac{F_T}{F_t} - \mathbb{E} \left(\ln \frac{F_T}{F_t} \right) \right]^2, \quad f_{\text{vv}}(F_t) = 0.$$

and “adjustment for risk” relates to the fact that the expectations in the definition of Skew are taken under the risk-neutral probability. Note that $f_{\text{vv}}(F_T)$ is the de-meanned version of $f_v(F_T)$, and its expectation can be easily found as,

$$\mathbb{E}[f_{\text{vv}}(F_T)] = \mathbb{E}[f_v(F_T)] - \mathbb{E}[f_{\log}(F_T)]^2,$$

where $f_{\log}(F_T) = \ln \frac{F_T}{F_t}$, the log-contract, which is priced in the usual VIX fashion,

$$-\mathbb{E}[f_{\log}(F_T)] = e^{r(T-t)} \left(\int_0^{F_t} \frac{1}{K^2} \cdot \text{Put}_t(K, T) dK + \int_{F_t}^{\infty} \frac{1}{K^2} \cdot \text{Call}_t(K, T) dK \right).$$

Likewise,

$$\mathbb{E}[f_{\text{sk}}(F_T)] = e^{r(T-t)} \left(\int_0^{F_t} \omega_{\text{sk},t}(K) \cdot \text{Put}_t(K, T) dK + \int_{F_t}^{\infty} \omega_{\text{sk},t}(K) \cdot \text{Call}_t(K, T) dK \right)$$

$$\omega_{\text{sk},t}(K) \equiv f_{\text{sk}}''(F_T) = \frac{3}{K^2} \left(\ln \frac{K}{F_t} - \mathbb{E}[f_{\log}(F_T)] \right)$$

10.10 American options

10.10.1 Real options theory

The option can be exercised at any time before the expiry date, T . When the option is exercised, it yields a payoff equal to a function of the underlying asset price, say $\psi(S(t))$. Let \mathcal{C}_t be the price of an American option as of time t . In discrete time, we have:

$$\mathcal{C}_t = \max \{ \psi(S_t), e^{-r\Delta t} \mathbb{E}[\mathcal{C}_{t+\Delta t}] \}.$$

We assume that the nature of the option, summarized by the payoff $\psi(S_t)$, is such that there are two regions, a stopping region and a continuation region, defined as follows:

- (i) *Stopping region*, where time-to-maturity and the price of the asset underlying the option are such that it is optimal to exercise, $\mathcal{C}_t = \max \{ \psi(S_t), e^{-r\Delta t} \mathbb{E}[\mathcal{C}_{t+\Delta t}] \} = \psi(S_t)$, in which case, of course, $\mathcal{C}_t \geq e^{-r\Delta t} \mathbb{E}[\mathcal{C}_{t+\Delta t}]$. By rearranging terms

$$0 \geq e^{-r\Delta t} \frac{\mathbb{E}[\mathcal{C}_{t+\Delta t}] - \mathcal{C}_t}{\Delta t} - \frac{1 - e^{-r\Delta t}}{\Delta t} \mathcal{C}_t. \quad (10.85)$$

The expected return on the option under the risk-neutral probability is less than that on a bank deposit, which further clarifies why it is optimal to exercise early. Naturally, the fact the option is yielding less than the safe interest rate is not an arbitrage. We could simply not short the derivative, as no one else is willing to buy it, as it is not optimal to do so.

- (ii) *Continuation region*, where time-to-maturity and the price of the asset underlying the option are such that it is optimal to wait, $\mathcal{C}_t = \max \{ \psi(S_t), e^{-r\Delta t} \mathbb{E}[\mathcal{C}_{t+\Delta t}] \} = e^{-r\Delta t} \mathbb{E}[\mathcal{C}_{t+\Delta t}]$, or

$$0 = e^{-r\Delta t} \frac{\mathbb{E}[\mathcal{C}_{t+\Delta t}] - \mathcal{C}_t}{\Delta t} - \frac{1 - e^{-r\Delta t}}{\Delta t} \mathcal{C}_t. \quad (10.86)$$

The expected return on the option under the risk-neutral probability is the same as that on a bank deposit.

Note that the existence of these two regions is not guaranteed. For example, we shall see that it is never optimal to exercise early American calls written on assets that do not distribute dividends. When the two regions are, instead, well-defined, they define an exercise “envelope,” a function of the asset price underlying the option and time-to-maturity. It is a “free boundary” problem: we need to find a boundary that triggers some action, in this case, exercising the option, and the boundary is free in that it is not given in advance as in the case of, say, the barrier options of the following section.

This problem can be quite complex, but sometimes, simplifies for those derivatives with an infinite expiry date, T . This simplification arises as in this case, the option price and, hence, the envelope, only depends on the underlying asset price. Under this assumption, and the assumption that the price of the asset underlying the option is a geometric Brownian motion with volatility parameter σ , we have that the option price satisfies, in the limit $\Delta t \rightarrow 0$:

$$\text{Stopping region:} \quad \mathcal{L}[\mathcal{C}] - r\mathcal{C} \leq 0 \text{ and } \mathcal{C} = \psi(S) \quad (10.87)$$

$$\text{Continuation region:} \quad \mathcal{L}[\mathcal{C}] - r\mathcal{C} = 0 \quad (10.88)$$

where $\mathcal{L}[\mathcal{C}] = \frac{1}{2}\sigma^2 S^2 \mathcal{C}_{SS} + rS\mathcal{C}_S$. Eqs. (10.87)-(10.88) are the continuous time counterparts to Eqs. (10.85)-(10.86). To these two equations, we need to add a number of conditions, discussed in the two examples in the subsections below.

10.10.2 Perpetual puts

Consider a perpetual put, where $\psi(S) = (K - S)^+$, and the price p is, accordingly, a function of the underlying asset price S only. This price satisfies Eqs. (10.87)-(10.88), with some additional conditions and qualifications. First, we assume, and later verify, that there exists a value for the asset price, the free boundary, S_* say, such that, it is optimal to exercise the option whenever $S < S_*$. In other terms, Eqs. (10.87)-(10.88) can be written as:

$$\text{Stopping region } (S \leq S_*): \quad p(S_*) = K - S_* \quad (10.89)$$

$$\text{Continuation region } (S > S_*): \quad \mathcal{L}[p] - rp = 0 \quad (10.90)$$

where K is the strike price of the option. Eq. (10.89) is a “value-matching” condition, as explained in Chapter 4 in a related context. It ensures that the pricing function p is continuous as we move from the continuation region towards the stopping region.

Second, we require the following boundary condition:

$$\lim_{S \rightarrow \infty} p(S) = 0. \quad (10.91)$$

That is, as the asset price gets large, the value of the put option needs to approach zero, as the probability the derivative is ever exercised becomes negligible.

Finally, the pricing function, $p(S)$, satisfies the following “smooth-pasting” condition, obtained after taking the derivative in Eq. (10.89), as also explained in Chapter 4:

$$p'(S_*) = -1. \quad (10.92)$$

We conjecture that in the continuation region, the pricing function p that solves Eq. (10.90) has the form $p(S) = AS^\gamma$, for two constants A and γ . Plugging this guess into Eq. (10.90) reveals that actually, the pricing function satisfying it, has the following form:

$$p(S) = A_+ S^{\gamma_+} + A_- S^{\gamma_-}, \quad (10.93)$$

where A_+ and A_- are two constants, to be pinned down, $\gamma_+ = 1$ and $\gamma_- = -\frac{2r}{\sigma^2}$. To satisfy the boundary condition in Eq. (10.91), we need that $A_+ = 0$, which leaves $p(S) = A_- S^{\gamma_-}$. Evaluating this function at S_* , as in Eq. (10.89), and using the smooth pasting condition in Eq. (10.92), yields:

$$\begin{cases} p(S_*) = A_- S_*^{\gamma_-} = K - S_* \\ p'(S_*) = \gamma_- A_- S_*^{\gamma_- - 1} = -1 \end{cases} \quad (10.94)$$

The endogenous variables of this system are the two constants A_- and S_* . We have:

$$S_* = \frac{r}{r + \frac{1}{2}\sigma^2}K, \quad (10.95)$$

and $A_- = (K - S_*)S_*^{-\gamma_-}$, such that

$$p(S) = (K - S_*) \left(\frac{S}{S_*} \right)^{\gamma_-}.$$

A few comments are in order. First, Eq. (10.95) shows that the value to wait increases with σ^2 . Second, when the short-term rate is zero, $S_* = 0$, meaning it is never optimal to exercise, and the option is worthless. Intuitively, in the stopping region, the expected return on the option under the risk-neutral probability is less than that on a bank deposit. When $r = 0$, this expected return is negative, which destroys the time-value of money argument underpinning early exercise.

10.10.3 Perpetual calls

As anticipated, not any payoff gives rise to well-defined stopping and continuation regions, such as those in Eqs. (10.87)-(10.88). For call options, where $\psi(S) = (S - K)^+$, it is never optimal to exercise early, when the underlying assets do not pay dividends. To illustrate, we may literally follow the same arguments of the previous subsection, and find that the call price, $c(S)$, satisfies,

$$\text{Stopping region } (S \geq S_*): c(S_*) = S_* - K \quad (10.96)$$

$$\text{Continuation region } (S < S_*): \mathcal{L}[c] - rc = 0 \quad (10.97)$$

The solution to Eq. (10.97) has the same functional form as that in Eq. (10.93), with the same values of γ_- and γ_+ . However, due to the obvious conjectures about the location of the stopping and continuation regions in Eqs. (10.96)-(10.97), it satisfies the boundary condition $\lim_{S \rightarrow 0} c(S) = 0$, rather than $\lim_{S \rightarrow \infty} c(S) = 0$, as the put price does in Eq. (10.91). Therefore, and because $\gamma_+ = 1$, we must have that $c(S) = A_+S$, or

$$c(S_*) = A_+S_* = S_* - K,$$

where the second equality follows by the value matching condition in Eq. (10.96). Solving for S_* yields,

$$S_* = \frac{K}{1 - A}.$$

As for the smooth-pasting condition we have that $c_S(S_*) = A_+ = 1$, such that $S_* = \infty$. It is never optimal to exercise. In other words, there are no solutions to the counterparts to the two Eqs. (10.94). The call option price fails to simultaneously satisfy the value matching and smooth pasting condition.

If the underlying asset has payouts during the life of the option, say due to storage costs, or even dividends, the problem has, instead, a well-defined solution. Assume that under the risk-neutral probability, the underlying asset price satisfies,

$$\frac{dS_t}{S_t} = (r - q) dt + \sigma \hat{W}_t,$$

where q is the constant payout ratio, $q \equiv \frac{D_t}{S_t}$, with D_t denoting the instantaneous payout (e.g., a dividend), and as usual, \hat{W}_t denotes a Brownian motion under the risk-neutral probability.

For this problem, Eqs. (10.87)-(10.88) reduce to,

$$\text{Stopping region } (S \geq S_*): c(S_*) = S_* - K \quad (10.98)$$

$$\text{Continuation region } (S < S_*): \mathcal{L}_q[c] - rc = 0 \quad (10.99)$$

where now, the infinitesimal generator is $\mathcal{L}_q[c] = \frac{1}{2}\sigma^2 S^2 c_{SS} + (r - q)Sc_S$. We can show the solution to Eq. (10.99) is,

$$c(S) = A_+ S^{\gamma^+} + A_- S^{\gamma^-},$$

where, $\gamma^+ = \frac{1}{\sigma^2} (\alpha + \sqrt{\alpha^2 + 2\sigma^2 r})$, $\gamma^- = \frac{1}{\sigma^2} (\alpha - \sqrt{\alpha^2 + 2\sigma^2 r})$, and $\alpha \equiv q - r + \frac{1}{2}\sigma^2$. Clearly, we have that $\gamma^- < 0$, and $\gamma^+ > 0$, such that the conjectures about the location of the stopping and continuation regions in Eqs. (10.98)-(10.99) deliver the boundary condition $\lim_{S \rightarrow 0} c(S) = 0$, and then,

$$c(S) = A_+ S^{\gamma^+}. \quad (10.100)$$

The value matching condition and the smooth pasting conditions equivalent to the two Eqs. (10.94) are, now

$$\begin{cases} c(S_*) = A_+ S_*^{\gamma^+} = S_* - K \\ c'(S_*) = \gamma_+ A_+ S_*^{\gamma^+ - 1} = 1 \end{cases}$$

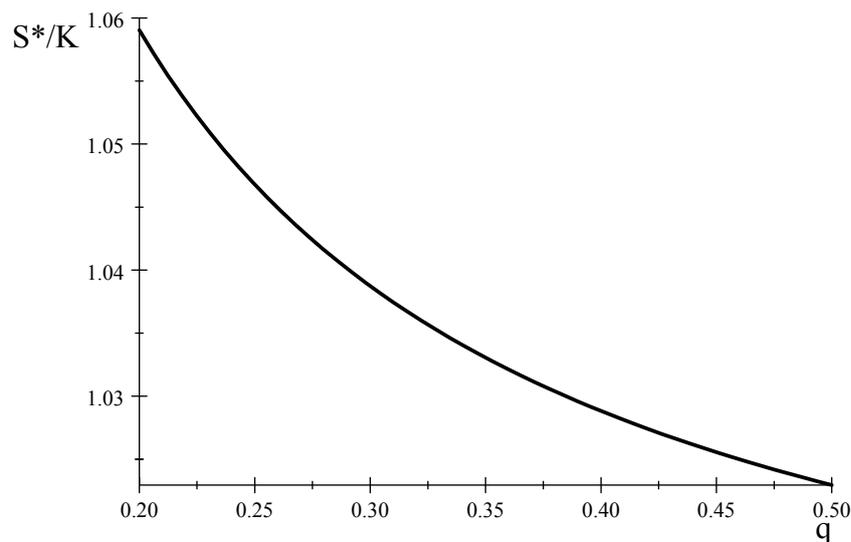
The solution to this system is,

$$A_+ = (S_* - K) S_*^{-\gamma^+}, \quad S_* = \frac{\gamma^+}{\gamma^+ - 1} K$$

such that the call option price in Eq. (10.100) can be written as:

$$c(S) = (S_* - K) \left(\frac{S}{S_*} \right)^{\gamma^+}.$$

The following picture depicts the triggering ratio, $\frac{S^*}{K}$, after which exercise is triggered as a function of q , when $r = 1\%$ and $\sigma = 15\%$.



The triggering ratio blows up as the payout ratio shrinks to zero. In general, the expected optimal stopping time is inversely related to the payout ratio q .

10.11 A few exotics

10.12 Market imperfections

10.13 Appendix 1: The original arguments underlying the Black & Scholes formula

The arguments in Black and Scholes (1973) and Merton (1973) rely on the assumption the option is traded. Accordingly, create a self-financed portfolio of \bar{n}_S units of the underlying asset and n_C units of the European call option, where \bar{n}_S is an arbitrary number. Such a portfolio is worth $V = \bar{n}_S S + n_C C$ and since it is self-financed it satisfies:

$$\begin{aligned} dV &= \bar{n}_S dS + n_C dC \\ &= \bar{n}_S dS + n_C \left[C_S dS + \left(C_\tau + \frac{1}{2} \sigma^2 S^2 C_{SS} \right) d\tau \right] \\ &= (\bar{n}_S + n_C C_S) dS + n_C \left(C_\tau + \frac{1}{2} \sigma^2 S^2 C_{SS} \right) d\tau \end{aligned}$$

where the second line follows by Itô's lemma. Therefore, the portfolio is locally riskless whenever

$$n_C = -\bar{n}_S \frac{1}{C_S},$$

in which case V must appreciate at the r -rate

$$\frac{dV}{V} = \frac{n_C (C_\tau + \frac{1}{2} \sigma^2 S^2 C_{SS}) d\tau}{\bar{n}_S S + n_C C} = \frac{-\frac{1}{C_S} (C_\tau + \frac{1}{2} \sigma^2 S^2 C_{SS})}{S - \frac{1}{C_S} C} d\tau = r d\tau.$$

The last equality, and the boundary condition, lead to the Black-Scholes partial differential equation (10.13).

10.14 Appendix 2: Black (1976)

We want to evaluate the following expectation:

$$\mathbb{E} [x(T) - K]^+,$$

where

$$x(T) = x(t)e^{-\frac{1}{2} \int_t^T \gamma^2(\tau) d\tau + \int_t^T \gamma(\tau) d\tilde{W}(\tau)}.$$

Let \mathbb{I}_{exe} be the indicator of all events s.t. $x(T) \geq K$. We have

$$\begin{aligned} \mathbb{E}_t [x(T) - K]^+ &= \mathbb{E}_t [x(T) \cdot \mathbb{I}_{\text{exe}}] - K \cdot \mathbb{E}_t [\mathbb{I}_{\text{exe}}] \\ &= x(t) \cdot \mathbb{E}_t \left[\frac{x(T)}{x(t)} \cdot \mathbb{I}_{\text{exe}} \right] - K \cdot \mathbb{E}_t [\mathbb{I}_{\text{exe}}] \\ &= x(t) \cdot \mathbb{E}_{Q^x} [\mathbb{I}_{\text{exe}}] - K \cdot \mathbb{E}_t [\mathbb{I}_{\text{exe}}] \\ &= x(t) \cdot Q^x (x(T) \geq K) - K \cdot Q (x(T) \geq K). \end{aligned}$$

where the probability Q^x is defined as:

$$\frac{dQ^x}{dQ} = \frac{x(T)}{x(t)} = e^{-\frac{1}{2} \int_t^T \gamma^2(\tau) d\tau + \int_t^T \gamma(\tau) d\tilde{W}(\tau)},$$

a Q -martingale starting at one. Under Q^x , $dW^x(\tau) = d\tilde{W}(\tau) - \gamma d\tau$ is a Brownian motion, such that

$$x(T) = x(t) e^{\frac{1}{2} \int_t^T \gamma^2(\tau) d\tau + \int_t^T \gamma(\tau) dW^x(\tau)}.$$

Note that $Q(x(T) \geq K) = \Phi(d_2)$ and $Q^x(x(T) \geq K) = \Phi(d_1)$, where

$$d_{2/1} = \frac{\ln \left(\frac{x(t)}{K} \right) \mp \frac{1}{2} \int_t^T \gamma^2(\tau) d\tau}{\sqrt{\int_t^T \gamma^2(\tau) d\tau}}.$$

10.15 Appendix 3: Stochastic volatility

10.15.1 Hull & White equation

We assume that the asset price and its volatility solve Eqs. (10.34), but that the two Brownian motions W and W_σ are uncorrelated. We use the Law of Iterated Expectations to elaborate on (10.42), and arrive to Eq. (10.44) as follows:

$$\begin{aligned}
C(S(t), \sigma^2(t), T-t) &= e^{-r(T-t)} \mathbb{E}_t \left[(S(T) - K)^+ \mid S(t), \sigma^2(t) \right] \\
&= \mathbb{E}_t \left[\mathbb{E} \left(e^{-r(T-t)} (S(T) - K)^+ \mid S(t), \{\sigma^2(\tau)\}_{\tau \in [t, T]} \right) \mid S(t), \sigma^2(t) \right] \\
&= \mathbb{E}_t [C_{\text{BS}}(S(t), T-t; K, \sqrt{\tilde{V}_{t,T}}) \mid S(t), \sigma^2(t)] \\
&= \mathbb{E}_t [C_{\text{BS}}(S(t), T-t; K, \sqrt{\tilde{V}_{t,T}}) \mid \sigma^2(t)] \\
&= \int C_{\text{BS}}(S(t), T-t; K, \sqrt{\tilde{V}_{t,T}}) \phi(\tilde{V}_{t,T} \mid \sigma^2(t)) d\tilde{V}_{t,T} \\
&\equiv \mathbb{E}_t^{\tilde{V}} [C_{\text{BS}}(S(t), T-t; K, \sqrt{\tilde{V}_{t,T}})], \tag{10A.1}
\end{aligned}$$

where $\phi(\tilde{V}_{t,T} \mid \sigma^2(t))$ denotes the density of $\tilde{V}_{t,T}$ conditional upon the current level of the variance, $\sigma^2(t)$. The third and fourth equalities follow by the assumption W and W_σ are uncorrelated, such that

$$\phi(\tilde{V}_{t,T} \mid S(t), \sigma^2(t)) = \phi(\tilde{V}_{t,T} \mid \sigma^2(t)),$$

for otherwise the current level of the index, $S(t)$, would help predict $\tilde{V}_{t,T}$. In other terms, conditionally on the variance path $\{\sigma^2(\tau)\}_{\tau \in [t, T]}$, $\ln\left(\frac{S(T)}{S(t)}\right)$ is normally distributed under the risk-neutral probability,

$$\ln\left(\frac{S(T)}{S(t)}\right) = r(T-t) - \frac{1}{2} \int_t^T \sigma^2(\tau) d\tau + \int_t^T \sigma(\tau) dW(\tau),$$

with:

$$\begin{aligned}
\mathbb{E}_t \left[\ln\left(\frac{S(T)}{S(t)}\right) \mid \{\sigma^2(\tau)\}_{\tau \in [t, T]} \right] &= r(T-t) - \frac{1}{2} (T-t) \tilde{V}_{t,T} \\
\text{var}_t \left[\ln\left(\frac{S(T)}{S(t)}\right) \mid \{\sigma^2(\tau)\}_{\tau \in [t, T]} \right] &= \int_t^T \sigma^2(\tau) d\tau = (T-t) \tilde{V}_{t,T}
\end{aligned}$$

Therefore, the Black & Scholes formula can be safely exploited to evaluate the inner expectation of the second line of Eq. (10A.1).

10.15.2 Extensions

Romano and Touzi (1997) extend the Hull & White equation to the case where asset returns and volatility are correlated. Consider the following model:

$$\begin{cases} \frac{dS(t)}{S(t)} &= rdt + \sigma(t) \left(\rho(t) dW_\sigma(t) + \sqrt{1-\rho^2(t)} dZ(t) \right) \\ \frac{d\sigma^2(t)}{d\sigma^2(t)} &= b(\sigma(t)) dt + a(\sigma(t)) dW_\sigma(t) \end{cases}$$

where the correlation process, $\rho(t)$, does not depend on $S(t)$, and W_σ and Z are two independent Brownian motions, such that,

$$S(T) = S(t) L_{t,T} \cdot e^{r(T-t) - \frac{1}{2} \int_t^T \sigma^2(\tau) (1-\rho^2(\tau)) d\tau + \int_t^T \sigma(\tau) \sqrt{1-\rho^2(\tau)} dZ(\tau)},$$

where $L_{t,T}$ is as in Eq. (10.45) of the main text. We have, using the Law of Iterated Expectations,

$$\begin{aligned} C(S(t), \sigma^2(t), \rho(t), T-t) &= e^{-r(T-t)} \mathbb{E}_t[(S(T) - K)^+ | S(t), \sigma^2(t), \rho(t)] \\ &= \mathbb{E}_t \left[\mathbb{E} \left(e^{-r(T-t)} (S(T) - K)^+ | S(t), \{\sigma^2(\tau), \rho(\tau)\}_{\tau \in [t, T]} \right) \middle| S(t), \sigma^2(t), \rho(t) \right] \\ &= \mathbb{E}_t [C_{\text{BS}}(S(t), L_{t,T}, T-t; K, \sqrt{\tilde{V}_{t,T}^\rho}) | S(t), \sigma^2(t), \rho(t)], \end{aligned}$$

where $\tilde{V}_{t,T}^\rho$ is as in Eq. (10.45) of the main text. The third equality follows because by assumption, both the variance and correlation processes are independent of $\{Z(\tau)\}_{\tau \in [t, T]}$, such that conditionally upon the variance and the correlation paths, $\{\sigma^2(\tau)\}_{\tau \in [t, T]}$ and $\{\rho(\tau)\}_{\tau \in [t, T]}$, $\ln \left(\frac{S(T)}{S(t)L_{t,T}} \right)$ is normally distributed under the risk-neutral probability,

$$\ln \left(\frac{S(T)}{S(t)L_{t,T}} \right) = r(T-t) - \frac{1}{2} \int_t^T \sigma^2(\tau) (1 - \rho^2(\tau)) d\tau + \int_t^T \sigma(\tau) \sqrt{1 - \rho^2(\tau)} dZ(\tau),$$

with

$$\begin{aligned} \mathbb{E}_t \left[\ln \left(\frac{S(T)}{S(t)L_{t,T}} \right) \middle| \{\sigma^2(\tau)\}_{\tau \in [t, T]} \right] &= r(T-t) - \frac{1}{2} (T-t) \tilde{V}_{t,T}^\rho \\ \text{var}_t \left[\ln \left(\frac{S(T)}{S(t)L_{t,T}} \right) \middle| \{\sigma^2(\tau)\}_{\tau \in [t, T]} \right] &= \int_t^T \sigma^2(\tau) d\tau = (T-t) \tilde{V}_{t,T}^\rho \end{aligned}$$

The fourth line follows by the same arguments leading to Eq. (10A.1).

10.15.3 Smile analytics

We develop the approximation stated in Eq. (10.35). The assumption is that the asset price and its volatility solve Eqs. (10.34), with the Brownians W and W_σ being uncorrelated, such that the market price of the option is that generated by the Hull & White equation (10.44). By definition, the Black & Scholes implied volatility, IV , satisfies,

$$C_t^\$ = \mathbb{E}_t [C_{\text{BS}}(S, T-t, K; \sqrt{\tilde{V}_{t,T}})] = C_{\text{BS}}(S, T-t; \text{IV}(X, T-t)). \quad (10A.2)$$

Let $\mu_V(t) = \mathbb{E}_t(\sqrt{\tilde{V}_{t,T}})$ the expected average volatility, and consider a Taylor's second order expansion of the Black & Scholes function about $\mu_V(t)$,

$$\begin{aligned} \mathbb{E}_t [C_{\text{BS}}(S, T-t, K; \sqrt{\tilde{V}_{t,T}})] &\approx C_{\text{BS}}(S, T-t, K; \sqrt{\mu_V(t)}) + \frac{1}{2} \frac{\partial^2 C_{\text{BS}}(S, T-t, K, \sqrt{V})}{\partial \sqrt{V}^2} \bigg|_{\sqrt{V}=\mu_V(t)} \cdot \text{var}_t(\sqrt{\tilde{V}_{t,T}}), \end{aligned}$$

and,

$$\begin{aligned} C_{\text{BS}}(S, T-t; \text{IV}(X, T-t)) &\approx C_{\text{BS}}(S, T-t, K; \sqrt{\mu_V(t)}) + \frac{\partial C_{\text{BS}}(S, T-t, K, \sqrt{V})}{\partial \sqrt{V}} \bigg|_{\sqrt{V}=\mu_V(t)} \cdot (\text{IV}(X, T-t) - \mu_V(t)). \end{aligned}$$

By plugging these two approximations into Eq. (10A.2) leaves:

$$\text{IV}(X, T-t) \approx \mu_V(t) + \frac{1}{2} \frac{\partial^2 C_{\text{BS}}(S, T-t; K, \sigma) / \partial \sigma^2}{\partial C_{\text{BS}}(S, T-t; K, \sigma) / \partial \sigma} \Bigg|_{\sigma=\mu_V(t)} \cdot \text{var}_t(\sqrt{\tilde{V}_{t,T}}). \quad (10A.3)$$

The vega, $\partial C_{\text{BS}} / \partial \sigma$, and the volga, $\partial^2 C_{\text{BS}} / \partial \sigma^2$, for the Black-Scholes model are well-known. They are:

$$\begin{aligned} \frac{\partial C_{\text{BS}}(S, T-t; \sigma)}{\partial \sigma} &= S\sqrt{T-t} \phi\left(\frac{X + \frac{1}{2}\sigma^2(T-t)}{\sigma\sqrt{T-t}}\right) \\ \frac{\partial^2 C_{\text{BS}}(S, T-t; K, \sigma)}{\partial \sigma^2} &= \left(\frac{X^2 - (\frac{1}{2}\sigma^2(T-t))^2}{\sigma^3(T-t)}\right) \frac{\partial C_{\text{BS}}(S, T-t; K, \sigma)}{\partial \sigma}. \end{aligned}$$

Replacing these expressions into Eq. (10A.3) yields the approximation in Eq. (10.35) of the main text.

10.16 Appendix 4: Local volatility

In all the proofs to follow, all expectations are taken to be expectations conditional on \mathcal{F}_t . However, to simplify notation, we simply write $\mathbb{E}(\cdot|\cdot) \equiv \mathbb{E}(\cdot|\cdot, \mathcal{F}_t)$.

PROOF OF EQS. (10.60) AND (10.67). We first derive Eq. (10.60), a result encompassing Eq. (10.57). By assumption,

$$\frac{dS_t}{S_t} = rdt + \sigma_t d\hat{W}_t,$$

where σ_t is some \mathcal{F}_t -adapted process. For example, $\sigma_t \equiv \sigma(S_t, t) \cdot v_t$, all t , where v_t is solution to the second equation in (10.59). Next, by assumption we are observing a set of option prices $C(K, T)$ with a continuum of strikes K and maturities T . We have,

$$C(K, T) = e^{-r(T-t)} \mathbb{E}(S_T - K)^+, \quad (10A.4)$$

and

$$\frac{\partial}{\partial K} C(K, T) = -e^{-r(T-t)} \mathbb{E}(\mathbb{I}_{S_T \geq K}). \quad (10A.5)$$

For fixed K ,

$$d_T(S_T - K)^+ = \left[\mathbb{I}_{S_T \geq K} r S_T + \frac{1}{2} \delta(S_T - K) \sigma_T^2 S_T^2 \right] dT + \mathbb{I}_{S_T \geq K} \sigma_T S_T d\hat{W}_T,$$

where δ is the Dirac's delta. Hence, by using the identity, $(S_T - K)^+ + K \mathbb{I}_{S_T \geq K} = S_T \mathbb{I}_{S_T \geq K}$,

$$\frac{d\mathbb{E}(S_T - K)^+}{dT} = r \left[\mathbb{E}(S_T - K)^+ + K \mathbb{E}(\mathbb{I}_{S_T \geq K}) \right] + \frac{1}{2} \mathbb{E}[\delta(S_T - K) \sigma_T^2 S_T^2].$$

By multiplying throughout by $e^{-r(T-t)}$, and using Eqs. (10A.4)-(10A.5),

$$e^{-r(T-t)} \frac{d\mathbb{E}(S_T - K)^+}{dT} = r \left[C(K, T) - K \frac{\partial C(K, T)}{\partial K} \right] + \frac{1}{2} e^{-r(T-t)} \mathbb{E}[\delta(S_T - K) \sigma_T^2 S_T^2]. \quad (10A.6)$$

We have,

$$\begin{aligned} \mathbb{E}[\delta(S_T - K) \sigma_T^2 S_T^2] &= \iint \delta(S_T - K) \sigma_T^2 S_T^2 \underbrace{\phi_T(\sigma_T | S_T) \phi_T(S_T)}_{\equiv \text{joint density of } (\sigma_T, S_T)} dS_T d\sigma_T \\ &= \int \sigma_T^2 \left[\int \delta(S_T - K) S_T^2 \phi_T(S_T) \phi_T(\sigma_T | S_T) dS_T \right] d\sigma_T \\ &= K^2 \phi_T(K) \int \sigma_T^2 \phi_T(\sigma_T | S_T = K) d\sigma_T \\ &\equiv K^2 \phi_T(K) \mathbb{E}[\sigma_T^2 | S_T = K]. \end{aligned}$$

By replacing this result into Eq. (10A.6), and using the famous relation in Eq. (10.27) of the main text,

$$\frac{\partial^2 C(K, T)}{\partial K^2} = e^{-r(T-t)} \phi_T(K), \quad (10A.7)$$

we obtain

$$e^{-r(T-t)} \frac{d\mathbb{E}(S_T - K)^+}{dT} = r \left[C(K, T) - K \frac{\partial C(K, T)}{\partial K} \right] + \frac{1}{2} K^2 \frac{\partial^2 C(K, T)}{\partial K^2} \mathbb{E}[\sigma_T^2 | S_T = K]. \quad (10A.8)$$

We also have,

$$\frac{\partial}{\partial T} C(K, T) = -rC(K, T) + e^{-r(T-t)} \frac{\partial \mathbb{E}(S_T - K)^+}{\partial T}.$$

Therefore, by replacing the previous equality into Eq. (10A.8), and by rearranging terms,

$$\frac{\partial}{\partial T} C(K, T) = -rK \frac{\partial C(K, T)}{\partial K} + \frac{1}{2} K^2 \frac{\partial^2 C(K, T)}{\partial K^2} \mathbb{E}[\sigma_T^2 | S_T = K].$$

This is,

$$\mathbb{E}[\sigma_T^2 | S_T = K] = 2 \frac{\frac{\partial C(K, T)}{\partial T} + rK \frac{\partial C(K, T)}{\partial K}}{K^2 \frac{\partial^2 C(K, T)}{\partial K^2}} \equiv \sigma_{\text{loc}}^2(K, T). \quad (10A.9)$$

That is, if Eq. (10A.4) holds true, volatility must be restricted to satisfy Eq. (10A.9). As an example, let $\sigma_t \equiv \sigma(S_t, t) \cdot v_t$, where v_t is solution to the second of Eqs. (10.59). Then,

$$\begin{aligned} \sigma_{\text{loc}}^2(K, T) &= \mathbb{E}[\sigma_T^2 | S_T = K] \\ &= \mathbb{E}[\sigma^2(S_T, T) \cdot v_T^2 | S_T = K] \\ &= \sigma^2(K, T) \mathbb{E}[v_T^2 | S_T = K] \\ &\equiv \tilde{\sigma}_{\text{loc}}^2(K, T) \mathbb{E}[v_T^2 | S_T = K], \end{aligned}$$

which proves Eq. (10.60).

10.17 Appendix 5: Volatility contracts

We provide proofs of results relating to volatility contracts.

PROOF OF EQ. (10.67). We have,

$$\begin{aligned}\mathbb{E}(\sigma_T^2) &= \int_0^\infty \mathbb{E}[\sigma_T^2 | S_T = K] \phi_T(K) dK \\ &= 2 \int_0^\infty \frac{\frac{\partial C(K,T)}{\partial T} + rK \frac{\partial C(K,T)}{\partial K}}{K^2 \frac{\partial^2 C(K,T)}{\partial K^2}} \phi_T(K) dK \\ &= 2e^{r(T-t)} \int_0^\infty \frac{\frac{\partial C(K,T)}{\partial T} + rK \frac{\partial C(K,T)}{\partial K}}{K^2} dK,\end{aligned}$$

where the second line follows by Eq. (10A.9), and the third line follows by Eq. (10A.7). This proves Eq. (10.67). ■

PROOF OF EQ. (10.69). By a Taylor expansion with remainder, we have that for any function f smooth enough,

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \int_{x_0}^x (x - t) f''(t) dt. \quad (10A.11)$$

Let F_t be the forward rate, $F_t = e^{r(T-t)} S_t$. By applying this formula to $\ln F_T$,

$$\begin{aligned}\ln F_T &= \ln F_t + \frac{1}{F_t} (F_T - F_t) - \int_{F_t}^{F_T} (F_T - t) \frac{1}{t^2} dt \\ &= \ln F_t + \frac{1}{F_t} (F_T - F_t) - \int_0^{F_t} (K - F_T)^+ \frac{1}{K^2} dK - \int_{F_t}^\infty (F_T - K)^+ \frac{1}{K^2} dK \\ &= \ln F_t + \frac{1}{F_t} (F_T - F_t) - \int_0^{F_t} (K - S_T)^+ \frac{1}{K^2} dK - \int_{F_t}^\infty (S_T - K)^+ \frac{1}{K^2} dK,\end{aligned} \quad (10A.11)$$

where the second equality follows because $\int_{x_0}^x (x - t) \frac{1}{t^2} dt = \int_0^{x_0} (t - x)^+ \frac{1}{t^2} dt + \int_{x_0}^\infty (x - t)^+ \frac{1}{t^2} dt$, and the third equality follows because the forward price at T satisfies $F_T = S_T$. Hence, by $\mathbb{E}(F_T) = F_t$,

$$-\mathbb{E}\left(\ln \frac{F_T}{F_t}\right) = e^{r(T-t)} \left[\int_0^{F_t} \frac{P_t(K, T)}{K^2} dK + \int_{F_t}^\infty \frac{C_t(K, T)}{K^2} dK \right]. \quad (10A.12)$$

On the other hand, by Itô's lemma,

$$\mathbb{E}\left(\int_t^T \sigma_u^2 du\right) = -2\mathbb{E}\left(\ln \frac{F_T}{F_t}\right). \quad (10A.13)$$

By replacing Eq. (10A.13) this formula into Eq. (10A.12) yields Eq. (10.69). ■

REMARK A1. The previous proof results hold when the short-term rate is constant. The case of stochastic interest rates can actually be dealt with, although with some tools, which will be introduced more systematically in Chapter 12 (Section 12.2). We anticipate how these tools work in the present appendix, as they allow us to solve for the fair price of variance contracts even when interest rates are stochastic. Note that if interest rates are stochastic, Eq. (10A.12) generalizes to:

$$-\mathbb{E}\left(e^{-\int_t^T r_\tau d\tau} \ln \frac{F_T}{F_t}\right) = -\mathbb{E}\left[e^{-\int_t^T r_\tau d\tau} \left(\frac{F_T - F_t}{F_t}\right)\right] + \int_0^{F_t} \frac{P_t(K, T)}{K^2} dK + \int_{F_t}^\infty \frac{C_t(K, T)}{K^2} dK. \quad (10A.14)$$

The left hand side of Eq. (10A.14) can be written as

$$\mathbb{E} \left(e^{-\int_t^T r_\tau d\tau} \ln \frac{F_T}{F_t} \right) = P(t, T) \mathbb{E} \left(\frac{e^{-\int_t^T r_\tau d\tau}}{P(t, T)} \ln \frac{F_T}{F_t} \right) = P(t, T) \mathbb{E}^{Q^F} \left(\ln \frac{F_T}{F_t} \right), \quad (10A.15)$$

where \mathbb{E}^{Q^F} denotes the expectation taken under a new probability, known as the *forward probability*. Naturally, the first term on the right side of Eq. (10A.14) is zero, as a forward has no value at inception. But then, this zero value condition implies that:

$$F_t = \mathbb{E} \left[\frac{e^{-\int_t^T r_\tau d\tau}}{P(t, T)} F_T \right] = \mathbb{E}^{Q^F} (F_T).$$

That is, the forward price is a martingale under the forward probability. Therefore, Eq. (10A.13) is replaced with,

$$\mathbb{E}^{Q^F} \left(\int_t^T \omega_u^2 du \right) = -2 \mathbb{E}^{Q^F} \left(\ln \frac{F_T}{F_t} \right), \quad (10A.16)$$

where ω_t now denotes the instantaneous volatility of the forward price. By combining Eqs. (10A.14), (10A.15) and (10A.16), we get,

$$\mathbb{E}^{Q^F} \left(\int_t^T \omega_u^2 du \right) = \mathbb{E} \left(e^{-\int_t^T r_\tau d\tau} \int_t^T \omega_u^2 du \right) = \frac{2}{P(t, T)} \left[\int_0^{F_t} \frac{P_t(K, T)}{K^2} dK + \int_{F_t}^\infty \frac{C_t(K, T)}{K^2} dK \right].$$

That is, the fair price of a variance contract for a swap of *forward volatility* can be expressed in a model-free format. Note that it is the price of a variance contract that we can express in a model-free fashion, not the (undiscounted) expected realized variance. Indeed, the payoff of a variance contract for forward realized variance is:

$$\int_t^T \omega_u^2 du - \mathbb{P}_{\text{var}}(t, T),$$

such that the zero value condition at inception,

$$0 = \left[e^{-\int_t^T r_\tau d\tau} \left(\int_t^T \omega_u^2 du - \mathbb{P}_{\text{var}}(t, T) \right) \right],$$

leads to,

$$\mathbb{P}_{\text{var}}(t, T) = \frac{1}{P(t, T)} \left[e^{-\int_t^T r_\tau d\tau} \int_t^T \omega_u^2 du \right] = \mathbb{E}^{Q^F} \left(\int_t^T \omega_u^2 du \right).$$

PROOF THAT $(\hat{\theta}_\tau, \hat{\psi}_\tau)$ IN EQ. (10.81) IS SELF-FINANCED. For a portfolio strategy to be self-financed, we need to have $\psi_\tau M_\tau = V_\tau - \theta_\tau S_\tau$ and $dV_\tau = \theta_\tau dS_\tau + \psi_\tau dM_\tau$, or:

$$dV_\tau = \theta_\tau S_\tau \frac{dS_\tau}{S_\tau} + \psi_\tau M_\tau \frac{dM_\tau}{M_\tau} = \theta_\tau S_\tau \left(\frac{dS_\tau}{S_\tau} - r d\tau \right) + r V_\tau d\tau, \quad (10A.17)$$

where the second line follows by $\psi_\tau M_\tau = V_\tau - \theta_\tau S_\tau$. With $(\hat{\theta}_\tau, \hat{\psi}_\tau)$, we have that:

$$\begin{aligned} d\hat{V}_\tau &= \hat{\theta}_\tau dS_\tau + \hat{\psi}_\tau dM_\tau \\ &= \frac{M_\tau}{M_T} \frac{dS_\tau}{S_\tau} + \hat{\psi}_\tau M_\tau r d\tau \\ &= \frac{M_\tau}{M_T} \frac{dS_\tau}{S_\tau} + \left(\hat{V}_\tau - \frac{M_\tau}{M_T} \right) r d\tau \\ &= \frac{M_\tau}{M_T} \left(\frac{dS_\tau}{S_\tau} - r d\tau \right) + r \hat{V}_\tau d\tau, \end{aligned} \quad (10A.18)$$

where we have used the portfolio weights in Eq. (10.81) and the expression for the portfolio value \hat{V} in Eq. (10.82). Eq. (10A.18) is the same as Eq. (10A.17), once we use the portfolio weight $\hat{\theta}_\tau$ in Eq. (10.81). Therefore, $(\hat{\theta}_\tau, \hat{\psi}_\tau)$ is self-financed. ■

10.18 Appendix 6: Skewness contracts

By Eq. (10A.11), we have that, for any function f as many times differentiable as we might need,

$$\begin{aligned} f(F_T) &= f(F_t) + f'(F_t)(F_T - F_t) + \int_{F_t}^{F_T} (F_T - t) f''(t) dt \\ &= f(F_t) + f'(F_t)(F_T - F_t) + \int_0^{F_t} f''(K)(K - F_T)^+ dK + \int_{F_t}^{\infty} f''(K)(F_T - K)^+ dK, \end{aligned}$$

where $F_t = e^{r(T-t)}S_t$, the forward rate. Multiplying both sides of this equation by $e^{-r(T-t)}$, and taking expectations, yields Eq. (10.83) in the main text.

References

- Bakshi, G., N. Kapadia, D. Madan (2003): "Stock Return Characteristics, Skew Laws, and Differential Pricing of Individual Equity Options." *Review of Financial Studies* 16, 101-143.
- Ball, C.A. and A. Roma (1994): "Stochastic Volatility Option Pricing." *Journal of Financial and Quantitative Analysis* 29, 589-607.
- Bergman, Y. Z., B. D. Grundy, and Z. Wiener (1996): "General Properties of Option Prices." *Journal of Finance* 51, 1573-1610.
- Black, F. (1976a): "The Pricing of Commodity Contracts." *Journal of Financial Economics* 3, 167-179.
- Black, F. (1976b): "Studies of Stock Price Volatility Changes." *Proceedings of the 1976 Meeting of the American Statistical Association*, 177-81.
- Black, F. and M. Scholes (1973): "The Pricing of Options and Corporate Liabilities." *Journal of Political Economy* 81, 637-659.
- Bollerslev, T. (1986): "Generalized Autoregressive Conditional Heteroskedasticity." *Journal of Econometrics* 31, 307-327.
- Bollerslev, T., Engle, R. and D. Nelson (1994): "ARCH Models." In: McFadden, D. and R. Engle (Editors): *Handbook of Econometrics* (Volume 4), 2959-3038. Amsterdam, North-Holland
- Christie, A.A. (1982): "The Stochastic Behavior of Common Stock Variances: Value, Leverage, and Interest Rate Effects." *Journal of Financial Economics* 10, 407-432.
- Clark, P. K. (1973): "A Subordinated Stochastic Process Model with Fixed Variance for Speculative Prices." *Econometrica* 41, 135-156.
- Corradi, V. (2000): "Reconsidering the Continuous Time Limit of the GARCH(1,1) Process." *Journal of Econometrics* 96, 145-153.
- Cox, J. C., J. E. Ingersoll and S. A. Ross (1985): "A Theory of the Term Structure of Interest Rates." *Econometrica* 53, 385-407.
- Demeterfi, K., E. Derman, M. Kamal and J. Zou (1999): "More Than You Ever Wanted To Know About Volatility Swaps." Goldman Sachs Quantitative Strategies Research Notes.
- Duffie, D. and C-f. Huang (1985): "Implementing Arrow-Debreu Equilibria by Continuous Trading of Few Long-Lived Securities." *Econometrica* 53, 1337-1356.
- El Karoui, N., M. Jeanblanc-Picqué and S. Shreve (1998): "Robustness of the Black and Scholes Formula." *Mathematical Finance* 8, 93-126.
- Engle, R.F. (1982): "Autoregressive Conditional Heteroskedasticity with Estimates of the Variance of United Kingdom Inflation." *Econometrica* 50, 987-1008.

- Fama, E. (1965): “The Behaviour of Stock Market Prices.” *Journal of Business* 38, 34-105.
- Fornari, F. and A. Mele (2006): “Approximating Volatility Diffusions with CEV-ARCH Models.” *Journal of Economic Dynamics and Control* 30, 931-966.
- Gatheral, J. (2006): *The Volatility Surface: A Practitioner’s Guide*. New York: John Wiley and Sons.
- Harrison, J.M. and D.M. Kreps (1979): “Martingales and Arbitrage in Multiperiod Securities Markets.” *Journal of Economic Theory* 20, 381-408.
- Heston, S.L. (1993a): “Invisible Parameters in Option Prices.” *Journal of Finance* 48, 933-947.
- Heston, S.L. (1993b): “A Closed Form Solution for Options with Stochastic Volatility with Applications to Bond and Currency Options.” *Review of Financial Studies* 6, 327-344.
- Hull, J. and A. White (1987): “The Pricing of Options with Stochastic Volatilities.” *Journal of Finance* 42, 281-300.
- Mandelbrot, B. (1963): “The Variation of Certain Speculative Prices.” *Journal of Business* 36, 394-419.
- Mele, A. (1998): *Dynamiques non linéaires, volatilité et équilibre*. Paris: Editions Economica.
- Mele, A. and F. Fornari (2000): *Stochastic Volatility in Financial Markets. Crossing the Bridge to Continuous Time*. Boston: Kluwer Academic Publishers.
- Merton, R. (1973): “Theory of Rational Option Pricing.” *Bell Journal of Economics and Management Science* 4, 637-654.
- Nelson, D.B. (1990): “ARCH Models as Diffusion Approximations.” *Journal of Econometrics* 45, 7-38.
- Nelson, D.B. (1991): “Conditional Heteroskedasticity in Asset Returns: A New Approach.” *Econometrica* 59, 347-370.
- Renault, E. (1997): “Econometric Models of Option Pricing Errors.” In: Kreps, D., Wallis, K. (Editors): *Advances in Economics and Econometrics* (Volume 3), 223-278. Cambridge: Cambridge University Press.
- Romano, M. and N. Touzi (1997): “Contingent Claims and Market Completeness in a Stochastic Volatility Model.” *Mathematical Finance* 7, 399-412.
- Scott, L. (1987): “Option Pricing when the Variance Changes Randomly: Theory, Estimation, and an Application.” *Journal of Financial and Quantitative Analysis* 22, 419-438.
- SEC-CFTC (2010): “Findings Regarding the Market Events of May 6, 2010.” A joint report by the Securities and Exchanges Commission & the Commodity Futures Trading Commission, September.
- Tauchen, G. and M. Pitts (1983): “The Price Variability-Volume Relationship on Speculative Markets.” *Econometrica* 51, 485-505.

Taylor, S. (1986): *Modeling Financial Time Series*. Chichester, UK: Wiley.

Vasicek, O. (1977): “An Equilibrium Characterization of the Term Structure.” *Journal of Financial Economics* 5, 177-188.

Wiggins, J. (1987): “Option Values and Stochastic Volatility: Theory and Empirical Estimates.” *Journal of Financial Economics* 19, 351-372.

The engineering of fixed income securities

11.1 Introduction

This chapter is an introduction to the theory and practice of fixed income securities. Fixed income securities quite differ from equities and equity derivatives. Consider, for example, a simple pure discount bond, which is quite difficult to price, being tied down to the time value of money. The value of a bond reflects intertemporal preferences and beliefs of market participants, which are unobservable and, importantly, not traded. For this reason, it is impossible to relate the price of a bond to the current state of the world in a preference-free format, which we can do, instead, when it comes to relate option prices to the underlying in a complete market setting such as that in Black and Scholes (1973).

This chapter reviews models where we can still price fixed income securities in a preference-free setting. We rely on “no-arbitrage models,” which are the fixed income counterparts to the local volatility models reviewed in Chapter 10, so to speak. Within this framework, we give up modeling the current security prices in the first place. Rather, we take these prices as given, and exploit all the information embedded into them so as to extract risk-neutral probabilities of future price movements. Once risk-neutral probs are reverse-engineered, we can price any interest rate product in a preference-free format. “No-arbitrage models” means that the only assumption we are really making is absence of arbitrage.

The main model that illustrates this way to proceed through closed-form formulae is that of Ho and Lee (1986). The Ho and Lee approach is an elegant way through which models can be calibrated to data while ensuring absence of arbitrage. However, the model relies on unrealistic assumptions, and might lead to negative interest rates. We develop a calibration approach based on the extraction of Arrow-Debreu security prices, which can accommodate for more realistic interest rate developments. Arrow-Debreu securities are abstract securities that only pay off in mutually exclusive states of the world, and their value then naturally relates to the risk-neutral probability of the events where they specifically pay off, as first explained in Chapter 2. While these assets do not necessarily exist, we can extract their shadow value from the price of fixed income securities anyway, which we can use to price any interest rate derivative. We center around these themes and provide many numerical applications that include the pricing of interest rate derivatives such as options on bonds, swaps, caps, callable or convertible bonds,

emphasizing the joint behavior of derivative prices and the underlying—for example, the price of a derivative can tell us about whether the underlying is mispriced. The framework of analysis is in discrete time, and relies on “implied binomial trees,” avoiding as many conceptual intricacies as possible. Chapter 12 deals with more advanced topics in interest rate modeling and derivative evaluation, including the empirical motivation underlying them as well as a systematic analysis through continuous time methods. Finally, this chapter does not cover credit risk, which is the focus of Chapter 13. We now proceed with a number of basic pieces of motivation, explaining more in detail a few of the very issues arising in fixed income markets.

11.1.1 Relative pricing in fixed income markets

While bond prices cannot be given a preference-free representation in terms of the current state of the world, we can still aim to price interest rate derivatives in a preference-free fashion. “Relative pricing” is the keyword, which describes the situation where we price a number of assets given the price of some other assets, while ensuring that there are no arbitrage opportunities, as explained in many other junctures of these lectures, such as those in Chapters 2 and 10. Pricing an option on a bond is quite a different matter than pricing the bond in the first place. The ultimate goal in this specific example is to ensure that the option price is preference-free even if the bond price is not really. It is a challenging task. Consider, for example, the Black & Scholes formula. The reasoning leading to it cannot exactly be applied to evaluate fixed income securities. Indeed, the Black & Scholes model relies on the assumption of a constant volatility of the underlying price. In the context of interest rate derivatives, instead, the volatility of the underlying asset price depends on the maturity of the underlying, as it tends to zero as maturity goes to zero. More generally, pricing and hedging interest rate derivatives requires a model that describes the evolution of the entire term structure of interest rates. The general principles underlying the APT are still the same, though.

11.1.2 Many evaluation paradigms

After reading this introductory chapter, the reader will discover that there are so many methods and models we can use to price interest rate derivatives. It is indeed the case as derivatives houses do really have dozens of models, with houses possibly asking for different prices for the same product. While dozens of new methods are available to price fixed income products, we do not see the emergence of a “single” model to price all of the extant fixed income products. Typically, any bank has a battery of different models, with pieces of this battery possibly fighting for different goals. For example, a bank might display a preference for a certain type of models as a result of (i) its culture and history, or (ii) the particular business is pursuing. For example, in the next chapter, we shall see that to price interest rate options such as “caps,” we may use the market model, which relies on the “Black 76” formula. However, using this model implies that we do not have a closed-form solution for the price of “swaptions,” which can only be solved through numerical methods. If the swaptions business is not important for the bank then, we may safely adopt the market model. [In progress]

11.1.3 Plan of the chapter

The chapter is organized as follows. Section 11.2 through 11.4 develop the basics underlying fixed income securities, such as interest rate and market conventions, devices to smooth out the yield curve, duration, convexity, and an introduction to basic hedging and trading strategies.

Section 11.5 is the first section to deal with models that aim to fit the initial yield curve without errors, using binomial trees. We have two fundamental ways to achieve this goal. As for the first one, developed in Section 11.5, we freeze “scenarios,” i.e. the values of the short-term rate on the branches of a tree, and search for the risk-neutral probs such that the prices predicted by the model agree with the market. As for the second, we fix risk-neutral probs, and search for “scenarios” such that the model and the market are the same. Section 11.6 deals with the second approach, which is the essence of the Ho and Lee (1986) model. Naturally, we may consider situations where we might simultaneously search for probabilities and short-term rate scenarios that make models consistent with the markets. These situations are quite complex and necessitate a general framework of analysis, developed in Section 11.7, and hinging upon calibration through Arrow-Debreu securities. Section 11.8 concludes this chapter and provides numerical examples of how to evaluate bonds with callability and convertibility features, which will receive a more systematic theoretical treatment in the next chapters.

11.2 Markets and interest rate conventions

11.2.1 *Markets for interest rates*

There are three main types of markets for interest rates: (i) LIBOR; (ii) Treasury rate; (iii) Repo rate (or repurchase agreement rate).

11.2.1.1 *LIBOR* (London Interbank Offer Rate) and other interbank rates

Financial institutions trade with each other deposits for maturities ranging from just overnight to one year at a given currency. The LIBOR is the rate at which financial institutions are willing to lend in these markets, on average. It is an average indicative quote of the interbank lending market. It is calculated by Thomson Reuters for ten currencies, and published daily by the British Bankers Association. Instead, the LIBID (London Interbank Bid Rate) is the rate that these financial institutions are prepared to pay to borrow money, on average. Normally, LIBID < LIBOR. The LIBOR is a fundamental point of reference to financial institutions, which look at it as an opportunity cost of capital. Moreover, many fixed income instruments are indexed to the LIBOR: forward rate agreements, interest rate swaps, or variable mortgage rates.

The LIBOR is distinct from the US Federal Funds rate. Banks have to maintain reserves with the Federal Reserve to partially back deposits and to clear financial transactions, as further explained in Section 13.6 of Chapter 13. Transactions involve banks with excess reserves with the Fed, which earn no interest, to banks with reserve deficiencies. The Federal Funds rate is the overnight rate at which banks lend these reserves to each other. The Federal Funds rate is affected by the FDRBNY, which aims to make it lie within a range of the target rate decided by the governors at Federal Open Market Committee meetings. This range is “maintained” through open market operations.

11.2.1.2 *Treasury rate*

It is the rate at which a given Government can borrow at a given currency.

11.2.1.3 *Repo rate* (or repurchase agreement rate)

A Repo agreement is a contract by which one counterparty sells some assets to another one, with the obligation to buy these assets back at some future date. The assets act as collateral.

The rate at which such a transaction is made is the repo rate. One day repo agreements give rise to *overnight repos*. Longer-term agreements give rise to *term repos*.

11.2.1.4 Spreads

Interest rate spreads isolate interesting pieces of information, as they remove common components of the interest rates generating the spreads, which we might not be interested in. An important example stems from the overnight interest swap rate (OIS), which is the swap rate in a swap agreement of fixed against variable interest rate payments, where the variable interest rate payments are made of an overnight reference, typically an average, unsecured interbank overnight rate, such as the Federal Funds rate in the US, SONIA in the UK or EONIA in the Euro area. (See the next chapter, Section 12.8.5 for definitions of swaps and swap rates.) An interesting indicator, then, is the “3-month LIBOR – 3-month OIS” spread, also known as the LIBOR-OIS spread. Because payments relating to overnight rates are not subject to default risk, and the overnight rate is “anchored” to monetary policy, the LIBOR-OIS spread is capable of isolating credit views about financial institutions. It is generally flat, although then it reached high record levels during the 2007 subprime crisis (see Figure 11.1). Instead, the so-called TED (Treasury bill rate minus Eurodollar LIBOR) spread, also captures “flight to quality” effects occurring during times of crisis, when Treasury bonds are considered particularly valuable. For this “flight to quality” reason, the TED spread might fail isolate views about developments in the interbank market.

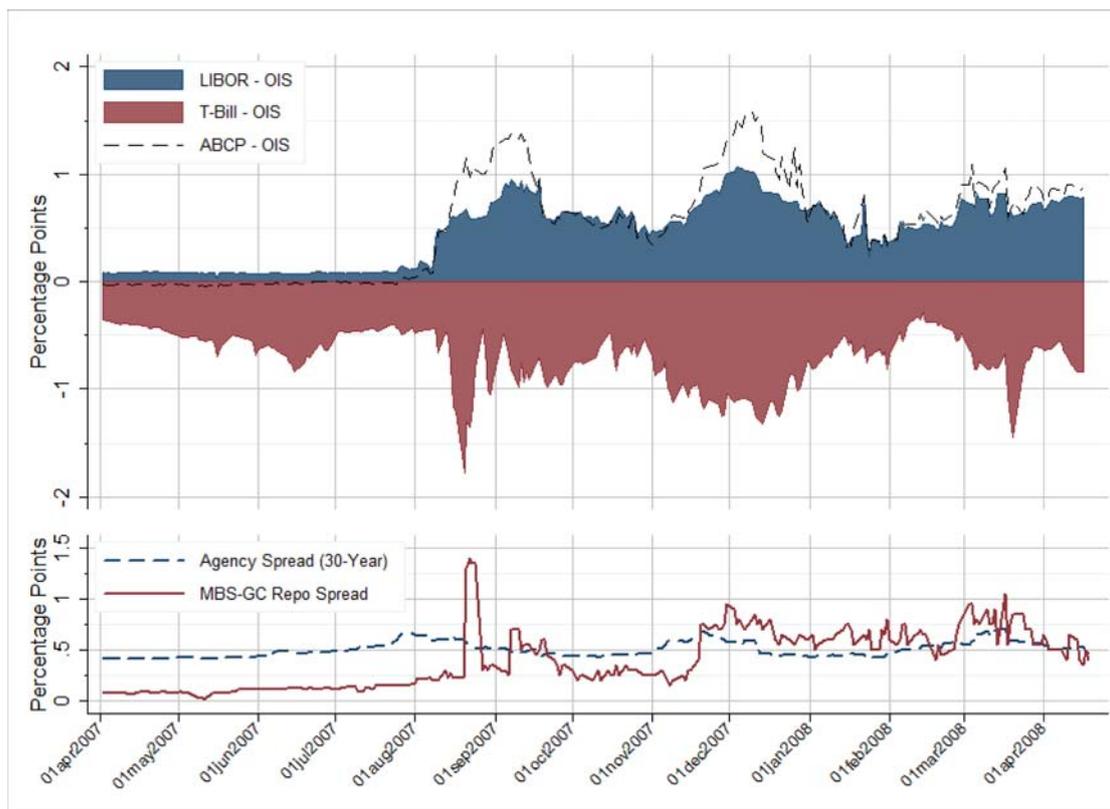


FIGURE 11.1. Antonio Mele does not claim any copyright on this picture, which is taken from Brunnermeier (2009). The picture has been put here for illustrative purposes only, and permission to the author shall be duly asked before the book will be published.

On a historical note, the Federal Funds rate has been the object of much empirical research. In an attempt to explain how the “credit view” contributes to growth more than Friedman’s monetary view, Bernanke and Blinder (1992) show that the Federal Funds rate makes the predicting power of M1 growth insignificant, as we further review in Section 13.6 of Chapter 13. This finding initially spread enthusiasm about the ability of this rate to explain short-run aggregate fluctuations. However, as surveyed for example by Stock and Watson (2003), the explanatory power of the Federal Funds rate evaporizes, once we condition on the term spread, a fact we comment in Section 12.2.2 of the next chapter.

11.2.2 Mathematical definitions of interest rates

11.2.2.1 Simply compounded interest rates

A simply compounded interest rate at time τ , for the time interval $[\tau, T]$, is defined as the solution L to the following equation:

$$P(\tau, T) = \frac{1}{1 + (T - \tau)L(\tau, T)}. \quad (11.1)$$

This definition is intuitive, and is the most widely used in the market practice. As an example, LIBOR rates are computed in this way. In this case, $P(\tau, T)$ is generally interpreted as the initial amount of money to invest at time τ to obtain £ 1 at time T .

11.2.2.2 Yield curves

The *yield-to-maturity*, or *spot rate*, for some maturity date T is the yield on the zero maturing at T , denoted as $r(T)$. This spot rate $r(T)$ is the solution to the following equation,

$$P(t, T) = \frac{1}{(1 + r(t, T))^T}.$$

With semi-annual compounding, we have that $P(t, T) = (1 + \frac{r(t, T)}{2})^{-2T}$. In general, we have that if the interest is compounded m times in a year, at the annual rate r , then, investing for T years gives $(1 + \frac{r}{m})^{mT}$. Continuous compounding is obtained by letting $m \rightarrow \infty$ in the previous expression, leaving e^{rT} . Therefore, the *continuously compounded* spot rate is obtained as:

$$R(t, T) = -\frac{1}{T} \ln P(t, T).$$

It is a sort of “average rate” for investing from time t to time $T > t$. The function, $T \mapsto R(t, T)$, is called the *yield curve*, or the *term structure of interest rates*.

A related and widely used concept is the *par yield curve*. Let $B(t, T)$ be the time t price of a bond that pays off the principal of £1 at expiry T , and a known sequence of constant coupons $C(t, T)$ at $t + 1, t + 2, \dots, T$, such that, in the absence of arbitrage and any other frictions, its price is:

$$B(t, T) = C(t, T) \cdot \sum_{i=1}^{T-t} P(t, t+i) + P(t, T).$$

Note that $C(t, T)$ is fixed at time t . A par bond is one that quotes at parity, $B(t, T) = 100\%$. The par yield curve is the sequence of coupon rates $C(t, T)$, for T varying, that correspond to the par bonds:

$$C(t, T) = \frac{B(t, T) - P(t, T)}{\sum_{i=1}^{T-t} P(t, t+i)}, \quad B(t, T) = 1. \quad (11.2)$$

In other words, the coupon rates $C(t, T)$ have to “adjust” to make the market happy to have the coupon bearing bond quote at par, $B(t, T) = 1$. An interesting interpretation of par-yield is the following. Rewrite Eq. (11.2) as follows: $1 - P(t, T) = C(t, T) \cdot \sum_{i=1}^{T-t} P(t, t+i)$. The right-hand side of this equation is the present value of the flow of known coupons, $C(t, T)$, receivables at the dates $t+1, t+2, \dots, T$. The left-hand side is the present value of the flow of future, and unknown, LIBOR rates, $L(t+1-i, t+i)$ for $i = 1, 2, \dots, T-1$, receivable at the dates $t+1, t+2, \dots, T$,

$$\sum_{i=1}^{T-t} \text{Val} \left(\frac{1}{P(t+i-1, t+i)} - 1 \right) = \sum_{i=1}^{T-t} (P(t, t+i-1) - P(t, t+i)) = 1 - P(t, T),$$

where $\text{Val}(x)$ denotes the current price of receiving a random amount of x dollars at the maturity date where these are due, and where we have used (i) the definition of the LIBOR rate, as defined in Eq. (11.1) and, (ii) a no-arbitrage relation, which we shall show in Section 12.1 of Chapter 12, namely that the present value of $1/P(t+i-1, t+i)$ paid at $t+i$ is simply $P(t, t+i-1)$. Therefore, we can interpret the par yield as the fixed rate in a swap contract that costs nothing at origination, where one counterparty pays another counterparty, the fixed rate against the variable LIBOR—a *spot* swap rate. These swap contracts, of which Section 11.5.4.2 gives a numerical example, are analyzed in detail in Section 12.7.5 of Chapter 12.

11.2.2.3 Forward rates

In a *forward rate agreement* (FRA, henceforth), two counterparties agree that the interest rate on a given principal in a future time-interval $[T, S]$ will be fixed at some level K . Let the principal be normalized to one. The FRA works as follows: at time T , the first counterparty receives $\mathcal{L}1$ from the second counterparty; at time $S > T$, the first counterparty pays back $\mathcal{L}[1 + 1 \cdot (S - T)K]$ to the second counterparty. The amount K is agreed upon at time t . Therefore, the FRA makes it possible to lock-in future interest rates. We consider simply compounded interest rates because this is the standard market practice.

The amount K for which the current value of the FRA is zero is called the *simply-compounded forward rate* as of time t for the time-interval $[T, S]$, and is usually denoted as $F(t, T, S)$. We can use absence of arbitrage to express $F(t, T, S)$ in terms of bond prices, as follows:

$$\frac{P(t, T)}{P(t, S)} = 1 + (S - T) F(t, T, S). \quad (11.3)$$

Indeed, an investor in a zero from time t to time S is one who simultaneously makes (i) a *spot* loan from t to T , and (ii) a *forward* loan from T to S . In the absence of arbitrage, it must be the case that,

$$\underbrace{[1 + r(t, S)]^{S-t}}_{\text{zero loan}} = \underbrace{[1 + r(t, T)]^{T-t}}_{\text{spot loan}} \times \underbrace{[1 + (S - T) F(t, T, S)]}_{\text{forward loan}},$$

where $r(t, S)$ is the spot rate at time t , defined as the solution to, $P(t, S) = 1/[1 + r(t, S)]^{S-t}$. Eq. (11.3) follows by the definition of $r(t, S)$, and by rearranging terms of the previous equality. Alternatively, consider the following portfolio implemented at time t . Go long one bond maturing at T and short $P(t, T)/P(t, S)$ bonds maturing at S , for the time period $[t, S]$. The initial cost of this portfolio is zero because,

$$-P(t, T) + \frac{P(t, T)}{P(t, S)} P(t, S) = 0.$$

At time T , the portfolio yields $\mathcal{L}1$, originating from the bond purchased at time t . At time S , the $P(t, T)/P(t, S)$ bonds shorted at t , and maturing at S , must be purchased. But at time S , the cost of purchasing $P(t, T)/P(t, S)$ bonds maturing at S is obviously $\mathcal{L} P(t, T)/P(t, S)$. The portfolio, therefore, is acting as a FRA: it pays $\mathcal{L}1$ at time T , and $-\mathcal{L} P(t, T)/P(t, S)$ at time S . In addition, the portfolio costs nothing at time t . Therefore, the interest rate implicitly paid in the time-interval $[T, S]$ must be equal to the forward rate $F(t, T, S)$, as stated in Eq. (11.3).

11.2.3 Yields to maturity on coupon bearing bonds

Finally, the yield to maturity \hat{y} (YTM, henceforth) on a bond is simply its rate of return. It is the discount rate that would equate the present value of the stream of payoffs with its market price,

$$\hat{y} : B(T) = \sum_{i=1}^n \frac{C_{t_i}}{(1 + \hat{y})^{t_i}} + \frac{1}{(1 + \hat{y})^T}, \quad (11.4)$$

where $T \equiv t_n$. This formula differs from the price formula $B(T) = \sum_{i=1}^n \frac{C_{t_i}}{(1+r(t_i))^{t_i}} + \frac{1}{(1+r(T))^T}$, as Eq. (11.4) uses the same discount rate \hat{y} to discount the future payments. Clearly, for zeros we have that spot rates coincide with YTM, i.e. $\hat{y} = R(T)$.

Next, suppose that coupon payments are the same for each i , $C_{t_i} = \bar{C}$ say, and the payment dates are set regularly. Eq. (11.4) then collapses to,

$$B(T) = (1 - Z^T) \frac{\bar{C}}{\hat{y}} + Z^T, \quad Z^T \equiv (1 + \hat{y})^{-T}.$$

That is, the price of the coupon bearing bond, $B(T)$, is a convex combination of that of a perpetuity, $\frac{\bar{C}}{\hat{y}}$, and that of a zero expiring at T , Z^T . For large maturities, the bond price gets closer to the perpetuity, whereas for low maturities, it is closer to the zero. If $\bar{C} > \hat{y}$, $B(T) \in \left(1, \frac{\bar{C}}{\hat{y}}\right)$, and if $\bar{C} < \hat{y}$, $B(T) \in \left(\frac{\bar{C}}{\hat{y}}, 1\right)$. In the special case where $\bar{C} = \hat{y}$, the bond would quote at par.

This property is a special case of a more general characteristics of *floating rate bonds*. Floating rate bonds pay coupons equal to the LIBOR, and would quote at par at their first reset date (see Section 12.7.5 in the next chapter). Mathematically, $B(T)$ can be understood as the price of a “floating” rate bond in a market without uncertainty, where the same coupon, \bar{C} , would always be paid. If this coupon is the same as the interest rate we use to discount future cash flows, our coupon bearing bond is, in fact, a “floating” rate bond, and would therefore need to quote at par.

11.3 Curve fitting

11.3.1 Extracting zeros from bond prices

In principle, the zeros can be “extracted” from the market price of the bonds, provided there is a sufficient spread of bonds across maturities. As an example, consider three bonds. The first bond pays off at T_1 , the second bond pays off at T_1, T_2 , the third bond pays off at T_1, T_2, T_3 . By

no-arbitrage,

$$\begin{bmatrix} B(T_1) \\ B(T_2) \\ B(T_3) \end{bmatrix} = \begin{bmatrix} C_{11} + 1 & 0 & 0 \\ C_{21} & C_{22} + 1 & 0 \\ C_{31} & C_{32} & C_{33} + 1 \end{bmatrix} \begin{bmatrix} P(T_1) \\ P(T_2) \\ P(T_3) \end{bmatrix},$$

for some coupons C_{ij} . Therefore, we can use the observed prices $B(t, T_i)$ and the payments C_{ij} to calculate the zeros $P(t, T_i)$ as,

$$\begin{bmatrix} P(T_1) \\ P(T_2) \\ P(T_3) \end{bmatrix} = \begin{bmatrix} C_{11} + 1 & 0 & 0 \\ C_{21} & C_{22} + 1 & 0 \\ C_{31} & C_{32} & C_{33} + 1 \end{bmatrix}^{-1} \begin{bmatrix} B(T_1) \\ B(T_2) \\ B(T_3) \end{bmatrix}. \quad (11.5)$$

The previous procedure can be generalized to the case in which “some maturity is missing.” The resulting algorithm is known as the *bootstrap*, which is described next.

11.3.2 Bootstrapping

Bootstrapping proceeds as follows. Let B_i be the price of a bond paying off coupons at the sequence of dates t_1, t_2, \dots, t_i and a principal of £1 at t_i . Let P_i be the price of the zero maturing at t_i . Then,

- (i) The equation $B_1 = (C_{11} + 1)P_1$ implies that we can extract the zero P_1 as follows, $P_1 = \frac{B_1}{1+C_{11}}$.
- (ii) Given the equation $(C_{22} + 1)P_2 + C_{21}P_1 = B_2$, and the previously computed P_1 , we proceed to extract the zero P_2 as follows, $P_2 = \frac{B_2 - C_{21}P_1}{C_{22} + 1}$.
- (iii) In general, we extract the zero P_n as follows, $P_n = \frac{B_n - \sum_{i=1}^{n-1} C_{ni}P_i}{C_{nn} + 1}$.
- (iv) The previous steps work if we have an ordered number of bonds and all of the maturity dates. Indeed, the previous procedure boils down to the computation of the solution of Eq. (11.5). When some of the maturity dates are not available, we replace the required coupon rate C_{ni} at time t_i with a linear interpolation \hat{C}_{ni} between the coupon $C_{n,i-1}$ at time t_{i-1} and $C_{n,i+1}$ at time t_{i+1} , as follows,

$$\hat{C}_{ni} = \frac{t_{i+1} - t_i}{t_{i+1} - t_{i-1}} C_{n,i-1} + \frac{t_i - t_{i-1}}{t_{i+1} - t_{i-1}} C_{n,i+1}.$$

The effects of the interpolation should be “visible” near the missing maturities.

Next, consider a sequence of coupon bearing bonds maturing at n with fixed coupon streams C_n and define the par yield as in Eq. (11.2), as the fixed sequence C_n such that the price B_n is forced to equal 100%. We can extract the value of the zeros and, then, the yield curve, from step (iii) above, by using the recursive formula,

$$P_n = \frac{B_n - C_n \sum_{i=1}^{n-1} P_i}{C_n + 1}, \quad (11.6)$$

where $B_n = 100\%$. The following table provides a numerical example.

Coupon	Maturity, n	Zero price	$\sum_{i=1}^n P_i$	Yield curve*
6.00%	1	0.9434	0.9434	6.00%
7.00%	2	0.8728	1.8162	7.04%
8.00%	3	0.7914	2.6076	8.11%
9.50%	4	0.6870	3.2946	9.84%
9.00%	5	0.6454	3.9400	9.15%
10.50%	6	0.5306	4.4706	11.14%
11.00%	7	0.4579	4.9285	11.81%
11.25%	8	0.4005	5.3290	12.12%
11.50%	9	0.3472	5.6762	12.47%
11.75%	10	0.2980	not useful	12.87%

*Discretely compounded

11.3.3 Splines

We may use statistical techniques alternative to bootstrapping, to cope with situations in which the number of bonds does not equal the number of maturity dates. Suppose we observe N bonds, where the i -th bond entitles to receive the coupons C_{ij} , for $j = 1, \dots, M_i$. We assume that the bond prices are observed with errors, or

$$B(M_i) = \sum_{j=1}^{M_i} C_{ij}P(t_j) + P(t_{M_i}) + \epsilon_i, \quad i = 1, \dots, N,$$

where ϵ_i is the measurement error for the i -th bond.

We aim to find the curve $T \mapsto P(T)$ that minimizes the errors, in some statistical sense. The natural device is to “parametrize” the function $P(T)$, with a number of k parameters, where $k < N$. To parametrize the function $P(t_j)$ for a generic t_j , we can use *polynomials*, as originally suggested by McCulloch (1971, 1975),

$$P(t_j) = 1 + a_1 t_j + a_2 t_j^2 + \dots + a_k t_j^k,$$

where the a_i are the parameters. *Cubic splines* are polynomials up to the third order, and are very popular. The parameters a_i can be estimated by minimizing the sum of the squared errors, $\sum_{i=1}^N \epsilon_i^2$. A well-known pitfall of polynomials is that a high k might imply that while the polynomial approximation works reasonably well near the observed maturities, it may exhibit an erratic behavior in between. To avoid this problem, we can use *local polynomials*, which are low-order polynomials (typically splines) fitted to non-overlapping subintervals.

Naturally, we may also want to parametrize the spot rates, $R(T)$, as polynomials. Alternatively, Nelson and Siegel (1987) propose the following parametrization,

$$R(T) = \beta_1 + \beta_2 \left(\frac{1 - e^{-\lambda T}}{\lambda T} \right) + \beta_3 \left(\frac{1 - e^{-\lambda T}}{\lambda T} - e^{-\lambda T} \right),$$

where β_i and λ are the parameters. These coefficients may be given an interpretation, in terms of economic factors driving the yield curve, as reviewed in the next chapter. The coefficient β_1 governs the level of the yield curve. The coefficient β_2 relates to the slope, as an increase

in this coefficient increases short yields more than long yields. The coefficient β_3 shapes the curvature, as an increase in this coefficient has little effect on very short and very long yields, but increases the middle of the yield curve. Moreover, the coefficient λ controls the exponential decay of the yield curve: small values of λ translate to slow decay and can better fit the curve at long maturities; large values of λ , instead, lead to a fast decay, which helps fit the short-end of the yield curve. Finally, λ determines where the loading on β_3 achieves its maximum. Diebold and Li (2006) have used this setting to estimate β_i for each date, and then used these estimated time series of β_i to forecast future values of β_i through vector autoregressions and, then, the future yield curve.

11.3.4 Arbitrage

Bond prices need, naturally, to satisfy restrictions preventing arbitrage. We illustrate how an arbitrage opportunity can be exploited, using data in Tuckman (2002) (p. 8-12).

11.3.4.1 Data

Suppose that on some hypothetical date, say February 3009, we observe the bond prices in Table 11.1.

TABLE 11.1.

Treasury Bond Prices on February 15, 3009

Coupon	Maturity	Price
7.875%	8/15/09	101.40
14.250%	2/15/10	108.98
6.375%	8/15/10	102.16
6.250%	2/15/11	102.57
5.250%	8/15/11	100.84

We can bootstrap the price of the zeros implicit in Table 11.1, proceeding as described in Sections 11.3.1 and 11.3.2, obtaining the figures in Table 11.2.

TABLE 11.2.

Implicit zeros on February 15, 3009

Time to maturity	Implicit zero
0.5	$p(0, 0.5) = 0.97557$
1	$p(0, 1) = 0.95247$
1.5	$p(0, 1.5) = 0.93045$
2	$p(0, 2) = 0.90796$
2.5	$p(0, 2.5) = 0.88630$

Next, suppose to observe additional bond prices, those in Table 11.3:

TABLE 11.3.

Treasury Bond Prices on February 15, 3009

Coupon	Maturity	Market price
13.375%	8/15/09	104.080
10.750%	2/15/11	110.938
5.750%	8/15/11	102.020
11.125%	8/15/11	114.375

Are these additional bond prices, those in Table 11.3, compatible with the prices in Table 11.3, in terms of absence of arbitrage opportunities? How could we profit, in a frictionless world, of any arbitrage opportunities “left on the table”?

11.3.4.2 A basic no-arb condition

Let us cast the problem in a more general format. Suppose we observe a vector of N bond prices, with a $N \times N$ matrix C of coupons, where each row of C gives the stream of the coupons promised by a given asset. We know that the $N \times 1$ vector of zeros P , satisfies, $B = CP$. That is, assuming that the matrix C is invertible,

$$P = C^{-1}B. \quad (11.7)$$

Next, suppose there exists some asset that: (i) promises to pay:

$$c^* = [c_1^* \quad c_2^* \quad \cdots \quad c_N^* + 100],$$

and (ii) has a price, b^* , such that:

$$b^* < c^*P. \quad (11.8)$$

The right hand side of this inequality, c^*P , is the “no-arbitrage price” of the asset, which in this example is greater than the market price of the asset. The inequality gives rise to arbitrage opportunity, which can be exploited by going long the asset, and shorting a portfolio “synthesizing” it. To synthesize the asset to go long for, we solve the following system of N equations with N unknowns,

$$\pi C = c^*, \quad (11.9)$$

where the vector of unknowns, π , contains the number of assets in the synthesizing portfolio: by purchasing the portfolio π , one is entitled to receive πC in the future, which we want to equal c^* . The solution to Eq. (11.9) is:

$$\hat{\pi} = c^*C^{-1}. \quad (11.10)$$

Accordingly, the value of this portfolio, V say, is given by,

$$V = \hat{\pi}B = c^*C^{-1}B = c^*P > b^*,$$

where the last equality follows by the “zero pricing equation” (11.7), and the inequality holds by the inequality (11.8).

To summarize, we now have the following situation: (i) the asset we hold produces the cash flows that are needed to pay out the coupons of the “synthesizing” portfolio we sold, and (ii) the price of the asset we go long is less than the value of the portfolio we short. This situation is an arbitrage opportunity, as initially claimed. We now use these insights to check whether arbitrage opportunities exist and, maybe, exploited, using the data in Tables 11.1 through 11.3.

First step: detecting arbitrage opportunities in the market

In a first step, we compute the *no-arbitrage* prices of the bonds in Table 11.3, using the zeros extracted from Table 11.1, and reported in Table 11.2. Denote these prices with B_1 (for the six month 13.375%), B_2 (for the two year 10.750%), B_3 (for the 2.5 year 5.750%), and B_4 (for the 2.5 year 11.125%). For the 13.375% six month bond, we have:

$$B_1 = \left(\frac{13.375}{2} + 100 \right) p(0, 0.5) = 104.08,$$

which matches the market price in Table 11.3. As for the two 10.750% year bond,

$$B_2 = \frac{10.750}{2} [p(0, 0.5) + p(0, 1) + p(0, 1.5)] + \left(\frac{10.750}{2} + 100 \right) p(0, 2) = 111.04.$$

The no-arbitrage price of the 5.75% bond expiring in 2.5 years is:

$$B_3 = \frac{5.75}{2} [p(0, 0.5) + p(0, 1) + p(0, 1.5) + p(0, 2)] + \left(\frac{5.75}{2} + 100 \right) p(0, 2.5) = 102.007.$$

Finally, the no-arbitrage price of the 11.125% bond expiring in 2.5 years is:

$$B_4 = \frac{11.125}{2} [p(0, 0.5) + p(0, 1) + p(0, 1.5) + p(0, 2)] + \left(\frac{11.25}{2} + 100 \right) p(0, 2.5) = 114.511.$$

To summarize,

Treasury Bond Prices on February 15, 3009

Coupon	Maturity	Market price	No-arbitrage price
13.375%	8/15/09	104.080	104.080
10.750%	2/15/11	110.938	111.041
5.750%	8/15/11	102.020	102.007
11.125%	8/15/11	114.375	114.511

While there are no arbitrage opportunities for the 13.375% bond expiring in six months, the price of the 10.750% bond expiring in 2 years is less than its no-arbitrage price: this bond “trades cheap.” In contrast, the 2.5 year 5.750% bond “trades rich,” although the resulting arbitrage does not seem to be quite sensible.

Second step: implementing the arbitrage trade

We now proceed to exploit the mispricing related to the 10.750% bond expiring in 2 years. We use the insights developed in Section 11.3.4.2 to implement the arbitrage. We have, $N = 4$, and $c^* = \left[\frac{10.750}{2} \quad \frac{10.750}{2} \quad \frac{10.750}{2} \quad \frac{10.750}{2} + 100 \right]$, and we use the first four bonds in Table 11.1 to construct an arbitrage portfolio. In terms of the coupon matrix C , we have,

$$C = \begin{bmatrix} \frac{7.875}{2} + 100 & 0 & 0 & 0 \\ \frac{14.250}{2} & \frac{14.250}{2} + 100 & 0 & 0 \\ \frac{6.375}{2} & \frac{6.375}{2} & \frac{6.375}{2} + 100 & 0 \\ \frac{6.250}{2} & \frac{6.250}{2} & \frac{6.250}{2} & \frac{6.250}{2} + 100 \end{bmatrix}.$$

We implement the following trade: (i) buy x 10.750% bonds expiring in 2 years, which cost $110.938 \cdot x$; (ii) create x portfolios satisfying Eq. (11.10),

$$\hat{\pi} = c^* C^{-1} = \begin{bmatrix} 0.0189 & 0.0197 & 0.0212 & 1.0218 \end{bmatrix}.$$

If we short x of these portfolios, then, by construction, the coupons we need to pay are exactly matched by the coupons we receive from the x 10.750% bonds expiring in 2 years. However, the market value of the x portfolios we short equals,

$$x \cdot V = x \cdot \hat{\pi} B = x \cdot \begin{bmatrix} 0.0189 & 0.0197 & 0.0212 & 1.0218 \end{bmatrix} \begin{bmatrix} 101.40 \\ 108.98 \\ 102.16 \\ 102.57 \end{bmatrix} = x \cdot 111.041,$$

where the vector of the market prices, B , is taken from Table 11.1. Therefore, the gains from this trade are, $x \cdot (111.041 - 110.938) = 0.103 \cdot x$. For example, by trading £1,000,000 at face value, i.e. $x = 10000$, then, arbitrage profits equal £1030.

11.4 Duration and convexity hedging and trading

The risk of going long a default-free bond is that the future bond price is uncertain, due to the possibility that the spot interest rates could change in the future. Synthetically, we can say that the risk of a bond is related to the changes in the required bond return, or the YTM. Consider the definition of the YTM \hat{y} in Eq. (11.4). Next, consider the following function $B(y; T)$,

$$B(y; T) = \sum_{i=1}^n \frac{C_{t_i}}{(1+y)^{t_i}} + \frac{1}{(1+y)^T}.$$

This function aims to “mimic” how the market price $B(T)$ would behave if the YTM \hat{y} changed to some value y . Naturally,

$$B(\hat{y}; T) = B(T).$$

Motivated by the previous remarks, we can define a measure of risk of the bond based on the *sensitivity* of the bond price with respect to changes in y . Economically, we are trying to answer the following question: What happens to the bond price once we perturb the one rate \hat{y} that discounts all the payoffs? Mathematically, this sensitivity is the first partial of the “bond-pricing” formula $B(y; T)$ with respect to y ,

$$B_y(y; T) = -\frac{1}{1+y} \left[\sum_{i=1}^n \frac{t_i \cdot C_{t_i}}{(1+y)^{t_i}} + \frac{T \cdot 1}{(1+y)^T} \right]$$

where the subscript denotes a partial derivative, i.e. $B_y(y; T) = \frac{\partial}{\partial y} B(y; T)$. Graphically, this sensitivity measure $B_y(y; T)$ is the tangent to the price-yield relation, Figure 11.2 illustrates in the case of a zero-coupon bond with time to maturity equal to 10 years.

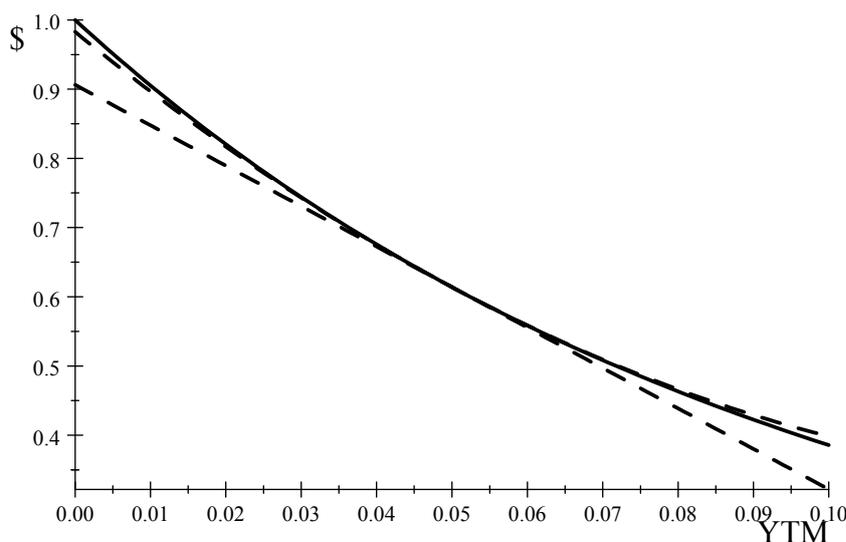


FIGURE 11.2. The relation between the YTM and the bond price, and its first-order (duration) and second-order (convexity) approximations. The solid line depicts the price of a zero coupon bond expiring in 10 years, as a function of the YTM, $(1 + \text{YTM})^{-10}$, and the two dashed lines are first-order and second-order Taylor's expansions around $\text{YTM} = 5\%$.

11.4.1 Duration

We define the “Macaulay duration” as,

$$D_{\text{Mac}} \equiv \frac{-B_y(y; T)}{B(y; T)} (1 + y) = \sum_{i=1}^n \omega_{t_i} \cdot t_i + \hat{\omega}_T \cdot T,$$

where

$$\omega_{t_i} = \frac{C_{t_i} / (1 + y)^{t_i}}{B(y; T)}, \quad \hat{\omega}_T = \frac{1 / (1 + y)^T}{B(y; T)}.$$

In words, the Macaulay duration is a weighted average of the payment dates. The weights ω_{t_i} are the discounted coupons at the various payment dates, $C_{t_i} / (1 + y)^{t_i}$, related to the current market value of these coupons, i.e. the bond price $B(y; T)$ when the YTM is y . That is, the weights are the proportions of the bond's present value that is attributable to the payoff at date t . The weights satisfy $\sum_{i=1}^n \omega_{t_i} + \hat{\omega}_T = 1$. Therefore, $D_{\text{Mac}} \leq T$. The Macaulay duration is a measure of how far in the future the bond pays off. For zeros, $D_{\text{Mac}} = T$.

For small y , $D_{\text{Mac}}(y)$ is simply the semi-elasticity of the bond price with respect to the YTM. This semi-elasticity is also referred to as “modified duration”:

$$D \equiv \frac{-B_y}{B} = \frac{D_{\text{Mac}}}{1 + y}.$$

A simple computation reveals that the modified duration, D , satisfies: $\frac{\partial D}{\partial y} = \frac{-B_{yy}}{B} + \left(\frac{B_y}{B}\right)^2$. Therefore, the modified duration is *decreasing* in the YTM when the bond price is sufficiently convex in the YTM, which is surely the case for long-term maturity dates.

Interestingly, the modified duration is *increasing* in the YTM when the bond price is concave in the YTM, a property that arises for callable bonds and mortgage-backed securities (MBS, henceforth), as explained in the next chapters (see also, Section 11.8.1 of this chapter, for a numerical exercise). Intuitively, the incentives to proceed to early repayments “kick in” as the YTM decreases, which makes the duration of the MBS decrease.

The Macaulay duration for *continuously compounded rates* is even simpler to compute. First, define the continuously compounded YTM as the single number \hat{x} such that

$$B(\hat{x}; T) = \sum_{i=1}^n c_{t_i} e^{-\hat{x} \cdot t_i} + e^{-\hat{x} \cdot T},$$

where $B(\hat{x}; T)$ is the market price of a bond paying off the principal of one at maturity and the stream of payoffs c_{t_i} . Next, consider, the function $x \mapsto B(x; T)$. Compute the semi-elasticity of the bond price $B(x; T)$ with respect to the continuously compounded YTM x ,

$$\frac{-B_x(x; T)}{B(x; T)} = \frac{\sum_{i=1}^n c_{t_i} t_i e^{-x \cdot t_i} + T \cdot e^{-x \cdot T}}{B(x; T)} = \sum_{i=1}^n w_{t_i} \cdot t_i + \hat{w}_T \cdot T,$$

where $B_x(x; T) = \frac{\partial B(x; T)}{\partial x}$, $w_{t_i} = \frac{c_{t_i} e^{-x \cdot t_i}}{B(x; T)}$ and $\hat{w}_T = \frac{e^{-x \cdot T}}{B(x; T)}$. Note, the weights are such that $\sum_{i=1}^n w_{t_i} + \hat{w}_T = 1$. Therefore, the “Macaulay duration” for continuously compounded rates is equal to the semi-elasticity of the bond price with respect to the continuously compounded YTM x .¹ This result may simplify some calculations.

11.4.2 Convexity

Convexity measures how the sensitivity, B_y , changes with y . Mathematically, convexity is related to the second partial of the bond price with respect to y , B_{yy} . If the second partial, B_{yy} , is positive, then, the interest rate sensitivity declines as y increases (see Figure 11.1). This is because $\frac{\partial}{\partial y} (-B_y) = -B_{yy} < 0$. Formally, convexity is defined as,

$$C \equiv \frac{B_{yy}}{B}.$$

We may, then, consider the following expansion of the bond price:

$$\frac{\Delta B}{B} \approx -D \cdot \Delta y + \frac{1}{2} C \cdot (\Delta y)^2. \quad (11.11)$$

That is, for very “convex securities”, duration may not be a safe measure of return, as Figure 11.2 illustrates.

11.4.3 Asset-liability management

11.4.3.1 Introductory issues

We can use duration to assess how exposed a bond portfolio is to movements in the interest rates. We can, then, “immunize” a portfolio of bonds to interest rates changes. Duration is relevant to asset-liability management. For example, pension funds have known streams of liabilities that must be matched by the assets they hold. In words, the duration of the assets must equal the duration of the liabilities. In the UK, pension funds must mark-to-market the liabilities. Therefore, one objective of these funds is to “immunize” their liabilities against movements in the interest rates.

Alternatively, consider the following basic example. A bank borrows £100 at 2% for a year and lends this money at 4% for 5 years, where the higher rate compensates for many things such as risk, the bank’s market power, etc. Assuming that the bank’s borrower does not default, in the first year, the bank generates profits equal to $\pounds(4\% - 2\%) \cdot 100 = 2$, according to its books. However, the right calculation to make should not relate to past market (interest rate) conditions, but to the current ones. Suppose, for example, that in one year, the interest rate for borrowing raises from 2% to 5%, and remains such for 4 additional years. This assumption is unrealistic, but it gives the idea of where the action is. In this case, the market value of the assets is: $\frac{100 \cdot 1.04^5}{1.05^4} = 100.09$. Note, we discount using the 5% interest rate, as this is the cost of capital for the bank.² The market value of the liabilities is $100 \cdot 1.02 = 102$. The bank’s problem is a duration mismatch.

¹Mathematically, we could have obtained this result in a straightforward manner, as follows. Define the bond price function as $B(y(x))$, where by definition, $y(x) = e^x - 1$. Hence, $B_x(y(x)) = B_y(y(x)) y'(x) = B_y(y(x)) e^x = B_y(y(x)) (1 + y)$. It follows that $D_{\text{Mac}} = \frac{-B_y(1+y)}{B} = \frac{-B_x}{B}$.

²Suppose, for example, that the bank wants to borrow £102 to pay off its liabilities, and for 4 additional years, then the profit at time 1 is $\frac{100(1.04)^5 - 102(1.05)^4}{(1.05)^4} = -1.9057$. Alternatively, the 5% interest rate is just an opportunity cost of capital, defined as $\max\{\text{borrowing cost, lending rate}\}$, where the borrowing cost is that the bank might obtain from other banks, for example.

Let us consider a second example, relating to the asset-liability management of pension funds. Consider the following extreme example. In 30 years from now, a pension fund is due to deliver £100,000 to some future retiree. Suppose the current market situation is such that the yield curve is flat at 4%, such that the market value of this liability is $£100,000 \cdot (1.04)^{-30} = £30,832$. Accordingly, the would-be retiree invests £30.832 in the pension fund. So we have the following situation:

Cash	Pensions
£30,832	£30,832

Suppose, now, that the pension fund does not invest this cash. This is of course inefficient, but it is precisely the point of this simple exercise to see why the strategy is inefficient.

Consider two extreme cases, occurring under two scenarios underlying developments in the fixed income market. In one week,

- (i) Scenario \uparrow : the yield curve shifts up parallelly to 5%. Accordingly, the value of the liability for the pension fund is: $£100,000 \cdot (1.05)^{-30} = 23,138$.

Cash	Profit
£30,832	£7,694
	Pensions
	£23,138

- (ii) Scenario \downarrow : the yield curve shifts down parallelly to 3%. Accordingly, the value of the liability for the pension fund is: $£100,000 \cdot (1.03)^{-30} = 41,199$.

Cash	Loss
£30,832	-£10,367
	Pensions
	£41,199

Therefore, a drop in the yield curve results in a loss for the pension fund: when interest rates go down, the pension fund faces a challenging situation as it has to honour its obligations in 30 years, but the financial market “yields less” than one week ago.

Naturally, the pension fund would face the opposite situation were interest rates to go up. In some countries, we do not like pension funds to experience volatility. The previous volatility arises simply because the pension fund, receives £30,832, and then it just puts this money “under the pillow.” The most efficient way to erase volatility is to invest £30,832 in a 30 bond as soon as we receive this money—at the market conditions of 4%. This is perfect hedging! But, we do not necessarily have access to such a bond. How do we proceed, then?

We now develop examples that illustrate how to deal systematically with issues relating to asset-liability management.

11.4.3.2 Hedging

Let us consider a portfolio of two bonds with different durations. Its value is given by,

$$V = B_1(\hat{y}_1)\theta_1 + B_2(\hat{y}_2)\theta_2,$$

where $B_1(\hat{y}_1)$ and $B_2(\hat{y}_2)$ are the market value of the bonds, \hat{y}_1 and \hat{y}_2 are the YTM on the bonds and, finally, θ_1 and θ_2 are the quantities of bonds in the portfolio. Let us consider a small change in the two YTM \hat{y}_1 and \hat{y}_2 . We have,

$$dV = - [D(\hat{y}_1) B_1(\hat{y}_1)\theta_1 d\hat{y}_1 + D(\hat{y}_2) B_2(\hat{y}_2)\theta_2 d\hat{y}_2].$$

The question is: How should we choose θ_1 and θ_2 so as to make the value of the portfolio remain constant after a change in \hat{y}_1 and \hat{y}_2 ?

Let us assume a *parallel shift* in the term structure of interest rates. In this case, $d\hat{y}_1 = d\hat{y}_2$. The portfolio is said to be *immunized* if its value V does not change as \hat{y}_1 and \hat{y}_2 change, i.e. $dV = 0$, which is true when,

$$\theta_1 = - \frac{D(\hat{y}_2) B_2(\hat{y}_2)}{D(\hat{y}_1) B_1(\hat{y}_1)} \theta_2. \quad (11.12)$$

A useful interpretation of this portfolio is that we may be holding a bond with some duration, say we hold θ_2 units of the second bond. Given these holdings, we may wish to sell another bond, possibly with a lower duration, to hedge against movements in the price of the bond we hold.

Alternatively, we can think of the second asset as a *liability* the value of which fluctuates after a change in the interest rates. Then, we may wish to purchase some *asset* to hedge against the liability. Mathematically, $\theta_2 < 0$ and $\theta_1 > 0$. Moreover, Eq. (11.12) reveals that the number of assets to hold to hedge against the liability is high if the ratio of the two durations of the assets, $D(\hat{y}_2)/D(\hat{y}_1)$, is large. In this case, the hedging position is obviously inefficient. Asset-liability management, and “immunization”, is costly when we hedge high-duration liabilities with low duration assets. We now illustrate these cases through a few basic examples.

11.4.3.3 A first example: hedging zeros with zeros

Suppose that we hold one bond, a zero with maturity equal to 5 years. We want to hedge the risk of this bond through another bond, a zero with maturity equal to 1 year. Let us assume that the term-structure is flat at 5%, discretely compounded. Then,

$$\begin{aligned} B_1(\hat{y}_1) &= \frac{1}{1 + \hat{y}_1} = \frac{1}{1 + 0.05} = 0.95238, & D(\hat{y}_1) &= \frac{D_{\text{Mac}}(\hat{y}_1)}{1 + \hat{y}_1} = \frac{1}{1 + 0.05} = 0.95238 \\ B_2(\hat{y}_2) &= \frac{1}{(1 + \hat{y}_2)^5} = \frac{1}{(1 + 0.05)^5} = 0.78353, & D(\hat{y}_2) &= \frac{D_{\text{Mac}}(\hat{y}_2)}{1 + \hat{y}_2} = \frac{5}{1 + 0.05} = 4.7619 \end{aligned}$$

and:

$$\theta_1 = - \frac{D(\hat{y}_2) B_2(\hat{y}_2)}{D(\hat{y}_1) B_1(\hat{y}_1)} \theta_2 = - \frac{4.7619 \cdot 0.78353}{0.95238 \cdot 0.95238} \cdot 1 = -4.1135.$$

That is, to hedge the 5Y zero, we need to short-sell approximately four 1Y zeros. The balance of this hedging position is,

$$B_1(\hat{y}_1)\theta_1 + B_2(\hat{y}_2)\theta_2 = (-4.1135) \cdot 0.95238 + 0.78353 = -3.1341, \quad (11.13)$$

a quite inefficient hedge. The reason this is inefficient is clear. Hedging high maturity bonds with short maturity ones implies we should rebalance too often. Moreover, as time goes on, the sensitivity of the short-term bonds to changes in the YTM is very small (at the extreme, the price equals face value plus coupon, at maturity), compared to that of long-term bonds. Therefore, rebalancing becomes increasingly severe as time unfolds.

Next, we study how the value of this portfolio changes after large changes in the YTM. By the assumption that the initial term-structure is flat at 5%, $\hat{y}_1 = \hat{y}_2 = 5\%$. Moreover, by rearranging Eq. (11.13),

$$B_2(y = 5\%) = 4.1135 \cdot B_1(y = 5\%) - 3.1341. \quad (11.14)$$

The left hand side of this equation is the price of the 5Y bond. The right hand side is the value of the “replicating” portfolio, which consists of (i) approximately 4 units of the 1Y bond, and (ii) the balance of the hedging position. Precisely, the right hand side is simply a net obligation: the value of the assets we need to purchase back (approximately 4 units of the 1Y bond), net of some cash we already have, which we can use to partially purchase these assets (£3.1341).

If interest rates do not change, then, approximately, and abstracting from passage of time, there will be no profits or losses, once we liquidate, or mark-to-market, this positioning. If interest rates change, $y \neq 5\%$, Eq. (11.14) can only approximately hold,

$$B_2(y) \approx 4.1135 \cdot B_1(y) - 3.1341.$$

Figure 11.3 plots the left hand side and the right hand side of this relation.

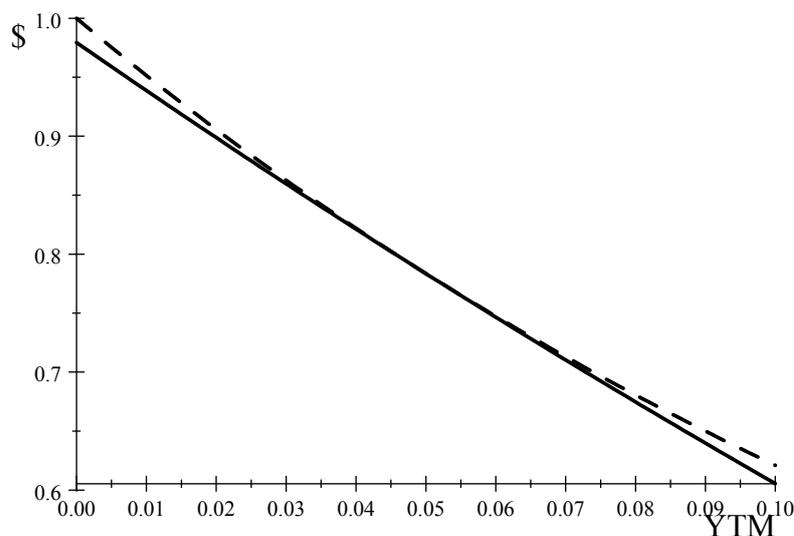


FIGURE 11.3. Dashed line (top): The price of the 5Y zero, $B_2(y) = \frac{1}{(1+y)^5}$, where y is the YTM. Solid line (bottom): The value of the “replicating” portfolio consisting of (i) 4.1135 units of the 1Y zero, and (ii) the balance of the hedging position, which is equal to $-\text{£}3.1341$, i.e. $4.1135 \cdot B_1(y) - 3.1341$, where $B_1(y) = \frac{1}{1+y}$ is the 1Y zero price.

What is going on? We are hedging the 5Y zero by selling approximately four 1Y zeros. In a neighborhood of $y = 5\%$, the value of the “synthetic” 5Y zero we sold, $4.1135 \cdot B_1(y) - 3.1341$, behaves as $B_2(y)$. However, the 5Y zero displays more convexity than the “synthetic” bond. This larger convexity implies that:

- If interest rates go down, the price of the 5Y zero bond we hold increases more than the value of the “synthetic” bond we sold. As a result, we make profits.
- If interest rates go up, the price of the 5Y zero bond we hold decreases less than the value of the “synthetic” bond we sold. As a result, we make profits.

In all cases, we make profits. Mathematically, profits are equal to $(B_2^+ - \theta_1 B_1^+) - (B_2 - \theta_1 B_1) \equiv \Delta B_2 - \theta_1 \Delta B_1$, where B_2^+ is the price of the 5Y bond, B_1 is the price of the 1Y bond. Then, by convexity, B_2 increases more than $\theta_1 B_1$ when interest rates go down, and B_2 decreases less than $\theta_1 B_1$ when interest rates go up, or $\Delta B_2 - \theta_1 \Delta B_1 \geq 0$. However, this is not an arbitrage opportunity! The previous reasoning hinges on the assumption of a parallel shift in the term-structure of interest rates, that is $d\hat{y}_1 = d\hat{y}_2$, where $\hat{y}_1 =$ spot rate for 1 year, and $\hat{y}_2 =$ spot rate for 5 years. While parallel shifts in the term-structure seem empirically relevant, they are not the only shifts that are likely to occur, as we shall explain in the next chapter.

To sum up, duration hedging is a useful tool, but with quite important limitations. As Eq. (11.11) makes clear, duration is only a first-order approximation to the price of a bond. Moreover, duration hedging obviously requires rebalancing, which might be substantial. As we now, a conventional bond is strictly convex in the YTM. Therefore, for large changes in the YTM, the duration-based hedging ratios should be updated. Re-adjustments are in order anyway, independently of whether YTM change or not, as the duration of conventional fixed income securities obviously decreases over time.

11.4.3.4 Duration trading: Barbell and bullet hedges

As a second example of duration hedging, consider the “barbell” trading, which is a way to hedge some liability (a “bullet”) with duration D_2 through two assets with durations D_1 and D_3 , where $D_1 < D_2 < D_3$ —that is a trade where we buy D_2 and sell D_1 & D_3 . This trade is expected to work when we expect the yield curve to flatten, with its short-end part not going too much high. Moreover, investing in the short-term segment of the yield curve, allows one to invest elsewhere relatively rapidly once the first asset expires, were the bond market to go down.

Let us consider the previous example, and suppose there is another bond available for trading, a zero with maturity equal to 10 years. We aim to hedge against movements in the price of the 5Y zero with a portfolio consisting of (i) one 1Y zero and (ii) the 10Y zero. We continue to assume that the yield-curve is flat at 5%, and only consider parallel shifts in the term-structure of interest rates.

Such a “butterfly” trade can be implemented as follows. We look for a portfolio of the 1Y and 10Y zero with the following properties: (i) the market value of the portfolio equals the market price of the 5Y zero,

$$B_2(\hat{y}_2) = B_1(\hat{y}_1)\theta_1 + B_3(\hat{y}_3)\theta_3; \quad (11.15)$$

and (ii) the local risk of the portfolio equals the local risk of the 5Y zero, $\partial B_2(\hat{y}_2)/\partial \hat{y}_2 = -D(\hat{y}_2)B_2(\hat{y}_2)$, i.e.:

$$D(\hat{y}_2)B_2(\hat{y}_2) = D(\hat{y}_1)B_1(\hat{y}_1)\theta_1 + D(\hat{y}_3)B_3(\hat{y}_3)\theta_3. \quad (11.16)$$

The solution to Eqs. (11.15) and (11.16) is given by,

$$\theta_1 = \frac{D(\hat{y}_3) - D(\hat{y}_2)B_2(\hat{y}_2)}{D(\hat{y}_3) - D(\hat{y}_1)B_1(\hat{y}_1)}, \quad \theta_3 = \frac{D(\hat{y}_2) - D(\hat{y}_1)B_2(\hat{y}_2)}{D(\hat{y}_3) - D(\hat{y}_1)B_3(\hat{y}_3)}. \quad (11.17)$$

By the same computations made in the previous example, we have that $B_3(\hat{y}_3) = 0.61391$ and $D(\hat{y}_3) = 9.5238$. By using the figures in the previous example, we compute θ_1 and θ_3 in Eqs. (11.17) to be

$$\theta_1 = \frac{9.5238 - 4.7619}{9.5238} \frac{0.78353}{0.95238} = 0.45706, \quad \theta_3 = \frac{4.7619 - 0.95238}{9.5238 - 0.95238} \frac{0.78353}{0.61391} = 0.56724.$$

Figure 11.4 depicts the behavior of the bullet price and the market value of the barbell as we change the YTM. Note that the barbell portfolio is more convex than the bullet! Moreover, the barbell trade is “self-financed.” By construction, the value of the bullet we sell equals the value of the barbell portfolio. So now, large movements in the YTM lead to profits, provided we maintain the assumption of parallel shifts in the term-structure of interest rates. Note that the direction of interest rate movements does not matter in value creation. This “convexity trading” resembles a standard hedge fund strategy where, say, we go long a number of “undervalued” stocks and short a number of “overvalued” stocks such that the initial value of the portfolio is zero. Then, we are likely to make profits: in good times, the undervalued stock should increase in value more than the overvalued, and in bad times, the drop in value of the undervalued stock should be less severe than that of the overvalued. Naturally, the value driver of the barbell is, again, convexity: as Eq. (11.11) illustrates, the convexity term, C , is, trivially, always positive, independently of the sign of Δy . Therefore, as soon as we hedge a bond with a portfolio that has the same duration as the given bond, but higher convexity, the position leads to profits, given the assumptions made so far.

A barbell trade does not lead to an arbitrage. The scenario underlying the P&L of Figure 11.4 relies on the assumption of a parallel shift in the term structure of interest rates. However, as explained in the next chapter (Section 12.3), it is not realistic to simultaneously assume large and parallel movements in the term-structure of interest rates. Historically, large interest rate shifts (that is, typically, shifts occurring over large horizons of time) are accompanied by the occurrence of a variety of shape modifications. Factors affecting parallel movements in the yield curve are frequent, but they are not the only ones. At least three factors are needed to explain the entire variation of the yield curve.

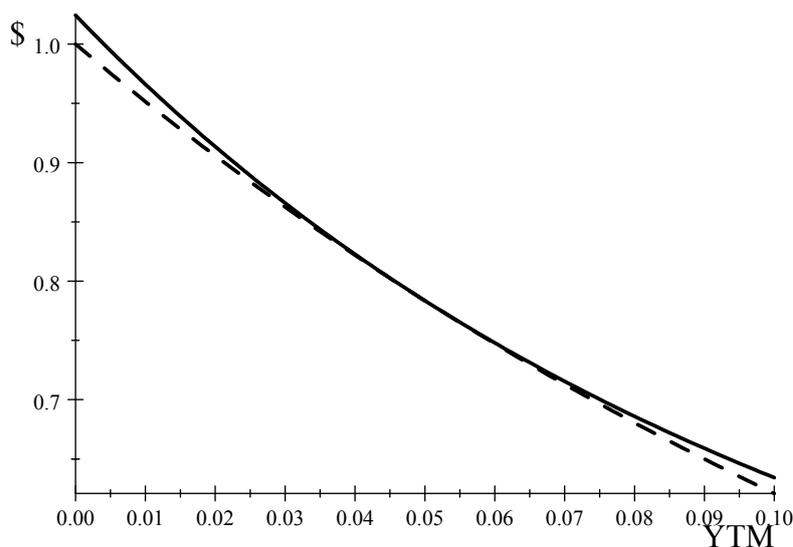


FIGURE 11.4. “Barbell trading.” Dashed line (bottom): The price of the 5Y zero, $B_2(y) = \frac{1}{(1+y)^5}$, where y is the YTM. Solid line (top): The value of the “barbell” portfolio consisting

of (i) 0.45706 units of the 1Y zero and (ii) 0.56724 of the 10Y zero, i.e. $B_1(y_1) \cdot 0.45706 + B_3(y_3) \cdot 0.56724$, where $B_1(y) = \frac{1}{1+y}$ is the 1Y zero price and $B_3(y) = \frac{1}{(1+y)^{10}}$ is the 10Y zero price.

Table 11.4 considers the case of *non-parallel* shifts in the term-structure. We assume that the initial term-structure is not flat. Then, we consider two scenarios: (i) A “twist” in the term-structure, i.e. long-term rates lower than short-term; (ii) a “steepening” of the term-structure.

TABLE 11.4.

	YTM	Bullet price	Mod. dur.	Barbell value = $\theta_1 B_1(\hat{y}_1) + \theta_3 B_3(\hat{y}_3)$
Initial term-structure				
1Y	$\hat{y}_1 = 4\%$	$B_1(\hat{y}_1) = 0.961$	$D(\hat{y}_1) = 0.961$	
5Y	$\hat{y}_2 = 5\%$	$B_2(\hat{y}_2) = 0.783$	$D(\hat{y}_2) = 4.762$	
10Y	$\hat{y}_3 = 6\%$	$B_3(\hat{y}_3) = 0.558$	$D(\hat{y}_3) = 9.434$	Barbell value = 0.783
“Twist”				
1Y	$\hat{y}_1 = 6\%$	$B_1(\hat{y}_1) = 0.943$	$D(\hat{y}_1) = 0.943$	
5Y	$\hat{y}_2 = 5\%$	$B_2(\hat{y}_2) = 0.783$	$D(\hat{y}_2) = 4.762$	
10Y	$\hat{y}_3 = 4\%$	$B_3(\hat{y}_3) = 0.675$	$D(\hat{y}_3) = 9.615$	Barbell value = 0.847
“Steepening”				
1Y	$\hat{y}_1 = 4\%$	$B_1(\hat{y}_1) = 0.961$	$D(\hat{y}_1) = 0.961$	
5Y	$\hat{y}_2 = 5\%$	$B_2(\hat{y}_2) = 0.783$	$D(\hat{y}_2) = 4.762$	
10Y	$\hat{y}_3 = 7\%$	$B_3(\hat{y}_3) = 0.508$	$D(\hat{y}_3) = 9.346$	Barbell value = 0.751

We use the portfolio in Eq. (11.17), and find that in correspondence of the initial term-structure ($\hat{y}_1 = 4\%$, $\hat{y}_2 = 5\%$, $\hat{y}_3 = 6\%$), $\theta_1 = 0.449$ and $\theta_3 = 0.629$. We keep this portfolio fixed, and compute the barbell value, $\theta_1 B_1(\hat{y}_1) + \theta_3 B_3(\hat{y}_3)$, occurring at the two scenarios “twist” and “steepening,” with $B_2(\hat{y}_2) = 0.783$ in all cases. The trade is as follows: at time zero, we sale short the five year bond, which we hedge through the barbell portfolio (θ_1, θ_3) , using the proceeds of the short-sale. Then, at some future date, we purchase back the five year bond and sell back the portfolio (θ_1, θ_3) . The convexity of the barbell trade is, in fact, a view about movements of long-term bond prices, and leads to profits in the “twist” scenario. That is, by convexity, the price B_3 varies more than the price of shorter maturity zeros, thus leading to profits. Note, however, that this strategy leads to losses in the “steepening” scenario.

We need to state an important caveat. The previous conclusions deserve further scrutiny. They rely on a static analysis, and abstract from the fact that term-structure movements should occur under the assumption of no-arbitrage. For example, the value of the zeros changes over the horizons we are designing scenarios for, even without any changes in the yield curve. Whether this effect is minor depends on the horizon and the model we use to generate scenarios! In Section 11.6.6, we shall revisit the example of this section and illustrate how passage of time and absence of arbitrage can be factored into the analysis, and change some results emanating from Table 11.4.

11.4.3.5 Fixed income arbitrage strategies

The previous “convexity trades” are examples of *yield curve arbitrage* strategies. They may purely rely on convexity or, as discussed in the previous section, on directional views about interest rate movements. For example, we have explained, we may short five year bonds, and go long two- and ten-year bonds, as we view that short-term interest raise will raise and medium term interest rates will lower. This “butterfly” strategy is somehow cheap, intellectually, and not necessarily rewarding, and will be further analyzed in Section 11.6.6. *Swap spread arbitrage* is a popular strategy. It was responsible of leading LTCM to a loss of about \$1.6 billion in 1997. The strategy works as follows: (i) enter a swap paying the floating LIBOR, L_t , and receiving a fixed rate \bar{C} ; (ii) short a par Treasury with the same maturity as the swap, thus paying the fixed coupon rate \bar{CT} , and invest the proceeds at the repo rate r_t . Thus, the payoff of the strategy is the fixed spread to be received, $F = \bar{C} - \bar{CT}$, and the floating spread to be paid, $S_t = L_t - r_t$. So we go long or short this strategy according to whether we view F to be larger or smaller than the average floating spread S_t over the strategy horizon. Historically, the spread S_t has certainly been volatile, but quite stable, so it is a reasonable strategy. The problem occasionally, though, S_t can attain quite large values. More sophisticated strategies rely on *models*, which identify which points of the yield curve are misaligned from those predicted by the model. The strategy, substantially, is: buy the cheap and short the model-based rich, where the model-based rich is replicated through a portfolio with cash and the bonds that are well-priced by the model, weighted with model-based delta, as in the derivation of the bond pricing formula in Section 12.4.2.2 of the next chapter.

11.4.3.6 Negative convexity, and market volatility

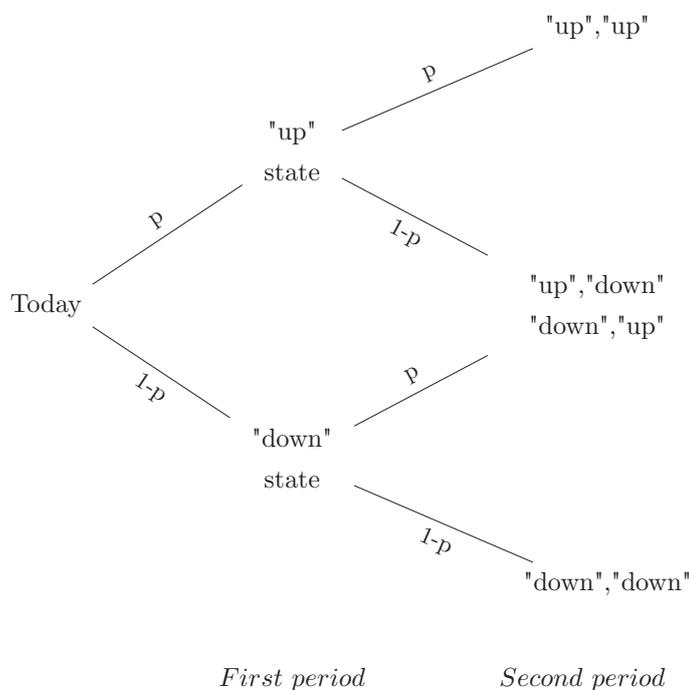
What happens when bond prices have “negative convexity”? In the next chapter, we shall see that the value of a callable bond can be *concave* in the short-term rate. A similar feature is displayed by mortgage-backed-securities (MBS, henceforth), which can now be concave in the YTM! The reason for this negative convexity is that early repayments are likely to occur as the YTM decreases, which entails two inextricable consequences: (i) the price of the MBS “increases less” than a conventional bond price after a decline in the YTM, especially when the YTM is low; (ii) the duration of the MBS decreases as the YTM decreases.

Hedging against MBS might lead to an increased volatility in rate markets. The mechanism is the following. Institutions that are long MBS would typically short conventional bonds for hedging purposes, consistently with the prediction of Eq. (11.12). However, the duration of MBS increases as interest rates increase, due tonegative convexity: $\frac{\partial \text{Duration}}{\partial r} = -\text{Convexity}$. Therefore, an interest rate increase can lead these institutions to short additional conventional bonds, which worsens liquidity and leads to a further increase in the interest rates, thereby feeding a vicious circle. Perli and Sack (2003) estimate that in 2002 and 2003, this mechanism may have amplified the volatility of long-term US rates by a factor between 15% and 30%. It is an instance of what is sometimes defined as “endogenous risk,” the circumstance that the trend of a certain economic variable triggers actions from market participants that in turn, reinforce the initial trend, as in the case of the 1987 crash discussed in Section 10.4.5 of Chapter 10, or in the case of assets sell-off in times of crisis, discussed in Section 13.5.3 of Chapter 13.

11.5 Foundational issues in interest rate modeling

In principle, we might utilize the classical binomial trees underlying contingent claim evaluation as analytical tools to price fixed income securities. To this end, however, we need to revise quite a few methodological details. Let us illustrate. It is instructive to review how binomial trees are built up, in general:

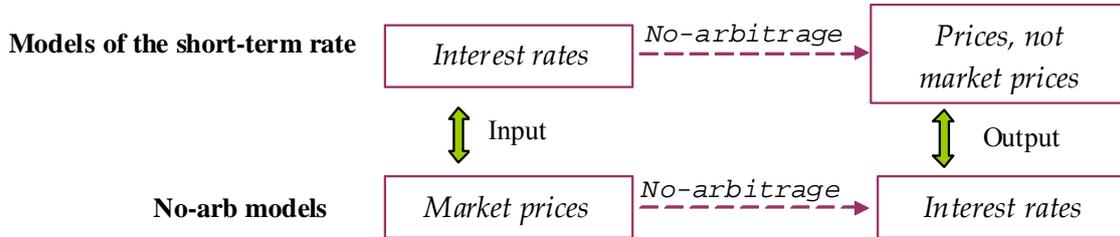
- (i) We begin with a probabilistic representation of how the price develops over time, using a tree-like information structure.
- (ii) For example, at the time of evaluation, we observe the state. In the next period, there can be two mutually exclusive states of the world: (a) the state “up,” occurring with probability p ; and (b) the state “down,” occurring with probability $1 - p$.
- (iii) After two periods, there can be three mutually exclusive states of the world, as in the following diagram. We label the tree in this diagram a “recombining” tree, to emphasize that the “up & down” and the “down & up” nodes are the same.



The previous diagram can be used to price options written on stocks. The stock price unfolds through the branches of the tree. Then, we figure out the no-arbitrage movements of the option price along the tree. Suppose, however, we wish to price an option written on a zero, a 3 Year zero say. Can we apply the same methodology to price the option? The answer is no, and the reason is that we cannot exogenously “track” the movements of the prices of the zero, as in the case of the stock price. Instead, after one year, the 3 Year zero becomes a 2 Year zero, i.e. quite a different asset.

These issues can be mitigated by modeling the movements of the entire yield curve. There are two approaches, as in the diagram below. In the first, we model the dynamics of the short-term rate, defined as the interest rate on a loan with maturity equal to the time intervals in the tree. The resulting model, referred to as *model of the short-term rate*, has implications in terms

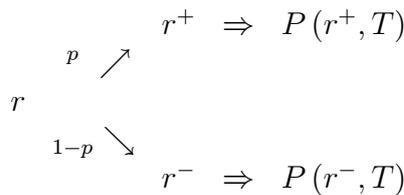
of the movements of the entire term-structure. This approach, developed in the next section, leads to evaluation formulae in which the current price of the zeros predicted by the model are not necessarily equal to the *market* prices. A second approach, based on *calibration*, leads to the so-called *no-arbitrage* models, where we model the dynamics of the entire term-structure. This approach gives rise to option evaluation formulae in which the current prices of the zeros predicted by the *model* are equal to the *market* prices. We describe this approach in the last sections of this chapter, using binomial trees, with the next chapter developing their continuous time counterparts.



11.5.1 Tree representation of the short-term rate

11.5.1.1 Recursive evaluation

Consider a two-period, two-state tree, where the current short-term rate is r . The development of the short-term rate is uncertain. That is, the future short-term rate, \tilde{r} , is random, and can take two values: either r^+ with probability p , or r^- with probability $1 - p$. We assume that $r^+ > r^-$.



We emphasize p is the physical probability. Suppose, also, that two zeros with distinct maturities are available for trading. A money market accounting technology is also available (MMA, in the sequel). Investing £1 in the MMA generates £1·(1 + r) in the second period. We derive an evaluation formula for the zero based on the previous probabilistic model for the short-term rate dynamics. The general idea is to build up a portfolio that contains one zero and the MMA. We shall make sure the value of this portfolio in the second period replicate the value of the zero we wish to price. By no-arbitrage, then, the value of the portfolio in the first period must equal the value of the zero we wish to price, and we shall be done. The appendix develops the arguments, and shows that in the absence of arbitrage, there is a λ , a function of r at most, such that the following relation holds true:

$$E_p [P(\tilde{r}, T)] - (1 + r) P(r, T) = \underbrace{\frac{\Delta P(\tilde{r}, T)}{\Delta \tilde{r}} \cdot \text{Vol}(\tilde{r} - r)}_{= \text{volatility of the price}} \cdot \underbrace{\lambda}_{= \text{unit risk premium}}, \quad (11.18)$$

where $E_p [P(\tilde{r}, T)]$ denotes the expectation of the bond price under the probability p , and $\text{Vol}(\tilde{r} - r) \equiv \Delta \tilde{r} = r^+ - r^-$, which is interpreted as the volatility of the short-term rate.

Eq. (11.18) is an APT relation, where λ is interpreted as a unit premium related to the risk of holding the bond over one period of time. It says that the expected excess return on the zero

equals the volatility of its price multiplied by the unit price of risk. We call the term,

$$\frac{\Delta P(\tilde{r}, T)}{\Delta \tilde{r}} \cdot \text{Vol}(\tilde{r} - r),$$

“price volatility” because it measures the amplitude of the price variation due to changes in the short-term rate in the future, $\frac{\Delta P(\tilde{r}, T)}{\Delta \tilde{r}}$, i.e. the “price-sensitivity”, where this price sensitivity is normalized by the volatility of the short-term rate, $\text{Vol}(\tilde{r} - r)$. We can elaborate on Eq. (11.18), so as to bridge to the continuous time APT relations seen in Chapter 4 of these Lectures. The key observation is that Eq. (11.18) relies on a tree with a trading period normalized to unity. When the trading period is equal to some Δt , the interest rate earned over that period is $r \cdot \Delta t$, and accordingly, $\lambda \cdot \Delta t$ is the unit premium to compensate for the risk of holding the bond over an amount of time equal to Δt . With these changes, Eq. (11.18) can be cast as,

$$E_p [P(\tilde{r}, T)] - (1 + r \cdot \Delta t) P(r, T) = \frac{\Delta P(\tilde{r}, T)}{\Delta \tilde{r}} \text{Vol}(\tilde{r} - r) \lambda \cdot \Delta t,$$

such that, by considering small Δt , and rearranging terms,

$$\frac{E_p(dP(r, T))}{dt} - rP(r, T) = \frac{\partial P(r, T)}{\partial r} \text{Vol}(dr) \lambda.$$

Once we assume r is solution to a stochastic differential equation, we can use Itô’s lemma, to turn the previous equation into a partial differential equation subject to a boundary condition that states the the bond price is one at expiration. Chapter 12 contains a more rigorous discussion of these topics.

Eq. (11.18) can now be cast in a format that we can use to make it easy to use. After rearranging terms, we obtain:

$$P(r, T) = \frac{(p - \lambda) P(r^+, T) + [1 - (p - \lambda)] P(r^-, T)}{1 + r} = \frac{E_q [P(\tilde{r}, T)]}{1 + r} \quad (11.19)$$

where $q \equiv p - \lambda$ is the risk-neutral probability.

A few considerations. We “expect” that $\lambda < 0$ because bond prices are decreasing in the short-term rate here. Then, $q \equiv p - \lambda > p$.³ Hence, the risk-neutral probability of an upward movement of the short-term rate, q , is higher than the true probability, p . An investor who goes long a bond, is concerned by an increase of the short-term rate in the future and, hence, “corrects” the true probability p by assigning a higher risk-adjusted probability to the “upward” state.

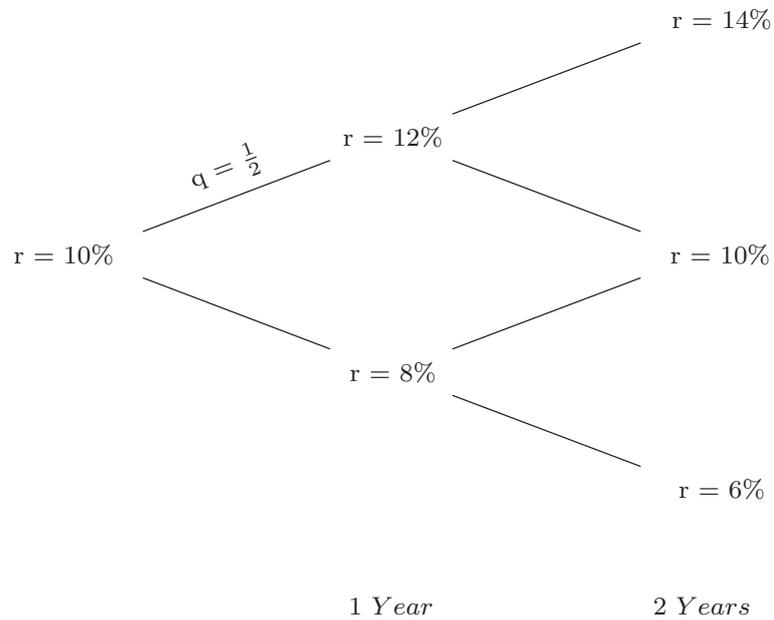
11.5.1.2 One example

Assume the current short-term rate equals 10%. We know that with probability p , the *physical* probability, the short-term rate as of the next year will increase by 2 percentage points, and with probability $1 - p$, it will decrease by 2 percentage points. Finally, with the same probability p , the short-term rate prevailing from the next year to two years time, will increase by 2 further percentage points from its previous value in one year time. We take the probability of an upward movement to be 20% and the absolute value of the Sharpe ratio to be 30%. Given these data, we use the formula, and obtain an estimate of the *risk-neutral probability* of an upward movement of the short-term rate, equal to $q = p - \lambda = 20\% - (-30\%) = 50\%$.

³To ensure q is a probability, we need to have that (i) $q \equiv p - \lambda > 0 \Leftrightarrow -\lambda > -p$ and (ii) $q \equiv p - \lambda < 1 \Leftrightarrow -\lambda < 1 - p$. That is, $-\lambda \in (-p, 1 - p)$

Pricing zeros

Let us price a zero maturing in two years, hinging upon the following tree:



We can use Eq. (11.19) to “fill-in” each node of the tree. We start from the end of the tree, where the price of the two year zero is £1, and then use Eq. (11.19) to fill every node, as illustrated in Figure 11.5.

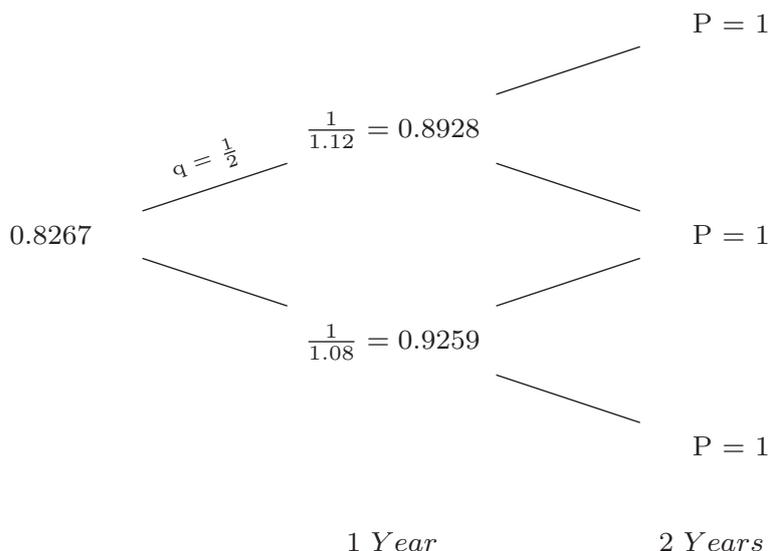


FIGURE 11.5.

The price of the zero, in one year, is simply one divided by the interest rate relevant at the beginning of the year, next year. The price we are looking for is obtained by applying Eq. (11.19) yielding,

$$\frac{E_q [P(\tilde{r}, 2)]}{1 + r} = \frac{qP(r^+, 2) + (1 - q)P(r^-, 2)}{1 + r} = \frac{\frac{1}{2}(0.8928) + \frac{1}{2}(0.9259)}{1.10} = 0.8267.$$

Convexity effects

What is the discretely compounded two-year spot rate? Does it equal 10%? It is a natural question, as the short-term rate is a martingale under the risk-neutral probability, q . But as it turns out, the answer to the previous question is in the negative. Let us elaborate. The two-year spot rate, $r(0, 2)$, satisfies,

$$0.8266 = \frac{1}{[1 + r(0, 2)]^2} \Leftrightarrow r(0, 2) = \sqrt{\frac{1}{0.8266}} - 1 = 9.98\%.$$

Even though $r = 10\%$ and $E_q(\tilde{r}) = 10\%$, we have that two years spot rate equals, 9.98%. That is,

$$0.8266 = \frac{1}{1 + r} E_q \left(\frac{1}{1 + \tilde{r}} \right) > \frac{1}{1 + r} \frac{1}{1 + E_q(\tilde{r})} = 0.8264.$$

Prices increase after activation of uncertainty. It's a convexity effect, similar to that we shall have to explain in the next chapter (Section 12.4.5.1, Figure 12.3).

11.5.2 Tree pricing

We can generalize the tree to a multiperiod case. We use Eq. (11.19) to evaluate zeros at all nodes of the tree and maturities. Given q , which can be estimated once we estimate p and λ , we use recursively Eq. (11.19). Then, we may price options on zeros. The weakness of the approach is that the initial term structure is predicted with error! Let us illustrate this approach with a concrete numerical example. Consider the following tree, where the current short-term rate for one year is $r = 4\%$.

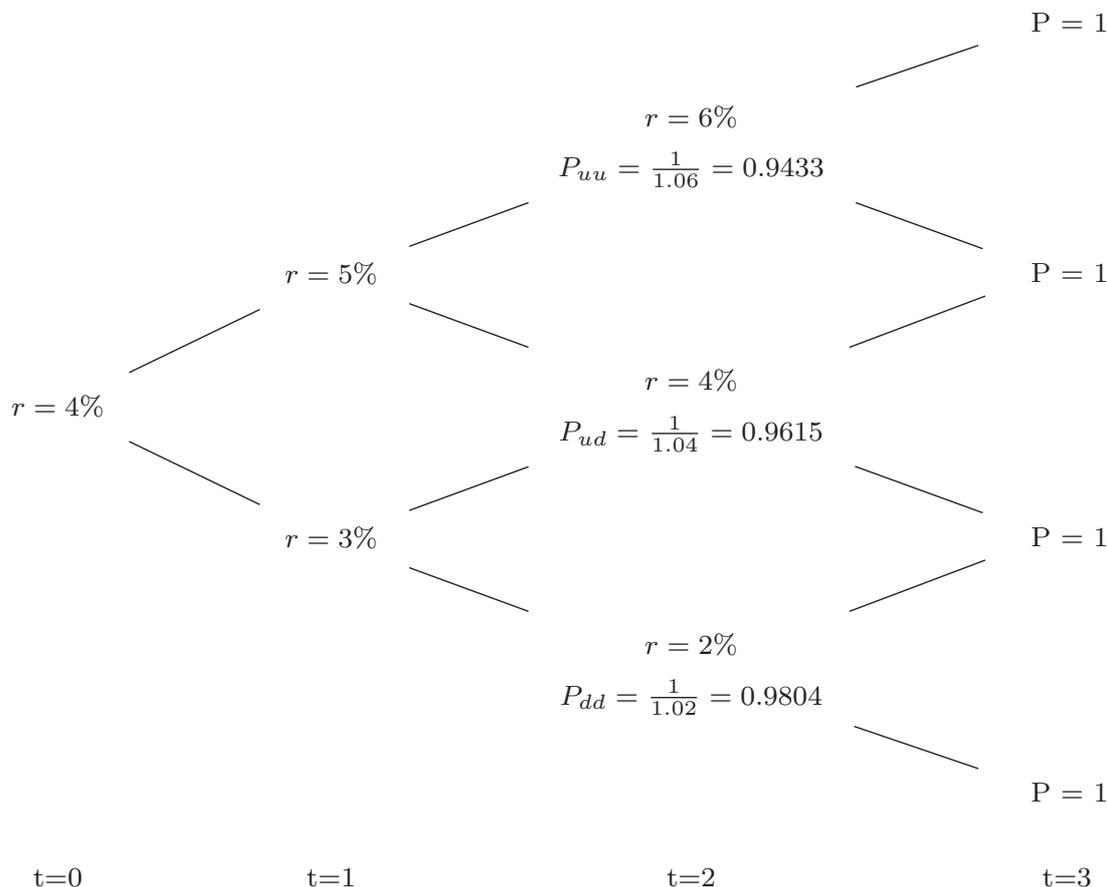
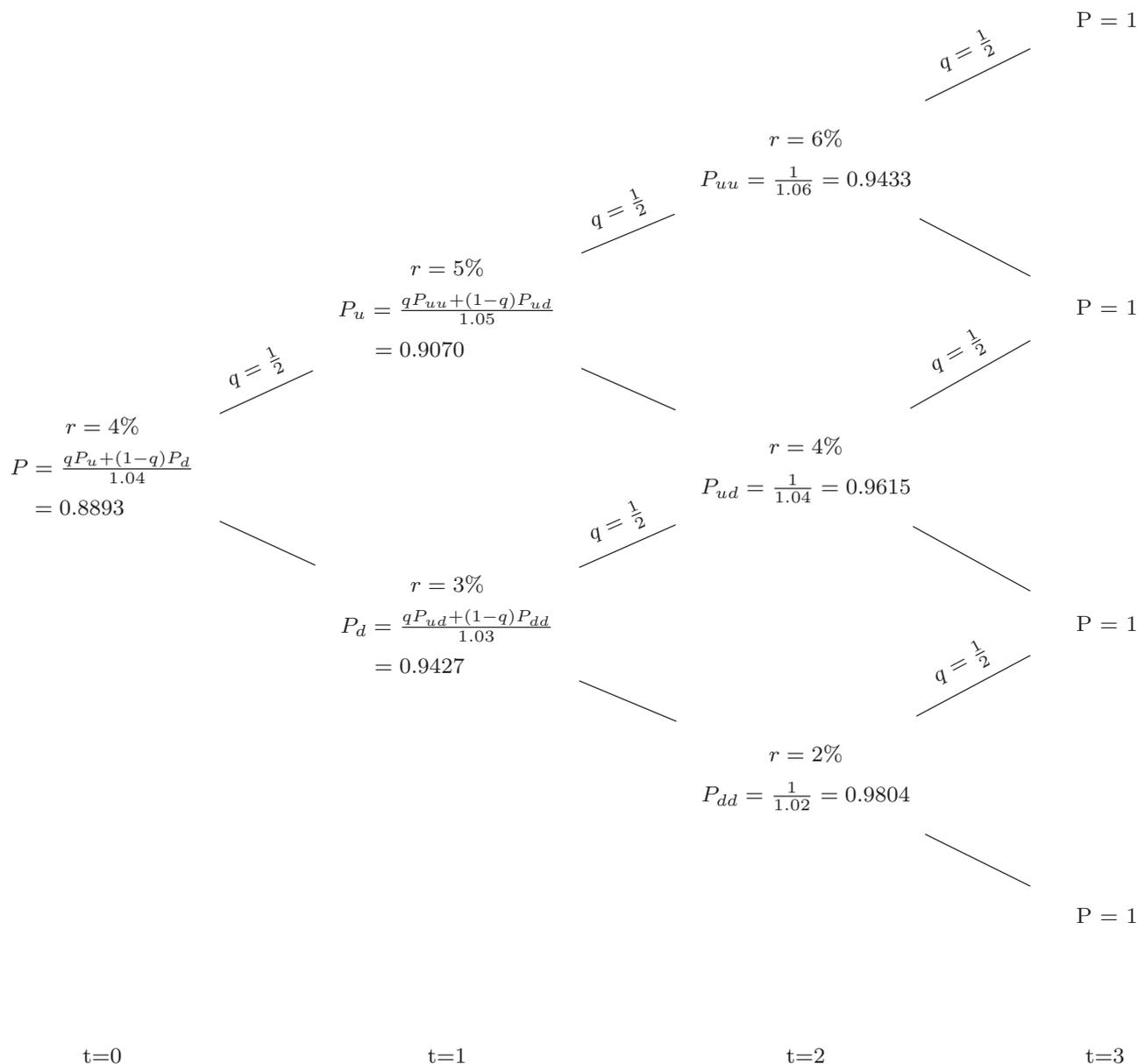


FIGURE 11.6. The dynamics of the short-term rate

At time $t = 1$, the short-term rate is either 5%, with probability p (the true probability) or 3%, with probability $1 - p$. At time $t = 2$, the short-term rate behaves as follows: (i) if at time $t = 1$, $r = 5\%$, then, at time $t = 2$, $r = 6\%$, with probability p , and $r = 4\%$ with probability $1 - p$; and (ii) if at time $t = 1$, $r = 3\%$, then, at time $t = 2$, $r = 4\%$, with probability p , and $r = 2\%$ with probability $1 - p$. Also shown in the previous diagram is the price of a hypothetical 3 Year zero, P , at time $t = 3$ and at time $t = 2$. At time $t = 3$, the expiration date, $P = 1$ in all states of nature. At time $t = 2$, the price P is $P(r, T) = E_q[P(\tilde{r}, T)] / (1 + r) = 1 / (1 + r)$, for $r = 6\%$, 4% and 2% . The issue, now, is how to compute the price of the zero in correspondence of the remaining nodes. We should use the formula, $P(r, T) = E_q[P(\tilde{r}, T)] / (1 + r)$ to populate the tree, but we do not know p , λ , and q . Suppose we “estimate” p and λ . In this case, we compute q as $q = p - \lambda$, as in Eq. (11.19). (For example, $p = 20\%$ and $\lambda = -30\%$, so that $q = 50\%$.) Suppose that we come up with $q = \frac{1}{2}$. Then, the following diagram gives the price of

the zero in all the nodes as of time $t = 1$, and at the evaluation time $t = 0$, leaving a the price of the 3 Year zero equal to 0.8893.



Next, consider a European call option written on the 3 Year zero, with expiration date equal to 2 and strike price $K = 0.95$. The following diagram gives the value of the option predicted by the model at each node of the tree. The model predicts that the current price of the call option is 0.0124.

the initial yield curve is fitted without error. These trees are called implied binomial trees—“implied” by the market prices. Let us consider the example in the previous section. To make the model-implied price of the 3 Year zero equal to the market price, $P_{\S} = 0.8700$, we cannot take the risk-neutral probability q as given, i.e. independent of the observed price $P_{\S} = 0.8700$, as we did before. Rather, we should *calibrate* the probability q , as follows,

$$P_{\S} = 0.8700 = \frac{1}{1.04} [q \cdot P_1(5\%) + (1 - q) \cdot P_1(3\%)] \quad (11.20)$$

where $P_1(5\%)$ and $P_1(3\%)$ are the prices of the zero at time $t = 1$, in the events that the short-term rate is up to 5% or down to 3%.

The previous equation follows, again, by Eq. (11.19). Note, now, that the unknown is not the price, which is instead given by the market price. Rather, we are looking for, or calibrating, the probability q that makes the RHS of Eq. (11.20) equal to its LHS. Naturally, we need to compute the prices of the zeros $P_1(5\%)$ and $P_1(3\%)$. These prices can be found by another application of Eq. (11.19), as follows,

$$P_1(5\%) = \frac{q \cdot 0.9433 + (1 - q) \cdot 0.9615}{1.05}, \quad P_1(3\%) = \frac{q \cdot 0.9615 + (1 - q) \cdot 0.9804}{1.03}.$$

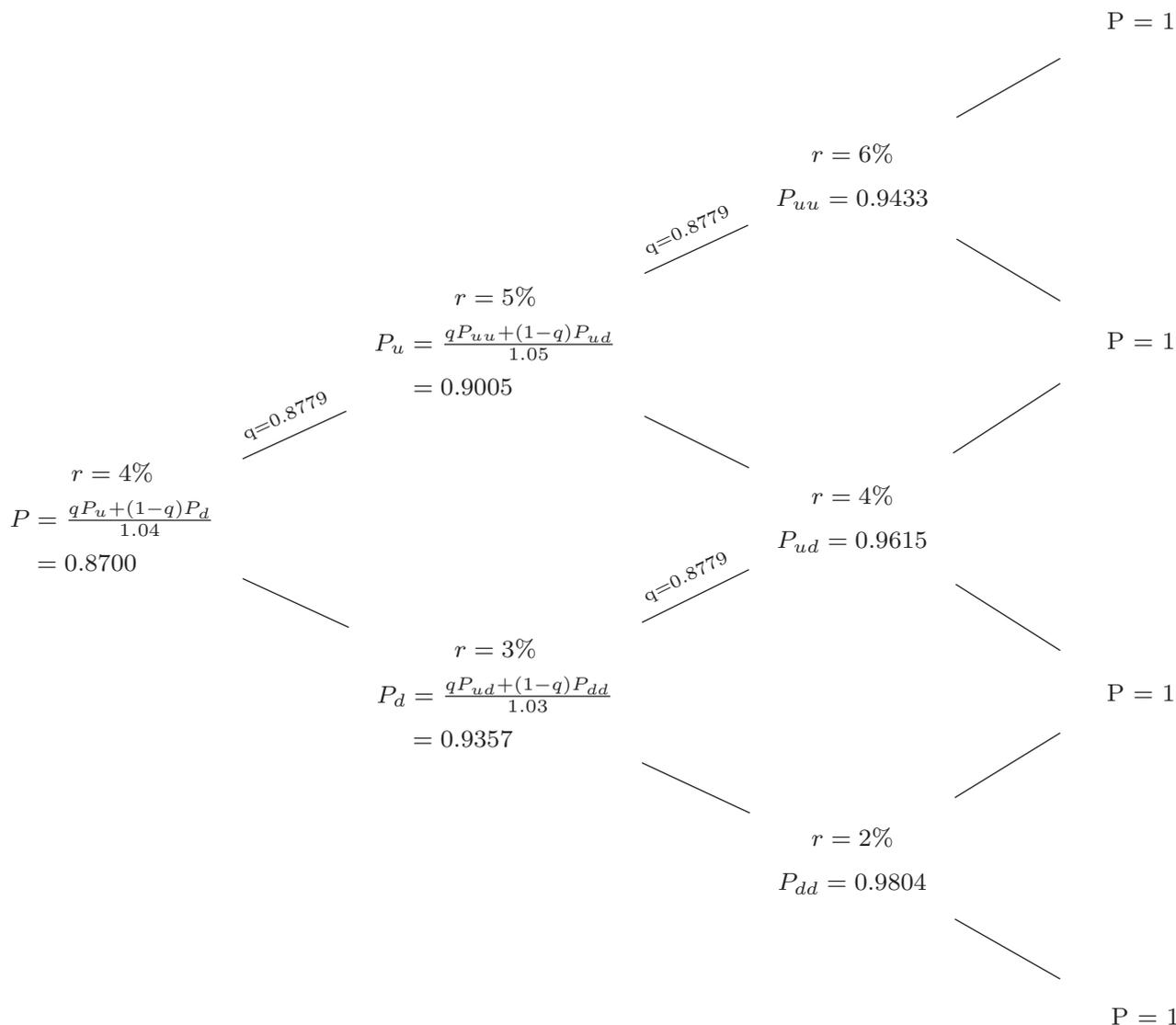
By replacing the previous expressions for $P_1(5\%)$ and $P_1(3\%)$ into Eq. (11.20), we obtain,

$$P_{\S} = 0.8700 = \frac{1}{1.04} \left(q \cdot \frac{q \cdot 0.9433 + (1 - q) \cdot 0.9615}{1.05} + (1 - q) \cdot \frac{q \cdot 0.9615 + (1 - q) \cdot 0.9804}{1.03} \right).$$

This is a nonlinear equation in q , which we can easily solve, to obtain, $q = 0.8779$. Hence, we find:

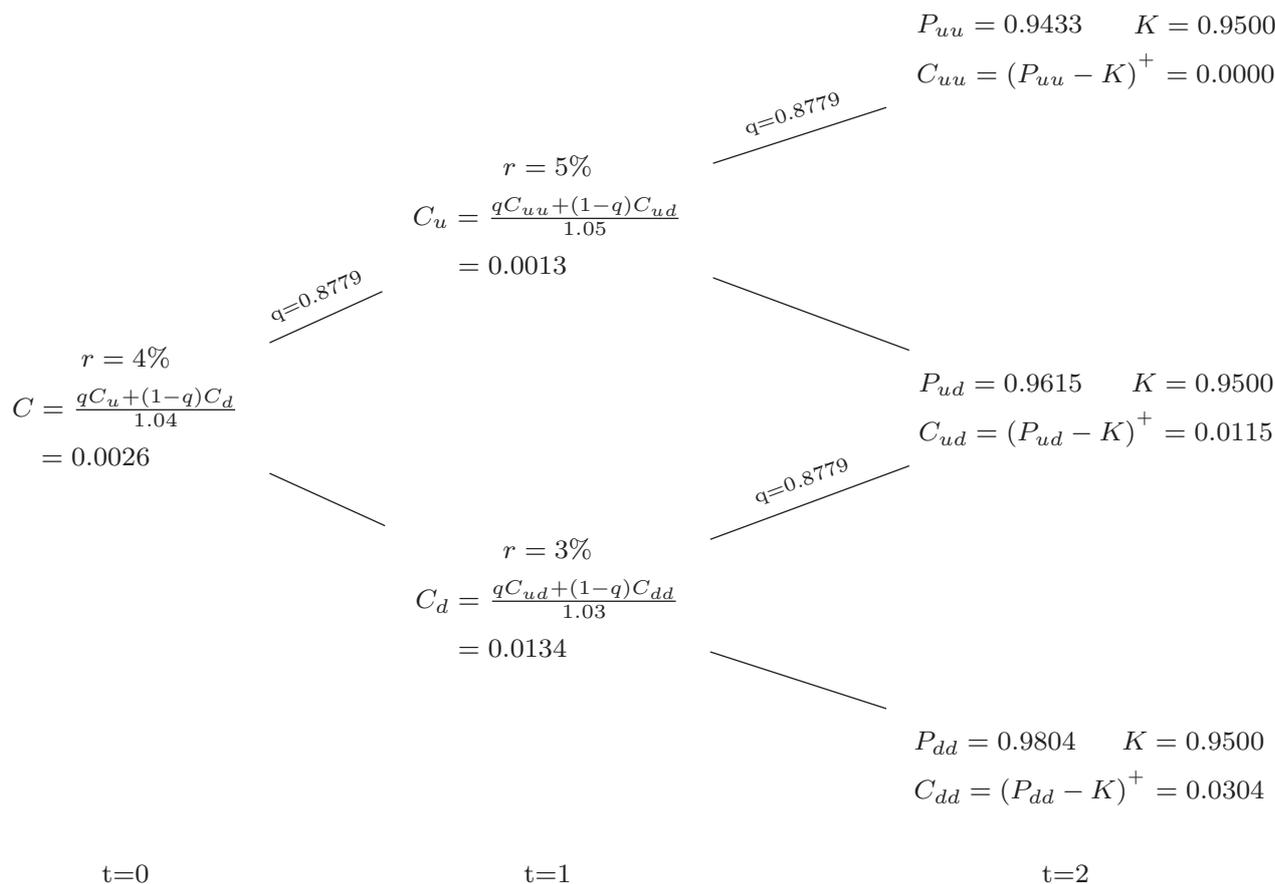
$$P_1(5\%) = 0.9005 \quad \text{and} \quad P_1(3\%) = 0.9357.$$

The next diagram depicts the *implied binomial tree*, i.e. the tree that we obtain after we match the model-implied price of the 3 Year zero to the market price, $P_{\S} = 0.8700$.



Note how different P_1 (5%) and P_1 (3%) are from the values we found earlier whilst imposing that $q = \frac{1}{2}$. In the “implied” tree, they are smaller than those obtained with $q = \frac{1}{2}$, state by state. This is because in the implied tree, $q = 0.8779$, such that the model can match a lower initial price, 0.8700. The implied tree puts more weight on those states of nature where the short-term rate is high or, equivalently, bond prices are low. We expect the price of the option in the implied binomial tree to be lower from that we found earlier, because option prices decrease with the underlying.

So let’s do the computations by utilizing the *implied binomial tree*:

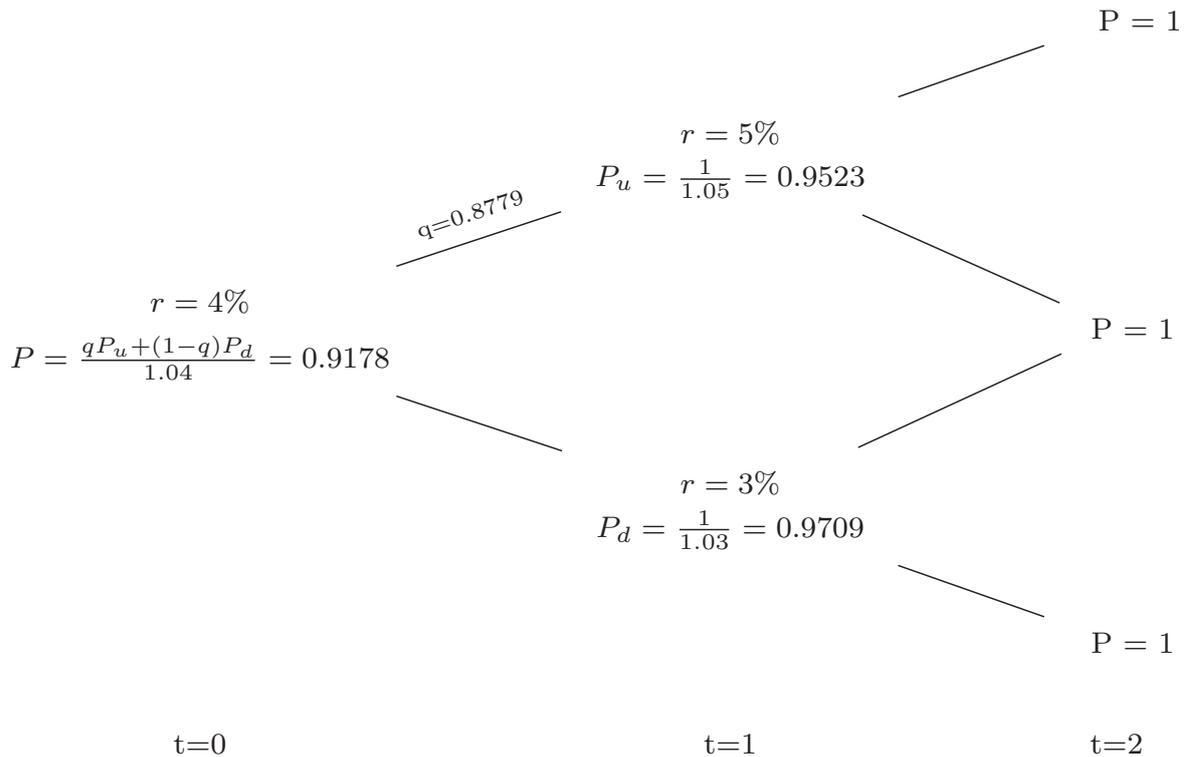


The calculations in the previous diagram reveal that the option price predicted by the implied binomial tree is 0.0026, which is one order of magnitude less than the option price we find earlier, 0.0124! The interpretation of this result relates, again, to the implied risk-neutral probability, which is much larger than $q = \frac{1}{2}$. The implied tree puts a relatively large weight on events where the short-term rate is high or bond prices are low, which makes the option price relatively so small.

11.5.3.3 Another zero

We are not done yet, really. Let us go back to the zero pricing problem, and assume we observe the price of a 2 Year zero, and that this price equals 0.9200, a reasonable figure. Is there any chance that the inputs to the pricing problem for the 3 Year zero could also lead to fit the 2 Year zero without errors? Of course there isn't. Indeed, in the next diagram, we use the inputs to the 3 Year zero, and Eq. (11.19), and find that the price of the 2 Year zero implied by the price of the 3 Year zero is equal to 0.9178. Unless the market price happens, by chance, to equal 0.9178, we cannot simultaneously fit the price of the 3 Year and the 2 Year zeros.

To simultaneously fit the price of the 3 Year and the 2 Year zeros, we should implement at least one of the two strategies: (i) to make the probabilities q time-varying; (ii) to calibrate the entire structure of the short-term movements in Figure 11.6. We implement the first of these two strategies in the next subsection. We develop the second strategy in Section 11.5.



11.5.3.4 Implementing implied binomial trees

We build up implied binomial trees in more general cases, arising in the presence of several bond prices to be matched. Suppose the time interval is six months, such that the short-term rate is for six months. The current short-term rate is 3.99%, annualized. It can change to either 4.50% or to 4.00%, with equal (physical) probability. Suppose that two zeros are available for trading: a 6M zero and a 1Y zero, where the current price of the 1Y zero is 0.95974. What is the risk-neutral probability implied by this tree? This probability must be such that, the price of all the zeros are matched exactly.

The tree we face is depicted below.

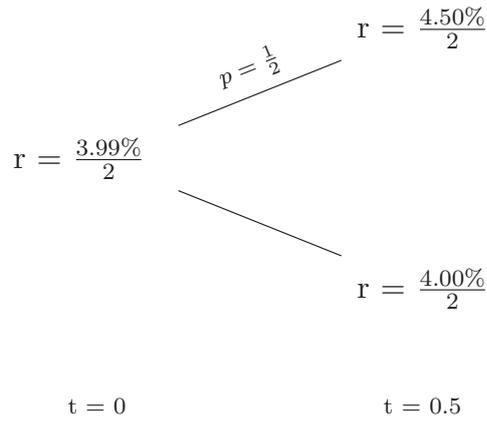
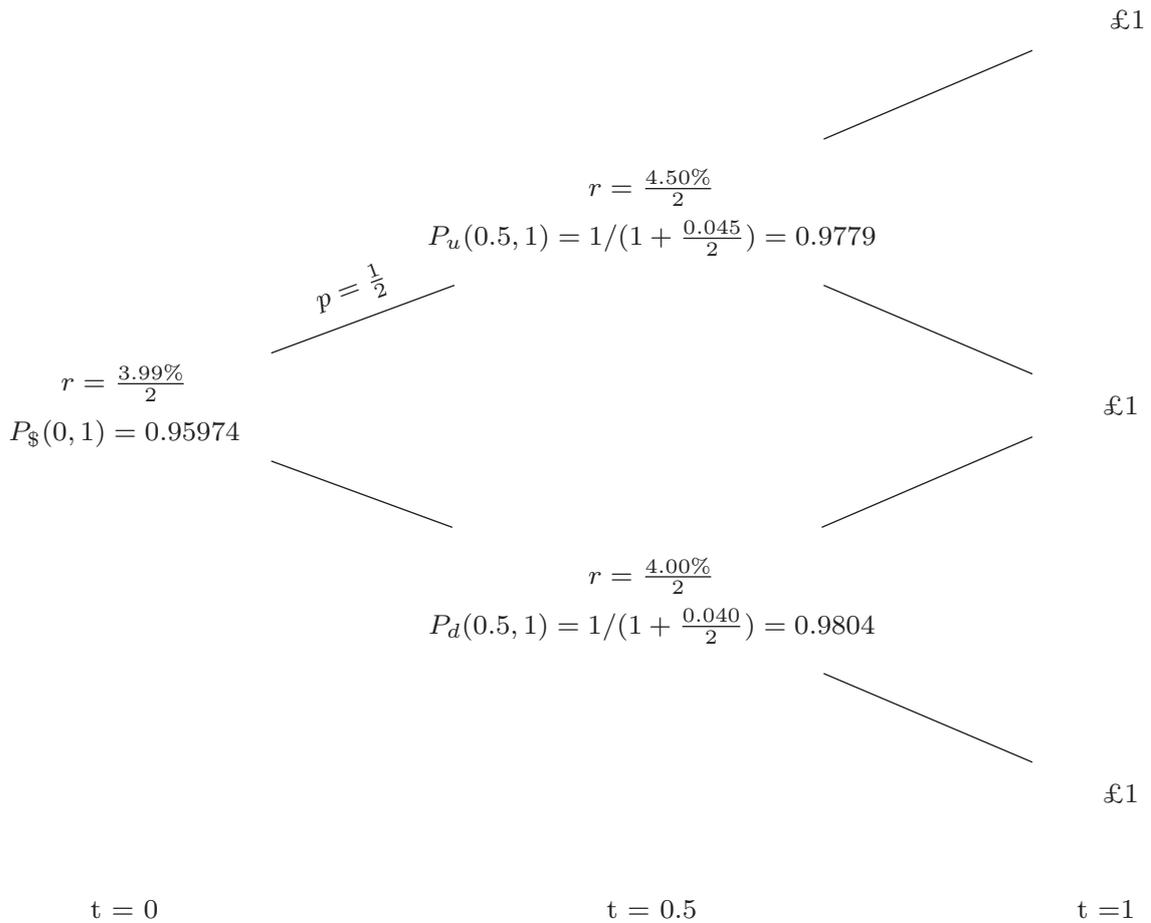


FIGURE 11.7. The dynamics of the short-term rate: high interest rate scenario

In this tree, $p = \frac{1}{2}$ denotes the *physical* probability. Naturally, the price of a 6M zero at $t = 0$, equals, $P_{\S}(0, 0.5) = 1 / (1 + \frac{0.0399}{2}) = 0.9804$. This price is actually observed. That is, the current short-term rate, 3.99%, is a mere definition. Next, we proceed to find the no-arbitrage movements of the 1Y zero, which are displayed below.



Note, the current market price, $P_{\S}(0, 1) = 0.95974$, is less than the *expected* price to prevail tomorrow, discounted at the current interest rate,

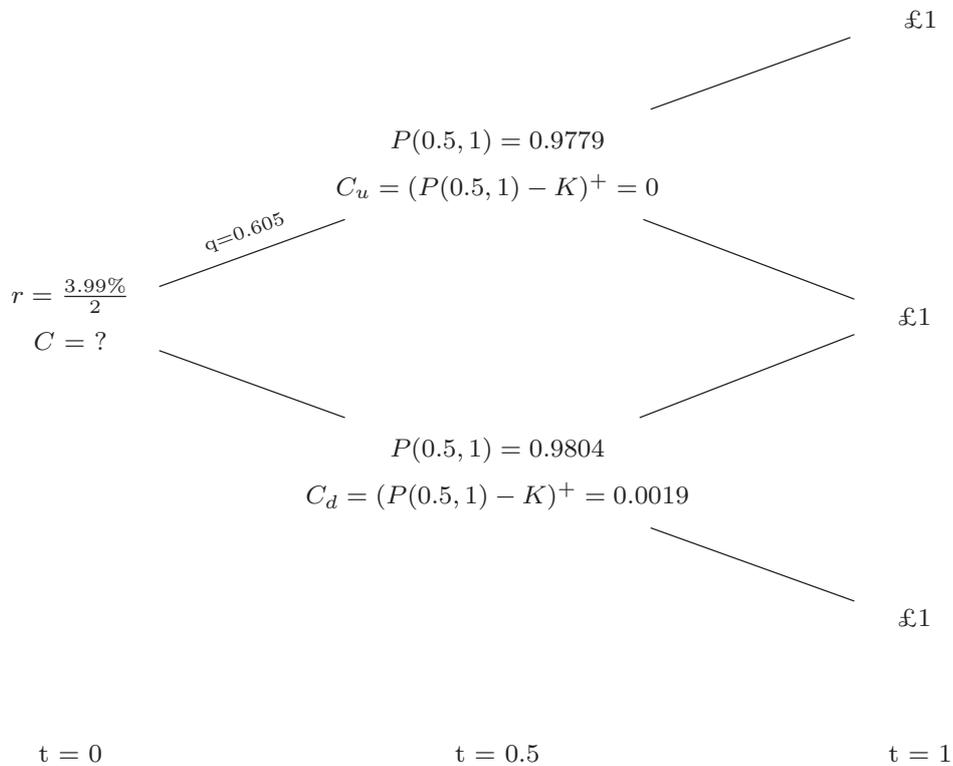
$$\frac{1}{1+r} E_p [P(0.5, 1)] = \frac{1}{1 + \frac{0.0399}{2}} \left(\frac{1}{2} 0.9779 + \frac{1}{2} 0.9804 \right) = 0.9599.$$

Hence, $p = \frac{1}{2}$ cannot be the risk-neutral probability. To find out the risk-neutral probability, we proceed as follows. In the absence of arbitrage opportunities,

$$\begin{aligned} P_{\S}(0, 1) &= 0.95974 \\ &= \frac{1}{1+r} [qP_{\text{up}}(0.5, 1) + (1-q)P_{\text{down}}(0.5, 1)] \\ &= \frac{1}{1 + \frac{0.0399}{2}} [q \cdot 0.9779 + (1-q) \cdot 0.9804] \end{aligned}$$

with obvious notation. This is one equation with one unknown, q , which is solved by $q = 0.605$.

We may now proceed with pricing derivatives. Consider a call option on the 1Y zero, with expiration date in six months and exercise price equal to 0.9785. Its payoff is as depicted below:

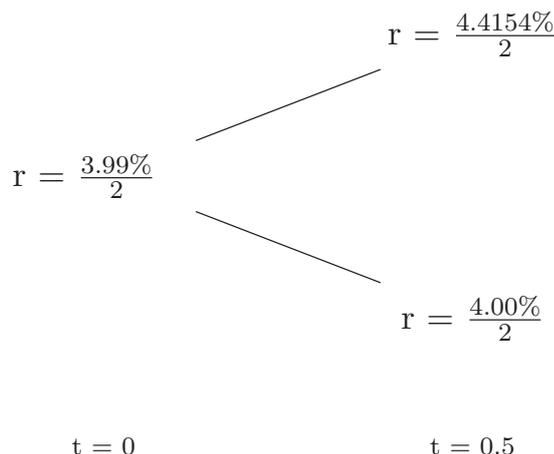


So the option price is, by risk-neutral evaluation,

$$C = \frac{1}{1 + \frac{0.0399}{2}} [q \cdot 0 + (1-q) \cdot 0.0019] = 0.9804 [0.395 \cdot 0.0019] = 7.3579 \times 10^{-4}. \quad (11.21)$$

What happens when the short-term rate does not evolve as in the diagram of Figure 11.7 but, instead, as in Figure 11.8?

FIGURE 11.8. The dynamics of the short-term rate: low interest rate scenario



The previous tree is one where the short-term rate in the upper state of the world equals $r = 4.4154\%$, not 4.50% , as in Figure 11.7. It implies that:

$$P_{\text{up}}(0.5, 1) = \frac{1}{1 + \frac{r}{2}} = \frac{1}{1 + \frac{4.4154\%}{2}} = 0.9784.$$

Then, the risk-neutral probability, q , solves the following pricing equation,

$$\begin{aligned} P_{\S}(0, 1) &= 0.95974 \\ &= \frac{1}{1 + r} [qP_{\text{up}}(0.5, 1) + (1 - q)P_{\text{down}}(0.5, 1)] \\ &= \frac{1}{1 + \frac{0.0399}{2}} [q \cdot 0.9784 + (1 - q) \cdot 0.9804]. \end{aligned}$$

The solution is, $q = 0.756$, which is higher than the solution we found earlier using the tree in Figure 11.7 (i.e., $q = 0.605$). The option price is, now,

$$C = \frac{1}{1 + \frac{0.0399}{2}} [q \cdot 0 + (1 - q) \cdot 0.0019] = 0.9804 [0.244 \cdot 0.0019] = 4.5451 \times 10^{-4}. \quad (11.22)$$

Why is this price smaller than that computed in Eq. (11.21)? In the tree of Figure 11.8, the up-state of the world is, so to speak, less severe than the up-state of the world in the tree of Figure 11.7. To be able to match the initial price $P_{\S}(0, 1) = 0.95974$, the model in Figure 11.8 must put more weight on the up-state of the world, i.e. a larger implied risk-neutral probability. This implies a larger risk-neutral probability that low bond prices will arise in the future and, hence, a lower option price.⁴

In a segmented market, two investment banks might have different views about developments in the short-term rate—the view in Figure 11.7 and that in Figure 11.8. The first bank favours a “high” interest rate scenario, but it is not too risk-averse to that scenario ($r_{\text{up}} = 4.5\%$, $q = 0.605$). The second bank favours a “mild” interest rate scenario, although it assigns a greater chance of this scenario to arise ($r_{\text{up}} = 4.4154\%$, $q = 0.756$). But then, naturally, both institutions need to agree on the initial bond price, $P_{\S}(0, 1) = 0.95974$. The segmentation could

⁴Mathematically, we have that $P_{\S}(0, 1) = \frac{1}{1+r} (P_{\text{down}} - q\Delta P)$, where $\Delta P \equiv P_{\text{down}} - P_{\text{up}} > 0$. While the price P_{down} in the tree of Figure 11.8 is the same as that in Figure 11.7, the “bond price volatility”, i.e. the difference ΔP , is lower in Figure 11.8 than in Figure 11.7. Therefore, the tree in Figure 11.8 is consistent with the given market price, $P_{\S}(0, 1)$, only when q increases from 0.605 to 0.756.

arise, for example, because the clientèle of the first bank and that of the second bank are unlikely to meet and, the prices for the option charged by the banks are not publicly known. In the absence of market imperfections (and arbitrage), however, the investment banks should agree on the option price too. Note that the price in Eq. (11.22) is almost half of that in Eq. (11.21). Derivatives can be quite nonlinear object, due to their optionality. A small deviation in the assumptions on the short-term rate developments can lead to dramatic option pricing implications.

Next, let us add another period to the tree in Figure 11.7, assuming that the short-term rate is as in the following diagram:

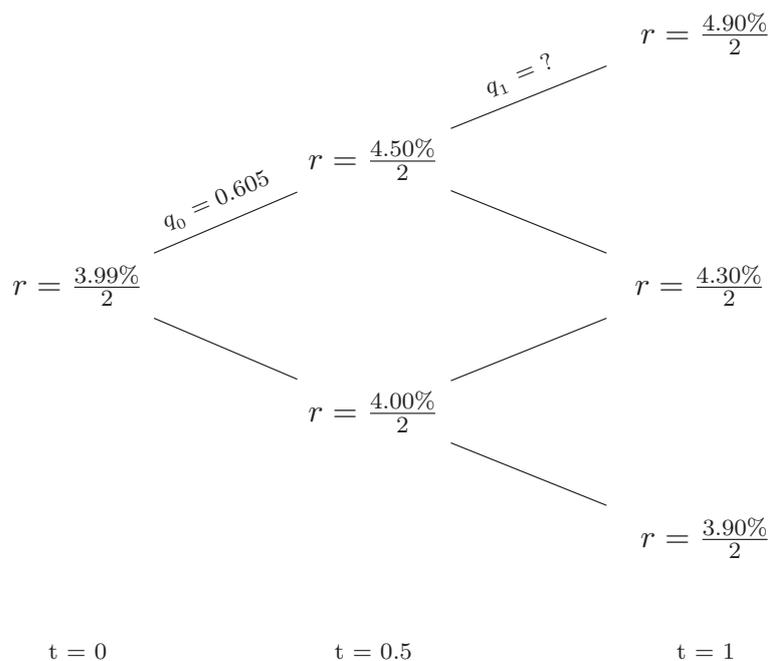


FIGURE 11.9.

In this tree, q_0 is the risk-neutral probability for the first period, and q_1 is the risk-neutral probability for the second period. We already know that $q_0 = 0.605$. The probability q_1 is the risk-neutral probability for the time-period $(0.5, 1)$, and can differ from q_0 . Suppose, also, that an additional zero is available for trading, a 1.5Y zero. The current price of this zero is $P_{\S}(0, 1.5) = 0.9382$. To derive the risk-neutral probability q_1 , we proceed to model the implied tree for the 1.5Y zero, as follows.

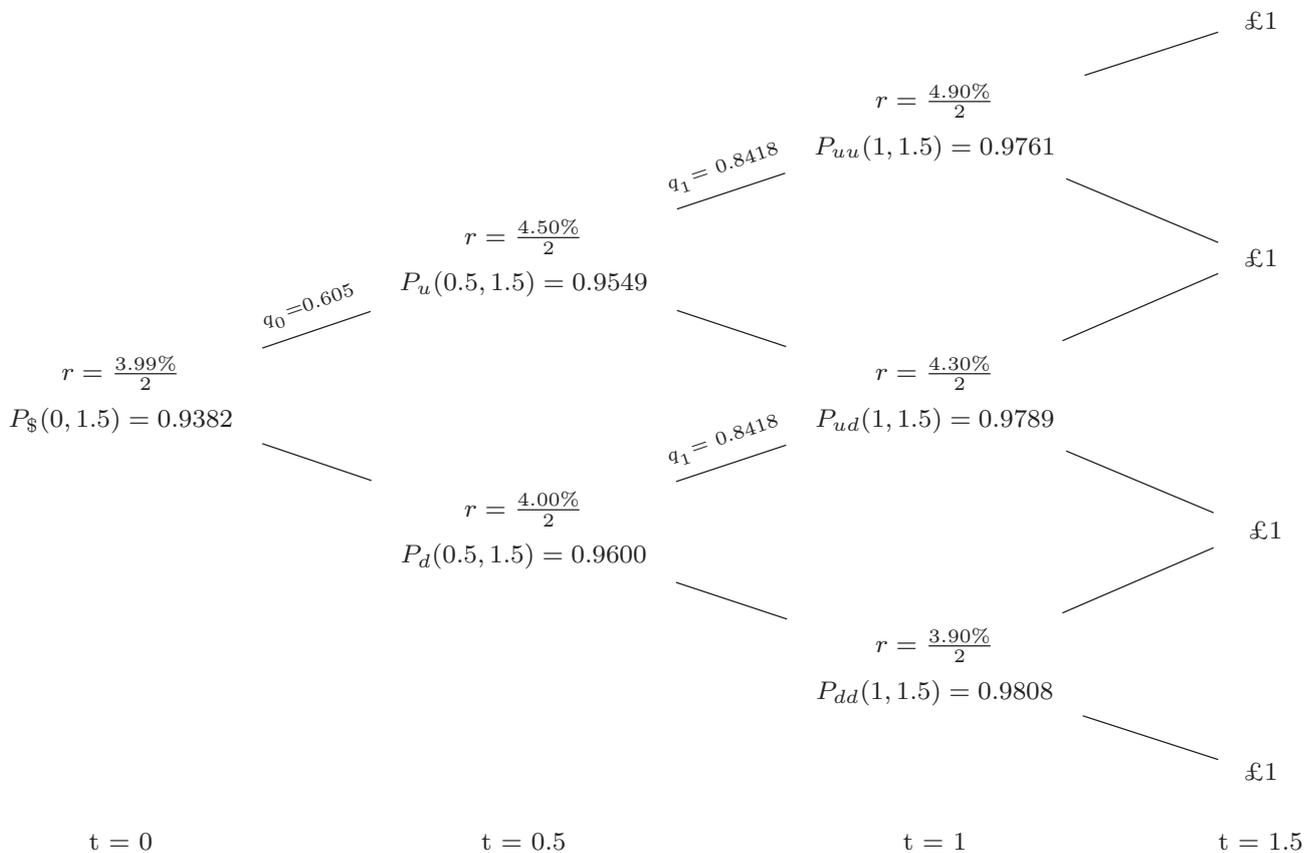
where $P_U(0.5, 1.5)$ and $P_D(0.5, 1.5)$ are as in Eqs. (11.23)-(11.24), and where $q_0 = 0.605$. So we have,

$$0.9382 = \frac{1}{1 + \frac{0.0399}{2}} [0.605 \cdot P_U(0.5, 1.5) + 0.395 \cdot P_D(0.5, 1.5)], \quad (11.25)$$

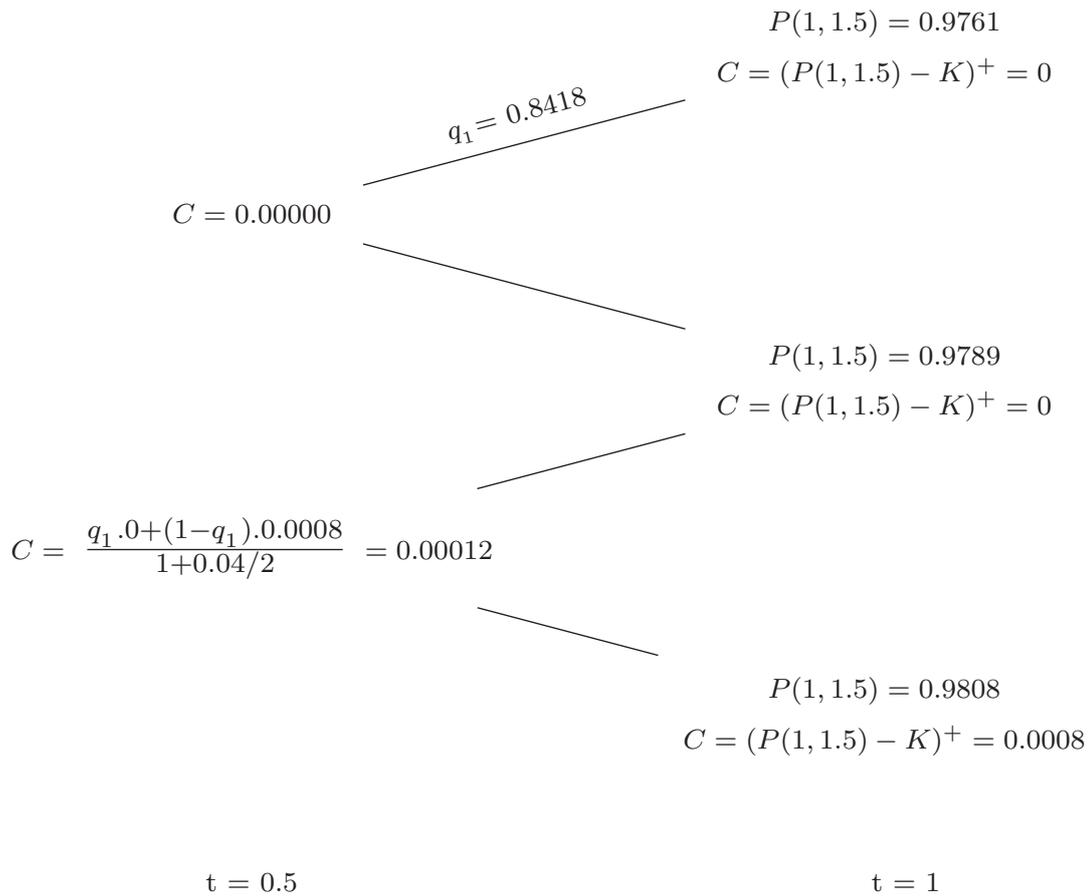
where $P_U(0.5, 1.5)$ and $P_D(0.5, 1.5)$ are as in Eqs. (11.23)-(11.24). Hence, by replacing Eqs. (11.23)-(11.24) into Eq. (11.25) leaves one equation with one unknown, q_1 . Solving, yields, $q_1 = 0.8412$, which implies that,

$$P_U(0.5, 1.5) = 0.9549, \quad P_D(0.5, 1.5) = 0.9600.$$

So, to sum up, we have the tree below.



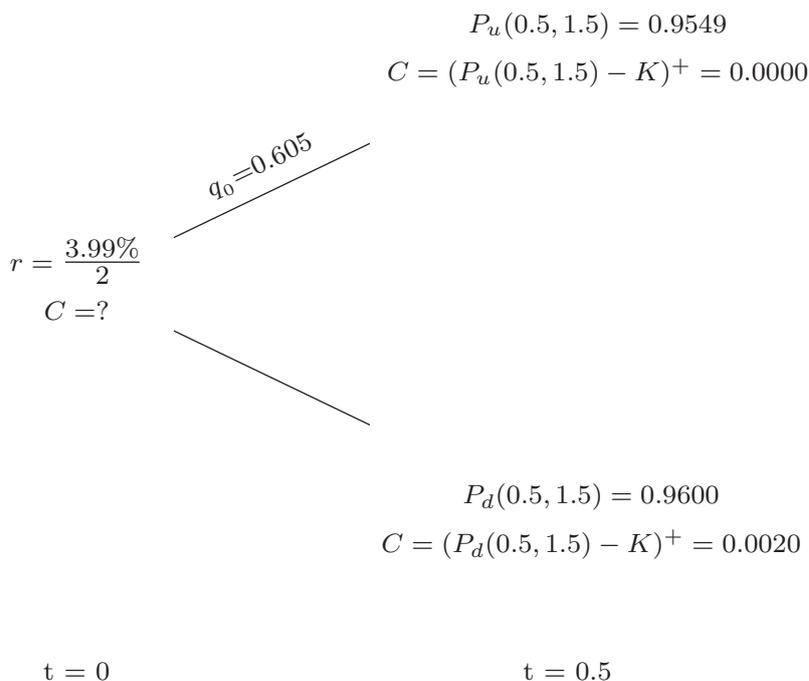
We are now ready to compute the no-arbitrage price of derivatives. Consider, for example, a call option on the 1.5Y zero, with expiration date in 1Y and exercise price equal to 0.9800. The price of the option at time $t = 0.5$, is either zero or $C = 0.00012$, as illustrated below.



We can now calculate the no-arbitrage price of the 1Y call option on the 1.5Y zero, struck at $K = 0.9800$. It is,

$$C = \frac{1}{1 + \frac{0.0399}{2}} [0 \cdot q_0 + 0.00012 \cdot (1 - q_0)] = 0.9804 [0.00012 \cdot (1 - 0.605)] = 4.647 \times 10^{-5}.$$

We can use Figure 11.9 to price additional derivatives, such as, say, a call option on the 1.5Y zero, with expiration date in six months, and exercise price equal to 0.9580. We have the following tree.

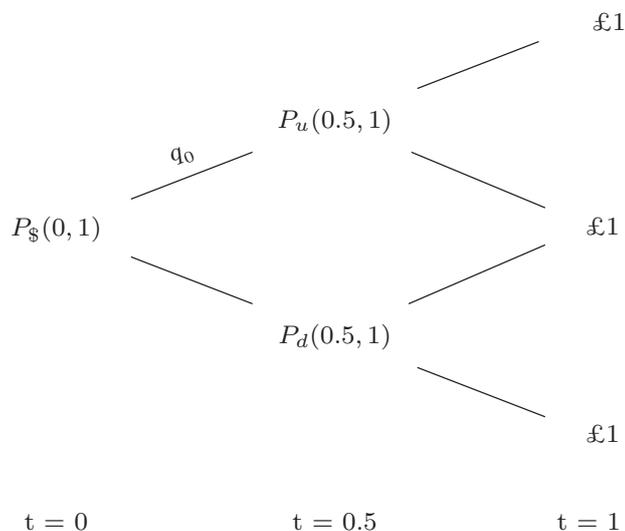


Therefore, the no-arbitrage price of the option is,

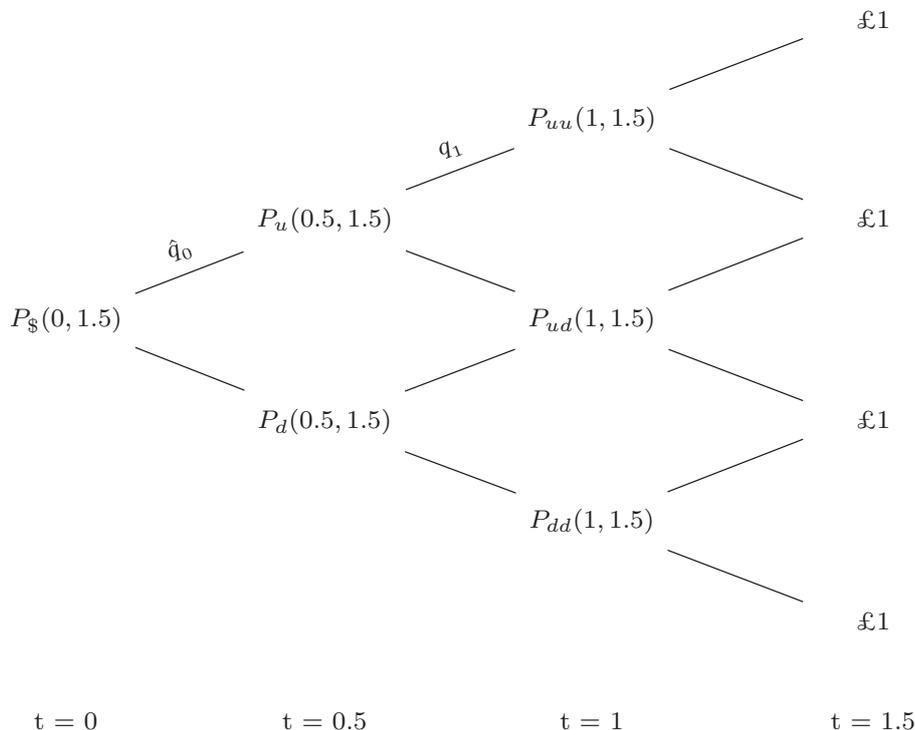
$$C = \frac{1}{1 + \frac{0.039}{2}} [q_0 \cdot 0 + (1 - q_0) \cdot 0.0020] = 0.9804 [0.395 \cdot 0.0020] = 7.745 \times 10^{-4}.$$

11.5.3.5 Summing up

What have we done? Note, the starting point is the probabilistic representation of the movements of the short-term rate in Figure 11.9, which we use to recover the two risk-neutral probabilities q_0 (for the time span $(0, 0.5)$) and q_1 (for the time span $(0.5, 1)$), using the information about the market price of two zeros, the 1Y and the 1.5Y. Precisely, given $P_{\S}(0, 1)$, the price of the 1Y zero, we recover q_0 , as illustrated below:



This is possible as $P_U(0.5, 1)$ and $P_D(0.5, 1)$ do not “depend” on q_0 and so they are obtained in a straightforward manner. Given q_0 , then, we compute q_1 , using $P_{\S}(0, 1.5)$, the price of the 1.5Y zero, as illustrated below:



Again, the risk-neutral probability, q_1 , can be recovered because $P_{UU}(1, 1.5)$, $P_{UD}(1, 1.5)$ and $P_{DD}(1, 1.5)$ do not “depend” on q_1 , as they are next to expiration, and are thus easily obtained. Given $P_{UU}(1, 1.5)$, $P_{UD}(1, 1.5)$ and $P_{DD}(1, 1.5)$, we can express $P_U(0.5, 1.5)$ and $P_D(0.5, 1.5)$ as two linear functions of q_1 . Finally, we impose the no-arbitrage property to $P_{\S}(0, 1.5)$, which forces the market price, $P_{\S}(0, 1.5)$, to be a linear function of $P_U(0.5, 1.5)$ and $P_D(0.5, 1.5)$ and, hence, q_1 , thereby allowing us to recover q_1 .

We can continue, by adding one more time period, as in the tree in Figure 11.10 below. We can recover q_2 , once we are given the market price of a 2Y zero, $P_{\S}(0, 2)$, as follows:

- The prices of the 2Y zero at time $t = 1.5$ (the filled nodes in Figure 11.10) (say $P(1.5, 2)$) are easily computed, given an assumption about the numerical values of the short-term rate in those nodes.
- Given the prices $P(1.5, 2)$ at time $t = 1.5$, and the previously calibrated probabilities \hat{q}_0 and \hat{q}_1 , we impose no-arbitrage thereby expressing the current market price $P_{\S}(0, 2)$ as a linear function of q_2 . Then, we solve for q_2 .

The calibration can continue. We extend the tree, by considering one more period. Then, we use the price of one additional zero to “recover” time varying risk-neutral probabilities. An alternative procedure consists in: (i) *fixing* the risk-neutral probabilities q to some value at all times (e.g., $q = \frac{1}{2}$), and (ii) figuring out the “implied” values for the short-term rate in each node of the tree. The next section develops a systematic approach for implementing this procedure.

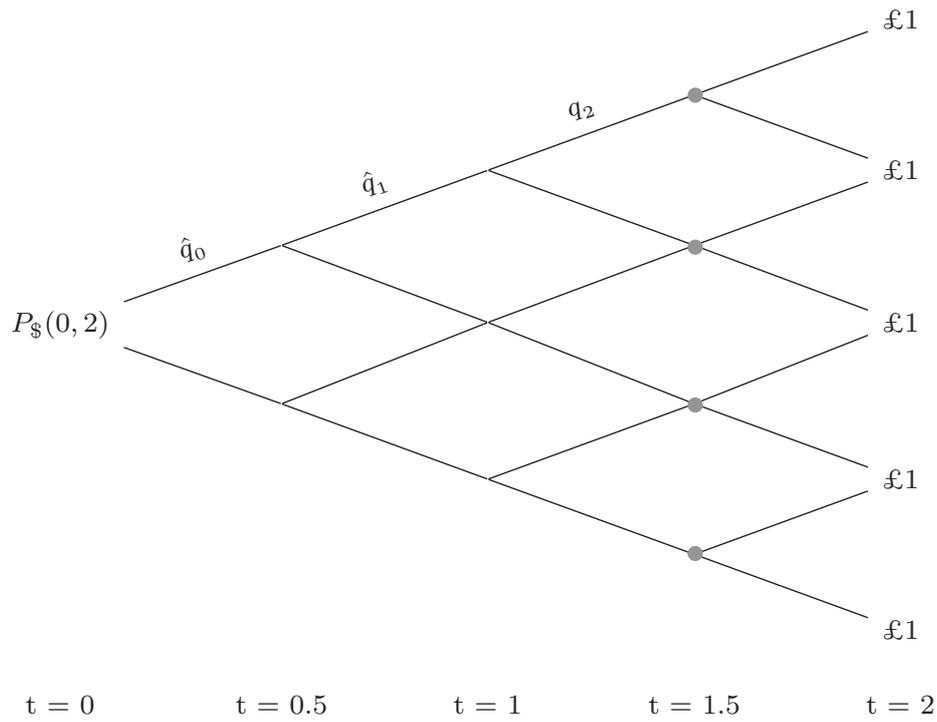


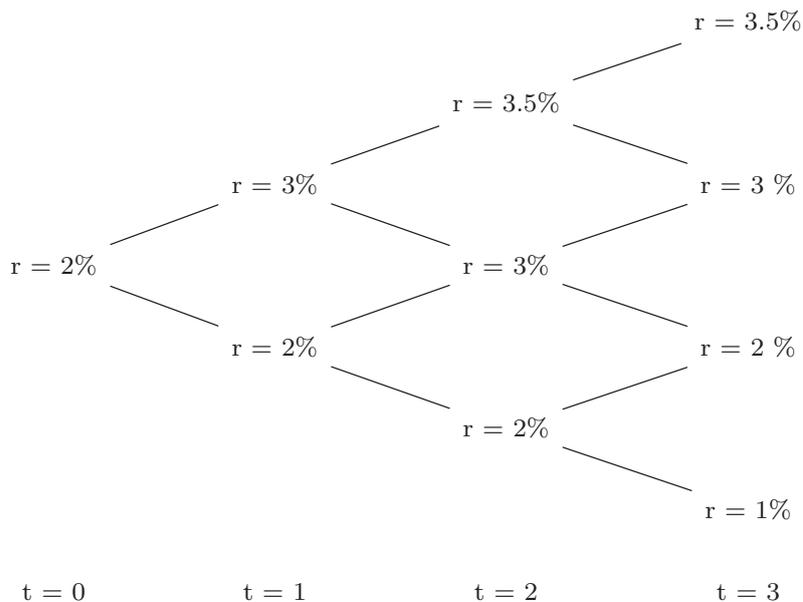
FIGURE 11.10.

11.5.4 Calibrating probabilities through derivative data

This section deals with two numerical examples where we exploit information from derivative data to say something about the assets underlying the very same derivative contracts. Namely, we shall use derivative data to calibrate risk-neutral probabilities.

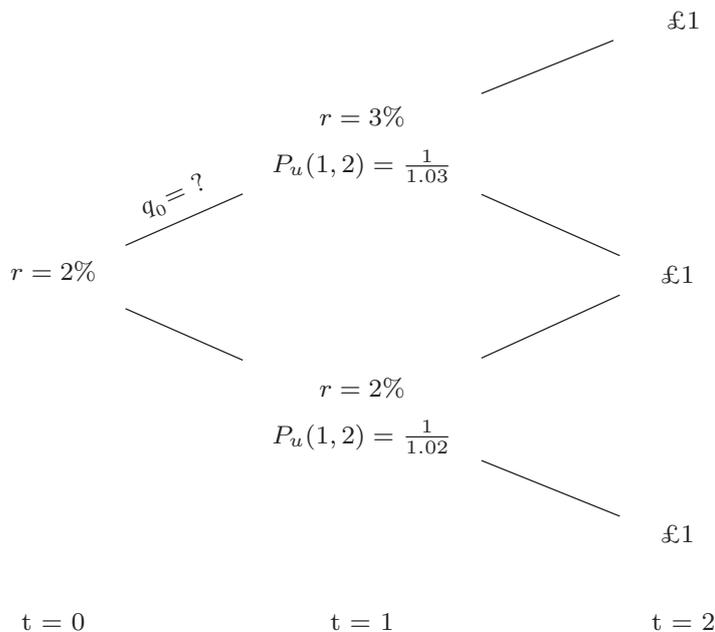
11.5.4.1 Options

Suppose that a two year zero coupon bond is traded for a price equal to $P_{\$}(0, 2) = 0.95500$. We assume that the short-term rate evolves over time according to the tree described in the following diagram.



Suppose that a European call option written on a three year zero coupon bond is traded. This option has a strike price equal to 0.97000, expires in two years, and quotes for $C_{\S}(0, 2) = 1.0141 \cdot 10^{-3}$. We can use the price of this derivative, to find the no-arbitrage price of a three year bond which, every year, pays off 3% of the principal of 1.00. Precisely, we use the price of the two year zero coupon bond to recover the risk-neutral probability applying for the first year, and the price of the option to recover the risk-neutral probability applying for the second year. With these probabilities, we determine the no-arbitrage price of the three year bond. (We assume the two probabilities are state-independent, for otherwise we would need the price of additional assets to reverse-engineer state independent risk-neutral probabilities.)

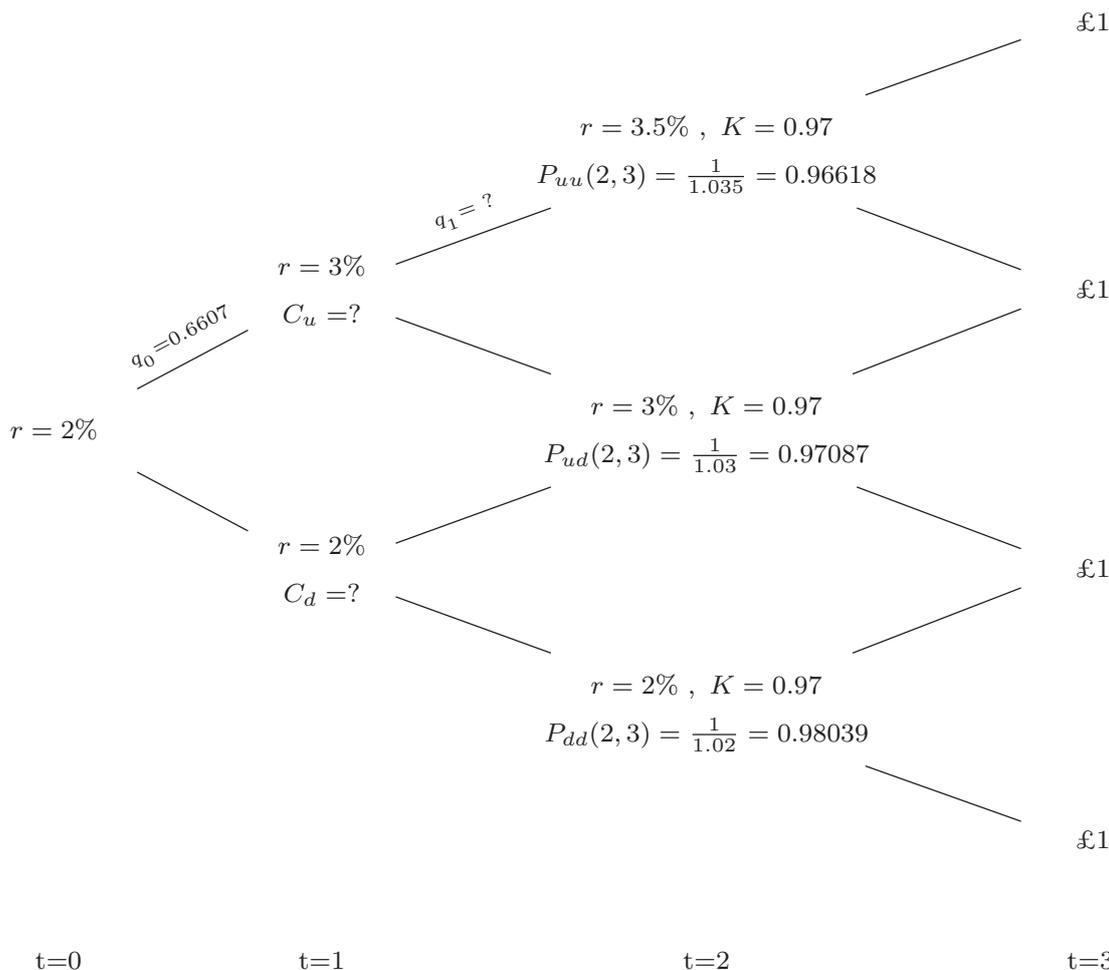
So we know that $P_{\S}(0, 2) = 0.95500$. Moreover, as illustrated below, we can extract the price of the 2Y bond in the up- and down- states of the world at time $t = 1$.



We have $P_U(1, 2) = \frac{1}{1.03} = 0.97087$, and $P_D(1, 2) = \frac{1}{1.02} = 0.98039$. We can now solve for the risk-neutral probability. We have,

$$\begin{aligned} P_{\S}(0, 2) &= 0.95500 \\ &= \frac{1}{1.02} (q_0 \cdot P_U(1, 2) + (1 - q_0) P_D(1, 2)) \\ &= \frac{1}{1.02} (q_0 \cdot 0.97087 + (1 - q_0) 0.98039). \end{aligned}$$

Solving for q_0 , yields, $q_0 = 0.6607$. We use this probability, and the price of the option, $C_{\S}(0, 2)$, to solve for the risk-neutral probability for the second period of the tree, as illustrated below.



In this tree, $K = 0.97000$ is the strike price of the option. The option price at time $t = 1$, in the two states, can be either C_U or C_D , where:

$$\begin{aligned} C_U &= \frac{1}{1.03} [q_1 \max \{P_{UU}(2, 3) - K, 0\} + (1 - q_1) \max \{P_{UD}(2, 3) - K, 0\}] \\ &= \frac{1}{1.03} (1 - q_1) \cdot 0.00087. \\ C_D &= \frac{1}{1.02} [q_1 \max \{P_{UD}(2, 3) - K, 0\} + (1 - q_1) \max \{P_{DD}(2, 3) - K, 0\}] \\ &= \frac{1}{1.02} [q_1 \cdot 0.00087 + (1 - q_1) \cdot 0.01039]. \end{aligned}$$

Hence, the option price satisfies,

$$\begin{aligned}
 C_{\S}(0, 2) &= 1.0141 \cdot 10^{-3} \\
 &= \frac{1}{1.02} (q_0 C_U + (1 - q_0) C_D) \\
 &= \frac{1}{1.02} (0.6607 C_U + 0.3393 C_D) \\
 &= \frac{1}{1.02} \left(0.6607 \frac{1}{1.03} (1 - q_1) \cdot 0.00087 + 0.3393 \frac{1}{1.02} [q_1 \cdot 0.00087 + (1 - q_1) \cdot 0.01039] \right).
 \end{aligned}$$

Solving for q_1 yields, $q_1 = 0.8000$.

Next, we compute the price of the zero maturing at time 3, $P(0, 3)$. We use the diagram in Figure 11.11.

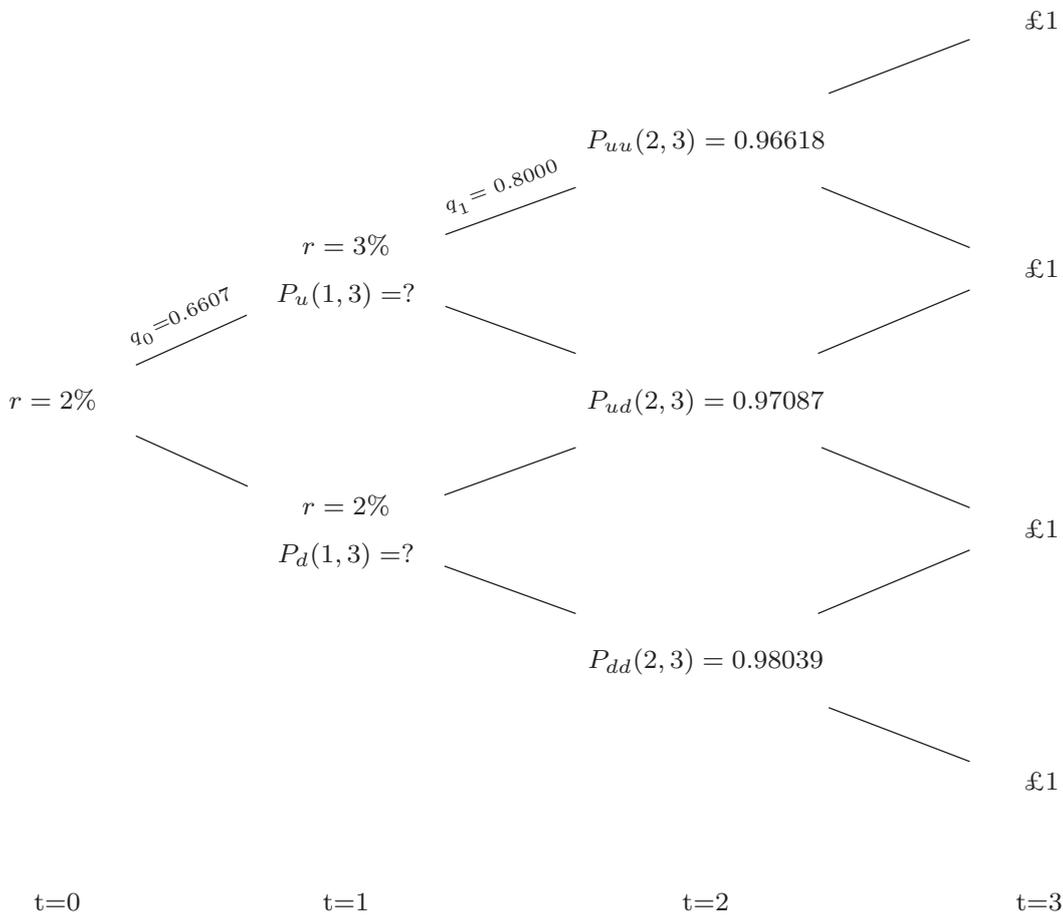


FIGURE 11.11.

We have,

$$\begin{aligned}
 P_U(1, 3) &= \frac{1}{1.03} [q_1 P_{UU}(2, 3) + (1 - q_1) P_{UD}(2, 3)] \\
 &= \frac{1}{1.03} (0.80 \cdot 0.96618 + 0.20 \cdot 0.97087) = 0.93895. \\
 P_D(1, 3) &= \frac{1}{1.02} [q_1 P_{UD}(2, 3) + (1 - q_1) P_{DD}(2, 3)] \\
 &= \frac{1}{1.02} (0.80 \cdot 0.97087 + 0.20 \cdot 0.98039) = 0.95370.
 \end{aligned}$$

The price of a 3Y zero coupon bond, embedded in the market prices, $P_{\S}(0, 2)$ and $C_{\S}(0, 2)$, is therefore:

$$\begin{aligned}
 P(0, 3) &= \frac{1}{1.02} [q_0 P_U(1, 3) + (1 - q_0) P_D(1, 3)] \\
 &= \frac{1}{1.02} (0.6607 \cdot 0.93895 + 0.3393 \cdot 0.95370) = 0.92545.
 \end{aligned}$$

We are now ready to evaluate the 3Y bond with 3% coupon rate. It is,

$$\begin{aligned}
 P_{\text{coupon}=3\%}(0, 3) &= 0.03 \cdot [P_{\S}(0, 1) + P_{\S}(0, 2) + P(0, 3)] + P(0, 3) \\
 &= 0.03 \cdot (0.98039 + 0.95500 + 0.92545) + 0.92545 = 1.0113.
 \end{aligned}$$

The discretely compounded yield curve implied by the previous calculations is given by $r_{0,1} = 2.00\%$ (1Y); $r_{0,2} : 0.95500 = (1 + r_{0,2})^{-2}$, or $r_{0,2} = 2.328\%$ (2Y); and $r_{0,3} : 0.92545 = (1 + r_{0,3})^{-3}$, or $r_{0,3} = 2.616\%$ (3Y). Note, we are capable of computing the yield curve, without knowing all bond data, but inverting some of them from the price of an option! We can go further. Suppose the price of the “missing” bond becomes available, so to speak. We want to make sure this price is consistent with absence of arbitrage. Suppose, for example, that the market price is $P_{\S}(0, 3) > P(0, 3) = 0.92545$, say. Then, we can sell short the 3Y zero, and set up a dynamic, self-financing strategy aiming to replicate the 3Y zero, i.e. capable of delivering £1 at maturity.

We would proceed as follows. Consider the tree in Figure 11.11. We build up a portfolio, which is long the option and a MMA. We assume the 3Y bond “converges” to the values P_{UU} , P_{UD} and P_{DD} in Figure 11.11, for otherwise we might implement a trivial arbitrage from time $t = 2$ to $t = 3$. At time $t = 0$, we go long Δ_0 options and M_0 units of the MMA, so as to make sure that the portfolio delivers $P_U(1, 3)$ in the upstate of $P_D(1, 3)$ in the downstate—thereby ensuring that the price of the bond is replicated at time $t = 1$. The value of this replicating strategy is, of course, $P(0, 3)$, so by short-selling the 3Y bond at $t = 0$, we obtain an initial profit, equal to $P_{\S}(0, 3) - P(0, 3)$. Suppose, then, that at time $t = 1$, we are in the up-node, such that the bond price is $P_U(1, 3)$. In this node, we can build up another portfolio long Δ_1 options and M_1 units of the MMA, aiming to replicate the price of the bond at time $t = 2$ —either $P_{UU}(2, 3)$ or $P_{UD}(2, 3)$. The value of this replicating portfolio would be just $P_U(1, 3)$, which is what is obtained by the replicating strategy at implemented at time $t = 0$. The strategy is clearly self-financed, as the following calculations reveal. By construction, $\Delta_0 C_U(1) + M_0(1 + r) = P_U(1, 3) = \Delta_1 C_U(1) + M_1$ (with $r = 2\%$), and $\Delta_1 \tilde{C}_U(2) + M_1(1 + r) = P_U(2, 3)$, where $r = 3\%$, and: $\tilde{C}_U(2), \tilde{P}_U(2, 3)$ are either $C_{UU}(2), P_{UU}(2, 3)$, or $C_{UD}(2), P_{UD}(2, 3)$, at time

$t = 2$, with straight forward notation. Therefore, we have that,

$$\begin{aligned} \tilde{P}_U(2, 3) - P_U(1, 3) &= \Delta_1 \left(\tilde{C}_U(2) - C_U(1) \right) + rM_1 \\ &= \Delta_1 \left(\tilde{C}_U(2) - C_U(1) - rC_U(1) \right) + rP_U(1, 3). \end{aligned}$$

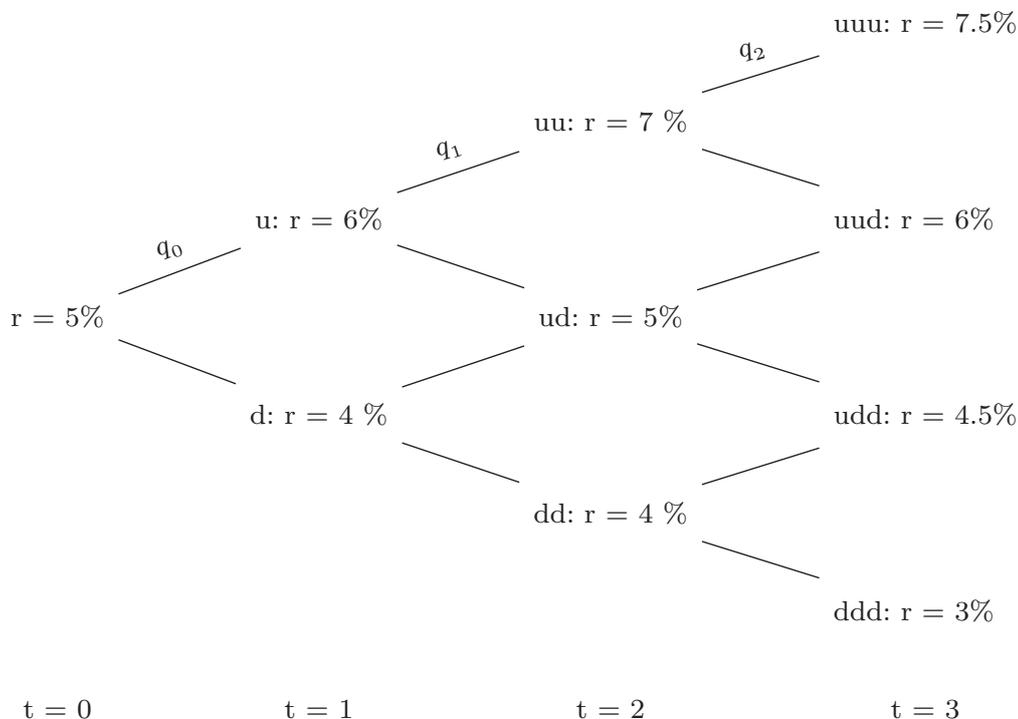
Likewise, if, instead, at time $t = 1$, we end up in the down-node, where the bond price (and the value of the strategy implemented at time $t = 0$) is $P_D(1, 3)$, we can invest $P_D(1, 3)$ in options and MMA so as to replicate the price of the bond at time $t = 2$ —either $P_{UD}(2, 3)$ or $P_{DD}(2, 3)$. The presence of dynamically complete markets allows us to implement an arbitrage.

11.5.4.2 Arrow-Debreu securities, and the pricing of interest rate derivatives

Arrow-Debreu securities are assets that only pay over a specific state of the world, as explained in Chapter 2. We shall deal with these securities in more detail in Section 11.7, because they allow us to implement perfectly fitting models quite elegantly. This section is a first introduction to them. We make use of Arrow-Debreu securities to, firstly, extract risk-neutral probabilities, and, secondly, to price quite basic interest rate derivatives, such as a “caplet” or a “forward rate agreement.” Likewise, the pricing of interest rate derivatives covered in this section is a preliminary introduction, as the next chapter will systematically deal with it, within a continuous time setting.

Extracting risk-neutral probabilities from Arrow-Debreu securities

Assume that the discretely compounded one-year rate, or the “short-term rate,” evolves over time as described by the following tree:



Assume three securities are available for trading: (i) a zero coupon bond expiring in two years, quoting for 0.91000; (ii) a zero coupon bond expiring in three years, quoting for 0.86500;

(iii) an Arrow-Debreu security, paying off £1 only at time 3 in the state “uuu” of the previous diagram, where the short-term rate equals 7.5%, quoting for 0.10000.

Assume that the risk-neutral probabilities of upward movements in the short-term rate change over time and take three values, q_0 , q_1 and q_2 , but are independent of the state of nature, as illustrated in the previous diagram. We can calibrate these probabilities, through the previously given available market data. First, we derive q_0 using the price of the two year bond, $P_{\$}(0, 2)$ say, as follows:

$$P_{\$}(0, 2) = 0.91000 = \frac{1}{1.05} (q_0 P_U(1, 2) + (1 - q_0) P_D(1, 2)),$$

where

$$P_U(1, 2) = \frac{1}{1.06} = 0.94340, \quad P_D(1, 2) = \frac{1}{1.04} = 0.96154.$$

Solving for q_0 yields $q_0 = 0.33284$. Next, we calibrate q_1 to match the price of the three year bond, $P_{\$}(0, 3)$ say. We have,

$$\begin{aligned} P_{\$}(0, 3) &= 0.86500 = \frac{1}{1.05} (q_0 P_U(1, 3) + (1 - q_0) P_D(1, 3)) \\ &= \frac{1}{1.05} (0.33284 * P_U(1, 3) + 0.66716 * P_D(1, 3)), \end{aligned}$$

where

$$\begin{aligned} P_U(1, 3) &= \frac{1}{1.06} (q_1 P_{UU}(2, 3) + (1 - q_1) P_{UD}(2, 3)), \\ P_D(1, 3) &= \frac{1}{1.04} (q_1 P_{UD}(2, 3) + (1 - q_1) P_{DD}(2, 3)), \end{aligned}$$

and

$$P_{UU}(2, 3) = \frac{1}{1.07} = 0.93458, \quad P_{UD}(2, 3) = \frac{1}{1.05} = 0.95238, \quad P_{DD}(2, 3) = \frac{1}{1.04} = 0.96154.$$

Solving for q_1 yields $q_1 = 0.66507$. Finally, we calibrate q_2 through the price of the Arrow-Debreu security paying in state “uuu.” This price, denoted as $p_{UUU}^{\$}(3)$, is given by:

$$p_{UUU}^{\$}(3) = 0.10000 = \frac{1}{1.05} \frac{1}{1.06} \frac{1}{1.07} q_0 q_1 q_2 = \frac{1}{1.05} \frac{1}{1.06} \frac{1}{1.07} 0.33284 * 0.66507 * q_2,$$

i.e. $q_2 = 0.53798$.

Pricing interest rate derivatives

Next, we use the previously calibrated probabilities to price some interest rate derivatives. First, consider a “caplet” contingent on the rates prevailing at time $t = 3$, paying off at time $t = 4$, with strike rate equal to 5%, and notional value equal to £100. The payoff of this derivative instrument at $t = 4$ is $\max\{L - 5\%, 0\}$, where L is the rate at $t = 3$. Therefore, the discounted payoffs at time $t = 3$ are:

$$\text{uuu: } \frac{1}{1.075} \max\{7.5 - 5, 0\} = 2.32560$$

$$\text{uud: } \frac{1}{1.06} \max\{6 - 5, 0\} = 0.94340$$

$$\text{udd: } \frac{1}{1.045} \max \{4.5 - 5, 0\} = 0$$

$$\text{ddd: } \frac{1}{1.03} \max \{3 - 5, 0\} = 0$$

As for time $t = 2$ and $t = 1$, we have:

$$t = 2 \quad \text{uu: } A_1 \equiv \frac{1}{1.07} (q_2 2.32560 + (1 - q_2) 0.94340) = 1.5766$$

$$\text{ud: } A_2 \equiv \frac{1}{1.05} (q_2 0.94340 + (1 - q_2) 0) = 0.48336$$

$$\text{dd: } A_3 \equiv 0$$

$$t = 1 \quad \text{u: } B_1 \equiv \frac{1}{1.06} (q_1 A_1 + (1 - q_1) A_2) = 1.1419$$

$$\text{d: } B_2 \equiv \frac{1}{1.04} (q_1 A_2 + (1 - q_1) A_3) = 0.30910$$

Therefore, the price of the caplet is:

$$\text{Caplet Price} = \frac{1}{1.05} (q_0 B_1 + (1 - q_0) B_2) = 0.55837.$$

Next, consider a forward rate agreement, whereby at time $t = 0$, two counterparties agree that at time $t = 4$, they will exchange with each other the variable short-term rate prevailing at time $t = 3$, against a fixed interest rate equal to K . We can use the previously calibrated probabilities to determine the *forward rate*, i.e. the level of K that makes the value of this agreement equal to zero at time $t = 0$. Take the case of a “payer” forward agreement, one for which the discounted payoffs at time $t = 3$ are:

$$\text{uuu: } \pi_1 \equiv \frac{1}{1.075} (7.5 - K)$$

$$\text{uud: } \pi_2 \equiv \frac{1}{1.06} (6 - K)$$

$$\text{udd: } \pi_3 \equiv \frac{1}{1.045} (4.5 - K)$$

$$\text{ddd: } \pi_4 \equiv \frac{1}{1.03} (3 - K)$$

At time $t = 2$ and $t = 1$, the payoffs are:

$$t = 2 \quad \text{uu: } \tilde{A}_1 \equiv \frac{1}{1.07} (q_2 (\pi_1 - \pi_2) + \pi_2) = 5.9519 - 0.87506 * K$$

$$\text{ud: } \tilde{A}_2 \equiv \frac{1}{1.05} (q_2 (\pi_2 - \pi_3) + \pi_3) = 4.7950 - 0.90443 * K$$

$$\text{dd: } \tilde{A}_3 \equiv \frac{1}{1.04} (q_2 (\pi_3 - \pi_4) + \pi_4) = 3.5215 - 0.92632 * K$$

$$t = 1 \quad \text{u: } \tilde{B}_1 \equiv \frac{1}{1.06} (q_1 \tilde{A}_1 + (1 - q_1) \tilde{A}_2) = 5.2495 - 0.83481 * K$$

$$\text{d: } \tilde{B}_2 \equiv \frac{1}{1.04} (q_1 \tilde{A}_2 + (1 - q_1) \tilde{A}_3) = 4.2004 - 0.87669 * K$$

We can now express the value of the contract as a function of the fixed rate K , as follows:

$$\text{Fwd}(K) \equiv \frac{1}{1.05} (q_0 \tilde{B}_1 + (1 - q_0) \tilde{B}_2) = 4.3329 - 0.82167 * K. \quad (11.26)$$

The forward rate is, simply, the value of K such that $\text{Fwd}(K) = 0$, i.e. $K = 5.27330\%$. More generally, we can determine the value of the forward rate agreement in Eq. (11.26), for any

value of K . For example, we have that $\text{Fwd}(K = 6) = 4.3329 - 0.82167 * 6 = -0.59712$, in percentage terms.

Finally, we can derive the price of a bond expiring at time $t = 4$, $P(0, 4)$, through the relation, $\frac{P_{\$}(0,3) - P(0,4)}{P(0,4)} = \frac{P_{\$}(0,3)}{P(0,4)} - 1 = 5.27330\%$, leading to $P(0, 4) = 0.82170$, which is indeed the same figure we can obtain by solving the tree for $P(0, 4)$, as we now show. We have, for time $t = 3$,

$$\text{uuu: } P_{UUU}(3, 4) = \frac{1}{1.075} = 0.93023$$

$$\text{uud: } P_{UUD}(3, 4) = \frac{1}{1.06} = 0.94340$$

$$\text{udd: } P_{UDD}(3, 4) = \frac{1}{1.045} = 0.95694$$

$$\text{ddd: } P_{DDD}(3, 4) = \frac{1}{1.03} = 0.97087$$

Then, we can solve, recursively, as usual:

$$t = 2 \quad \text{uu: } P_{UU}(2, 4) = \frac{1}{1.07} (q_2 P_{UUU}(3, 4) + (1 - q_2) P_{UUD}(3, 4)) = 0.87506$$

$$\quad \text{ud: } P_{UD}(2, 4) = \frac{1}{1.05} (q_2 P_{UUD}(3, 4) + (1 - q_2) P_{UDD}(3, 4)) = 0.90443$$

$$\quad \text{dd: } P_{DD}(2, 4) = \frac{1}{1.04} (q_2 P_{UDD}(3, 4) + (1 - q_2) P_{DDD}(3, 4)) = 0.92632$$

$$t = 1 \quad \text{u: } P_U(1, 4) = \frac{1}{1.06} (q_1 P_{UU}(2, 4) + (1 - q_1) P_{UD}(2, 4)) = 0.83481$$

$$\quad \text{d: } P_D(1, 4) = \frac{1}{1.04} (q_1 P_{UD}(2, 4) + (1 - q_1) P_{DD}(2, 4)) = 0.87669$$

Finally, we compute the price $P(0, 4) = \frac{1}{1.05} (q_0 P_U(1, 4) + (1 - q_0) P_D(1, 4)) = 0.82167 \approx 0.82170$, by rounding.

11.6 The Ho and Lee model

Ho and Lee (1986) develop a revolutionary approach to modeling yield curve movements. This approach does not rely on any economic theory meant to *explain* the yield curve that we observe. Rather, the objective is to take the yield curve as given, with a focus shifted towards modeling the *movements* of the entire yield curve. As explained, we need to “match” model prices to market prices, to avoid having derivatives with underlyings deviating from market prices. The next chapter derives the Ho and Lee model in continuous time, as this derivation allows us to illustrate a general approach to interest rate modeling, developed later by Heath, Jarrow and Morton (1992). The original derivation of the model is, however, in discrete time, and it is quite instructive to follow this approach here, so as to compare with alternative calibration methods.

The main idea underlying Ho and Lee is that the movements of the yield curve can be modeled through a binomial tree, much in the spirit of the Cox, Ross and Rubenstein (1979) tree representation of Black and Scholes (1973). However, in Black and Scholes (1973) and Cox, Ross and Rubenstein (1979), the asset underlying the option contract is a traded risk, such that the underlying price satisfies the martingale condition. Instead, interest rate derivatives generally depend on non-traded risks, which are not martingales. Moreover, the mere presence of boundary conditions induce bond return volatility to be time-varying.

Ho and Lee address these issues by modeling the movements of the entire collection of bond prices. We have three ways to achieve this task: (i) by making risk-neutral probs time-varying, for a given tree with predetermined values of the short-term rate (as in the previous sections);

(ii) by assuming a constant risk-neutral probability, and searching for the values of the short-term rate on the tree; (iii) by a combination of (i) and (ii). The Ho and Lee model relies on the second way. The key element of this model is the determination of the no-arbitrage ups and down of the entire yield curve, through modeling bond prices.

11.6.1 The tree

The key element of the model is the determination of the no-arbitrage ups and down of the entire yield curve, obtained by directly modeling bond prices of arbitrary expiration. Note that once bond prices are obtained, forward rates are obtained as a result. Therefore, the Ho and Lee model is a model of forward rate movements. It is a simple but powerful remark, because the key point of the model is, then, to re-express bond prices again, as a function of future forward rates. In this sense, the Ho and Lee model is a representation of current bond prices (in terms of forward rates), rather than a model of current bond prices. The key element of the Ho and Lee model relates to modeling future forward rate movements.

Assume that the price of any zero evolves according to a binomial tree. Let $P_j(t, T)$ be the price of a pure discount bond as of time t , with time to maturity $T - t$, after j upstate price movements of the bond price. Let $j \sim B(t, q)$, a binomial random variable,

$$E(j) = (1 - q)t, \quad Var(j) = q(1 - q)t,$$

where q is the risk-neutral probability of a single upstate movement. Therefore we have,

$$\begin{array}{ccc} & & P_{j+1}(t+1, T) \\ & \nearrow^{1-q} & \\ P_j(t, T) & & \\ & \searrow_q & \\ & & P_j(t+1, T) \end{array}$$

That is, if at time t , the number of upstate movements is equal to j then, at time $t + 1$, the number of upstate movements can either jump to $j + 1$, with probability $1 - q$, or stay at j , with probability q . (Therefore, we are now following the convention to have high values of the bond prices in the upper parts of the tree.) Note, further, that after one period, the price of any zero is one period closer to maturity. At maturity, $t = T$, the price of any zero is worth one unit of *numéraire*, viz

$$P_j(T, T) = 1, \quad \text{for all } j \text{ and } T.$$

Note, in the previous tree, it shall not necessarily hold that $P_j(t + 1, T) < P_j(t, T)$. On the contrary, we would expect that especially when the maturity approaches, $P_j(t + 1, T) > P_j(t, T)$, as the price of the zero needs to converge to par.

11.6.2 The price movements and the martingale restriction

In the absence of arbitrage opportunities, the expected return on the zero at t must equal the short-term rate, viz $P_j(t, T) = e^{-r_j(t)} E_q(P_j(t + 1, T))$, or

$$P_j(t, T) = P_j(t, t + 1) [(1 - q) P_{j+1}(t + 1, T) + q P_j(t + 1, T)], \quad (11.27)$$

where $P_j(t, t + 1) = e^{-r_j(t)}$, and $r_j(t)$ is the continuously compounded short-term rate at time t after j upward movements. We call this condition the *martingale restriction*.

Let us introduce notation for the movements of the price of any zero along the tree,

$$\underbrace{\frac{P_{j+1}(t+1, T)}{P_j(t, T)} = u(T-t) \frac{1}{P_j(t, t+1)}}_{\text{up at } t} \quad \text{and} \quad \underbrace{\frac{P_j(t+1, T)}{P_j(t, T)} = d(T-t) \frac{1}{P_j(t, t+1)}}_{\text{down at } t}. \quad (11.28)$$

The two functions $u(\cdot)$ and $d(\cdot)$, also called “perturbation functions,” are taken to be state-independent. They capture the fact that in the case of uncertainty, the price of the zero can either go up or down with respect to the risk-free of return. In other words, Eqs. (11.28) tell us that the discounted gross return from going long a bond is:

$$\underbrace{\frac{P(t+1, T)}{P(t, T)}}_{\text{Gross return}} \cdot \underbrace{P(t, t+1)}_{\text{Discount}} = \begin{cases} u(T-t) & \text{with probability } 1-q \\ d(T-t) & \text{with probability } q \end{cases}$$

where the two functions $u(T-t)$ and $d(T-t)$ have to be determined endogenously. If there was no uncertainty, we would have $u(T-t) = d(T-t) = 1$, for all $t \leq T$. In general, we have that $d(T-t) \leq 1 \leq u(T-t)$, as we shall now demonstrate.

One period before the expiration date, i.e. at $t = T - 1$, our price is certain to jump to one, with jump size equal to the short-term rate $r_j(t)$. Hence, the following boundary condition for the two functions $u(\cdot)$ and $d(\cdot)$ holds:

$$u(1) = d(1) = 1. \quad (11.29)$$

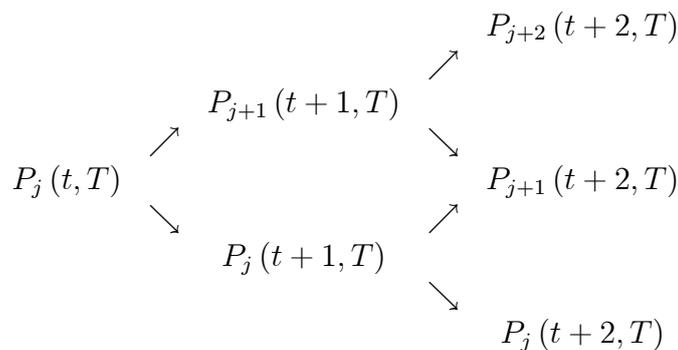
In terms of the two functions $u(\cdot)$ and $d(\cdot)$, the martingale restriction in Eq. (11.27) is,

$$1 = (1-q)u(T-t) + qd(T-t), \quad t \leq T. \quad (11.30)$$

This relation quite matches the standard risk-neutral relation for stock prices in which the short-term rate is tied down to the up and down movements of the stock price. However, in this context the up and down movements of the zero price depend on the maturity of the price itself through the two functions $u(T-t)$ and $d(T-t)$, which are endogenous, which makes the evaluation problem more intricate.

11.6.3 The recombining condition

Ho and Lee consider a recombining tree: the price $P_j(t, T)$ we are looking for depends only on j , not on the exact sequence of up and down movements leading to j upstate movements. To summarize, we are looking for two functions $u(T-t)$ and $d(T-t)$ such that (i) the no-arbitrage condition in Eq. (12.17) holds true and (ii) the tree is recombining. We now elaborate the arguments that lead to the recombining property of the tree.



The recombining property of the tree implies that the bond price at time $t + 2$ in the event of $j + 1$ jumps, i.e. $P_{j+1}(t + 2, T)$, can be generated by one of the two paths:

(i) The path $P_j(t, T) \rightarrow P_{j+1}(t + 1, T) \rightarrow P_{j+1}(t + 2, T) \rightarrow$ “up & down”

(ii) The path $P_j(t, T) \rightarrow P_j(t + 1, T) \rightarrow P_{j+1}(t + 2, T) \rightarrow$ “down & up”

We can use the two relations in Eqs. (11.28), to figure out the two paths leading to the bond price at time $t + 2$ in the event of $j + 1$ jumps, i.e. $P_{j+1}(t + 2, T)$. We have that along the first path,

$$\underbrace{\frac{P_{j+1}(t + 1, T)}{P_j(t, T)} = u(T - t) \frac{1}{P_j(t, t + 1)}}_{\text{up at } t}, \quad \underbrace{\frac{P_{j+1}(t + 2, T)}{P_{j+1}(t + 1, T)} = d(T - t - 1) \frac{1}{P_{j+1}(t + 1, t + 2)}}_{\text{down at } t+1},$$

and along the second path,

$$\underbrace{\frac{P_j(t + 1, T)}{P_j(t, T)} = d(T - t) \frac{1}{P_j(t, t + 1)}}_{\text{down at } t}, \quad \underbrace{\frac{P_{j+1}(t + 2, T)}{P_j(t + 1, T)} = u(T - t - 1) \frac{1}{P_j(t + 1, t + 2)}}_{\text{up at } t+1}.$$

To sum up:

$$\begin{aligned} P_{j+1}(t + 2, T) &= d(T - t - 1) \frac{1}{P_{j+1}(t + 1, t + 2)} \cdot \overbrace{u(T - t) \frac{1}{P_j(t, t + 1)} P_j(t, T)}^{\equiv P_{j+1}(t+1, T)} \quad (\text{up \& down}) \\ P_{j+1}(t + 2, T) &= u(T - t - 1) \frac{1}{P_j(t + 1, t + 2)} \cdot \underbrace{d(T - t) \frac{1}{P_j(t, t + 1)} P_j(t, T)}_{\equiv P_j(t+1, T)} \quad (\text{down \& up}) \end{aligned}$$

By equating the previous two equations, we obtain,

$$\frac{u(T - t)}{d(T - t)} = \frac{u(T - t - 1) P_{j+1}(t + 1, t + 2)}{d(T - t - 1) P_j(t + 1, t + 2)} \quad (11.31)$$

By evaluating Eq. (11.31) at $T = t + 2$,

$$\frac{u(2)}{d(2)} = \frac{u(1) P_{j+1}(t + 1, t + 2)}{d(1) P_j(t + 1, t + 2)} = \frac{P_{j+1}(t + 1, t + 2)}{P_j(t + 1, t + 2)} \equiv \delta^{-1},$$

where we assume that δ is constant. Clearly, $0 \leq \delta \leq 1$. Substituting back into Eq. (11.31),

$$\frac{u(T - t)}{d(T - t)} = \frac{u(T - t - 1)}{d(T - t - 1)} \delta^{-1}.$$

Therefore, given that $u(1) = d(1) = 1$,

$$\frac{u(T - t)}{d(T - t)} = \delta^{-(T-t-1)}. \quad (11.32)$$

Eq. (11.32) gives us the condition under which the tree is recombining. To rule out arbitrage opportunities, the martingale restriction in Eq. (12.17) must also hold true. Therefore, we have

to solve the following system of two equations (Eq. (11.32) and Eq. (12.17)) with two unknowns ($u(\cdot)$ and $d(\cdot)$),

$$\begin{cases} u(T-t) = \delta^{-(T-t-1)} d(T-t) \\ (1-q)u(T-t) + qd(T-t) = 1 \end{cases}$$

The solution to this system is,

$$u(T-t) = \frac{1}{(1-q) + q\delta^{T-t-1}}, \quad d(T-t) = \frac{\delta^{T-t-1}}{(1-q) + q\delta^{T-t-1}}. \quad (11.33)$$

So we have solved the problem. We know how to “populate” the tree. Suppose we know how to assign values to q and δ . Given q and δ , and an initial bond price $P(t, T)$, we can use Eqs. (11.28) to populate the tree, using the solution for $u(T-t)$ and $d(T-t)$ given in Eqs. (11.33). In this way, we can figure out the exact bond prices to insert in each node of the tree. Once we have computed the bond prices in each node, we can price interest rate derivatives, i.e. the asset the payoff of which depend on the particular value taken by the bond price on a given set of nodes. Below, we provide the closed-form solution for the bond price in this model.

What is the interpretation of δ ? We have defined δ to be, $\delta^{-1} \equiv \frac{P_{j+1}(t+1, t+2)}{P_j(t+1, t+2)}$, or,

$$\ln \delta^{-1} = \ln \left(\frac{P_{j+1}(t+1, t+2)}{P_j(t+1, t+2)} \right) = -[r_{j+1}(t+1) - r_j(t+1)]. \quad (11.34)$$

But we know that conditionally upon time t and (price) jumps equal to $j \leq t$, the short-term rate is binomially distributed, and can take on two values: (i) $r_{j+1}(t+1)$ with probability $1-q$ and $r_j(t+1)$ with probability q . Then, the conditional variance of the short-term rate is,

$$\text{var}_t[\tilde{r}(t+1)] = q(1-q)[r_{j+1}(t+1) - r_j(t+1)]^2,$$

where $\text{var}_t[\tilde{r}(t+1)]$ is the conditional variance at time t , of the short-term rate one-period ahead. Then, we may use Eq. (11.34), and the previous equation, to obtain,

$$\sqrt{\text{var}_t[\tilde{r}(t+1)]} = \sqrt{q(1-q)} \cdot \ln \delta^{-1}.$$

That is, δ is a parameter related to the *volatility of the short-term rate*, which in this basic model, is constant. In general, δ could be time-varying, which might lead to models without closed-form solutions.

Let $F_S^j(t)$ be the forward rate as of time t after the occurrence of j upward movements in the bond price, and let the continuously compounded forward rate $\hat{F}_S^j(t)$ be defined as,

$$\hat{F}_S^j(t) \equiv \ln(1 + F_S^j(t)), \quad j \leq t.$$

In Appendix 2, we show that,

$$P_j(t, T) = \frac{P(0, T)}{P(0, t)} e^{-\sum_{S=t}^{T-1} (\hat{F}_S^j(t) - \hat{F}_S(0))}. \quad (11.35)$$

That is, the key element that remains to be identified relates to the development of forward interest rates, $\hat{F}_S^j(t) - \hat{F}_S(0)$. In Appendix 2, we show that,

$$\hat{F}_S^j(t) = \hat{F}_S(0) + \ln \frac{u(S+1-t)}{u(S+1)} - (t-j) \ln \delta, \quad j \leq t, \quad (11.36)$$

such that by replacing Eq. (11.36) into Eq. (11.35), and using the solution for the perturbation function $u(\cdot)$ in Eqs. (11.33), leaves:

$$P_j(t, T) = \frac{P(0, T)}{P(0, t)} \delta^{(T-t)(t-j)} \prod_{S=t}^{T-1} \frac{(1-q) + q\delta^{S-t}}{(1-q) + q\delta^S}. \quad (11.37)$$

From the perspective of time 0, the price of the zero at t , in each state j is only a function of the initial yield curve, the volatility parameter δ , and the risk-neutral probability q .

11.6.4 Calibration of the model

We need to “estimate” the value of δ . We can proceed as follows. Consider Eq. (11.37), and let $T = t + 1$. We have,

$$P_j(t, t+1) = \frac{P(0, t+1)}{P(0, t)} \delta^{t-j} \frac{1}{(1-q) + q\delta^t}.$$

The continuously compounded short-term rate predicted by the model is,

$$r_j(t) \equiv -\ln P_j(t, t+1) = \hat{F}_t(0) + \ln((1-q) + q\delta^t) - (t-j) \ln \delta, \quad j \leq t, \quad (11.38)$$

where $\hat{F}_t(0) \equiv \ln P(0, t) - \ln P(0, t+1)$. We also have,

$$r_j(1) - r(0) = \hat{F}_1(0) - \hat{F}_0(0) + \ln((1-q) + q\delta) + \ln \delta^{-1} \cdot (1-j).$$

Hence, the parameter δ can be chosen so that the volatility of the short-term rate predicted by the model matches exactly the volatility of the short-term rate that we see in the data. Concretely, we can take $\hat{\delta} = \exp(-\text{Std}(\Delta r) / \sqrt{q(1-q)})$, where $\text{Std}(\Delta r)$ is the standard deviation of the short-term rate in the data.

Note, then, the interesting feature of the model. The Ho and Lee model doesn’t take any a priori stance on the dynamics of the short-term rate. Rather, it imposes: (i) the martingale restriction on bond prices, an economic restriction, Eq. (12.17); and (ii) the simplifying assumption the tree is recombining, a technical condition, Eq. (11.28). These two conditions suffice to tell what to expect from the dynamics of the short-term rate. While deliberately simple, the Ho and Lee model is quite powerful. The modern approach to interest rate modeling simply aims to make the Ho and Lee methodology more accurate for practical purposes.

11.6.5 An example

Assume that three zero coupon bonds are available for trading, with current market prices: (i) $P_{\S}(0, 1) = 0.9851$ (the price of a 6M zero), (ii) $P_{\S}(0, 2) = 0.9685$ (the price of a 1Y zero), and (iii) $P_{\S}(0, 3) = 0.9445$ (the price of the 1.5Y zero). We know that the price of one-period zero at time t , in the event of j upward *price*-jumps from the current date to t , is:

$$P_j(t, t+1) = \frac{P_{\S}(0, t+1)}{P_{\S}(0, t)} \delta^{t-j} \frac{1}{(1-q) + q\delta^t}, \quad j \leq t, \quad (11.39)$$

where $P_{\S}(0, t)$ is the current market price of a zero expiring at time t , with t equal to six months, one year and eighteen months, in this example. We assume that $q = \frac{1}{2}$ and $\delta = 0.9802$.

11.6.5.1 The dynamics of the short-term rate

We want to determine the evolution of the short-term rate on a recombining tree for as many periods as we can, given the market price of the zeros we observe. We use Eq. (11.39) to find the one-period zeros in each node.

- $t = 0$. We have, trivially, $P(0, 1) = P_{\S}(0, 1) = 0.9851$.

- $t = 1$. We have three cases:

$$- j = 0: P_0(1, 2) = 2 \frac{P_{\S}(0, 2)}{P_{\S}(0, 1)} \delta \frac{1}{1+\delta} = 0.9733$$

$$- j = 1: P_1(1, 2) = 2 \frac{P_{\S}(0, 2)}{P_{\S}(0, 1)} \frac{1}{1+\delta} = 0.9930$$

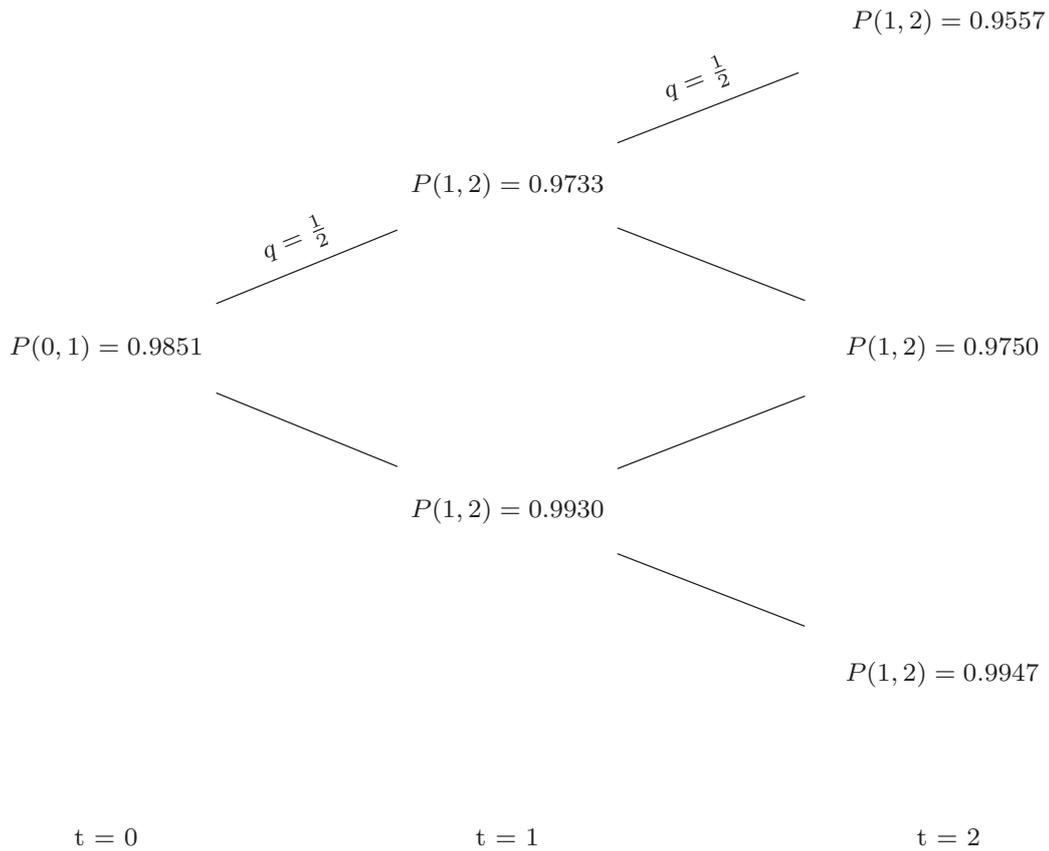
- $t = 2$. We have three cases:

$$- j = 0: P_0(2, 3) = 2 \frac{P_{\S}(0, 3)}{P_{\S}(0, 2)} \delta^2 \frac{1}{1+\delta^2} = 0.9557$$

$$- j = 1: P_1(2, 3) = 2 \frac{P_{\S}(0, 3)}{P_{\S}(0, 2)} \delta \frac{1}{1+\delta^2} = 0.9750$$

$$- j = 2: P_2(2, 3) = 2 \frac{P_{\S}(0, 3)}{P_{\S}(0, 2)} \frac{1}{1+\delta^2} = 0.9947$$

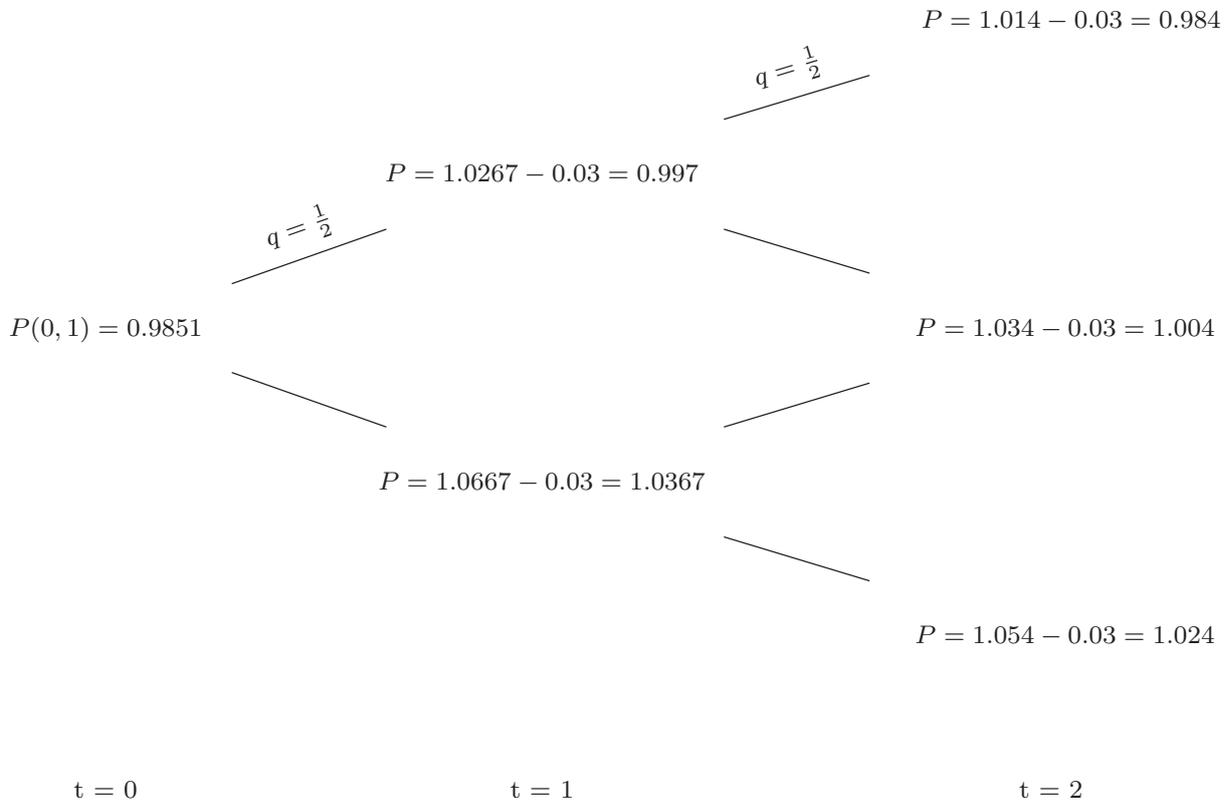
So we face the tree below.



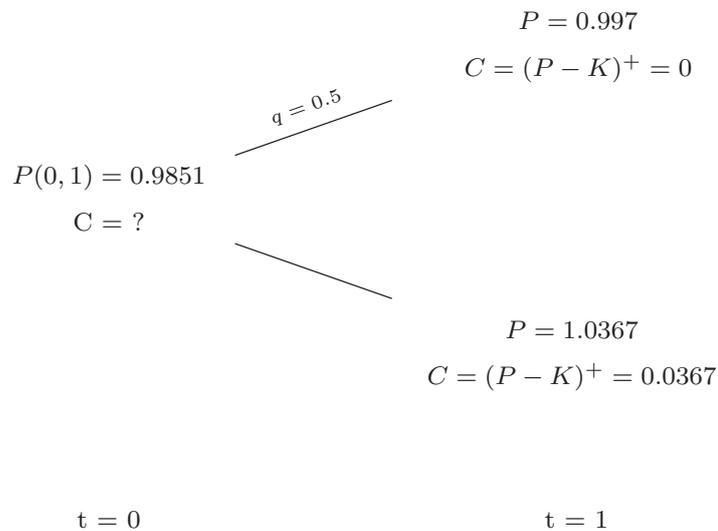
11.6.5.2 Pricing a coupon bearing bond

Suppose, now, that we want to find the price of some additional bond, e.g., a 1.5Y bond which pays (semiannually) coupons at 3% of the principal of £1. First, we need to find the value of this bond in each node of the tree. Note, at each node, the price equals (i) the discounted expectation of its future value (including coupons), and (ii) the current coupons, as illustrated in the tree below. That is, the convention, here, is that the bond purchased at time t doesn't give the owner the right to receive any coupon at time t , only from time $t + 1$ onwards.

the *ex-coupons* bond price. (This is because if we purchase the bond today, we are not entitled to receive any coupon, today. The flow of coupons we are entitled to receive starts from the next period.) We easily obtain the tree below. We must just subtract the coupon, 0.03, from each cum-coupons price in each node of the tree. Then, we obtain:



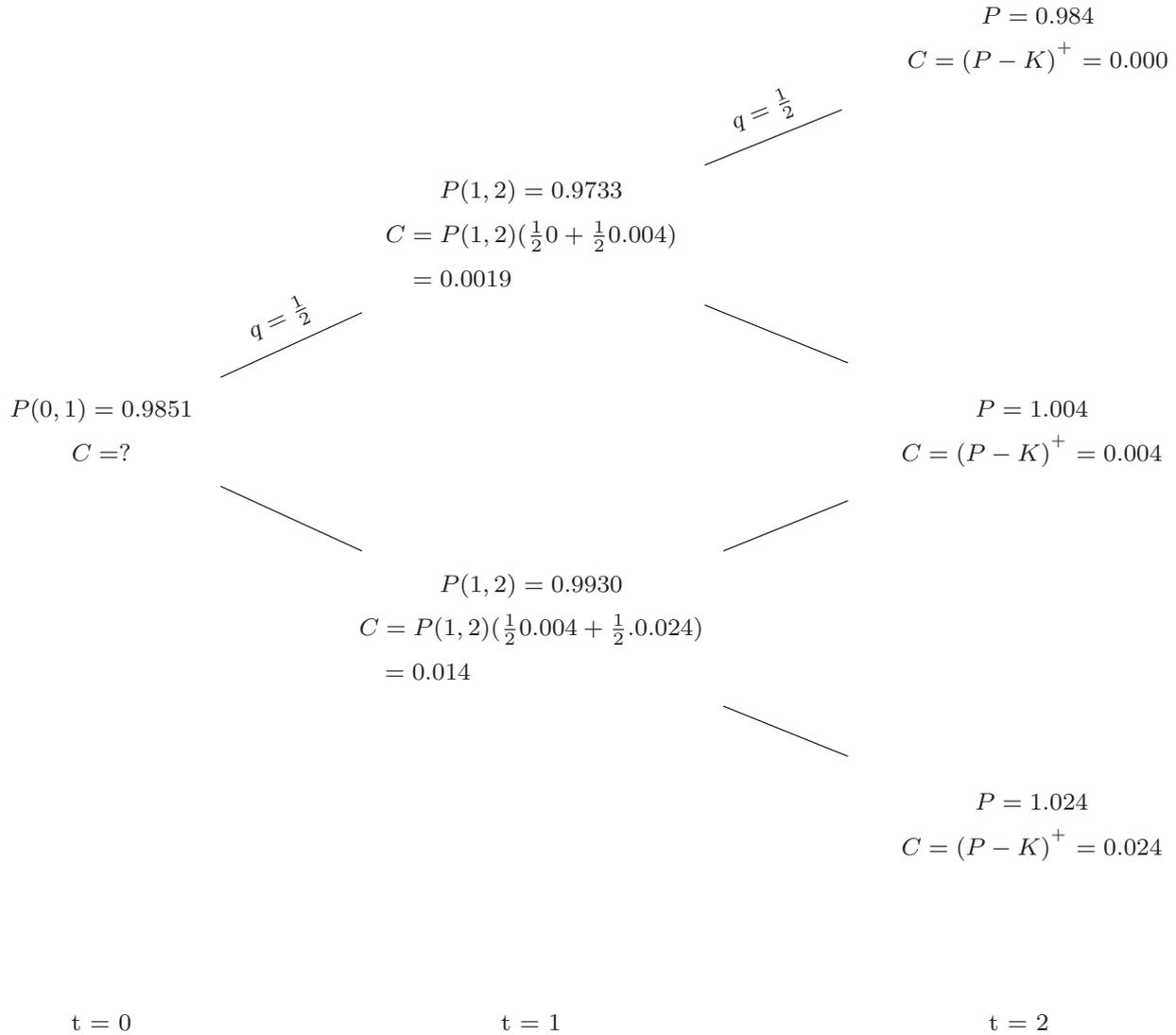
We are ready to price the two options. As for the call option on the 1.5Y bond, with 6 months maturity, and strike price $K = \pounds 1$, we have the following tree:



Therefore,

$$C = 0.9851 \left(\frac{1}{2} \cdot 0 + \frac{1}{2} \cdot 0.0367 \right) = 1.808 \times 10^{-2}.$$

The call option on the 1.5Y bond with 1 year maturity, and strike price $K = \pounds 1$, is dealt with similarly. We have the following tree:



Therefore, the price of the option is,

$$C = P(0, 1) \left(\frac{1}{2} \cdot 0.0019 + \frac{1}{2} \cdot 0.014 \right) = 0.9851 \left(\frac{1}{2} \cdot 0.0019 + \frac{1}{2} \cdot 0.014 \right) = 7.831 \times 10^{-3}.$$

11.6.6 Continuous-time approximations, with an application to barbell trading

11.6.6.1 The approximation

Consider Eq. (11.38), and define $r = -\Delta t^{-1} \ln P$, and $r_j^\Delta(t) \equiv r_j(t) \Delta t$, such that:

$$r_j^\Delta(t) = \zeta_t - (t(1 - q) - j) \ln \delta, \quad \zeta_t \equiv \hat{F}_t(0) + \ln \left(\delta^{-qt} (1 - q) + \delta^{(1-q)t} q \right).$$

We have,

$$\mathbb{E}_0 [r_j^\Delta(t)] = \zeta_t, \quad \text{and} \quad \sigma^2 t \equiv \mathbb{V}_0 [r_j^\Delta(t)] = (\ln \delta^{-1})^2 q(1-q)t,$$

such that we may define $\ell \equiv \sigma/\sqrt{q(1-q)}$, and, then, $\delta = e^{-\ell}$. Replacing this into the definition of ζ_t , yields, after expanding terms up to the second order,

$$\begin{aligned} \zeta_t &= \hat{F}_t(0) + \ln(e^{\ell q t}(1-q) + e^{-\ell(1-q)t}q) \\ &\approx \hat{F}_t(0) + \ln\left(1 + \frac{1}{2}\ell^2 q(1-q)t^2\right) \\ &\approx \hat{F}_t(0) + \frac{1}{2}\ell^2 q(1-q)t^2 \\ &= \hat{F}_t(0) + \frac{1}{2}\sigma^2 t^2. \end{aligned}$$

Note, this expansion is accurate when ℓt is small, which empirically is indeed, as we have that, typically, $\ell t \approx 10^{-2}t$, which then works for t up to at least 50 years! However, these calculations might also be considered as the starting point for the initial drift of the short-term rate from zero to time t . So, we have, approximately, that,

$$\mathbb{E}_0 [r_j^\Delta(t)] = \hat{F}_t(0) + \frac{1}{2}\sigma^2 t^2, \quad \text{and} \quad \mathbb{V}_0 [r_j^\Delta(t)] = \sigma^2 t. \quad (11.40)$$

In the next chapter (Section 12.4.2), we shall show, consistently with the previous calculations, that in continuous time, the Ho and Lee model predicts the short-term rate to be the solution to:

$$dr(t) = \left[\frac{\partial}{\partial t} f_{\S}(0, t) + \sigma^2 t \right] dt + \sigma d\tilde{W}(t),$$

where $f_{\S}(0, t)$ is the instantaneous continuously compounded forward rate, and $\tilde{W}(t)$ is a Brownian motion defined under the risk-neutral probability. In fact, in the next chapter, it will be shown that the instantaneous forward rate predicted by the Ho & Lee model is:

$$f(s', t) - f(s, t) = \sigma^2 \int_s^{s'} (t - \tau) d\tau + \sigma (\tilde{W}(t) - \tilde{W}(s)), \quad (11.41)$$

such that, for $r(s') = f(s', s')$,

$$r(s') - f(s, s') = \sigma^2 \int_s^{s'} (s' - \tau) d\tau + \sigma (\tilde{W}(s') - \tilde{W}(s)), \quad (11.42)$$

the continuous time counterpart to the two conditions in Eqs. (11.40). By combining Eqs. (11.41)-(11.42), we obtain, after simple computations, that:

$$f(s', t) - f(s, t) = \sigma^2 (s' - s)(t - s') + r(s') - f(s, s'). \quad (11.43)$$

As the next chapter shows (see Section 12.5.1), we have that for any model, including Ho & Lee,⁵ the following representation holds true:

$$P(t, T) = \frac{P(0, T)}{P(0, t)} \cdot e^{-\int_t^T [f(t, u) - f(0, u)] du}. \quad (11.44)$$

⁵For example, Eq. (11.35) in the Appendix provides the discrete time counterpart to Eq. (11.44).

Using the expression for $f(s', t) - f(s, t)$ in Eq. (11.43), and integrating,

$$\int_t^T [f(t, u) - f(0, u)] du = \frac{1}{2}\sigma^2 (T-t)^2 t + (T-t)(r(t) - f(0, t)).$$

Replacing this expression into Eq. (11.44) leaves:

$$P(t, T) = \frac{P(0, T)}{P(0, t)} \cdot \exp\left(-\frac{1}{2}\sigma^2 (T-t)^2 t - (r(t) - f(0, t))(T-t)\right). \quad (11.45)$$

It is a neat expression, which we may use, for a variety of purposes, such as option pricing. The next section develops an example relating to barbell trading.

11.6.6.2 Application to barbell trading

We revisit the barbell trading strategy of Section 11.4.3.4, where we argued that this strategy leads to positive profits due to “convexity,” as summarized by Figure 11.4. The key point in this argument is that it abstracts from passage of time, and may, in fact, lead us to misinterpret what is a merely static analysis. We may use the Ho and Lee model to analyze the profit and losses of a barbell trade, in a dynamic context free from arbitrage. We consider two situations: one, where the initial yield curve is flat, and a second, where the initial yield curve is upward sloping.

As for the flat yield curve, we use the continuously compounded rate corresponding to the flat 5% of Section 11.4.3.5, delivering $r = \ln 1.05 = 0.04879$. The number of assets to include into the portfolio, θ_1 and θ_2 , are as in Eq. (11.17), i.e. $\theta_1 = 0.45706$ and $\theta_3 = 0.56724$. Instantaneous forward rates are $f(0, T) = \lim_{S \downarrow T} F(0, T, S) = -\partial \ln P(0, T) / \partial T = r$. Using Eq. (11.45), with volatility parameter $\sigma = 0.03$, we calculate the value of the strategy a few months later, as follows:

$$\text{Barb}(t) = 100 * (\theta_1 P(t, 1) + \theta_3 P(t, 10) - P(t, 5)). \quad (11.46)$$

Figure 11.12 depicts the value of the barbell, $\text{Barb}(t)$, for investment horizons equal to 1 month, 3 months, 6 months and one year.

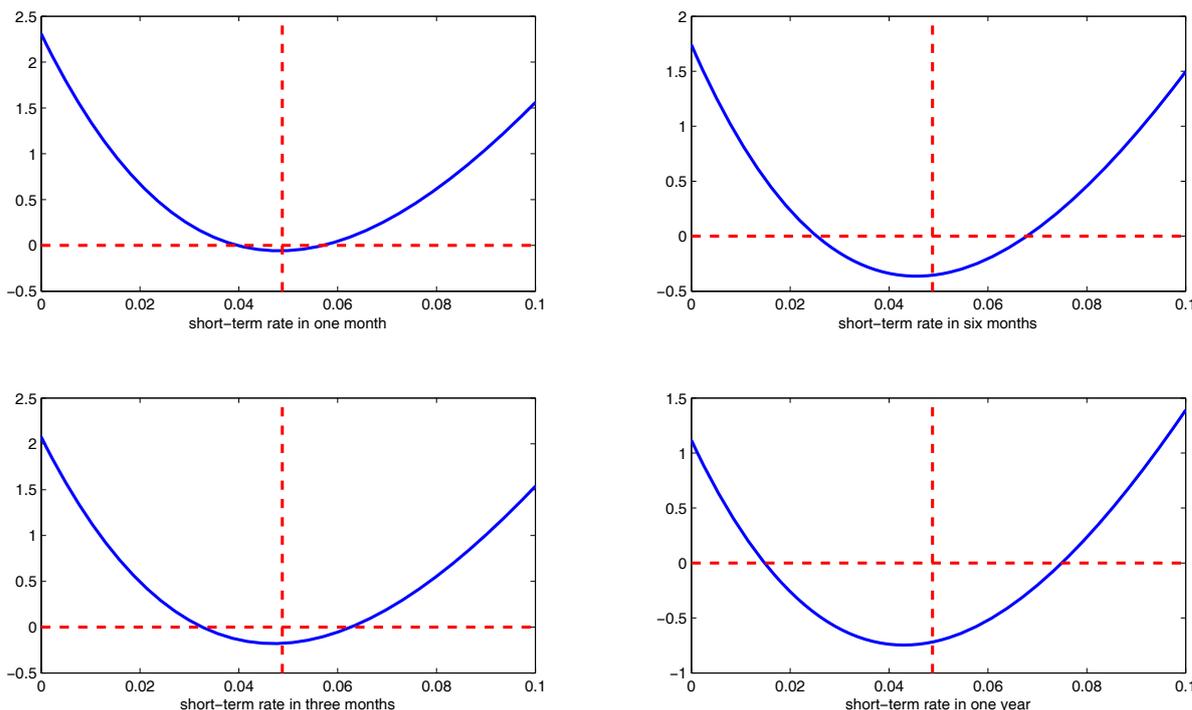


FIGURE 11.12. Profit and losses arising from barbell trading, $\text{Barb}(t)$ in Eq. (11.46), under the assumption the yield curve is driven by the Ho and Lee model, Eq. (11.45). The initial yield curve is assumed to be flat at $r = 4.8790\%$. Investment horizons are $t = 1/12$ (NW quadrant), $t = 3/12$ (SW quadrant), $t = 6/12$ (NE quadrant) and, $t = 1$ (SE quadrant). The vertical dashed lines pass through $r = 4.8790\%$, and the horizontal dashed lines pass through zero.

This trade is quite risky. For long investment horizons, it pays off when the short-term rate fluctuates significantly away from the initial value, $r = 4.8790\%$. The amount of fluctuations in the short-term rate diminishes as we shrink the investment horizon. Nevertheless, this amount appears to be considerable: for example, at one-month horizon, we should require the short-term rate to move from $r = 4.8790\%$ to either values larger than $r = 6\%$ or lower than $r = 4\%$, in order to claim for positive profits. Actually, these results suggest that a short position in the barbell trade (i.e., sell the barbell portfolio and go long the 5Y bond) should be an interesting strategy to implement in periods where we do not expect high volatility of interest rates. For example, for investment horizons of 6 months, the profits from a short position in the barbell trade are positive within a quite significant range of variation of the short-term rate, $[2.5\%, 6.8\%]$.

Finally, we consider a scenario where the initial yield curve is upward sloping, and generate prices as $P(0, t) = e^{-tY(t)}$, where $Y(t) \equiv 0.01(1 + \ln t)$. We still compute the portfolio according to Eq. (11.17), i.e. we rely on the self-financing condition in Eq. (11.15) and both (i) the locally riskless condition in Eq. (11.16), $dB_2(\hat{y}_2) = \theta_1 dB_1(\hat{y}_1) + \theta_3 dB_3(\hat{y}_3)$, and (ii) the (generically

incorrect) assumption of parallel shifts in the yield curve, $\frac{\partial B_2(\hat{y}_2)}{\partial \hat{y}_2} = \frac{\partial B_1(\hat{y}_1)}{\partial \hat{y}_1} \theta_1 + \frac{\partial B_3(\hat{y}_3)}{\partial \hat{y}_3} \theta_3$. Figure 11.13 depicts the profit and losses arising from the trade.

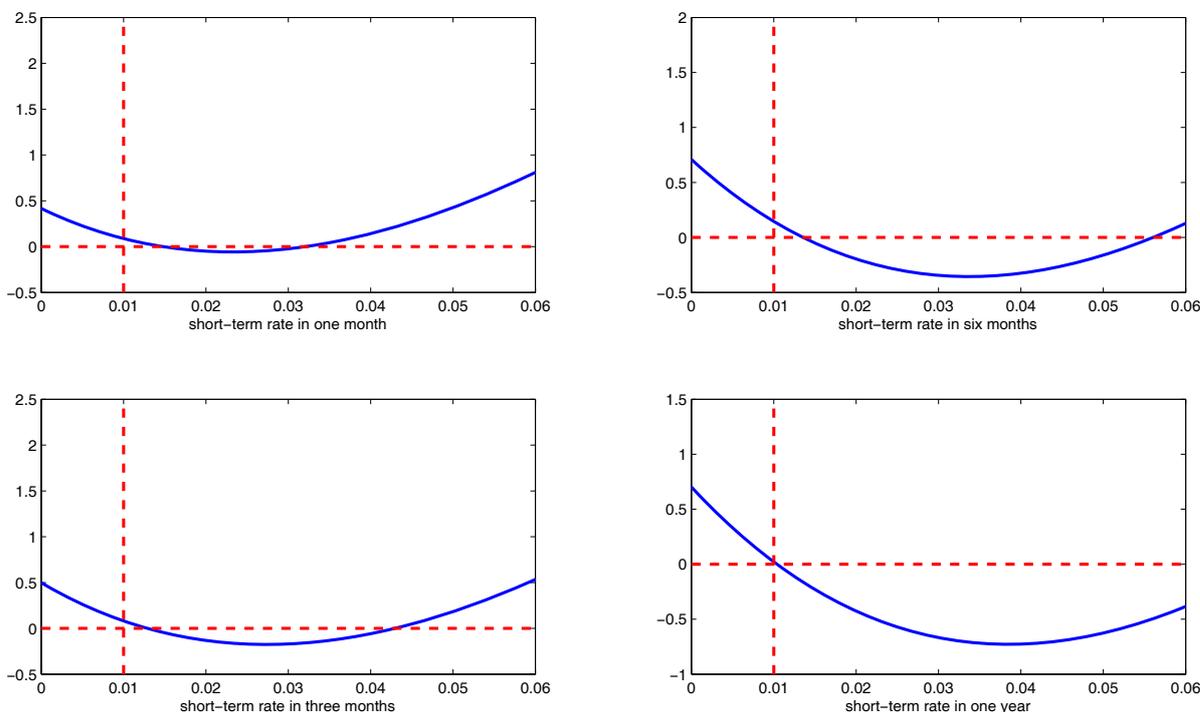


FIGURE 11.13. Profit and losses arising from barbell trading, $\text{Barb}(t)$ in Eq. (11.46), under the assumption the yield curve is driven by the Ho and Lee model, Eq. (11.45). The initial yield curve is assumed to be upward sloping, generated by the equation, $Y(t) = 0.01(1 + \ln t)$, with prices given by $P(0, t) = e^{-tY(t)}$. Investment horizons are $t = 1/12$ (NW quadrant), $t = 3/12$ (SW quadrant), $t = 6/12$ (NE quadrant) and, $t = 1$ (SE quadrant). The vertical dashed lines pass through the current short-term rate, $r = 1.0\%$, and the horizontal dashed lines pass through zero.

Similarly as for the profit and losses summarized in Figure 11.12, the trade leads to profits only when the short-term rate increases, and significantly, from the initial value $r = 1\%$. In particular, when r moves around 1%, profits increase as r lowers, and decrease, as r goes up. This effect relates to that arising within the static exercise described in Table 11.4: long term bonds benefit from a decrease in r more than short-term, and lose their value more than short-term bonds as r increases. However, as the interest rate increases significantly, the barbell generates profits because the convexity of 10 year bonds dominates overall.

11.7 Beyond Ho and Lee: Calibration

The approach in the previous sections imposes no-arbitrage conditions to bond prices, which have implications on forward rates, thereby ultimately determining an *implied* stochastic process of the short-term rate. In this section, we determine no-arbitrage dynamics for the short-term rate in the first place. The advantage of the approach in this section is that it is general. The

Ho and Lee (1986) model relies on a number of assumption we can easily drop, through the “calibration” perspective developed in this section. We illustrate how this calibration works by developing three points. First, we review how to use Arrow-Debreu securities in the very applied context of fixed income security evaluation. We show that Arrow-Debreu securities allow us to turn the martingale restriction of the previous sections to a set of conditions that are quite easy to use, even when we would face complex models without a closed-form expression. Second, we use these same Arrow-Debreu securities and implement an algorithm aimed to “populate” the short-term rate tree, while ensuring that the initial term-structure is perfectly fitted. Finally, we apply these ideas and illustrate how to solve two models, in practice: (i) the Ho and Lee model, and (ii) a model developed by Black, Derman and Toy (1990).

11.7.1 Arrow-Debreu securities

We know, from Chapter 2, that an Arrow-Debreu security is an asset that pays off £1 in some prespecified state of the nature, and zero otherwise. Consider, for example, the diagram in Figure 11.14.

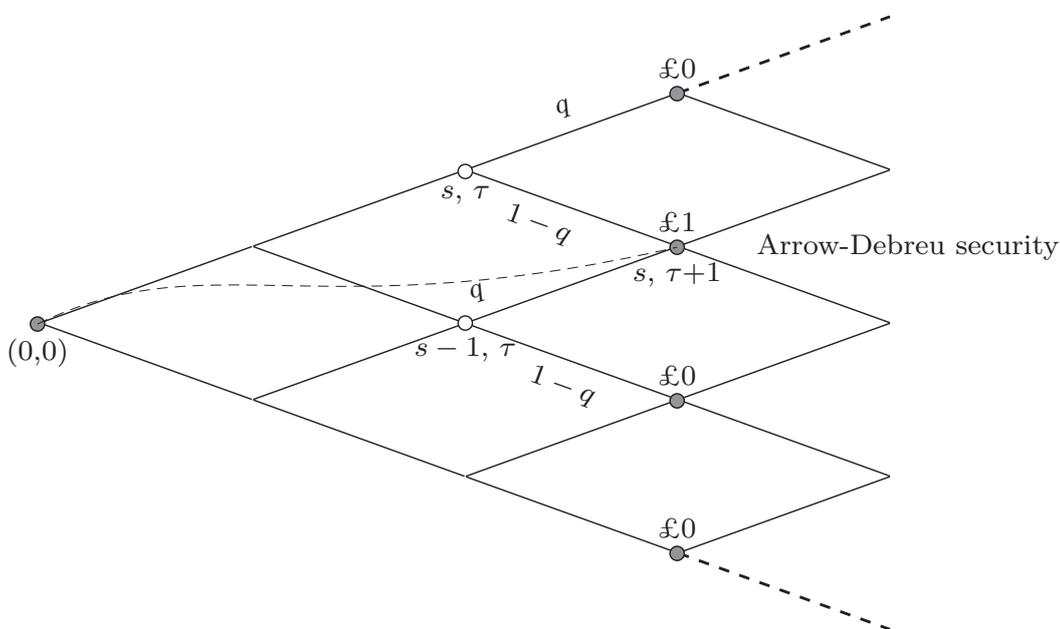


FIGURE 11.14. In the binomial tree of this section, an Arrow-Debreu security for state s at time $\tau + 1$ is a security that pays £1 at time $\tau + 1$ in state s , and zero otherwise. This section aims to show how to recover Arrow-Debreu prices from the price of fixed income securities.

In this diagram, q is the risk-neutral probability of an upward movement of the short-term rate. A generic pair (s, τ) at each node tracks the number of upward movements of the short-term rate, s , and calendar time, τ , where $s \leq \tau$, as there can only be one possible short-term rate movement in each period. From now on, let us focus on the Arrow-Debreu security for the state s at time $\tau + 1$.

Let $p_s(\tau)$ denote the *current* price of an Arrow-Debreu security that pays off £1 in state s at time τ , and zero otherwise. Then, the current market price of a zero that matures at time T is necessarily,

$$P_{\S}(0, T) = \sum_{s=0}^T p_s(T).$$

More generally, consider a derivative that pays off $D_s(\tau)$ in node (s, τ) , meaning a dividend equal to $D_1(\tau)$ in state $s = 1$, equal to $D_2(\tau)$ in state $s = 2, \dots$, and equal to $D_\tau(\tau)$ in state $s = \tau$. The price of this asset, denoted as $C_{\S}(0, T)$, is given by,

$$C_{\S}(0, T) = \sum_{\tau=1}^T \sum_{s=0}^{\tau} p_s(\tau) D_s(\tau). \quad (11.47)$$

Our objective, now, is to “recover” the price of the Arrow-Debreu securities $p_s(\tau)$ for all s and τ , where $\tau \in \{1, \dots, T\}$, from the observation of the initial term-structure of interest rates.

Consider the Arrow-Debreu security that promises to pay £1 in node $(s, \tau + 1)$ (see Figure 11.14). Let its value at time τ in state j ($j \leq \tau$) be denoted as $\pi_{j,\tau}[s, \tau + 1]$. What is this value at time τ in all states? A key observation is that in this tree, the node $(s, \tau + 1)$ (the filled circle) can only be “accessed to” through the nodes (s, τ) and the nodes $(s - 1, \tau)$ occurring at time τ (the two empty circles in Figure 11.14). At time τ then, the value $\pi_{j,\tau}[s, \tau + 1]$ is zero in all the nodes (j, τ) except the empty circles (s, τ) and $(s - 1, \tau)$. Indeed, if we do not happen to be at one of those nodes denoted with empty circles, we cannot reach the node $(s, \tau + 1)$ (the filled circle), where the Arrow-Debreu security pays off.

So, we are left with finding the values $\pi_{j,\tau}[s, \tau + 1]$ in the nodes corresponding to the empty circles (s, τ) and $(s - 1, \tau)$, i.e. $\pi_{s,\tau}[s, \tau + 1]$ and $\pi_{s-1,\tau}[s, \tau + 1]$. Let $r_s(\tau)$ be the continuously compounded short-term rate in node (s, τ) . Consider the upper node (s, τ) . We have,

$$\pi_{s,\tau}[s, \tau + 1] = e^{-r_s(\tau)} [0 \cdot q + 1 \cdot (1 - q)] = e^{-r_s(\tau)} (1 - q).$$

Similarly, in the lower node, $(s - 1, \tau)$,

$$\pi_{s-1,\tau}[s, \tau + 1] = e^{-r_{s-1}(\tau)} [1 \cdot q + 0 \cdot (1 - q)] = e^{-r_{s-1}(\tau)} q.$$

We can think of our Arrow-Debreu security for $(s, \tau + 1)$ as a derivative that at time τ , delivers the following “payoffs”

$$\begin{cases} \pi_{s,\tau}[s, \tau + 1] = e^{-r_s(\tau)} (1 - q) \\ \pi_{s-1,\tau}[s, \tau + 1] = e^{-r_{s-1}(\tau)} q \\ \pi_{j,\tau}[s, \tau + 1] = 0, \quad \text{for all } j < s \end{cases} \quad (11.48)$$

These “payoffs” are simply the market value of the Arrow-Debreu security for $(s, \tau + 1)$, in the various states occurring at time τ , i.e. the money the holder can make by selling the asset at time τ , in the various states. Therefore, we can apply Eq. (11.47) to obtain,

$$\begin{aligned} p_s(\tau + 1) &= \sum_{j=0}^{\tau} p_j(\tau) \pi_{j,\tau}[s, \tau + 1] \\ &= p_s(\tau) \pi_{s,\tau}[s, \tau + 1] + p_{s-1}(\tau) \pi_{s-1,\tau}[s, \tau + 1]. \end{aligned}$$

By replacing the Arrow-Debreu prices in (11.48) into the previous equation, we obtain the so-called *forward equation* for the Arrow-Debreu prices,

$$p_s(\tau + 1) = p_s(\tau) e^{-r_s(\tau)} (1 - q) + p_{s-1}(\tau) e^{-r_{s-1}(\tau)} q \quad (11.49)$$

11.7.2 The algorithm in two examples

The algorithm aims to populate the interest rate tree by making a repeated use of the forward equation (11.49) and the zero pricing equation,

$$P_{\S}(0, \tau + 1) = \sum_{s=0}^{\tau} p_s(\tau) e^{-r_s(\tau)}.$$

The input to the algorithm are a number of zeros equal to the largest maturity date the tree extends to. We illustrate the algorithm by developing two examples.

11.7.2.1 Two model examples

We begin with Ho and Lee. We assume continuous compounding, for analytical reasons clarified below. By Eq. (11.38), the short-term rate predicted by the Ho and Lee model is:

$$r_j(\tau) = \hat{F}_{\tau}(0) + \ln((1 - q) + q\delta^{\tau}) - (\tau - j) \ln \delta. \quad (11.50)$$

where $\hat{F}_{\tau}(0)$ is the continuously compounded forward rate, at time zero, for maturity $[\tau, \tau + 1]$, and j is the number of *upward* movements of the entire set of *bond prices*. Naturally, $s \equiv (t - j)$ is the number of downward movements of the bond prices or, equivalently, the number of *upward* movements of the *short-term rate*. Hence, we may equivalently index the short-term rate by s , instead than by j , and rewrite Eq. (11.50) as follows:

$$r_s(\tau) = \underbrace{\hat{F}_{\tau}(0) + \ln((1 - q) + q\delta^{\tau})}_{\equiv r_0(\tau)} + \ln \delta^{-1} \cdot s, \quad (11.51)$$

such that $r_0(\tau)$ is the short-term rate at time τ , in the event of zero upward movements in the short-term rate, and δ is a volatility parameter, i.e. such that $\ln \delta^{-1} = \frac{\text{Std}(\Delta r)}{\sqrt{q(1-q)}}$, with $\text{Std}(\Delta r)$ denoting the standard deviation of the short-term rate in the data. Incidentally, note that the short-term rate movements do depend on the value of the risk-neutral probability q used in the calibration.

At time zero, the price of a zero maturing at time $\tau + 1$ is:

$$P_{\S}(0, \tau + 1) = \sum_{s=0}^{\tau} p_s(\tau) e^{-r_s(\tau)} = e^{-r_0(\tau)} \sum_{s=0}^{\tau} \delta^s p_s(\tau),$$

where the second equality follows by the assumption that the short-term rate is solution to Eq. (11.51).

By rearranging terms in the previous equation, we obtain a closed-form expression for the future short-term rate at time τ , in the event of zero upward movements,

$$r_0(\tau) = \ln \left(\frac{\sum_{s=0}^{\tau} \delta^s p_s(\tau)}{P_{\S}(0, \tau + 1)} \right). \quad (11.52)$$

We use Eq. (11.52) and the forward equation (11.49) to populate the interest rate tree, under the assumption that $q = \frac{1}{2}$. Precisely, the algorithm proceeds as follows:

- (i) Given the boundary condition for the Arrow-Debreu price, $p_0(0) = 1$, compute the initial value of the short-term rate, $r_0(0)$, using Eq. (11.52), as $r_0(0) = \ln(1/P_{\S}(0, 1))$.

- (ii) Suppose we know the *future* value of the short-term rate at time $\tau - 1$, in the event of no upward movements, i.e. $r_0(\tau - 1)$. Then, given the value of $r_0(\tau - 1)$, and the price of the Arrow-Debreu securities $p_s(\tau - 1)$ for $s \leq \tau - 1$, compute $p_s(\tau)$ for $s \leq \tau$, through the forward equation (11.49),

$$p_s(\tau) = p_s(\tau - 1) \delta^s e^{-r_0(\tau-1)} (1 - q) + p_{s-1}(\tau - 1) \delta^{s-1} e^{-r_0(\tau-1)} q, \quad q = \frac{1}{2},$$

where the last equation follows by plugging Eq. (11.51) into Eq. (11.49).

- (iii) Given the Arrow-Debreu prices $p_s(\tau)$ for $s \leq \tau$, use Eq. (11.52) to compute the *future* value of the short-term rate at time τ , in the event of no upward movements, i.e. $r_0(\tau)$.
- (iv) If $\tau = T$, stop. Otherwise, go to (ii).

As a second example, consider the Black, Derman and Toy (1990) model. In this model, the short-term rate is solution to,

$$r_s(\tau) = \delta^s r_0(\tau), \quad (11.53)$$

where δ is, once again, a volatility parameter.⁶ For computational convenience, this model assumes that the short-term rate in Eq. (11.53) is discretely compounded. Accordingly, we rewrite the forward equation (11.49) in terms of discretely compounded rates,

$$p_s(\tau + 1) = p_s(\tau) \frac{1}{1 + r_s(\tau)} (1 - q) + p_{s-1}(\tau) \frac{1}{1 + r_{s-1}(\tau)} q. \quad (11.54)$$

The algorithm proceeds as follows:

- (i) Compute the initial value of the short-term rate, $r_0(0)$, as the solution to,

$$P_{\S}(0, 1) = \frac{1}{1 + r_0(0)}.$$

- (ii) Suppose we know the *future* value of the short-term rate at time $\tau - 1$, in the event of no upward movements, i.e. $r_0(\tau - 1)$. Then, given the value of $r_0(\tau - 1)$, and the price of the Arrow-Debreu securities $p_s(\tau - 1)$ for $s \leq \tau - 1$, compute $p_s(\tau)$ for $s \leq \tau$, through the forward equation (11.54),

$$p_s(\tau) = p_s(\tau - 1) \frac{1}{1 + \delta^s r_0(\tau - 1)} (1 - q) + p_{s-1}(\tau - 1) \frac{1}{1 + \delta^{s-1} r_0(\tau - 1)} q, \quad q = \frac{1}{2},$$

where the last equation follows by plugging Eq. (11.53) into Eq. (11.54).

- (iii) Given the boundary condition $p_0(0) = 1$, and the Arrow-Debreu prices, $p_s(\tau)$ for $s \leq \tau$, use the pricing equation for the zero,

$$P_{\S}(0, \tau + 1) = \sum_{s=0}^{\tau} p_s(\tau) \frac{1}{1 + \delta^s r_0(\tau)},$$

to solve, numerically, for the *future* value of the short-term rate at time τ , in the event of no upward movements, i.e. $r_0(\tau)$. Note, we did not need this additional step for the solution of the Ho and Lee model, as the short-term rate $r_0(\tau)$ is known in closed form in the Ho and Lee model (see Eq. (11.52)).

⁶In its most general form, this model assumes that $r_s(\tau) = \delta_{\tau}^s r_0(\tau)$, where δ_{τ} is a volatility parameter that varies deterministically over time. This more general formulation leads to more flexibility, which is useful to fit the term structure of volatility.

(iv) If $\tau = T$, stop. Otherwise, go to (ii).

11.7.2.2 A numerical example

Consider, again the Ho and Lee model example in Section 11.5.5, where three zeros were traded: (i) one zero maturing in 6 months, (ii) one zero maturing in 1 year, and (iii) one zero maturing in 1.5 years, with market prices $P_{\S}(0, 1) = 0.9851$, $P_{\S}(0, 2) = 0.9685$, $P_{\S}(0, 3) = 0.9445$. By Eq. (11.51), the Ho and Lee model assumes that,

$$r_s(\tau) = r_0(\tau) + (\ln \delta^{-1}) \cdot s. \quad (11.55)$$

We use Eq. (11.55) and find the values of the short-term rate $r_s(\tau)$ in each node, under the assumption that $q = \frac{1}{2}$, and that the standard deviation of the short-term rate is 0.014, annualized. To find δ , we may use the relation, $\ln \delta^{-1} = \frac{\text{Std}(\Delta r)}{\sqrt{q(1-q)}}$, where $q = \frac{1}{2}$ and $\text{Std}(\Delta r)$ is the standard deviation of the short-term rate, which equals $\text{Std}(\Delta r) = 0.014$, annualized. Therefore $\ln \delta^{-1} = \frac{0.014}{\sqrt{2}} / \frac{1}{2} = 0.02$ or $\delta = 0.9802$.

For the Ho & Lee model, we know the closed-form expression for $r_0(\tau)$,

$$r_0(\tau) = \ln \left(\frac{\sum_{s=0}^{\tau} \delta^s p_s(\tau)}{P_{\S}(0, \tau + 1)} \right), \quad (11.56)$$

where $p_s(\tau)$ denotes the price of an Arrow-Debreu security which pays of £1 in state s at time τ , and zero otherwise. Given the term-structure of prices $P_{\S}(0, \tau + 1)$, $\tau = 0, 1, 2$, we “populate” the tree using Eq. (11.56) and the forward equation for the Arrow-Debreu prices developed in the lecture notes,

$$p_s(\tau) = \frac{1}{2} e^{-r_0(\tau-1)} [\delta^s p_s(\tau-1) + \delta^{s-1} p_{s-1}(\tau-1)], \quad (11.57)$$

with the appropriate boundary conditions.

So we have to compute interest rates and Arrow-Debreu prices for $\tau = 0, 1, 2$.

- $\tau = 0$. Eq. (11.56) is trivial. It leads to,

$$r_0(0) = \ln \left(\frac{1}{P_{\S}(0, 1)} \right) = 0.015.$$

The forward equation for the Arrow-Debreu prices, Eq. (11.57), is also trivial,

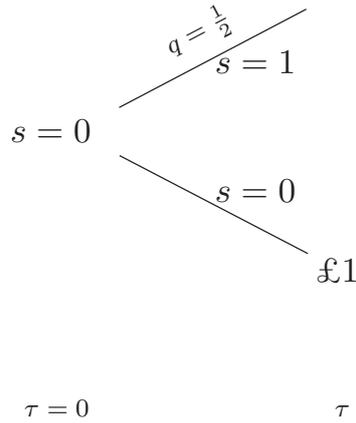
$$p_0(0) = 1.$$

- $\tau = 1$. Let us use Eq. (11.57), the forward equation for the Arrow-Debreu prices, to find $p_0(1)$ and $p_1(1)$. We have two cases:

– $s = 0$. We have:

$$p_0(1) = \frac{1}{2} e^{-r_0(0)} [p_0(0) + 0] = \frac{1}{2} e^{-r_0(0)} = 0.4925.$$

The previous relation holds because $p_0(1)$ is the current price of the Arrow-Debreu security which pays off £1 in state 0 at time 1, as illustrated by the tree in the Figure 1 below,



– $s = 1$. By a similar reasoning,

$$p_1(1) = \frac{1}{2}e^{-r_0(0)} [0 + p_0(0)] = \frac{1}{2}e^{-r_0(0)} = 0.4925.$$

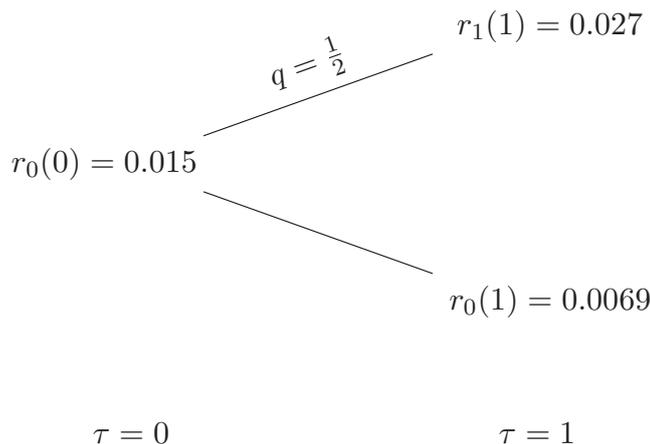
Eq. (11.56) is, now,

$$r_0(1) = \ln \left(\frac{p_0(1) + \delta p_1(1)}{P_{\S}(0, 2)} \right) = \ln \left(\frac{0.4925 \cdot (1 + 0.9802)}{0.9685} \right) = 0.0069.$$

Hence, by Eq. (11.55),

$$r_1(1) = r_0(1) + (\ln \delta^{-1}) = 0.0069 + 0.02 = 0.0270.$$

So, to sum up, we have the tree below,



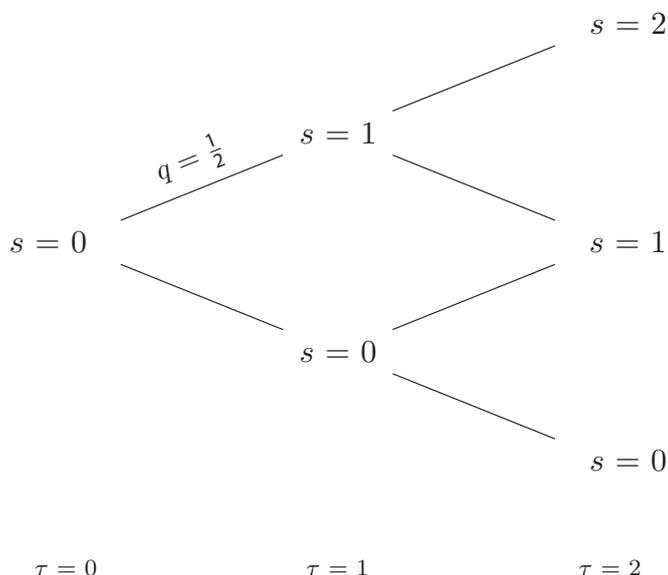
where $p_0(1) = p_1(1) = 0.4925$. Of course, the value of the two securities needs to be the same, because the risk-neutral probability is 50%.

We now proceed to compute the values of the short-term rate for one further period.

- $\tau = 2$. By Eq. (11.57), the forward equation for the Arrow-Debreu prices, we have the following three cases:

$$\begin{aligned}
 (s = 0) \quad p_0(2) &= \frac{1}{2}e^{-r_0(1)} [p_0(1) + 0] = 0.2446 \\
 (s = 1) \quad p_1(2) &= \frac{1}{2}e^{-r_0(1)} [\delta p_1(1) + p_0(1)] = 0.4843 \\
 (s = 2) \quad p_2(2) &= \frac{1}{2}e^{-r_0(1)} [0 + \delta p_1(1)] = 0.2397
 \end{aligned}$$

The tree below further illustrates how to obtain these prices.



Consider, for example, $p_0(2)$. It is the price of the Arrow-Debreu security for time 2, under two consecutive downward movements of the short-term rate. This state can only be accessed to through the state $s = 0$ at time $\tau = 1$. But at state $s = 0$ at time $\tau = 1$, the value of the Arrow-Debreu asset is $\frac{1}{2}e^{-r_0(1)}$. Hence, $p_0(2) = p_0(1) \cdot \frac{1}{2}e^{-r_0(1)}$. By a similar reasoning, we have that $p_2(2) = p_1(1) \cdot \frac{1}{2}e^{-r_1(1)} = p_1(1) \cdot \frac{1}{2}e^{-r_0(1)}\delta$. Note, there

is some symmetry in the distribution of the Arrow-Debreu prices, with $p_1(2)$ being the largest, being the price of the security that pays off with the highest likelihood. However, $p_0(2) > p_2(2)$, even if q is constant and equal to 50%, because discounting is more severe whilst crossing the nodes leading to $s = 2$, compared to the nodes leading to $s = 0$.

We can now compute the values of the short-term rate for each node. Eq. (11.56) is, now,

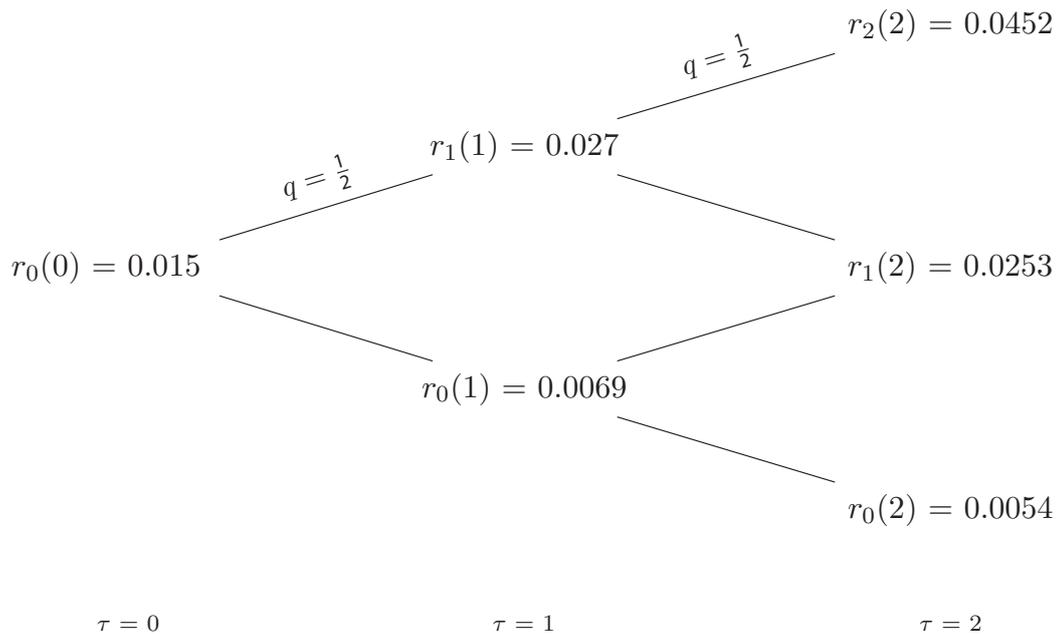
$$\begin{aligned} r_0(2) &= \ln \left(\frac{p_0(2) + \delta p_1(2) + \delta^2 p_2(2)}{P_{\S}(0, 3)} \right) \\ &= \ln \left(\frac{0.2446 + 0.9802 \cdot 0.4843 + (0.9802)^2 \cdot 0.2397}{0.9445} \right) = 0.0054. \end{aligned}$$

Hence, by Eq. (11.55),

$$r_s(2) = r_0(2) + (\ln \delta^{-1}) \cdot s = 0.0054 + 0.02 \cdot s, \quad s = 0, 1, 2.$$

This yields the following values for the short-term rate: $r_0(2) = 0.0054$, $r_1(2) = 0.0253$, and $r_2(2) = 0.0452$.

The diagram below summarizes the implied tree for the short-term rate in this model.



Naturally, the prices $P = e^{-r}$ in the nodes of the previous tree match those calculated in Section 11.6.5, apart from discrepancies arising due to rounding errors.

11.7.2.3 A second numerical example

Assume that the spot yield curve is 2.5% for $t = 1$ year, 4.5% for $t = 2$ years, and 6% for $t = 3$ years, continuously compounded and annualized. Consider the following model:

$$r_s(t) = r_0(t) + a * s, \tag{11.58}$$

where a is a constant and equal to 0.01, $r_s(t)$ is the continuously compounded short-term rate as of time t , after s upward movements and, finally, the unit period of time is taken to be one year. As we know, the Ho & Lee model predicts that the price as of time zero of an Arrow-Debreu security paying off in state s at time t , denoted as $p_s(t)$, satisfies the following forward equation, for $t > 1$, $s \geq 0$ and $s \leq t$:

$$p_s(t) = e^{-r_0(t-1)} \left[(1-q) (e^{-a})^s p_s(t-1) + q (e^{-a})^{s-1} p_{s-1}(t-1) \right], \quad (11.59)$$

where q is the risk-neutral probability of an upward movement in the short-term rate. Furthermore, according to this model, the price of a zero coupon bond, paying £1 at time t , $P_{\S}(0, t)$, equals,

$$P_{\S}(0, t) = e^{-r_0(t-1)} \sum_{s=0}^{t-1} (e^{-a})^s p_s(t-1). \quad (11.60)$$

Suppose, next, that the risk-neutral probability of an upward movement at any time t is not a constant q , but a function of calendar time, say q_t : q_t is, then, the probability of an upward movement in the short-term rate from time t to time $t+1$. Naturally, the assumption that q_t is time-varying, makes this model markedly distinct from Ho & Lee model. To calibrate this model, we consider the recursive equation for the Arrow-Debreu security prices:

$$p_s(t) = e^{-r_0(t-1)} \left[(1-q_{t-1}) (e^{-a})^s p_s(t-1) + q_{t-1} (e^{-a})^{s-1} p_{s-1}(t-1) \right], \quad (11.61)$$

where q_{t-1} denotes the risk-neutral probability of an upward movement in the short-term rate from time $t-1$ to time t . The boundary conditions are the usual ones: $p_0(0) = 1$, $p_s(t) = 0$, for $s \geq t$ and $s < 0$. Eq. (11.61) can be derived through the same arguments in Section 11.7.1.

Next, suppose the risk-neutral probability of an upward movement in the short-term rate in the first period equals $\frac{1}{2}$. Suppose, further, that available for trading is a derivative, which pays off an amount of £1 in state $s=2$ and an amount of £1 in state $s=0$, both at time $t=2$. The current price of this derivative equals 0.45514. The interpretation of the derivative is that of a contract that pays off when the interest rate experiences extreme movements (up-up or down-down)—a raw volatility contract. Its price can be expressed as the sum of the two Arrow-Debreu securities for these extreme interest rate movements. Let us set the nominal values of the zero coupon bonds to £1. To populate the interest rate tree, we need to compute the three zero prices, which are:

$$P_{\S}(0, 1) = e^{-0.025} = 0.97531, \quad P_{\S}(0, 2) = e^{-0.045*2} = 0.91393, \quad P_{\S}(0, 3) = e^{-0.06*3} = 0.83527.$$

We can start populating the tree. Eq. (11.60) can be rewritten as:

$$r_0(t) = \ln \left(\frac{\sum_{s=0}^t (e^{-a})^s p_s(t)}{P_{\S}(0, t+1)} \right). \quad (11.62)$$

We have,

- $t=0$. In this case, Eq. (11.62) is:

$$r_0(0) = \ln \left(\frac{1}{0.97531} \right) = 0.025.$$

- $t = 1$. We have two nodes to fill: $s = 0$ & $s = 1$. We use Eq. (11.61), as follows:

– $s = 0$: We have,

$$p_0(1) = e^{-r_0(0)}(1 - q_0)p_0(0) = 0.97531 * 0.5 * 1 = 0.48766.$$

– $s = 1$: We have,

$$p_1(1) = e^{-r_0(0)}q_0p_0(0) = 0.97531 * 0.5 * 1 = 0.48766.$$

Then, Eq. (11.62) is,

$$r_0(1) = \ln\left(\frac{p_0(1) + (e^{-a})p_1(1)}{P_{\S}(0, 2)}\right) = \ln\left(\frac{0.48766 * (1 + e^{-0.01})}{0.91393}\right) = 0.06$$

$$r_1(1) = r_0(1) + 0.01 = 0.07.$$

- $t = 2$. There are now three nodes to fill, corresponding to $s = 0$, $s = 1$ and $s = 2$. We use Eq. (11.61), as follows:

– $s = 0$: We have,

$$p_0(2) = e^{-r_0(1)}(1 - q_1)p_0(1) = e^{-0.06}(1 - q_1)0.48766.$$

– $s = 1$: We have,

$$p_1(2) = e^{-r_0(1)}[(1 - q_1)e^{-a}p_1(1) + q_1p_0(1)] = e^{-0.06}[(1 - q_1)e^{-0.01} + q_1]0.48766.$$

– $s = 2$: We have,

$$p_2(2) = e^{-r_0(1)}q_1e^{-a}p_1(1) = e^{-0.06}q_1e^{-0.01}0.48766.$$

We do not know yet q_1 . Yet the “rate volatility” asset, which quotes for 0.45514, can be used to extract q_1 . At time $t = 1$, its price is either $Z_u = e^{-0.07}q_1$ (in the up state of the world), or $Z_d = e^{-0.06}(1 - q_1)$ (in the down state of the world). So by no-arbitrage, its current price, satisfies

$$0.45514 = \frac{1}{2}e^{-0.025}(Z_u + Z_d) = \frac{1}{2}e^{-0.025}[e^{-0.07}q_1 + e^{-0.06}(1 - q_1)].$$

Solving for q_1 yields, $q_1 = 0.90$. Naturally, the same result is obtained by calibrating q_1 so as to make the price of the derivative, 0.45514, match the sum of the prices of the Arrow-Debreu securities paying off in states 0 and 2 at $t = 2$, viz $q_1 : 0.45514 = p_0(2) + p_2(2) = e^{-0.06}(1 - q_1)0.48766 + e^{-0.06}q_1e^{-0.01}0.48766$. So now, we can use $q_1 = 90\%$ and calculate the Arrow-Debreu prices, obtaining:

$$p_0(2) = e^{-0.06}(1 - .9)0.48766 = 0.04592$$

$$p_1(2) = e^{-0.06}[(1 - .9)e^{-0.01} + .9]0.48766 = 0.4588$$

$$p_2(2) = e^{-0.06}.9e^{-0.01}0.48766 = 0.40922$$

Note, there is no symmetry at all in the distribution of these Arrow-Debreu security prices. The price $p_0(2)$ is very low, due to the fact that q_1 is very high, such that the probability of reaching the lowest node of the tree at time $t = 2$ is quite low.

Next, by Eq. (11.62),

$$\begin{aligned} r_0(2) &= \ln \left(\frac{p_0(2) + e^{-a}p_1(2) + e^{-2a}p_2(2)}{P_{\text{\$}}(0,3)} \right) \\ &= \ln \left(\frac{0.04592 + e^{-0.01}0.4588 + e^{-2*0.01}0.40922}{0.83527} \right) = 0.07605, \end{aligned}$$

and,

$$\begin{aligned} r_1(2) &= r_0(2) + 0.01 = 0.07605 + 0.01 = 0.08605 \\ r_2(2) &= r_0(2) + 2 * 0.01 = 0.07605 + 2 * 0.01 = 0.09605 \end{aligned}$$

Finally, we wish to evaluate a European call option at time zero, written on the three year zero coupon bond with nominal value equal to £1. This option expires at $t = 2$ and has a strike price equal to £ 0.91000. At expiry, the option pays off:

$$\begin{aligned} C_2(2) &\equiv (e^{-r_2(2)} - 0.91)^+ = (e^{-0.09605} - 0.91)^+ = 0 \\ C_1(2) &\equiv (e^{-r_1(2)} - 0.91)^+ = (e^{-0.08605} - 0.91)^+ = 0.00755 \\ C_0(2) &\equiv (e^{-r_0(2)} - 0.91)^+ = (e^{-0.07605} - 0.91)^+ = 0.01677 \end{aligned}$$

Then,

$$\begin{aligned} C_u &= e^{-r_1(1)} (q_1 C_2(2) + (1 - q_1) C_1(2)) \\ &= e^{-0.07} * (0.9 * 0 + 0.1 * 0.00755) = 7.0396 \times 10^{-4} \\ C_d &= e^{-r_0(1)} (q_1 C_1(2) + (1 - q_1) C_0(2)) \\ &= e^{-0.06} * (0.9 * 0.00755 + 0.1 * 0.01677) = 7.9786 \times 10^{-3} \end{aligned}$$

which leads to $C = e^{-0.025} \frac{1}{2} (C_u + C_d) = 4.2341 \times 10^{-3}$.

Finally, using all the market data so far, we wish to evaluate a second European call option written on the three year bond, expiring in one year, and struck at £ 0.85. Its no-arbitrage price, denoted with C_T , is:

$$C_T = e^{-0.025} \frac{1}{2} ((P_u(1,3) - 0.85)^+ + (P_d(1,3) - 0.85)^+),$$

where

$$\begin{aligned} P_u(1,3) &= e^{-r_1(1)} (q_1 e^{-r_2(2)} + (1 - q_1) e^{-r_1(2)}) \\ &= e^{-0.07} * (0.90 * e^{-0.09605} + 0.10 * e^{-0.08605}) = 0.84786 \\ P_d(1,3) &= e^{-r_0(1)} (q_1 e^{-r_1(2)} + (1 - q_1) e^{-r_0(2)}) \\ &= e^{-0.06} * (0.90 * e^{-0.08605} + 0.10 * e^{-0.07605}) = 0.86498 \end{aligned}$$

That is, $C_T = e^{-0.025} \frac{1}{2} (0.01498) = 0.00730 \times 10^{-3}$. Suppose, now, that the market value of this option diverges from C_T , i.e. $C_T \neq C_{\text{\$}}$, where $C_{\text{\$}}$ is the market value of the option. For example, $C_T < C_{\text{\$}}$. To implement this arbitrage opportunity, we can sell the option, and use the proceeds to build up a portfolio comprising the bond expiring in three years and a money market account, with initial value:

$$V_0 = \Delta P_{\text{\$}}(0,3) + M,$$

where Δ and M are chosen to match the payoffs promised by the option at time 1:

$$(\Delta, M) : \begin{cases} \Delta P_u(1, 3) + Me^{r_0} = \pi_u \\ \Delta P_d(1, 3) + Me^{r_0} = \pi_d \end{cases}$$

where π_u and π_d are the payoffs of the one year option. The solution is,

$$\hat{\Delta} = \frac{\pi_u - \pi_d}{P_u(1, 3) - P_d(1, 3)}, \quad \hat{M} = e^{-r_0} \frac{\pi_d P_u(1, 3) - \pi_u P_d(1, 3)}{P_u(1, 3) - P_d(1, 3)}.$$

Using the numerical values obtained so far, $\pi_u = 0$, $\pi_d = 0.01498$, $P_u(1, 3) = 0.84786$, $P_d(1, 3) = 0.86498$, we have:

$$\hat{\Delta} = \frac{-0.01498}{0.84786 - 0.86498} = 0.875, \quad \hat{M} = e^{-0.025} \frac{0.01498 * 0.84786}{0.84786 - 0.86498} = -0.72356.$$

The current value of this portfolio is,

$$\hat{V}_0 = \hat{\Delta} P_{\$}(0, 3) + \hat{M} = 0.875 * 0.83527 - 0.72356 = 0.00730 = C_T.$$

11.8 Callables, puttable and convertibles with trees

This section provides an introductory discussion about the pricing of callable, puttable and convertible bonds, with and without credit risk, and develops basic pricing examples for callable and convertibles, relying on binomial trees. Chapter 12 develops a continuous time evaluation framework for callable and puttable bonds, while Chapter 13, contains a continuous time model to evaluate convertible bonds.

Callable bonds are assets that can be called back by the issuer at a pre-specified strike price, either at a fixed maturity date or at any fixed date before the expiration. The rationale behind this optionality is that at the date of issuance, the market might not, perhaps, share the same optimism as the issuer as regards the issuer's future creditworthiness. By adding the indenture to call the bonds, the issuer gives itself the option to refinance at some future date, at hopefully better market conditions. Although this specific example might link to agency problems or difference in beliefs between the bond issuer and market participants, the indenture to call the bond is an option that might generally arise as a result of pure hedging motives, arising by a concern that future interest rates might lower.

Naturally, the right to call the bonds rises the cost of capital, to the extent of the value of this (call) option to redeem the bonds. Mathematically, for each point in time τ say, when the option to redeem the bonds can be exercised at strike K , the value of the bond is:

$$\min \{D_\tau, K\} = D_\tau - \max \{D_\tau - K, 0\},$$

where D_τ is the time τ present value of the future expected discounted cash flows promised at time τ , by a callable bond with the same strike price K . Indeed, suppose that at τ , interest rates have decreased to an extent to make $D_\tau > K$. In this case, the issuer may proceed to redeem the bonds for K , and issue new callable debt, exercisable at any fixed date before the expiration, for a price D_τ , thereby cashing in the difference $D_\tau - K$. In doing so, the bond-issuer is left with the same optionalities he would have by not exercising the option to call, but with

the additional “money-shower,” $D_\tau - K$. It is, therefore, in the interest of the bond-issuer to exercise at τ , when $D_\tau > K$, and it is obviously not otherwise.

Puttable bonds, instead, are assets that give the holder the right to sell the bonds back to the issuer at some exercise price, either at a fixed maturity date or any fixed date before the expiration. The bondholders would exercise their option to tender the bonds to the issuer when market conditions improve from their perspective, i.e. when interest rates are high enough, so as to make bond prices lower than the exercise price. Issuing puttable bonds, therefore, lowers the cost of capital, to the extent of the value of the (put) option given to the bondholders to tender the bonds at the strike K . Suppose for example, that the bondholders can exercise their option at some pre-specified date. In this case, the payoff of the puttable bond is given by:

$$\max\{P, K\} = P + \max\{K - P, 0\},$$

where P is the price of a non-puttable bond. Indeed, suppose that this price, P , lowers to a level less than the strike price K . The bondholders, then, will find convenient to tender the bonds at K , buying conventional bonds at P , thereby cashing in $K - P$, and then wait until maturity. This trade would provide bondholders with a “money-shower” of $K - P$, at the exercise date. Alternatively, the bondholders would not exercise, and wait until maturity, in which case they would not receive the profit $K - P$, at the exercise date. Therefore, it is optimal to exercise when $K > P$, and it is obviously not when $K < P$.

Convertible bonds are assets that give the holder the right to convert them into a prespecified number of shares of the firm. Their value at each date when the conversion can take place is $\max\{CV, P\} = P + \max\{CV - P, 0\}$, where CV denotes the conversion value of the bonds, expressed in terms of the value of the firm’s shares: issuing convertible bonds now lowers the cost of capital to the extent of the option given to the bondholders to convert the bonds into shares. Convertible bonds can be made callable by the bond-issuers, at a strike K . Usually, if the bonds are called, the convertible bondholders have the option to either tender the bonds to the firm, or to convert them. On the other hand, the only reason the bond-issuers might call is that the price of the convertibles is up, compared to the strike price. Therefore, the option to make convertible bonds also callable puts a ceiling to the price of the convertibles bonds, given by the exercise price, K . Mathematically, in the presence of callability, the value of a convertible bond at each potential conversion date is $\max\{CV, \min\{P, K\}\}$: the option to call back takes away some of the optionality from the bondholders, who are, in effect, forced to convert, as soon as the price P increases to a level beyond K .

11.8.1 Callable bonds

11.8.1.1 Copying with credit risk

To evaluate callable bonds through trees, we may simply follow the methodology in this chapter, corrected for the presence of credit risk.

- (i) First, we “populate” a short-term rate tree through one of the models described in this chapter (say, for example, through the Black, Derman and Toy (1990) model).
- (ii) Second, we use this tree to find the value of some coupon bearing bond of interest, by just using the short-term rate process of the previous step.
- (iii) Third, we use the results obtained in the second step and build up a tree for the callable bond. In each node immediately preceding the maturity, we compare the strike price with

the non-callable coupon bearing bond price (ex-coupon) and take the minimum of the two. We add the coupon to this minimum and find, then, the payoff of the non-callable bond at the relevant node. This gives us $V = \min\{K, B^{\text{rolled-back}}(\text{ex-coupon})\} + \text{coupon}$, where K is the call price, and $B^{\text{rolled-back}}(\text{ex-coupon})$ is the ex-coupon bond price, found from the values of the bond V in the next nodes (by using, as usual, recursive, backward solution, i.e. the risk-neutral expectation of the future payoffs).

- (iv) Fourth, we go backward, discounting the values obtained in the previous steps, V say, obtaining, for each node, $V_- = \min\{K, V\} + \text{coupon}$, etc. Hence, we find the price. If the callable bond is not subject to default risk, we stop. Otherwise, we proceed to the next step.
- (v) Fifth, we correct for credit risk. The price we found in the fourth step is typically different than the market price. One issue is that the market price reflects the credit risk of the firm, and should be typically less than the price obtained in the fourth step. The trick, here, is to search for an *additional spread to add to the short-term rate* process obtained in the first step, such that the theoretical bond price equals the market price of the bond. This is done numerically, and alters the results obtained in steps 3 and 4.

At this point, we may price options written on callable bonds. Ho and Lee (2004) (Chapter 8, Section 8.3 p. 274-278) develop a number of useful exercises on the pricing of options on callable bonds, through tree methods.

11.8.1.2 A numerical example: without credit

To illustrate, let us consider a simpler situation, relating to the pricing of a callable bond without credit risk. Assume that the discretely compounded six-month rate, or the “short-term rate,” evolves over time according to the tree described in the following diagram:

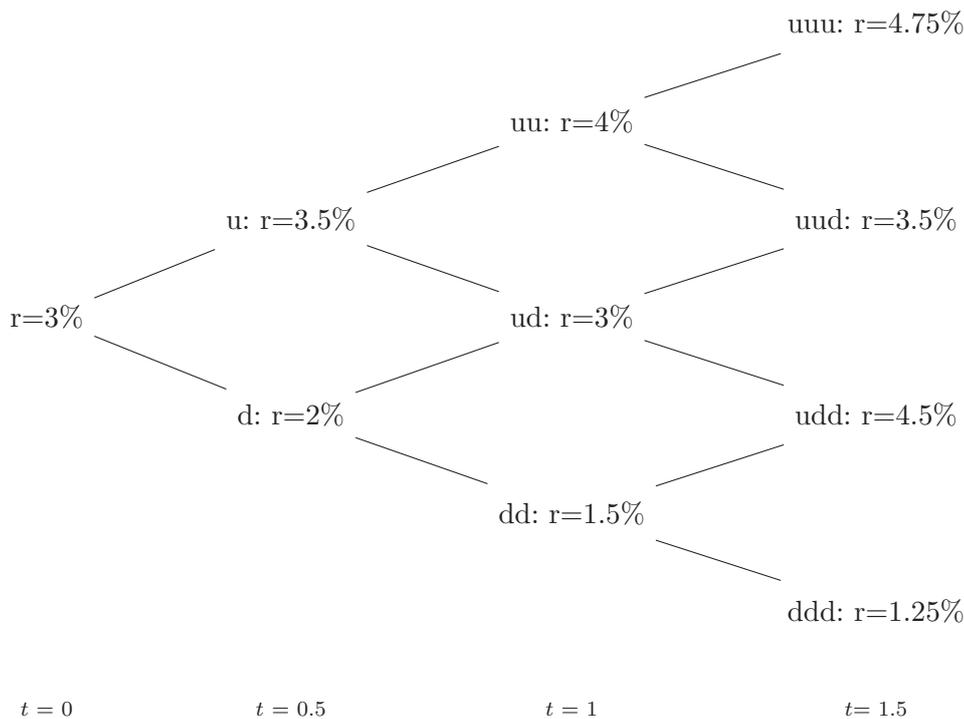


FIGURE 11.10.

Next, consider a bond expiring in two years, paying off coupon rates of 3% of the principal of £1 every six months, and callable at any time by the issuer, at par value. Let this bond be labeled “BCX.” Suppose that the prices of three zero coupon bonds expiring in one year, eighteen months and two years are, respectively, 0.94632, 0.91876 and 0.89166. We can use these market data to calibrate the risk-neutral probabilities of upward movements in the short-term rate implied by the binomial tree in Figure 11.15, provided these risk-neutral probabilities depend only on calendar time t , not on the specific state of nature at time t .

We assume that available for trading is also a conventional, (i.e. non-callable) bond maturing in two years and paying coupons semiannually, at 3% of the principal of £1. We wish to calculate the price movements of the non-callable coupon-bearing two year bond. We have, $P_{\S}(0, 0.5) = \frac{1}{1.03} = 0.97087$. Furthermore, as regards the zero expiring in one year:

$$P_{\S}(0, 1) = 0.94632 = P_{\S}(0, 0.5) * \left(q_0 \frac{1}{1.035} + (1 - q_0) \frac{1}{1.02} \right),$$

which solved for q_0 delivers $q_0 = 0.40$. As for the zero expiring in 1.5 years,

$$\begin{aligned}
 P_{\S}(0, 1.5) &= 0.91876 = P_{\S}(0, 0.5) * (q_0 P_U(0.5, 1.5) + (1 - q_0) P_D(0.5, 1.5)) \\
 &= 0.97087 * (0.40 * P_U(0.5, 1.5) + 0.60 * P_D(0.5, 1.5)),
 \end{aligned}$$

where

$$P_U(0.5, 1.5) = \frac{1}{1.035} \left(q_1 \frac{1}{1.04} + (1 - q_1) \frac{1}{1.03} \right), \quad P_D(0.5, 1.5) = \frac{1}{1.02} \left(q_1 \frac{1}{1.03} + (1 - q_1) \frac{1}{1.015} \right)$$

Solving for q_1 leaves $q_1 = 0.70$. As for the zero expiring in 2 years:

$$\begin{aligned}
 P_{\S}(0, 2) &= 0.89166 = P_{\S}(0, 0.5) * (q_0 P_U(0.5, 2) + (1 - q_0) P_D(0.5, 2)) \\
 &= 0.97087 * (0.40 * P_U(0.5, 2) + 0.60 * P_D(0.5, 2)),
 \end{aligned}$$

where

$$\begin{aligned} P_U(0.5, 2) &= \frac{1}{1.035} (q_1 P_{UU}(1, 2) + (1 - q_1) P_{UD}(1, 2)) \\ &= \frac{1}{1.035} (0.70 * P_{UU}(1, 2) + 0.30 * P_{UD}(1, 2)) \\ P_D(0.5, 2) &= \frac{1}{1.02} (q_1 P_{UD}(1, 2) + (1 - q_1) P_{DD}(1, 2)) \\ &= \frac{1}{1.02} (0.70 * P_{UD}(1, 2) + 0.30 * P_{DD}(1, 2)) \end{aligned}$$

and:

$$\begin{aligned} P_{UU}(1, 2) &= \frac{1}{1.04} \left(q_2 \frac{1}{1.0475} + (1 - q_2) \frac{1}{1.035} \right) \\ P_{UD}(1, 2) &= \frac{1}{1.03} \left(q_2 \frac{1}{1.035} + (1 - q_2) \frac{1}{1.02} \right) \\ P_{DD}(1, 2) &= \frac{1}{1.015} \left(q_2 \frac{1}{1.02} + (1 - q_2) \frac{1}{1.0125} \right) \end{aligned}$$

Solving for q_2 leaves $q_2 = 0.60$.

The price of a coupon bearing bond yielding 3% of the principal every six months is easy to calculate,

$$B(0, 2) = 0.03 * (0.97087 + 0.94632 + 0.91876 + 0.89166) + 0.89166 = 1.0035. \quad (11.63)$$

Given the market data and the previously calibrated risk-neutral probabilities, we now proceed with the calculation of the price of the callable coupon bearing bond. We discount the expected cash flows, through the evaluation formula, $\min\{D, 1\} + 0.03$, where D is the present value of the future expected discounted cash flows promised at each node by a callable bond with the same strike price K . We have:

(i) At $t = 1.5$ years,

- uu: $\frac{1.03}{1.0475} = 0.98329$ vs 1 \Rightarrow wait, and the value of the callable bond is 0.98329.
- uud: $\frac{1.03}{1.035} = 0.99517$ vs 1 \Rightarrow wait, and the value of the callable bond is 0.99517.
- udd: $\frac{1.03}{1.02}$ vs 1 \Rightarrow exercise, and the value of the callable bond is 1.
- ddd: $\frac{1.03}{1.0125}$ vs 1 \Rightarrow exercise, and the value of the callable bond is 1.

(ii) At $t = 1$ year, we have that $q = 60\%$, and, then:

- uu: $\frac{1}{1.04} [0.6 (\frac{1.03}{1.0475} + 0.03) + 0.4 (\frac{1.03}{1.035} + 0.03)] = 0.97889$ vs 1 \Rightarrow wait, and the value of the callable bond is 0.97889.
- ud: $\frac{1}{1.03} [0.6 (\frac{1.03}{1.035} + 0.03) + 0.4 (1 + 0.03)] = 0.99719$ vs 1 \Rightarrow wait, and the value of the callable bond is 0.99719.
- dd: $\frac{1}{1.0015} [0.6 (1 + 0.03) + 0.4 (1 + 0.03)] = 1.0285$ vs 1 \Rightarrow exercise, and the value of the callable bond is 1.

(iii) At $t = 0.5$ years, we have that $q = 70\%$, and, then:

- u: $\frac{1}{1.035} [(0.7 * 0.97889 + 0.3 * 0.99719) + 0.03] = 0.98008$ vs $1 \Rightarrow$ wait, and the value of the callable bond is 0.98008.
- d: $\frac{1}{1.02} [(0.7 * 0.99719 + 0.3 * 1) + 0.03] = 1.0079$ vs $1 \Rightarrow$ exercise, and the value of the callable bond is 1.

Finally, at the time of evaluation, we have that $q = 40\%$, and, then, the price of the callable bond is:

$$P^c = \frac{1}{1.03} (0.40 * 0.98008 + 0.60 * 1 + 0.03) = 0.99226.$$

Naturally, the callable bond is valued less than the conventional bond $B(0, 2)$ in Eq. (11.63): the difference is the value of the option given to the issuer to redeem these bonds, and arises when the interest rates go sufficiently down—negative convexity.

How would one proceed to price the BCX bond, if we the previous market data were unavailable? Assume that: (i) the risk-neutral probabilities of upward movements in the short-term rate are: (i.a) unknown from time zero to 0.5 years; (i.b) 70%, from 0.5 to one year; and (i.c) 60%, from one to 1.5 years; (ii) available for trading is a European call option written on the BCX bond; (iii) this option, which quotes for $\pounds 1.7226 \times 10^{-3}$, expires in 1.5 years, is struck at $\pounds 0.99000$, and becomes worthless as soon as the underlying callable bond is called back by the issuer? First, note that at the expiration, $t = 1.5$ years, the payoffs of the option are:

$$C_{uuu} = 0, \quad C_{uud} = 0.00517,$$

and, because of the sudden death assumption,

$$C_{udd} = C_{ddd} = 0.00000.$$

At $t = 1$ year, we have that $q = 60\%$, and, then:

$$\begin{aligned} C_{uu} &= \frac{1}{1.04} (0.6 * 0 + 0.4 * 0.00517) = 1.9885 \times 10^{-3} \\ C_{ud} &= \frac{1}{1.03} (0.6 * 0.00517 + 0.4 * 0) = 3.0117 \times 10^{-3} \\ C_{dd} &= 0 \end{aligned}$$

At $t = 0.5$ years, we have that $q = 70\%$, and, then:

$$\begin{aligned} C_u &= \frac{1}{1.035} (0.70 * C_{uu} + 0.30 * C_{ud}) \\ &= \frac{1}{1.035} (0.70 * 1.9885 \times 10^{-3} + 0.30 * 3.0117 \times 10^{-3}) = 2.2178 \times 10^{-3}. \\ C_d &= 0, \text{ by the sudden death assumption.} \end{aligned}$$

At the time of evaluation, the price of the call is,

$$C = 1.7226 \times 10^{-3} = \frac{1}{1.03} (qC_u + (1 - q)C_d) = \frac{1}{1.03} * q * 2.2178 \times 10^{-3},$$

where q is the risk-neutral probability of an upward movement in the short-term rate during the first six months. We can solve for this q , obtaining $q = 80\%$. Finally, given this probability, we can calculate the price of the callable bond. We have:

$$P^c = \frac{1}{1.03} (0.80 * 0.98008 + 0.20 * 1 + 0.03) = 0.98453.$$

It is lower than the price calculated earlier, because the price of the option is giving more weight (80%) than before (40%) to the occurrence of the state of the world where the interest rate goes up.

11.8.2 Convertible bonds

11.8.2.1 Evaluation issues

Consider the following convertible and callable bond. Let K be the strike at which the bond can be called by the bond-issuer, and let the parity, or *conversion value*, be $CV = CR \times S$, where S is the price of the common share. To evaluate this bond through a binomial tree, we may proceed through the following three steps:

- (i) First, we set the life of the tree equal to the life of the callable convertible bond.
- (ii) Second, we assess the evolution of the stock price along the tree, under the risk-neutral probability. This is done following the standard Cox, Ross and Rubinstein (1979) approach.
- (iii) Third, in each node, we compute the value of the bond as $\max\{CV, \min\{B, K\}\}$, where B is the value of the bond, “rolled-back” from the values of the bond in the next nodes through the usual recursive, backward method—relying on calculating the present value of the risk-neutral expectation of the future payoffs. That is, assuming the bondholder does not convert, the value is $B^* = \min\{B, K\}$, where B is the “rolled-back” value of the bond. Then, the value is $\max\{CV, B^*\}$.

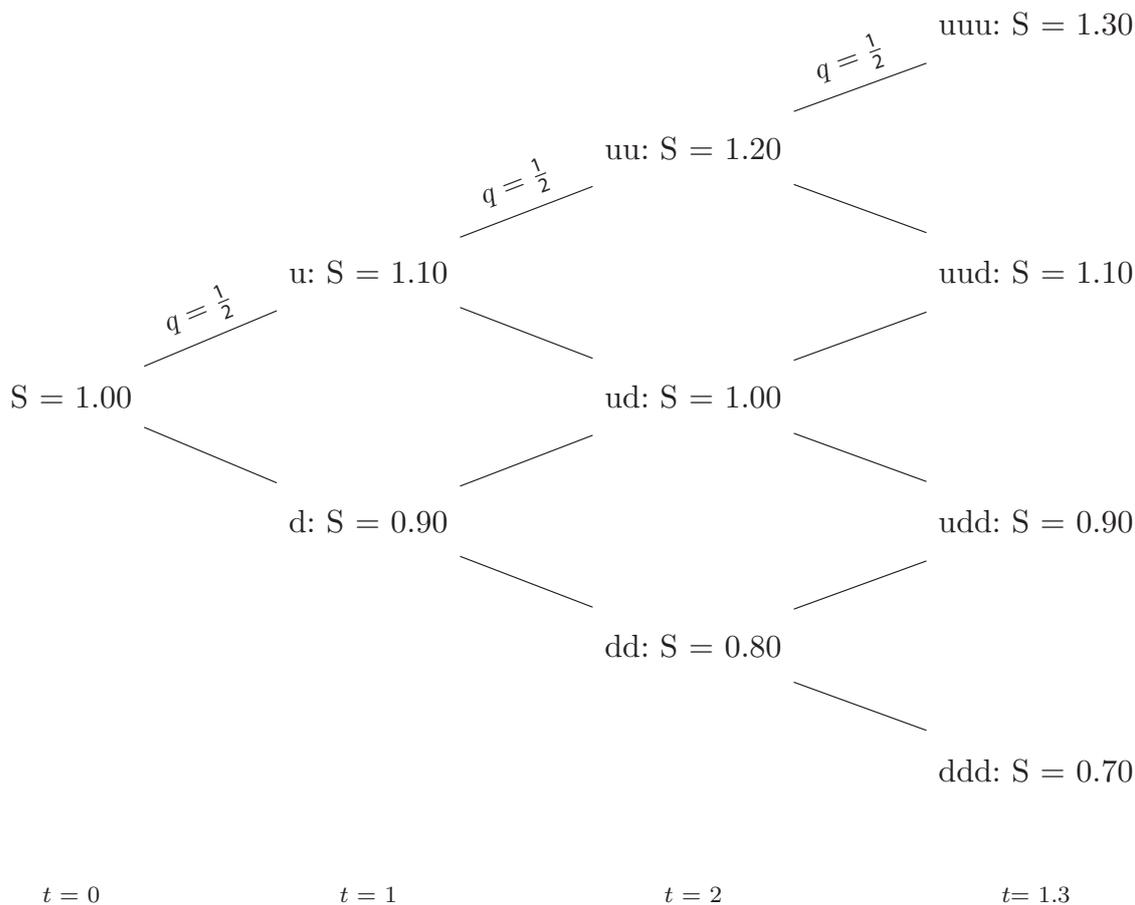
Note, this procedure leads to fill in the nodes, once we know the appropriate interest rate. If the firm was not subject to default risk, we would simply use the riskless interest rate. However, the firm is obviously subject to default risk. In practice, we proceed as follows. In each node, the value of the bond is decomposed into two parts. One part, related to the “pure debt component,” discounted at the defaultable interest rate; and one part related to the “pure equity component,” discounted at the default-free interest rate. Exercise 25.7 in Hull (2003) (p. 653-654) illustrates a specific example.

11.8.2.2 A numerical example: without credit

Consider a three year convertible bond, which can be converted at any time into one share of the underlying firm’s stock. The bond has a face value equal to 1, it is default-free, and it pays off a coupon of 3% of the face value every year, except the time at which it is issued. Moreover, in each period, it pays off the coupon, regardless of whether it will be converted or not.

The price of the share is assumed to be unaffected by any decision relating to the conversion of the bond, and evolves over time as described by the following tree:

In the previous diagram, each period corresponds to one year, S denotes the price of the share, and $q = \frac{1}{2}$ is the constant risk-neutral probability of price movements. Assume, finally,



that the yield curve is flat at 3%, discretely compounded, and that it will remain such over the next three periods, and in each state of the world.

We proceed to calculate the conversion value of the convertible bond at each node of the tree. We shall identify, then, the nodes where it is optimal for the bond-holder to convert. Finally, we shall determine the value of the convertible bond at time $t = 0$, as well as the value of the option to convert. As for the conversion value, we know this is simply the product of the conversion ratio times the current value of the outstanding stock, and equals $CV_t = CR \cdot S_t = S_t$, as the conversion ratio is one. To find the current value of the convertible bond, we proceed recursively, as explained earlier, and calculate, for each date t and each node, $\max\{CV_t, B_t\}$, where B_t denotes the present value of the future cash flows of the convertible, in case of no conversion at time t . The payoffs at time $t = 3$ are:

uuu: $CV = S = 1.30$, $B_{UUU} \equiv \max\{CV, 1\} + 0.03 = 1.33$, convert

uud: $CV = S = 1.10$, $B_{UUD} \equiv \max\{CV, 1\} + 0.03 = 1.13$, convert

udd: $CV = S = 0.90$, $B_{UDD} \equiv \max\{CV, 1\} + 0.03 = 1.03$

ddd: $CV = S = 0.70$, $B_{DDD} \equiv \max\{CV, 1\} + 0.03 = 1.03$

We have:

$t = 2$ uu: $CV = S = 1.20$, $B_{UU} \equiv \max\{CV, A_1\} + 0.03$, $A_1 \equiv \frac{1}{1.03} \frac{1}{2} (B_{UUU} + B_{UUD}) = \frac{1}{1.03} \frac{1}{2} (1.33 + 1.13) = 1.19420$. Hence $B_{UU} = 1.23000$, convert

ud: $CV = S = 1.00$, $B_{UD} \equiv \max\{CV, A_2\} + 0.03$, $A_2 \equiv \frac{1}{1.03} \frac{1}{2} (B_{UUD} + B_{UDD}) = \frac{1}{1.03} \frac{1}{2} (1.13 + 1.03) = 1.04850$. Hence $B_{UD} = 1.07850$

dd: $CV = S = 0.80$, $B_{DD} \equiv \max\{CV, A_3\} + 0.03$, $A_3 \equiv \frac{1}{1.03} \frac{1}{2} (B_{UDD} + B_{DDD}) = 1$. Hence $B_{DD} = 1.03000$

$t = 1$ u: $CV = S = 1.10$, $B_U \equiv \max\{CV, L_1\} + 0.03$, $L_1 \equiv \frac{1}{1.03} \frac{1}{2} (B_{UU} + B_{UD}) = \frac{1}{1.03} \frac{1}{2} (1.23 + 1.07850) = 1.1206$. Hence $B_U = 1.15060$

d: $CV = S = 0.90$, $B_D \equiv \max\{CV, L_2\} + 0.03$, $L_2 \equiv \frac{1}{1.03} \frac{1}{2} (B_{UD} + B_{DD}) = \frac{1}{1.03} \frac{1}{2} (1.07850 + 1.03000) = 1.0235$. Hence $B_D = 1.0535$

Finally, we have that:

$$B^{\text{convertible}} = \frac{1}{0.03} \frac{1}{2} (B_U + B_D) = \frac{1}{1.03} \frac{1}{2} (1.15060 + 1.0535) = 1.0700.$$

Instead, the value of a non-convertible three year coupon bearing bond is

$$\left(\frac{1}{1.03}\right)^3 + 0.03 * \left(\frac{1}{1.03} + \left(\frac{1}{1.03}\right)^2 + \left(\frac{1}{1.03}\right)^3\right) = 1.00000.$$

Therefore, the option to convert is worth 0.07000.

Next, assume that the convertible bond is also callable by the issuer, at any time, and at a strike value of 1.02000, and if it is called, the bond-holder has the option to tender the bond or to convert it into one share. This convertible, and callable, bond can be evaluated as in the previous calculations, although the formula to use in each node is, now, $\max\{CV_t, \min\{B_t, K\}\}$, with $K = 1.02000$. The payoffs at time $t = 3$ are, now:

uuu: $CV = S = 1.30$, $B_{UUU} \equiv \max\{CV, \min\{B = 1, 1.02\}\} + 0.03 = 1.33$, convert

uud: $CV = S = 1.10$, $B_{UUD} \equiv \max\{CV, \min\{B = 1, 1.02\}\} + 0.03 = 1.13$, convert

udd: $CV = S = 0.90$, $B_{UDD} \equiv \max\{CV, \min\{B = 1, 1.02\}\} + 0.03 = 1.03$

ddd: $CV = S = 0.70$, $B_{DDD} \equiv \max\{CV, \min\{B = 1, 1.02\}\} + 0.03 = 1.03$

We have:

$t = 2$ uu: $CV = S = 1.20$, $B_{UU} \equiv \max\{CV, \min\{A_1, 1.02\}\} + 0.03$, $A_1 \equiv \frac{1}{1.03} \frac{1}{2} (B_{UUU} + B_{UUD}) = \frac{1}{1.03} \frac{1}{2} (1.33 + 1.13) = 1.19420$. Hence $B_{UU} = \max\{CV, 1.02\} + 0.03 = 1.23$. The bond is called, but then converted into a share

ud: $CV = S = 1.00$, $B_{UD} \equiv \max\{CV, \min\{A_2, 1.02\}\} + 0.03$, $A_2 \equiv \frac{1}{1.03} \frac{1}{2} (B_{UUD} + B_{UDD}) = \frac{1}{1.03} \frac{1}{2} (1.13 + 1.03) = 1.04850$. Hence $B_{UD} = \max\{CV, 1.02\} + 0.03 = 1.05$. The bond is called, but not converted into a share

dd: $CV = S = 0.80$, $B_{DD} \equiv \max\{CV, \min\{A_3, 1.02\}\} + 0.03$, $A_3 \equiv \frac{1}{1.03} \frac{1}{2} (B_{UDD} + B_{DDD}) = 1$. Hence $B_{DD} = \max\{CV, 1\} + 0.03 = 1.03000$

$t = 1$ u: $CV = S = 1.10$, $B_U \equiv \max\{CV, \min\{L_1, 1.02\}\} + 0.03$, $L_1 \equiv \frac{1}{1.03} \frac{1}{2} (B_{UU} + B_{UD}) = \frac{1}{1.03} \frac{1}{2} (1.23 + 1.07850) = 1.1206$. Hence $B_U = \max\{CV, 1.02\} + 0.03 = 1.13000$. The bond is called, but then converted into a share

d: $CV = S = 0.90$, $B_D \equiv \max\{CV, \min\{L_2, 1.02\}\} + 0.03$, $L_2 \equiv \frac{1}{1.03} \frac{1}{2} (B_{UD} + B_{DD}) = \frac{1}{1.03} \frac{1}{2} (1.07850 + 1.03000) = 1.0235$. Hence $B_D = \max\{CV, 1.02\} + 0.03 = 1.0500$.
The bond is called, but not converted into a share

Therefore, we have that,

$$B^{\text{convertible, callable}} = \frac{1}{1.03} \frac{1}{2} (B_U + B_D) = \frac{1}{1.03} \frac{1}{2} (1.13000 + 1.0500) = 1.0583.$$

As expected, the value of a convertible callable is less than that of the convertible, due to the option given to the bond-issuers to call the bond.

11.9 Appendix 1: Proof of Eq. (11.18)

Let $P(r, T_i)$ denote the price of a zero with maturity T_i , $i = 1, 2$, when the interest rate is equal to r . We wish to replicate a zero with maturity T_2 by means of a portfolio that includes a zero with maturity T_1 . Consider the following portfolio: (i) Go long Δ zeros with maturity T_1 and (ii) invest M in the MMA. Let V_0 be the current value of this portfolio. V_0 is clearly a function of the current short-term rate r , and equals,

$$V_0(r) = \Delta \cdot P(r, T_1) + M.$$

In the second period, the value of the portfolio is random, as it depends on the development of the short-term rate \tilde{r} . Precisely, the value of the portfolio in the second period, is

$$V(\tilde{r}) = \begin{cases} V(r^+) = \Delta \cdot P(r^+, T_1) + M \cdot (1 + r), & \text{with probability } p \\ V(r^-) = \Delta \cdot P(r^-, T_1) + M \cdot (1 + r), & \text{with probability } 1 - p \end{cases}$$

We also know that in the second period, the value of the second zero is,

$$P(\tilde{r}, T_2) = \begin{cases} P(r^+, T_2), & \text{with probability } p \\ P(r^-, T_2), & \text{with probability } 1 - p \end{cases}$$

Next, we select Δ and M to make the value of the portfolio equal the value of the second zero, in each state of nature, viz

$$V(\tilde{r}) = P(\tilde{r}, T_2), \quad \text{in each state.}$$

Mathematically, this is tantamount to solving the following system of two equations with two unknowns (Δ and M),

$$\begin{cases} V(r^+) = \Delta \cdot P(r^+, T_1) + M \cdot (1 + r) = P(r^+, T_2) \\ V(r^-) = \Delta \cdot P(r^-, T_1) + M \cdot (1 + r) = P(r^-, T_2) \end{cases} \quad (11A.1)$$

The solution is,

$$\hat{\Delta} = \frac{P(r^+, T_2) - P(r^-, T_2)}{P(r^+, T_1) - P(r^-, T_1)}, \quad \hat{M} = \frac{P(r^-, T_2)P(r^+, T_1) - P(r^+, T_2)P(r^-, T_1)}{[P(r^+, T_1) - P(r^-, T_1)](1 + r)}.$$

By construction, the previous portfolio, $(\hat{\Delta}, \hat{M})$, replicates the value of the second zero in the second period. But if two assets (the portfolio, and the second zero) yield the same payoffs in each state of the nature, they must be worth the same, in the absence of arbitrage. Therefore, we must have,

$$V_0(r)|_{\Delta=\hat{\Delta}, M=\hat{M}} = \hat{\Delta} \cdot P(r, T_1) + \hat{M} = P(r, T_2),$$

or,

$$(1 + r) \hat{M} = (1 + r) P(r, T_2) - (1 + r) \hat{\Delta} \cdot P(r, T_1). \quad (11A.2)$$

Next, let us figure out the prediction of the model in terms of the expected return it generates for the price of the bond maturing at T_1 , when $(\Delta, M) = (\hat{\Delta}, \hat{M})$. To do this, multiply the first equation in (11A.1) by p , and multiply the second equation in (11A.1) by $1 - p$. Add the result for $\Delta = \hat{\Delta}$, $M = \hat{M}$ to obtain,

$$\hat{\Delta} \cdot [pP(r^+, T_1) + (1 - p)P(r^-, T_1)] + \hat{M} \cdot (1 + r) = pP(r^+, T_2) + (1 - p)P(r^-, T_2).$$

Replacing Eq. (11A.2) into the previous equation yields,

$$\begin{aligned} & \hat{\Delta} \cdot [(pP(r^+, T_1) + (1 - p)P(r^-, T_1)) - (1 + r)P(r, T_1)] \\ & = [pP(r^+, T_2) + (1 - p)P(r^-, T_2)] - (1 + r)P(r, T_2). \end{aligned}$$

Finally, replacing the solution for $\hat{\Delta}$ into the previous equation leaves,

$$\frac{[pP(r^+, T_1) + (1-p)P(r^-, T_1)] - (1+r)P(r, T_1)}{P(r^+, T_1) - P(r^-, T_1)} = \frac{[pP(r^+, T_2) + (1-p)P(r^-, T_2)] - (1+r)P(r, T_2)}{P(r^+, T_2) - P(r^-, T_2)}.$$

The previous equation is easy to interpret. The numerators are the expected *excess* returns from holding the assets. They equal $E_p [P(\tilde{r}, T_i)] - (1+r)P(r, T_i)$, where $E_p [P(\tilde{r}, T_i)]$ is what the investors expect to receive, the next period, by investing $\mathcal{L}P(r, T_i)$ today, in the bond; and $(1+r)P(r, T_i)$ is what the investors expect to receive, the next period, by investing $\mathcal{L}P(r, T_i)$ today, in the MMA. The denominators constitute a measure of volatility related to holding the assets. Then, the previous equation tells us that the Sharpe ratios, or the unit risk premiums, on the two zeros agree.

Let the Sharpe ratio on any zero be equal to some function λ of the short-term rate r only (and possibly of calendar time). This function, λ , does not clearly depend on the maturity of the zeros. Then, we have,

$$\begin{aligned} [pP(r^+, T_1) + (1-p)P(r^-, T_1)] - (1+r)P(r, T_1) &= [P(r^+, T_1) - P(r^-, T_1)] \lambda \\ &= \frac{P(r^+, T_1) - P(r^-, T_1)}{r^+ - r^-} \cdot [(r^+ - r^-)\lambda]. \end{aligned} \quad (11A.3)$$

We can interpret $(r^+ - r^-)$ as a measure of interest rate volatility, and define $\text{Vol}(\tilde{r} - r) \equiv (r^+ - r^-)$. Eq. (11.18) follows by rewriting Eq. (11A.3) for a generic maturity date $T > 2$.

11.10 Appendix 2: The Ho and Lee price representation

Define the discretely compounded forward rate, as the number $F_T(\tau) \equiv F(\tau, T, T+1)$, satisfying: $\frac{P(\tau, T+1)}{P(\tau, T)} = \frac{1}{1+F_T(\tau)}$, as in Eq. (11.3) of the main text. Iterating this equation leaves:

$$P(\tau, T) = \prod_{S=\tau}^{T-1} \frac{1}{1+F_S(\tau)} = \frac{P(t, T)}{P(t, \tau)} \frac{P(t, \tau)}{P(t, T)} \prod_{S=\tau}^{T-1} \frac{1}{1+F_S(\tau)}.$$

Therefore, for any $t : t < \tau < T$, we have that,

$$P(\tau, T) = \frac{P(t, T)}{P(t, \tau)} \prod_{S=\tau}^{T-1} \frac{1+F_S(t)}{1+F_S(\tau)}. \quad (11A.4)$$

Eq. (11A.4) is a convenient representation of the bond price at a future date τ : it is the ratio of the two *current* prices $P(t, T)$ and $P(t, \tau)$, and a factor relating to the development of *forward rates* from the current time t to time τ , i.e. $\frac{1+F_S(t)}{1+F_S(\tau)}$, for $S = \tau, \dots, T-1$. Hence, once we model forward rates, we have implications for bond price movements, which we can use to price, at the evaluation time t , interest rate derivatives, with payoffs depending on the realization of the bond price $P(\tau, T)$ at time τ .

We normalize the time-line and set $t = 0$. Redefining $\tau = t$, Eq. (11A.4) reduces to,

$$P(t, T) = \frac{P(0, T)}{P(0, t)} \prod_{S=t}^{T-1} \frac{1+F_S(0)}{1+F_S(t)}. \quad (11A.5)$$

Eq. (11.35) in the main text follows by Eq. (11A.5).

Next, we search for the model's predictions about forward rates, i.e. we prove Eq. (11.36). The proof is by induction. Eq. (11.36) holds true for $t = 0$. Next, suppose that it holds at time t . We wish to show that in this case, Eq. (11.36) would also hold at time $t+1$. At time $t+1$, we have two cases.

Case 1: A positive price jump occurs between time t and time $t+1$. In this case,

$$\begin{aligned} \hat{F}_S^{j+1}(t+1) &= \ln \frac{P_{j+1}(t+1, S)}{P_{j+1}(t+1, S+1)} \\ &= \ln \left[u(S-t) \frac{P_j(t, S)}{P_j(t, t+1)} \right] - \ln \left[u(S+1-t) \frac{P_j(t, S+1)}{P_j(t, t+1)} \right] \\ &= \ln \frac{u(S-t)}{u(S+1-t)} + \hat{F}_S^j(t) \\ &= \ln \frac{u(S+1-(t+1))}{u(S+1)} + \hat{F}_S(0) - [(t+1) - (j+1)] \ln \delta, \end{aligned}$$

where the first equality and the third follow by the definition of $\hat{F}_S^{j+1}(t)$, the second equality holds by the definition of the jump in Eq. (11.28), the fourth equality follows by using Eq. (11.36). Hence, Eq. (11.36) holds at time $t+1$ in the occurrence of a *positive* price jump between time t and time $t+1$.

Case 2: A negative price jump occurs between time t and time $t + 1$. In this case,

$$\begin{aligned}
\hat{F}_S^j(t+1) &= \ln \frac{P_j(t+1, S)}{P_j(t+1, S+1)} \\
&= \ln \left[d(S-t) \frac{P_j(t, S)}{P_j(t, t+1)} \right] - \ln \left[d(S+1-t) \frac{P_j(t, S+1)}{P_j(t, t+1)} \right] \\
&= \ln \frac{d(S-t)}{d(S+1-t)} + \hat{F}_S^j(t) \\
&= \ln \frac{d(S-t) \delta^{-(S-t)+1}}{d(S+1-t) \delta^{-(S+1-t)+1}} \delta^{-1} + \hat{F}_S(0) + \ln \frac{u(S+1-t)}{u(S+1)} - (t-j) \ln \delta \\
&= \ln \frac{u(S-t)}{u(S+1)} + \hat{F}_S(0) - [(t+1) - j] \ln \delta,
\end{aligned}$$

where the first four equalities follow by the same arguments produced in Case 1, the fifth equality holds by the relation $u(T) = d(T) \delta^{-(T-1)}$ in Eq. (11.32) and the last equality follows by rearranging terms. Hence, Eq. (11.36) holds at time $t + 1$ in the occurrence of a *negative* price jump between time t and time $t + 1$.

These two cases reveal that if Eq. (11.36) holds at time t for any $j \leq t$, it also holds at time $t + 1$, in each state of nature. By induction, Eq. (11.36) is therefore true.

References

- Bernanke, B. S. and A. Blinder (1992): “The Federal Funds Rate and the Channels of Monetary Transmission.” *American Economic Review* 82, 901-921.
- Black, F. and M. Scholes (1973): “The Pricing of Options and Corporate Liabilities.” *Journal of Political Economy* 81, 637-659.
- Black, F., E. Derman and W. Toy (1990): “A One Factor Model of Interest Rates and its Application to Treasury Bond Options.” *Financial Analysts Journal* (January-February), 33-39.
- Cox, J. C., S. A. Ross and M. Rubinstein (1979): “Option Pricing: A Simplified Approach.” *Journal of Financial Economics* 7, 229-263.
- Diebold, F. X. and C. Li (2006): “Forecasting the Term Structure of Government Bond Yields.” *Journal of Econometrics* 130, 337-364.
- Heath, D., R. Jarrow and A. Morton (1992): “Bond Pricing and the Term-Structure of Interest Rates: a New Methodology for Contingent Claim Valuation.” *Econometrica* 60, 77-105.
- Ho, T. S. Y. and S.-B. Lee (1986): “Term Structure Movements and the Pricing of Interest Rate Contingent Claims.” *Journal of Finance* 41, 1011-1029.
- Ho, T. S. Y. and S.-B. Lee (2004): *The Oxford Guide to Financial Modeling*. Oxford University Press.
- Hull, J. C. (2003): *Options, Futures, and Other Derivatives*. Prentice Hall. 5th edition (International Edition).
- Hull, J. C. and A. White (1990): “Pricing Interest Rate Derivative Securities.” *Review of Financial Studies* 3, 573-592.
- McCulloch, J. (1971): “Measuring the Term Structure of Interest Rates.” *Journal of Business* 44, 19-31.
- McCulloch, J. (1975): “The Tax-Adjusted Yield Curve.” *Journal of Finance* 30, 811-830.
- Nelson, C.R. and A.F. Siegel (1987): “Parsimonious Modeling of Yield Curves.” *Journal of Business* 60, 473-489.
- Tuckman, B. (2002): *Fixed Income Securities*. Wiley Finance.
- Vasicek, O. (1977): “An Equilibrium Characterization of the Term Structure.” *Journal of Financial Economics* 5, 177-188.

12

Interest rates

12.1 Introduction

This chapter surveys models and empirical facts underlying the term structure of interest rates and derivatives based thereon, which largely rely on continuous time methods. Its innovation against Chapter 11 is to provide a systematic approach to the many facets relating to fixed income securities, from the stylized facts pertaining to the factors driving the yield curve, their business cycle components, predictability and volatility, to more conceptual aspects relating to how we would need to think about duration in a random environment, or the pricing details of interest rate derivatives such as bond options, puttable and callable bonds, swaps, caps, floors, swaptions, to mention a few.

We know from previous chapters that to price derivatives, we need to make sure that the price of the underlying assets is pinned down without errors. When it comes to interest rate derivatives, this task is challenging, because the yield curve relies on risks that are typically not traded. Consider, for example, a model where the price of a zero coupon bond is only driven by random movements of the short-term rate—a one-factor model. Let $P(r, \tau, T)$ be the time τ price of a zero coupon bond expiring at time T , when the short-term rate is equal to r . The exact functional form of the pricing function $P(r, \tau, T)$ depends on (i) the assumptions we make on the dynamics of the short-term rate, and (ii) the assumptions we make on risk-aversion corrections. Models of this kind, and generalizations to multi-state variables, are known as “models of the short-term rate,” and are discussed in Section 12.4. These models are very important because once they are made complex enough to cope with the many facts that we see in the data, they might perform a series of tasks. For example, they can provide us with interpretations of the empirical facts; furthermore, they can be used to forecast developments in fixed income markets. Finally, they might be used for trading purposes should they reasonably point to market inefficiencies. However, these models lead to pricing errors—it is actually the presence of these errors to justify a potential use of these models for trading purposes.

A second class of models that does not lead to pricing errors is that developed by Heath, Jarrow and Morton (1992), and generalizes the Ho and Lee (1986) model examined in Chapter 11. These “no-arbitrage models,” of which we have seen instances in the previous chapter, analyzed through trees, are given a systematic treatment in continuous time, in Section 12.5. A

principle underlying these models is that bond prices need not to be modeled in the first place. Rather, current bond prices are taken as primitives, with the modeling focus being forward rates, i.e. interest rates prevailing today for borrowing in the future. There is a relation linking bond prices to forward rates. No arbitrage then restricts the joint behavior of future bond prices and forward rates. [...] In the next section, we provide definitions of interest rates and markets. Section 12.2.2 develops the two basic representations of bond prices: one in terms of the short-term rate and the other, in terms of forward rates. Section 12.2.3 develops the foundations of the so-called forward martingale probability, which is a probability measure under which forward interest rates are martingales. [...] Except for Section 12.4.6, we assume no default risk. Default risk is, instead, more systematically dealt with in the next chapter. [In progress]

12.2 Prices and interest rates

12.2.1 Bond prices

12.2.1.1 A first representation of bond prices

Consider the relation linking bond prices, P , to discretely compounded interest rates, L for the time interval $[\tau, T]$, introduced in Section 11.2.2.1 of the previous chapter:

$$P(\tau, T) = \frac{1}{1 + (T - \tau)L(\tau, T)}. \quad (12.1)$$

Given $L(\tau, T)$, the *short-term rate* process r is obtained as:

$$r(\tau) \equiv \lim_{T \downarrow \tau} L(\tau, T).$$

Next, let Q be a risk-neutral probability, and $\mathbb{E}_\tau[\cdot]$ denote the time τ conditional expectation under Q . By the FTAP, there are no arbitrage opportunities if and only if $P(\tau, T)$ satisfies, for all $\tau \in [t, T]$,

$$P(\tau, T) = \mathbb{E}_\tau \left[e^{-\int_\tau^T r(\ell) d\ell} \right], \quad (12.2)$$

Appendix 1 provides the proof of the if-part—there is no arbitrage if bond prices are as in Eq. (12.2). This proof is quite standard, in fact similar to those encountered in the first part of these Lectures. It is provided here, as it is capable of highlighting specific issues relating to interest rate modeling.

12.2.1.2 Forward rates, and a second representation of bond prices

Forward rates are interest rates that make the value of a *forward rate agreement* (FRA, henceforth) equal to zero at origination. Section 11.2.2.3 of the previous chapter provides the definition of a forward rate agreement, although the very same definition is restated below, for reasons clarified in a moment. Forward rates as of time t , for a forward rate agreement relating to a future time-interval $[T, S]$, are denoted with $F(t, T, S)$, and link to bond prices through a precise relation, derived in Section 11.2.2.3 of the previous chapter:

$$\frac{P(t, T)}{P(t, S)} = 1 + (S - T) F(t, T, S). \quad (12.3)$$

Clearly, the forward rate agreed at T for the time interval $[T, S]$ is the short-term rate applying to the same period:

$$F(T, T, S) = L(T, S). \quad (12.4)$$

Consider, next, a more general FRA, where a first counterparty agrees: (i) to pay an interest rate on a given principal at time T , fixed at some $K \neq F(t, T, S)$, and (ii) to receive, in exchange, the future interest rate prevailing at time T for the time interval $[T, S]$, $L(T, S)$, from a second counterparty. The profit at T , arising from this “interest rate swap” is:

$$(S - T) [L(T, S) - K]. \quad (12.5)$$

It is the same as the profit to a party who is long a FRA, who therefore enters the FRA, at time t , for the time-interval $[T, S]$, as a future borrower. Come time T , the party shall honour the FRA by borrowing £1 for the time-interval $[T, S]$ at a cost of K . At the same time, the party can lend this very same £1 at the random interest rate $L(T, S)$. The time S payoff deriving from this trade is, of course, the same as that in Eq. (12.5).

The value of the FRA, which we denote as $\text{IRS}(t, T, S; K)$, is the current market value of this future, random payoff. By the FTAP,

$$\begin{aligned} \text{IRS}(t, T, S; K) &= \mathbb{E}_t \left[e^{-\int_t^S r(\tau) d\tau} (S - T) [L(T, S) - K] \right] \\ &= \mathbb{E}_t \left[e^{-\int_t^S r(\tau) d\tau} (S - T) L(T, S) \right] - (S - T) P(t, S) K \\ &= \mathbb{E}_t \left[\frac{e^{-\int_t^S r(\tau) d\tau}}{P(T, S)} \right] - [1 + (S - T) K] P(t, S) \\ &= P(t, T) - [1 + (S - T) K] P(t, S), \end{aligned} \quad (12.6)$$

where the third line holds by the definition of L and the third line follows by the following relation:

$$P(t, T) = \mathbb{E}_t \left[\frac{e^{-\int_t^S r(\tau) d\tau}}{P(T, S)} \right]. \quad (12.7)$$

The economic interpretation of Eq. (12.7) is simple. Suppose that at time t , $\mathcal{L}P(t, T)$ are invested in a bond maturing at time T . At time T , this investment will obviously pay off £1. And at time T , £1 can be further rolled over another bond maturing at time S , thus yielding $\mathcal{L} 1/P(T, S)$ at time S . In other words, an investment at t equal to $\mathcal{L}P(t, T)$, leads to a “payoff” at S equal to $\mathcal{L} 1/P(T, S)$, whence Eq. (12.6).¹ Alternatively, note that the LIBOR, $L(T, S)$, although known at time T , is only paid off at time S , such that the value of $1 + (S - T) L(T, S)$, to be delivered at time S , is simply one at time T and, obviously, $P(t, T)$ at time t . That is, the value of $(S - T) L(T, S)$, to be delivered at time S , is simply $P(t, T) - P(t, S)$ at time t , whence the fourth equality.²

Finally, by replacing Eq. (12.3) into Eq. (12.6),

$$\text{IRS}(t, T, S; K) = (S - T) [F(t, T, S) - K] P(t, S). \quad (12.8)$$

¹Mathematically, we have, by the Law of Iterated Expectations, that

$$\mathbb{E}_t \left[\frac{e^{-\int_t^S r(\tau) d\tau}}{P(T, S)} \right] = \mathbb{E}_t \left[\mathbb{E} \left(\frac{e^{-\int_t^T r(\tau) d\tau} e^{-\int_T^S r(\tau) d\tau}}{P(T, S)} \middle| \mathcal{F}(T) \right) \right] = P(t, T).$$

²We are assuming, as it is standard, that settlements occur at S . When settlements occur at $T < S$, the value of $1 + (S - T) L(T, S)$ at T is obviously higher than one, and the calculations underlying Eq. (12.6) would not go through anymore, the technical issue being a “convexity effect” by which a payoff of $1 + (S - T) L(T, S)$ at time T is equivalent to a payoff of $(1 + (S - T) L(T, S))^2$ at time S . Brigo and Mercurio (2006, Chapter 13) and Veronesi (2010, Chapter 21) explain the standard market practice to deal with this issue.

As is clear, IRS can take on any sign, and is exactly zero when $K = F(t, T, S)$, where $F(t, T, S)$ solves Eq. (12.3). The notation, $\text{IRS}(\cdot)$, is used to emphasize that we are dealing with interest rate swaps, although more rigorously, interest rate swaps are those where payment exchanges will occur, repeatedly, over a given time horizon—the tenor of the swap, as explained in Section 12.6.7.

A useful remark. Comparing the second line in Eq. (12.6) with Eq. (12.8) reveals that:

$$F(t, T, S) = \mathbb{E}_t \left[\frac{e^{-\int_t^S r(\tau) d\tau}}{P(t, S)} L(T, S) \right].$$

That is, forward rates are not unbiased expectations of future interest rates, not even under the risk-neutral probability. We shall return to this point in Section 12.2.3.2.

Bond prices can be expressed in terms of these forward interest rates, namely in terms of the “instantaneous” forward rates. First, rearrange terms in Eq. (12.3) so as to obtain:

$$F(t, T, S) = -\frac{P(t, S) - P(t, T)}{(S - T)P(t, S)}.$$

The *instantaneous forward rate* $f(t, T)$ is defined as

$$f(t, T) \equiv \lim_{S \downarrow T} F(t, T, S) = -\frac{\partial \ln P(t, T)}{\partial T}. \quad (12.9)$$

It can be interpreted as the marginal rate of return from committing a bond investment for an additional instant. To express bond prices in terms of f , integrate Eq. (12.9), $f(t, \ell) = -\frac{\partial \ln P(t, \ell)}{\partial \ell}$, with respect to the maturity date ℓ , use the condition that $P(t, t) = 1$, and obtain:

$$P(t, T) = e^{-\int_t^T f(t, \ell) d\ell}. \quad (12.10)$$

12.2.1.3 The marginal nature of forward rates

Consider the *yield-to-maturity* introduced in Section 11.2.2.2 of the previous chapter, defined to be the function $R(t, T)$ such that:

$$P(t, T) \equiv e^{-(T-t) \cdot R(t, T)}. \quad (12.11)$$

Comparing Eq. (12.11) with Eq. (12.3) yields:

$$R(t, T) = \frac{1}{T-t} \int_t^T f(t, \tau) d\tau. \quad (12.12)$$

By differentiating Eq. (12.12) with respect to T yields:

$$\frac{\partial R(t, T)}{\partial T} = \frac{1}{T-t} [f(t, T) - R(t, T)].$$

This relation underscores the “marginal nature” of forward rates: the yield-curve, $R(t, T)$, is increasing in, decreasing in, or stationary at T , according to whether $f(t, T)$ exceeds, is lower, or equal the spot rate for maturity T .

12.2.2 Forward martingale probabilities

12.2.2.1 Definition

Let $\varphi(t, T)$ be the T -forward price of a claim $S(T)$ at T . That is, $\varphi(t, T)$ is the price agreed at t , which will be paid at T for delivery of the claim at T . Nothing has to be paid at t . By the FTAP, there are no arbitrage opportunities if and only if:

$$0 = \mathbb{E}_t \left[e^{-\int_t^T r(u)du} \cdot (S(T) - \varphi(t, T)) \right].$$

But since $\varphi(t, T)$ is known at time t ,

$$\mathbb{E}_t \left[e^{-\int_t^T r(u)du} \cdot S(T) \right] = \varphi(t, T) \cdot \mathbb{E}_t \left[e^{-\int_t^T r(u)du} \right]. \quad (12.13)$$

For example, assume S is the price process of a traded asset. By the FTAP, $\mathbb{E}_t[e^{-\int_t^T r(u)du} S(T)] = S(t)$, such that Eq. (12.13) collapses to the well-known formula: $\varphi(t, T)P(t, T) = S(t)$. However, entering the forward contract originated at t , at a later date $\tau > t$, costs. To calculate the marking-to-market of the forward at time τ , note that the final payoff at time T is $S(T) - \varphi(t, T)$. Discounting this payoff at $\tau \in [t, T]$ delivers $P(\tau, T) \cdot [\varphi(\tau, T) - \varphi(t, T)]$.

Next, let us elaborate on Eq. (12.13). We can use the bond pricing equation (12.2), and rearrange terms in Eq. (12.13), to obtain:

$$\varphi(t, T) = \mathbb{E}_t \left[\frac{e^{-\int_t^T r(u)du}}{P(t, T)} \cdot S(T) \right] = \mathbb{E}_t [\eta_T(T) \cdot S(T)], \quad (12.14)$$

where

$$\eta_T(T) \equiv \frac{e^{-\int_t^T r(u)du}}{P(t, T)}.$$

Eq. (12.14) suggests that we can define a new probability Q_F^T , as follows,

$$\eta_T(T) = \frac{dQ_F^T}{dQ} \equiv \frac{e^{-\int_t^T r(u)du}}{\mathbb{E}_t \left[e^{-\int_t^T r(u)du} \right]}. \quad (12.15)$$

Naturally, $\mathbb{E}_t[\eta_T(T)] = 1$. Moreover, if the short-term rate process is deterministic, $\eta_T(T)$ equals one and Q and Q_F^T are the same.

In terms of this new probability Q_F^T , the forward price $\varphi(t, T)$ is:

$$\varphi(t, T) = \mathbb{E}_t [\eta_T(T) \cdot S(T)] = \int [\eta_T(T) \cdot S(T)] dQ = \int S(T) dQ_F^T = \mathbb{E}_{Q_F^T} [S(T)], \quad (12.16)$$

where $\mathbb{E}_{Q_F^T}[\cdot]$ denotes the time t conditional expectation taken under Q_F^T . For reasons that will be clear in a moment, Q_F^T is referred to as the *T -forward martingale probability*. The forward martingale probability is a useful tool, which helps price interest-rate derivatives, as we shall explain in Section 12.8. It was introduced by Geman (1989) and Jamshidian (1989), and further analyzed by Geman, El Karoui and Rochet (1995). The appendix provides additional details: Appendix 2 relates forward prices to their certainty equivalent, and Appendix 3 illustrates additional technicalities about the forward martingale probability.

12.2.2.2 Martingale properties

Forward prices

Clearly, $\varphi(T, T) = S(T)$. Therefore, Eq. (12.16) is, also,

$$\varphi(t, T) = \mathbb{E}_{Q_F^T} [\varphi(T, T)].$$

Forward rates, and the expectation theory

Forward rates display a similar property:

$$f(t, T) = \mathbb{E}_{Q_F^T} [r(T)] = \mathbb{E}_{Q_F^T} [f(T, T)]. \quad (12.17)$$

where the last equality holds as $r(t) = f(t, t)$. The proof is also simple. We have,

$$\begin{aligned} f(t, T) &= -\frac{\partial \ln P(t, T)}{\partial T} \\ &= -\frac{\partial P(t, T)}{\partial T} \bigg/ P(t, T) \\ &= \mathbb{E}_t \left[\frac{e^{-\int_t^T r(\tau) d\tau}}{P(t, T)} \cdot r(T) \right] \\ &= \mathbb{E}_t [\eta_T(T) \cdot r(T)] \\ &= \mathbb{E}_{Q_F^T} [r(T)]. \end{aligned}$$

Finally, the simply-compounded forward rate satisfies the same property: given a sequence of dates $\{T_i\}_{i=0,1,\dots}$,

$$F(\tau, T_i, T_{i+1}) = \mathbb{E}_{Q_F^{T_{i+1}}} [L(T_i, T_{i+1})] = \mathbb{E}_{Q_F^{T_{i+1}}} [F(T_i, T_{i+1})], \quad \tau \in [t, T_i], \quad (12.18)$$

where the second equality follows by Eq. (12.4). To show Eq. (12.18), note that by definition, the simply-compounded forward rate $F(t, T, S)$ satisfies:

$$\text{IRS}(t, T, S; F(t, T, S)) = 0,$$

where $\text{IRS}(t, T, S; K)$ is the value as of time t of a FRA struck at K for the time-interval $[T, S]$. By rearranging terms in the second equality of Eq. (12.6),

$$F(t, T, S)P(t, S) = \mathbb{E}_t \left[e^{-\int_t^S r(\tau) d\tau} L(T, S) \right].$$

By the definition of $\eta_S(S)$,

$$F(t, T, S) = \mathbb{E}_{Q_F^S} [L(T, S)].$$

A well-known hypothesis in empirical finance is that known as the *expectation theory*, which states that forward rates equal future expected short-term rates. The empirical evidence about the expectation theory is reviewed in Section 12.3.1, but the previous relation already points to a difficulty in this theory. It shows that it is under the forward martingale probability that

the expectation theory holds true. A similar result holds for the instantaneous forward rate. Consider Eq. (12.17). We have,

$$\begin{aligned}
 f(t, T) &= \mathbb{E}_{Q_T^F}(r(T)) \\
 &= \mathbb{E}_Q(\eta_T(T) r(T)) \\
 &= \underbrace{\mathbb{E}(\eta_T(T))}_{=1} \mathbb{E}(r(T)) + cov_Q(\eta_T(T), r(T)) \\
 &= E_t(r(T)) + cov_t(\text{Ker}(T), r(T)) + cov_{Q,t}(\eta_T(T), r(T)), \tag{12.19}
 \end{aligned}$$

where $\text{Ker}(T)$ denotes the pricing kernel in the economy. That is, forward rates in general deviate from future expected short-term rates because of risk-aversion corrections (the second term in the last equality) and because interest rates are stochastic (the third term in the last equality).

12.2.3 Stochastic duration

Cox, Ingersoll and Ross (1979) introduce the notion of stochastic duration, which generalizes that of modified duration discussed in Chapter 11. Suppose the bond price is a function of the short term rate only, $P(r, T - t)$. Duration is a measure of risk for fixed income instruments. Define the *basis risk* as the semi-elasticity of the bond price with respect to the short-term rate,

$$\Psi(r, T - t) \equiv -\frac{P_r(r, T - t)}{P(r, T - t)},$$

where the subscript r denotes a partial derivative. Naturally, we want to make sure that the measure of duration for a zero coupon bond equals time-to-maturity, such that Ψ cannot represent a measure of duration, since in general, it does not equal $T - t$, except in the trivial case the short-term rate, r , is constant.

The idea underlying “stochastic duration” is to search for an hypothetical zero coupon bond with basis risk equal to the basis risk of a coupon bearing bond, or in general, any other bond, say for instance a callable bond, as follows:

$$\Psi(r, T^* - t) = -\frac{B_r(r, S - t)}{B(r, S - t)},$$

where $B(r, S - t)$ is the price of any bond at time t , possibly different from a simple zero coupon bond, which delivers the face value at time S , if no events preventing this occur prior to time S , such as the exercise of the callability provision, or even default. Stochastic duration is defined as the time-to-maturity $T^* - t$ of the zero coupon bond:

$$T^* - t \equiv D(r, S - t) = \Psi^{-1}\left(-\frac{B_r(r, T - t)}{B(r, T - t)}\right), \tag{12.20}$$

where Ψ^{-1} is the inverse function of $\Psi(r, \tau)$ with respect to time to maturity τ . Naturally, $D(r, T - t) = T - t$, for a pure discount bond. Moreover, the stochastic duration, $D(r, T - t)$, collapses to the modified duration introduced in the previous chapter, Section 11.4, once the short-term rate is a constant. The reason is that if r is constant, then, $P(r, T - t) = e^{-r(T-t)}$, and $\Psi^{-1}(r, x) = x$.

12.3 Stylized facts

12.3.1 The expectation hypothesis, and bond returns predictability

The expectation theory holds that *forward rates equal expected future short-term rates*, or

$$f(t, T) = E_t(r(T)),$$

where $E_t(\cdot)$ denotes expectation under the physical probability. By Eq. (12.12), then, the expectation theory implies that,

$$R(t, T) = \frac{1}{T-t} \int_t^T E_t(r(\tau)) d\tau. \quad (12.21)$$

A natural question arises as to whether the forward rate for maturity T , $f(t, T)$, is higher than the short-term rate expected to prevail at time T , $E_t(r(T))$. It is quite an old issue. One possibility might be that in the presence of risk-averse investors,

$$f(t, T) \geq E_t(r(T)). \quad (12.22)$$

By Jensen's inequality,

$$e^{-\int_t^T f(t, \tau) d\tau} \equiv P(t, T) = \mathbb{E}_t \left[e^{-\int_t^T r(\tau) d\tau} \right] \geq e^{-\int_t^T \mathbb{E}_t[r(\tau)] d\tau} \implies \int_t^T \mathbb{E}_t[r(\tau)] d\tau \geq \int_t^T f(t, \tau) d\tau.$$

Therefore, in a risk-neutral market, the inequality in (12.22) cannot hold. The inequality in (12.22) relates to the Hicks-Keynesian *normal backwardation hypothesis*.³ According to the explanation of Hicks, firms demand long-term funds but fund suppliers prefer to lend at shorter maturity dates. The market is cleared by intermediaries, who require a *liquidity premium* to be compensated for their risky activity of borrowing at short and lending at long maturities.

A final definition. The *term-premium* is defined as the difference between the spot rate and the future expected average short-term rate, for the same horizon, as follows:

$$\text{TP}(t, T) \equiv R(t, T) - \frac{1}{T-t} \int_t^T E_t(r(\tau)) d\tau = \frac{1}{T-t} \int_t^T [f(t, \tau) - E_t(r(\tau))] d\tau,$$

where the second equality follows by Eq. (12.12).

What does the empirical evidence suggest about the expectation hypothesis? Denote the continuously compounded returns on a zero expiring at some date T as $r_{t+1}^T = \ln \frac{P(t+1, T)}{P(t, T)}$. Using the definition of spot rates, $R(t, T)$, the excess returns, \hat{r}_{t+1}^T say, can be expressed as:

$$\begin{aligned} \hat{r}_{t+1}^T &\equiv \ln \frac{P(t+1, T)}{P(t, T)} - \ln \frac{1}{P(t, t+1)} \\ &= \ln \frac{P(t+1, T)}{P(t, T)} - R(t, t+1) \\ &= -(T-t-1)R(t+1, T) + (T-t)R(t, T) - R(t, t+1), \end{aligned}$$

³Note, the *normal backwardation (contango) hypothesis* states that forward prices are lower (higher) than future expected spot prices. In the case of the inequality in (12.22), the normal backwardation hypothesis is stated in terms of interest rates.

such that the expected change in the yield curve relates negatively to the expected excess returns and positively to the slope of the yield curve:

$$E_t [R(t+1, T) - R(t, T)] = -\frac{1}{T-t-1} E_t (\hat{r}_{t+1}^T) + \frac{1}{T-t-1} [R(t, T) - R(t, t+1)].$$

The expectation hypothesis implies that the risk-premium, $E_t (\hat{r}_{t+1}^T)$, is, roughly, constant. Indeed, we have:

$$\begin{aligned} E_t (r(t+1)) &= f(t, t+1) \\ &= E_t (r(t+1)) + \text{cov}_t (\text{Ker}(t+1), r(t+1)) + \text{cov}_{Q,t} (\eta_T(t+1), r(t+1)), \end{aligned}$$

where the first equality holds by the expectation hypothesis, and the second equality is Eq. (12.19). Therefore, the sum of the last two terms in the last equality is zero, implying that $E_t (\hat{r}_{t+1}^T)$ is, roughly, constant. Veronesi (2010, Chapter 7) builds up an example where these relations hold exactly, within an affine model; in Section 12.4.3, we illustrate how these relations work, analytically, by hinging upon a simple and famous model—that of Vasicek (1977).

Empirically, we can test for the expectation theory, by running the following regression:

$$R(t+1, T) - R(t, T) = \alpha_T + \beta_T \frac{1}{T-t-1} [R(t, T) - R(t, t+1)] + \text{Residual}_t,$$

and test for the null of $\alpha_T = 0$ and $\beta_T = 1$. A widely known empirical feature of US data is that the estimates of β_T are typically negative for all maturities T , and somewhat increasing with T in absolute value. In fact, Fama and Bliss (1987) show that the risk-premium $E_t (\hat{r}_{t+1}^T)$ relates to the forward spreads, defined as $f_t^T - R(t, t+1)$, in that regressing

$$\hat{r}_{t+1}^T = \alpha_T + \beta_T (f_t^T - R(t, t+1)) + \text{Residual}_t,$$

delivers statistically significant and positive values of β_T for many maturities T .

Cochrane and Piazzesi (2005) go one step further and consider the following regressions:

$$\hat{r}_{t+1}^T = \alpha_T + \beta_{1T} R(t, t+1) + \sum_{j=2}^5 \beta_{j,T} f_t^j + \text{Residual}_t,$$

where f_t^T is the forward rate for maturity $T-1$, $f_t^T = -\ln \frac{P(t, T)}{P(t, T-1)}$. They document a “tent shape” for the estimates of the coefficients $(\beta_{j,T})_{j=1}^5$, for bond maturities $T \in \{1, \dots, 5\}$, and where t is in years so as to make returns calculated on a yearly basis. They document that this tent shape is robust to estimating a factor model, in that the shape persists in the estimates of the coefficients $(b_j)_{j=1}^5$ in:

$$\hat{r}_{t+1}^T = \alpha_T + \beta_{1T} Z_t + \text{Residual}_t, \quad Z_t = b_1 R(t, t+1) + \sum_{j=2}^5 b_j f_t^j,$$

where Z_t is the common factor among the bond maturities $T \in \{1, \dots, 5\}$. Moreover, they argue that using the traditional factors known to explain movements in the yield curve (see Section 12.3.4) does not destroy the predicting power of their factors, in sample.

12.3.2 The yield curve and the business cycle

There is a simple prediction about the shape the yield-curve that we can make. By Jensen's inequality, $e^{-(T-t)R(t,T)} \equiv P(t,T) = \mathbb{E}_t[e^{-\int_t^T r(\tau)d\tau}] \geq e^{-\int_t^T \mathbb{E}_t(r(\tau))d\tau}$. Therefore, the yield curve satisfies: $R(t,T) \leq \frac{1}{T-t} \int_t^T \mathbb{E}_t(r(\tau)) d\tau$. For example, suppose that the short-term rate is a martingale under the risk-neutral probability, viz $\mathbb{E}_t(r(\tau)) = r(t)$. Then, the yield curve is bound to be: $R(t,T) \leq r(t)$. That is, the yield curve is not increasing in time-to-maturity, T , at least for small maturities. Positively sloped yield curve, then, likely arise because the short-term rate is *not* a martingale under the risk-neutral probability, which happens because of two fundamental, and not necessarily mutually exclusive, reasons: (i) interest rates are expected to increase, (ii) investors are risk-averse. On average, the US yield curve is upward sloping at maturity from one up to ten years.

There exists strong empirical evidence since at least Kessel (1965) or, later, Laurent (1988, 1989), Stock and Watson (1989), Estrella and Hardouvelis (1991) and Harvey (1991, 1993), that inverted yield curves predict recessions with a lead time of about one to two years. Figure 12.1 illustrates these empirical facts through a plot of the the difference between long-term and short-term yields on Treasuries—in short, the “term spread.”

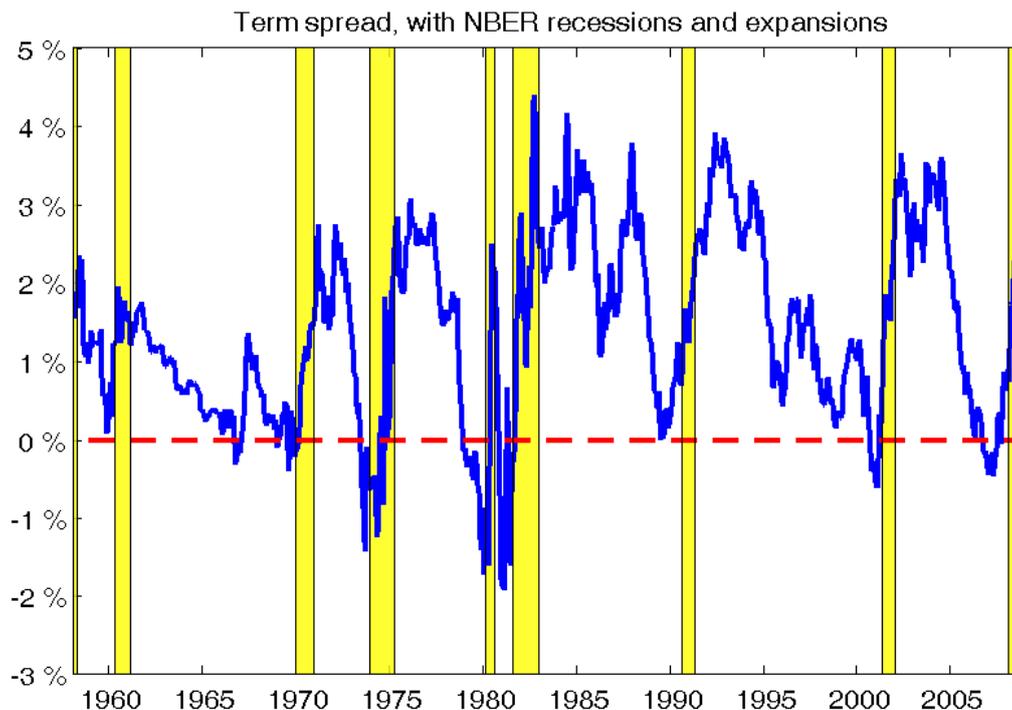


FIGURE 12.1. This picture depicts the time series of the term spread, defined as the difference between the 10 year yield minus the 3 month yield on US Treasuries. Sample data cover the period from January 1957 to December 2008. The shaded areas mark recession periods, as defined by the National Bureau of Economic Research. The end of the last recession was announced to have occurred on June 2009.

Naturally, there are recession episodes preceded by mild yield curve inversions. But the really striking empirical regularity is the sharp movements of the term spread towards a negative

territory, occurring prior to any recession episode. Note, it is not really important that the short-term rate goes up and the long-term segment of the yield curve goes down. The crucial empirical regularity, simply, relates to the term spread going down and eventually becoming negative prior to a recession. The explanations for these statistical facts are challenging, and might hinge upon both (i) the conduct of monetary policy and the expectations about it, and (ii) the risk-premiums agents require to invest in long-term bonds. We discuss these two points below.

(i) *The monetary channel:*

- (i.1) During expansions, monetary policy tends to be restrictive, to prevent the economy from heating up. At the height of an expansion, then, short-term yields go up.
- (i.2) Moreover, during recessions, monetary policy tends to keep interest rates low. At the height of an expansion, agents might be anticipating an incoming recession and, then, expecting central banks to lower future interest rates. Therefore, at the height of an expansion, future interest rates might be expected to lower. The expectation hypothesis in Eq. (12.21) would then predict the slope of the yield curve to decrease. Note that in the previous subsection, we have just learnt that the expectation hypothesis does not hold, empirically. Bond markets command risk-premiums. However, a risk-premium channel would reinforce the conclusion that the slope of the yield curve decreases during expansions, as argued in the next point.

(ii) *The risk-premium channel:* From Chapter 7, we know that risk-premiums are countercyclical, being high during recessions and low during expansions. The conditional equity premium is countercyclical, and so is the long-bond premium.⁴ In fact, long-term yields and equity expected returns are likely to be driven by the same state variables affecting the pricing kernel of the economy.⁵

Let's summarize. On the one hand, countercyclical monetary policy might be responsible of the negative price pressure on short-term bonds. On the other hand, expectations about countercyclical monetary policy as well as procyclical risk-appetite might be responsible for a positive price pressure on long-term bonds. These price pressures, we have argued, should occur at the height of an expansion. But the sample data we have are those where expansions are followed by recessions. Whence, the statistical facts about the predictive content of the yield curve, as we further formalize in Section 12.4.3 through a simple model.

Are these explanations plausible? It is interesting to note that these inversions did also use to occur prior to the creation of the Federal Reserve system. The creation of the US Central Bank might constitute a “Natural Experiment” to perform statistical inference about the importance of the gaming between central banks and the market expectations about the future conduct of monetary policy. Moreover, the inversion of the yield curve, which started to occur at around the beginning of 2006, might be attributable to a strong demand for long-term bonds, as warned

⁴An objection to this line of reasoning is that countercyclical risk-premiums might lead to expect future bond prices to decrease over a future recession, thereby destroying the effects of a procyclical short-term rate. In Section 12.3.3.2, we develop a model where these effects do not arise as soon as the effects of countercyclical risk-premiums are assumed to be bounded.

⁵That long term bonds and stock market are acknowledged to be tightly related is witnessed by a quite raw rule of thumb, whereby a stock market correction, such as a crash say, is deemed to be imminent when the spread 30 year bond yield minus the earning-price ratio is larger than 3%. This spread, which is usually around 1% or 2%—and on average, zero, once corrected for inflation—was indeed larger than 3% in 1987 and in 1997.

by some policy-makers at the time (see, e.g., the European Central Bank Monthly Bulletin, February 2006, p. 27). It is clearly challenging to quantify the extent of this demand pressure, arising, perhaps, from institutional investors such as Pension Funds whilst performing asset-liability management duties. It is undeniable, though, that the Federal Reserve at the time would target higher and higher interest rates, to cope with inflation concerns generated by a previous loose policy following the 2001 recession, Twin Towers attacks and maybe also the Corporate scandals in 2003. It is an open question as to whether the markets thought that this increased tightening was, maybe, marking the end of an expansion, thereby feeding an expectation future interest rates would drop again in the near future. Equally undeniably, the sharp tightening of the FED policy at the time would carry implications about financial developments such as the 2007 subprime crisis and, then, economic developments, as explained in the next chapter.

12.3.3 Additional stylized facts about the US yield curve

There are three additional features of data, which need to be noted.

- (i) Yields are highly correlated (say three year yields with four year yields, with five year yields, etc.), and suggest the existence of common factors driving all of them, discussed in Section 12.3.4 below.
- (ii) Yields are also highly persistent, and this persistence bears important consequences on derivative pricing, as explained in Section 12.8.4.
- (iii) The term-structure of unconditional volatility is downward sloping, a feature Section 12.8.4 attempts to rationalize.

12.3.4 Common factors affecting the yield curve

Which systematic risks affect the entire term-structure of interest rates? How many factors are needed to explain the variation of the yield curve? The standard “duration hedging” practice, reviewed in detail in Chapter 11, relies on the idea that most of the variation of the yield curve is successfully captured by a single factor that produces parallel shifts in the yield curve. How reliable is this idea, in practice?

Litterman and Scheinkman (1991) demonstrate that most of the variation (more than 95%) of the term-structure of interest rates can be attributed to the variation of three unobservable factors, which they label (i) a “level” factor, (ii) a “steepness” (or “slope”) factor, and (iii) a “curvature” factor. To disentangle these three factors, the authors make an unconditional analysis based on a *fixed-factor* model. Succinctly, this methodology can be described as follows.

Suppose that p returns computed from bond prices at p different maturities are generated by a linear factor structure, with a fixed number k of factors,

$$R_t = \bar{R} + B F_t + \epsilon_t, \quad (12.23)$$

$p \times 1$ $p \times 1$ $p \times k$ $k \times 1$ $p \times 1$

where R_t is the vector of returns, F_t is the zero-mean vector of common factors affecting the returns, assumed to be zero mean, \bar{R} is the vector of unconditional expected returns, ϵ_t is a vector of idiosyncratic components of the return generating process, and B is a matrix containing the factor loadings. Each row of B contains the factor loadings for all the common factors affecting a given return, i.e. the sensitivities of a given return with respect to a change of the factors. Each column of B contains the *term-structure of factor loadings*, i.e. how a change of a given factor affects the term-structure of excess returns.

12.3.4.1 Methodological details

Estimating the model in Eq. (12.23) leads to econometric challenges, mainly because the vector of factors F_t is unobservable.⁶ However, there exists a simple method, known as *principal components analysis* (PCA, henceforth), which leads to empirical results qualitatively similar to those holding for the general model in Eq. (12.23). We discuss these empirical results in the next subsection. We now describe the main methodological issues arising within PCA.

The main idea underlying PCA is to transform the original p correlated variables R into a set of new uncorrelated variables, the *principal components*. These principal components are linear combinations of the original variables, and are arranged in order of decreased importance: the first principal component accounts for as much as possible of the variation in the original data, etc. Mathematically, we are looking for p linear combinations of the demeaned excess returns,

$$Y_i = C_i^\top (R - \bar{R}), \quad i = 1, \dots, p, \quad (12.24)$$

such that, for p vectors C_i^\top of dimension $1 \times p$, (i) the new variables Y_i are uncorrelated, and (ii) their variances are arranged in decreasing order. The logic behind PCA is to ascertain whether a few components of $Y = [Y_1 \dots Y_p]^\top$ account for the bulk of variability of the original data. Let $C^\top = [C_1^\top \dots C_p^\top]$ be a $p \times p$ matrix such that we can write Eq. (12.24) in matrix format, $Y_t = C^\top (R_t - \bar{R})$ or, by inverting,

$$R_t - \bar{R} = C^{\top -1} Y_t. \quad (12.25)$$

Next, suppose that the vector $Y^{(k)} = [Y_1 \dots Y_k]^\top$ accounts for most of the variability in the original data,⁷ and let $C^{\top(k)}$ denote a $p \times k$ matrix extracted from the matrix $C^{\top -1}$ through the first k rows of $C^{\top -1}$. Since the components of $Y^{(k)}$ are uncorrelated and they are deemed largely responsible for the variability of the original data, it is natural to “disregard” the last $p - k$ components of Y in Eq. (12.25),

$$R_t - \bar{R} \underset{p \times 1}{\approx} \underset{p \times k}{C^{\top(k)}} \underset{k \times 1}{Y_t^{(k)}}.$$

If the vector $Y_t^{(k)}$ really accounts for most of the movements of R_t , the previous approximation to Eq. (12.25) should be fairly good.

Let us make more precise what the concept of variability is in the context of PCA. Suppose that the variance-covariance matrix of the returns, Σ , has p distinct eigenvalues, ordered from the highest to the lowest, as follows: $\lambda_1 > \dots > \lambda_p$. Then, the vector C_i in Eq. (12.24) is the eigenvector corresponding to the i -th eigenvalue. Moreover,

$$\text{var}(Y_i) = \lambda_i, \quad i = 1, \dots, p.$$

⁶Suppose that in Eq. (12.23), $F \sim N(0, I)$, and that $\epsilon \sim N(0, \Psi)$, where Ψ is diagonal. Then, $R \sim N(\bar{R}, \Sigma)$, where $\Sigma = BB^\top + \Psi$. The assumptions that $F \sim N(0, I)$ and that Ψ is diagonal are necessary to identify the model, but not sufficient. Indeed, any orthogonal rotation of the factors yields a new set of factors which also satisfies Eq. (12.23). Precisely, let T be an orthonormal matrix. Then, $(BT)(BT)^\top = BTT^\top B^\top = BB^\top$. Hence, the factor loadings B and BT have the same ability to generate the matrix Σ . To obtain a unique solution, one needs to impose extra constraints on B . For example, Jöreskog (1967) develop a maximum likelihood approach in which the log-likelihood function is, $-\frac{1}{2}N [\ln|\Sigma| + \text{Tr}(S\Sigma^{-1})]$, where S is the sample covariance matrix of R , and the constraint is that $B^\top \Psi B$ be diagonal *with elements arranged in descending order*. The algorithm is: (i) for a given Ψ , maximize the log-likelihood with respect to B , under the constraint that $B^\top \Psi B$ be diagonal with elements arranged in descending order, thereby obtaining \hat{B} ; (ii) given \hat{B} , maximize the log-likelihood with respect to Ψ , thereby obtaining $\hat{\Psi}$, which is fed back into step (i), etc. Knez, Litterman and Scheinkman (1994) describe this approach in their paper. Note that the identification device they describe at p. 1869 (Step 3) roughly corresponds to the requirement that $B^\top \Psi B$ be diagonal *with elements arranged in descending order*. Such a constraint is clearly related to principal component analysis.

⁷There are no rigorous criteria to say what “most of the variability” means in this context. Instead, a likelihood-ratio test is most informative in the context of the estimation of Eq. (12.23) by means of the methods explained in the previous footnote.

Finally, we have that

$$R_{\text{PCA}} = \frac{\sum_{i=1}^k \text{var}(Y_i)}{\sum_{i=1}^p \text{var}(R_i)} = \frac{\sum_{i=1}^k \lambda_i}{\sum_{i=1}^p \lambda_i}. \quad (12.26)$$

(Appendix 4 provides technical details and proofs of the previous formulae.) It is in the sense of Eq. (12.26) that in the context of PCA, we say that the first k principal components account for $R_{\text{PCA}}\%$ of the total variation of the data.

12.3.4.2 The empirical facts

The striking feature of the empirical results uncovered by Litterman and Scheinkman (1991) is that they have been confirmed to hold across a number of countries and sample periods. Moreover, the economic nature of these results is the same, independently of whether the statistical analysis relies on a rigorous factor analysis of the model in Eq. (12.23), or a more back-of-envelope computation based on PCA. Finally, the empirical results that hold for bond returns are qualitatively similar to those that hold for bond yields.

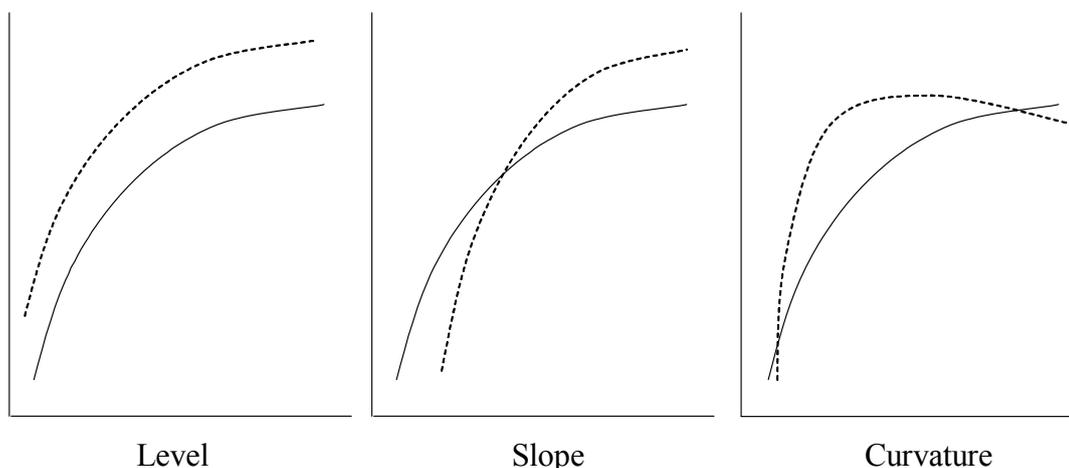


FIGURE 12.2. Changes in the term-structure of interest rates generated by changes in the “level,” “slope” and “curvature” factors.

Figure 12.2 visualizes the effects that the three factors have on the movements of the term-structure of interest rates.

- The first factor is called a “level” factor as its changes lead to parallel shifts in the term-structure of interest rates. Thus, this “level” factor produces essentially the same effects on the term-structure as those underlying the “duration hedging” portfolio practice. This factor explains approximately 80% of the total variation of the yield curve.
- The second factor is called a “steepness” factor as its variations induce changes in the slope of the term-structure of interest rates. After a shock in this steepness factor, the short-end and the long-end of the yield curve move in opposite directions. The movements of this factor explain approximately 15% of the total variation of the yield curve.
- The third factor is called a “curvature” factor as its changes lead to changes in the curvature of the yield curve. That is, following a shock in the curvature factor, the middle

of the yield curve and both the short-end and the long-end of the yield curve move in opposite directions. This curvature factor accounts for approximately 5% of the total variation of the yield curve.

Understanding the origins of these three factors is still a challenge to financial economists and macroeconomists. For example, macroeconomists explain that central banks affect the short-end of the yield curve, e.g. by inducing variations in Federal Funds rate in the US. However, the Federal Reserve decisions rest on the current macroeconomic conditions. Therefore, we should expect that the short-end of the yield-curve is related to the development of macroeconomic factors. Instead, the development of the long-end of the yield curve should largely depend on the market average expectation and risk-aversion surrounding future interest rates and economic conditions. Financial economists, then, should expect to see the long-end of the yield curve as being driven by expectations of future economic activity, and by risk-aversion. Indeed, Ang and Piazzesi (2003) demonstrate that macroeconomic factors such as inflation and real economic activity are able to explain movements at the short-end and the middle of the yield curve. Interestingly, they show that the long-end of the yield curve is driven by unobservable factors. However, it is not clear whether such unobservable factors are driven by time-varying risk-aversion or changing expectations. The compelling lesson, in general, is that models of the yield curve driven by only one factor are likely to be misspecified, due to the complexity of roles played by many institutions participating in the fixed income markets, and the links with the macroeconomy that decisions taken by these institutions have.

12.4 Models of the short-term rate

The short-term rate is simply the growth rate, or velocity, at which “locally” riskless investments appreciate, over the next instant. Naturally, this velocity is not a traded asset. Models where bond prices are tied up to interest rates are likely to be incomplete, in that to hedge against a bond, we cannot rely on anything underlying the bond price movements—what is traded is a money market account as well as the bond itself, not the interest rate. The evaluation framework in this contexts is one where bond prices can be replicated through other bond prices, as we explained in the discrete time setting of Chapter 11. It is the same issue we encountered in Chapter 10, where to price options in environments with random volatility, we needed to replicate options through other options.

12.4.1 Models versus representations

The fundamental relation in Eq. (12.2),

$$P(t, T) = \mathbb{E}_t \left[e^{-\int_t^T r(u) du} \right], \quad (12.27)$$

suggests to model the arbitrage-free price P , by assuming the short-term rate, r , is an exogenously given process. For example, we can rely on a Brownian information structure, and assume r be the solution to a stochastic differential equation such as:

$$dr(\tau) = b(r(\tau), \tau) d\tau + a(r(\tau), \tau) dW(\tau), \quad \tau \in (t, T], \quad (12.28)$$

for two functions b and a satisfying all the regularity conditions we need.

This approach was the first to emerge, after the seminal contributions of Merton (1973) (in a footnote!), Vasicek (1977) and Cox, Ingersoll and Ross (1985). This section illustrates the main modeling issues as well as the empirical challenges related to this approach. We examine one-factor “models of the short-term rate,” such as that in Eqs. (12.27)-(12.28), and also multifactor models, where the short-term rate is a function of a number of factors, $r(\tau) = R(y(\tau))$, where R is some function of a multidimensional diffusion process y .

Two fundamental issues to any model’s user are that the models he deals with be (i) fast to compute, and (ii) accurate. As for (i), we would look for models with closed form solutions, such as, for example, the so-called “affine” models (see Section 12.4.6). The second point is more subtle. A “perfect” accuracy cannot be achieved with models such as that in Eqs. (12.27)-(12.28), even when these models are given a multidimensional extension. After all, the model in Eqs. (12.27)-(12.28) is a *model of determination* of the observed yield curve. As such, it cannot *exactly fit* the observed term structure of interest rates. However, the model’s user might require an exact fit of the initial yield curve, whilst pricing interest rate derivatives. Interestingly, we can achieve a perfect fit can, once we augment Eq. (12.28) with an “infinite-dimensional” parameter, calibrated to the observed yield curve. However, this calibration might lead to intertemporal inconsistencies, explained in a moment.

Models leading to perfect accuracy are known as “no-arbitrage” models, and are the continuous-time counterparts of those relying on implied trees that we have dealt with in Chapter 11. In a sense, no-arbitrage models seem to be *representations*, rather than *models*, of bond prices. They work by forcing the short-term rate *process* to exactly pin down the yield curve we observe at a given instant. Intertemporal inconsistencies arise, because the parameters of the short-term rate that pin down the yield curve today, likely differ from those as of tomorrow. The philosophy underlying these models goes to the opposite extreme of the approach we describe in this section, where the short-term rate is the *input* of all subsequent movements of the term-structure of interest rates. As is clear, models of the short-term rate are very useful, when it comes to *explain* market behavior through a few inputs—a data reduction scientific principle. In these models, economically admissible, i.e. no-arbitrage, prices move as a result of random changes in the state variables, consistently with rational expectations. We wish to have models that make errors. We may end up trusting our models so much that a permanent deviation of market data from our predictions might lead us to implement trading strategies aimed to remove what we think could be pricing anomalies. Instead, no-arbitrage models are needed to deal with pricing problems where bond prices have to match market data, and being able to change over time as a result of a probability distribution determined by the current market data, even when these market data are mispriced, as we have explained in Chapter 11, and further develop in this chapter.

12.4.2 The bond pricing equation

12.4.2.1 A first derivation

Suppose bond prices are solutions to the following stochastic differential equation:

$$\frac{dP_i}{P_i} = \mu_{b_i} d\tau + \sigma_{b_i} dW, \quad (12.29)$$

where W is a standard Brownian motion in \mathbb{R}^d , μ_{b_i} and σ_{b_i} are some progressively measurable functions (σ_{b_i} is vector-valued), and $P_i \equiv P(\tau, T_i)$. The exact functional form of μ_{b_i} and σ_{b_i}

is *not* given, as in the BS case. Rather, it is endogenous and must be found as a part of the equilibrium.

As shown in Appendix 1, the price system in (12.29) is arbitrage-free if and only if

$$\mu_{bi} = r + \sigma_{bi}\lambda, \quad (12.30)$$

for some \mathbb{R}^d -dimensional process λ satisfying some basic regularity conditions. The meaning of (12.30) can be understood by replacing it into Eq. (12.29), and obtaining:

$$\frac{dP_i}{P_i} = (r + \sigma_{bi}\lambda) d\tau + \sigma_{bi}dW.$$

The previous equation tells us that the growth rate of P_i is the short-term rate plus a *term-premium* equal to $\sigma_{bi}\lambda$. In the bond market, there are no obvious economic arguments enabling us to sign term-premia. Empirical evidence suggests that term-premia did take both signs over the last twenty years. But term-premia would be zero in a risk-neutral world. In other terms, bond prices are solutions to:

$$\frac{dP_i}{P_i} = r d\tau + \sigma_{bi}d\tilde{W},$$

where $\tilde{W} = W + \int \lambda d\tau$ is a Q -Brownian motion and Q is the risk-neutral probability.

To derive Eq. (12.30) with the help of a specific version of theory developed in Appendix 1, we now work out the case $d = 1$. Consider two bonds, and the dynamics of the value V of a self-financed portfolio in these two bonds and a money market account:

$$dV = [\pi_1(\mu_{b1} - r) + \pi_2(\mu_{b2} - r) + rV] d\tau + (\pi_1\sigma_{b1} + \pi_2\sigma_{b2}) dW,$$

where π_i is wealth invested in bond maturing at T_i : $\pi_i = \theta_i P_i$. We can zero uncertainty by setting

$$\pi_1 = -\frac{\sigma_{b2}}{\sigma_{b1}}\pi_2.$$

By replacing this into the dynamics of V ,

$$dV = \left[-\frac{\mu_{b1} - r}{\sigma_{b1}}\sigma_{b2} + (\mu_{b2} - r) \right] \pi_2 d\tau + rV d\tau.$$

Notice that π_2 can always be chosen so as to make the value of this portfolio appreciate at a rate strictly greater than r . It is sufficient to set:

$$\text{sign}(\pi_2) = \text{sign} \left[-\frac{\mu_{b1} - r}{\sigma_{b1}}\sigma_{b2} + (\mu_{b2} - r) \right].$$

Therefore, to rule out arbitrage opportunities, it must be the case that:

$$\frac{\mu_{b1} - r}{\sigma_{b1}} = \frac{\mu_{b2} - r}{\sigma_{b2}}.$$

The previous relation tells us that the Sharpe ratio for any two bonds has to equal a process λ , say, and Eq. (12.30) immediately follows. Clearly, this function, λ , does not depend on none of the two maturity dates, T_1 or T_2 . Since T_1 and T_2 are arbitrary, then, λ is independent of time to maturity, T . It is natural, as λ is the unit price of risk agents require to be compensated for

the fluctuations of the short-term rate, and it must be independent of the assets they trade on, i.e. the maturity.

In models of the short-term rate such as that in Eq. (12.28), the two functions μ_{b_i} and σ_{b_i} in Eq. (12.29) can be determined through Itô's lemma. Let $P(r, \tau, T)$ be the rational bond price function, i.e., the price as of time τ of a bond maturing at T when the state at τ is r . Since r is solution to (12.28), Itô's lemma then implies that:

$$dP = \left(\frac{\partial P}{\partial \tau} + bP_r + \frac{1}{2}a^2P_{rr} \right) d\tau + aP_r dW,$$

where subscripts denote partial derivatives.

Comparing this equation with Eq. (12.29) then reveals that:

$$\mu_b P = \frac{\partial P}{\partial \tau} + bP_r + \frac{1}{2}a^2P_{rr}, \quad \sigma_b P = aP_r.$$

Now replace these functions into Eq. (12.30) to obtain the the bond price satisfies the following partial differential equation (PDE, henceforth):

$$\frac{\partial P}{\partial \tau} + bP_r + \frac{1}{2}a^2P_{rr} = rP + \lambda aP_r, \quad \text{for all } (r, \tau) \in \mathbb{R}_{++} \times [t, T), \quad (12.31)$$

with the boundary condition $P(r, T, T) = 1$ for all $r \in \mathbb{R}_{++}$.

Eq. (12.31) shows that the bond price, P , depends on both the drift of the short-term rate, b , and the risk-aversion correction, λ . This circumstance occurs as the initial asset market structure is incomplete, in the following sense. In the Black-Scholes model, the option is redundant, given the initial market structure. In the context we analyze here, the short-term rate r is *not* a traded asset. In other words, the initial market structure has one untraded risk (r) and zero assets: the factor generating uncertainty in the economy, r , is not traded. Therefore, the drift of the short-term rate cannot be equal to $r \cdot r = r^2$ under the risk-neutral probability, but rather $b - \lambda a$, thereby leading to Eq. (12.31). Therefore, the bond price depends on the specific functional forms b , a and λ .

While this kind of dependence might be seen as a kind of hindrance to practitioners, it can also be viewed as a good piece of news. Indeed, information about agents' risk-appetite λ can be backed out, after having estimated the two functions (b, a) . In turn, information about agents' risk-appetite can, for example, help central bankers to take decisions about the interest rates to set.

By specifying the drift and diffusion functions b and a , and by identifying the risk-premium λ , the PDE in Eq. (12.31) can explicitly be solved, either analytically or numerically. Choices concerning the exact functional form of b, a and λ are often made on the basis of either analytical or empirical reasons. In the next section, we will examine the first, famous short-term rate models where b, a and λ have a particularly simple form. We will discuss the analytical advantages of these models, but we will also highlight the major empirical problems associated with these models. In Section 12.4.4 we provide a very succinct description of models exhibiting jump (and default) phenomena. In Section 12.4.5, we introduce multifactor models: we will explain why do we need such more complex models, and show that even in this more complex case, arbitrage-free bond prices are still solutions to PDEs such as (12.31). In Section 12.4.6, we will present a class of analytically tractable multidimensional models, known as affine models. We will discuss their historical origins, and highlight their importance as regards the

econometric estimation of bond pricing models. Finally, Section 12.4.7 presents the “perfectly fitting” models, and Appendix 5 provides a few technical details about the solution of one of these models.

12.4.2.2 Derivation based on duration

The idea, here, is to replicate the price of a bond expiring at some time T_1 , say $P^1 \equiv P(r, \tau, T_1)$, with a self-financed portfolio comprising a money market account and a second bond expiring at time $T_2 > T_1$. The value of the self-financed portfolio is $V = \Delta \cdot P^2 + M$, where Δ is the number of bonds maturing at T_2 to be put in the portfolio, $P^2 = P(r, \tau, T_2)$, and M is the amount of resources put in the money market account. Since the portfolio is self-financed, we have, by the usual arguments, that,

$$dV = \Delta \cdot dP^2 + dM = (\Delta \cdot LP^2 + rM) d\tau + \Delta \cdot aP_r^2 dW, \quad (12.32)$$

where $LP^2 = \frac{\partial P^2}{\partial r} + bP_r^2 + \frac{1}{2}a^2P_{rr}^2$. And, obviously,

$$dP^1 = LP^1 d\tau + aP_r^1 dW. \quad (12.33)$$

Let the initial value of the portfolio match the bond price. Then, comparing the diffusive terms in Eq. (12.32) and Eq. (12.33), we find the delta to be:

$$\hat{\Delta} = \frac{\partial P(r, \tau, T_1) / \partial r}{\partial P(r, \tau, T_2) / \partial r}.$$

Comparing the drift terms in Eq. (12.32) and Eq. (12.33),

$$LP^1 = \Delta \cdot LP^2 + rM = \Delta \cdot LP^2 + r(V - \Delta P^2) = \Delta \cdot LP^2 + r(P^1 - \Delta P^2),$$

where the last line follows as we're using the values (Δ, M) such that the portfolio matches the value of the first bond. Rearranging terms yields, $LP^1 - rP^1 = \Delta \cdot (LP^2 - rP^2)$, and evaluating this for $\Delta = \hat{\Delta}$,

$$\frac{LP^1 - rP^1}{P_r^1} = \frac{LP^2 - rP^2}{P_r^2} \equiv \Lambda \equiv \lambda a,$$

for some Λ and λ independent of calendar time.

The delta, $\hat{\Delta}$, can be interpreted as the ratio of the durations of the two bonds, as explained in Chapter 11.

12.4.3 Some famous short-term rate models

12.4.3.1 Vasicek

Vasicek (1977) develops the seminal contribution, assuming that the short-term rate is a mean-reverting process, solution to:

$$dr(\tau) = \kappa(\bar{r} - r(\tau)) d\tau + \sigma dW(\tau), \quad \tau \in (t, T], \quad (12.34)$$

where \bar{r} , κ and σ are positive constants. This model is more sensible than that of Merton (1973), where the short-term rate is an arithmetic Brownian motion. The intuition underlying the importance of mean-reversion is as follows. Suppose, first, that $\sigma = 0$, in which case,

$$r(\tau) = \bar{r} + e^{-\kappa(\tau-t)} (r(t) - \bar{r}). \quad (12.35)$$

If the current level of the short-term rate $r(t) = \bar{r}$, it will be “locked-in” at \bar{r} forever. If, $r(t) < \bar{r}$, the short-term rate shall steadily increase, and converge to \bar{r} as $\tau \rightarrow \infty$. Likewise, the short-term rate shall converge to \bar{r} when $r(t) > \bar{r}$. The speed of convergence of r to the long-term value \bar{r} depends on the magnitude of κ : the higher κ , the higher the speed of convergence to \bar{r} .

In the general case, $\sigma \neq 0$, the solution to Eq. (12.34) is,

$$r(\tau) = \bar{r} + e^{-\kappa(\tau-t)} (r(t) - \bar{r}) + \sigma e^{-\kappa\tau} \int_t^\tau e^{\kappa s} dW(s),$$

where the integral has to be understood in the Itô's sense. The interpretation of this solution is similar to that given in the determinist case, in that the short-term rate now fluctuates around its “central tendency” \bar{r} . In other words, shocks are absorbed at a speed depending on the magnitude of κ , leading the short-term rate to display a mean-reverting behavior. Indeed, the conditional expectation of r is the same as that in Eq. (12.35),

$$E[r(\tau)|r(t)] = \bar{r} + e^{-\kappa(\tau-t)} (r(t) - \bar{r}). \quad (12.36)$$

Moreover, the conditional variance of r is:

$$\text{var}[r(\tau)|r(t)] = \frac{\sigma^2}{2\kappa} [1 - e^{-2\kappa(\tau-t)}].$$

Finally, it can be shown that r is normally distributed, with expectation and variance given by the two functions given above.

To solve for the entire term-structure of interest rates, we need to make assumptions about the risk-premium, λ . A closed-form expression for the bond price obtains, once we assume λ is a constant. Indeed, by replacing a constant risk-premium λ and the functions $b(r) = \kappa(\bar{r} - r)$ and $a(r) = \sigma$ into Eq. (12.31), and denoting $r^* \equiv \bar{r} - \frac{\lambda\sigma}{\kappa}$, we obtain that the bond price P is solution to:

$$0 = \frac{\partial P}{\partial \tau} + \kappa(r^* - r)P_r + \frac{1}{2}\sigma^2 P_{rr} - rP, \quad \text{for all } (r, \tau) \in \mathbb{R} \times [t, T], \quad (12.37)$$

with the usual boundary condition. Intuitively, $\kappa(r^* - r)$ is the drift of the short-term rate under Q , which is higher than under P for $\lambda < 0$, reflecting higher Arrow-Debreu state prices for the bad states of the world arising when interest rates are high. It is instructive to see how this partial differential equation can be solved. We guess a solution of the form:

$$P(r, \tau, T) = e^{A(\tau, T) - B(\tau, T) \cdot r}, \quad (12.38)$$

for two functions A and B to be determined. Now suppose the guess in Eq. (12.38) is true. By replacing the partial derivatives of P into Eq. (12.37) leaves the equation $\phi_1(\tau) + \phi_2(\tau)r = 0$, where for all τ ,

$$\phi_1(\tau) \equiv A_1(\tau, T) - \kappa r^* B(\tau, T) + \frac{1}{2}\sigma^2 B^2(\tau, T) \quad \text{and} \quad \phi_2(\tau) \equiv \kappa B(\tau, T) - B_1(\tau, T) - 1.$$

That is, $0 = \phi_1(\tau) = \phi_2(\tau)$, two ordinary differential equations subject to the boundary conditions $A(T, T) = 0$ and $B(T, T) = 0$. The solutions are:

$$A(\tau, T) = \left(\frac{1 - e^{-\kappa(T-\tau)}}{\kappa} - (T - \tau) \right) r_\infty - \frac{\sigma^2}{4\kappa^3} (1 - e^{-\kappa(T-\tau)})^2, \quad B(\tau, T) = \frac{1}{\kappa} (1 - e^{-\kappa(T-\tau)}),$$

where

$$r_\infty = r^* - \frac{1}{2} \left(\frac{\sigma}{\kappa} \right)^2.$$

The term-structure of spot rates predicted by this model is, by the definition in Eq. (12.11),

$$R(r, t, T) = -\frac{A(t, T)}{T-t} + \frac{B(t, T)}{T-t} r. \quad (12.39)$$

The quantity, r_∞ , is interpreted as the “asymptotic” spot rate, as it is the limit of $R(r, t, T)$ for large T .

The model displays a number of features that match some empirical facts—such as selected shapes of the yield-curve. Plot it.

12.4.3.2 A theoretical case study: expectations and business cycles in Vasicek

We can use the Vasicek’s model to illustrate a few issues pertaining to the expectation hypothesis and the interplay between long-term yields and the expectation of future short-term rates. First, we express the yield curve in Eq. (12.39) in a way that is more convenient to interpret. We mentioned earlier that the short-term rate is conditionally normally distributed. Now, by Lambertson and Lapeyre (1997, Chapter 6), the term $\int_t^T r(\tau) d\tau$ is conditionally normally distributed, and then, by Eq. (12.11),

$$R(r, t, T) = \mathbb{E}_t \left(\frac{1}{T-t} \int_t^T r(\tau) d\tau \right) - \frac{1}{2} \frac{1}{T-t} \text{var} \left(\int_t^T r(\tau) d\tau \right). \quad (12.40)$$

The second term in Eq. (12.40) reflects Jensen’s inequality effects. It equals,

$$\frac{1}{2} \frac{1}{T-t} \text{var} \left(\int_t^T r(\tau) d\tau \right) = \frac{1}{2} \left(1 - \frac{1 - e^{-\kappa(T-t)}}{(T-t)\kappa} \right) \left(\frac{\sigma}{\kappa} \right)^2 - \frac{\sigma^2}{4(T-t)\kappa^3} (1 - e^{-\kappa(T-t)})^2,$$

and is of second order importance, quantitatively, compared to the first term. We can approximate the yield curve with this first term, which is an expectation of the average short-term rate under the risk-neutral probability, and by Eq. (12.36) and Eq. (12.31), can be decomposed as the sum of the expectation under the physical probability plus risk-premiums terms, as follows:

$$\begin{aligned} R(r, t, T) &\approx \mathbb{E}_t \left(\frac{1}{T-t} \int_t^T r(\tau) d\tau \right) \\ &= E_t \left(\frac{1}{T-t} \int_t^T r(\tau) d\tau \right) + \frac{1}{T-t} \int_t^T \mathcal{E}(\tau) d\tau, \end{aligned} \quad (12.41)$$

where:

$$\begin{aligned} E_t \left(\frac{1}{T-t} \int_t^T r(\tau) d\tau \right) &= \bar{r} + (r(t) - \bar{r}) \frac{1 - e^{-\kappa(T-t)}}{\kappa(T-t)} \\ \mathcal{E}(t) &\equiv E_t \left(\frac{dP(r, t, T)}{P(r, t, T)} - r \right) = \frac{P_r(r, t, T)}{P(r, t, T)} \sigma \lambda = -\frac{1}{\kappa} (1 - e^{-\kappa(T-t)}) \sigma \lambda \end{aligned}$$

Eq. (12.41) says that long-term rates reflect expectations of the future short-term rate and risk-premiums terms, defined as the average expected return on the bond.

We can rely on this simple framework to describe some of the business cycle properties of the yield curve. We assume that a state variable y , is capable to track some business cycle conditions, and is solution to the following stochastic differential equation,

$$dy(\tau) = \kappa(\bar{y} - y(\tau)) d\tau + \sigma_y dW(\tau),$$

for three positive constants, b , \bar{y} and σ_y . Next, suppose that: (i) the *nominal* short-term rate is procyclical, in that $r(\tau) \equiv A + By(\tau)$, for two positive constants A and B , and that (ii) risk-premiums are countercyclical in that, $\lambda_y(y) \equiv \lambda_0 - \lambda_1 y$, for two additional constants, where $\lambda_1 > 0$, to ensure countercyclicity, and λ_0 might in principle take any sign, although it is reasonable to expect that $\lambda_0 > 0$, which would imply that the constant portion of the risk-premium is positive anyway. We shall return to the sign of λ_0 soon.

Given the assumptions made so far, the short-term rate is solution to Eq. (12.34), with parameters $\bar{r} \equiv A + B\bar{y}$ and $\sigma \equiv B\sigma_y$. While the risk-premium is time-varying, being equal to $\lambda(r) \equiv \lambda_0 - \frac{\lambda_1}{B}(r - A)$, the short-term rate is still conditionally normally distributed under the risk-neutral probability, and the yield-curve can be solved in closed form, with a solution like that in Eq. (12.38), and an approximation like that in Eq. (12.41). It is instructive to calculate the decomposition in Eq. (12.41) predicted by this model. We have,

$$R(r, t, T) \approx \mathbb{E}_t \left(\frac{1}{T-t} \int_t^T r(\tau) d\tau \right) = r^* + (r(t) - r^*) \frac{1 - e^{-\kappa^*(T-t)}}{\kappa^*(T-t)},$$

where $\kappa^* = \kappa - \sigma \frac{\lambda_1}{B}$ and $r^* = \frac{\kappa \bar{r} - \sigma \lambda_0 - \sigma \frac{\lambda_1}{B} A}{\kappa - \sigma \frac{\lambda_1}{B}}$. The model's parameters clearly collapse to Vasicek's, once we assume $\lambda_1 = 0$. A countercyclical risk-premium, $\lambda_1 > 0$, leads to a tilt in the unconditional short-term rate, r^* , and, importantly, an increased persistence of the short-term rate, due to the fact that $\kappa^* < \kappa$. Clearly, there are no solutions to the model for large T , when risk-premium effects are so large to make $\kappa^* < 0$. Finally, note that although it is reasonable to assume $\lambda_0 > 0$, as explained, we might also wish to make sure that $\lambda(r)$ is negative for most of the time, to ensure positive expected excess returns.

The term-spread over the short-term rate predicted by the model is,

$$R(A + By(t), t, T) - r \approx \left(\frac{1 - e^{-\kappa^*(T-t)}}{\kappa^*(T-t)} - 1 \right) (y(t) - y^*),$$

where $y^* = \frac{\kappa \bar{y} - \sigma_y \lambda_0}{\kappa - \sigma_y \lambda_1}$ is the unconditional expectation of the procyclical variable y , taken under the risk-neutral probability. According to this model, the term spread is the product of two terms. The first is negative when $\kappa^* > 0$, and in this case, the model formalizes explanations given in Section 12.3.2. Before a peak, i.e. when $y(t) < y^*$, the yield curve is upward sloping. After this peak is achieved and, then, $y(t) > y^*$, the probability the economy would enter into a recession becomes more likely, given the mean-reverting nature of $y(t)$, and the yield curve becomes inverted, as nominal rates are procyclical, and countercyclical risk-aversion is mild enough to guarantee that $\kappa^* > 0$. Note that if λ_1 had to be so large to make $\kappa^* < 0$, the model would generate the wrong predictions, with an inverted yield curve during the rising part of a boom, not the descending part. The mechanism would be that during a peak, we would expect that the future short-term rate would be low, but risk-aversion to be so high, to dwarf expectations effects and push future prices down, to an extent that would compensate for the procyclical effects generated by the short-term rate.

Naturally, note that this model is very simple, being driven by one factor only, the business cycle variable $y(t)$, thereby leading to a sharp prediction about the slope of the yield curve: a positive slope before a peak, $y(t) < y^*$, and a negative after the peak, $y(t) > y^*$, just as we observe in the data. This model thus isolates the business cycle component of the yield curve that relates to its inversions. The crucial point is that the model is silent as regards the business cycle variable $y(t)$. If we knew $y(t)$, we could use it to forecast the business cycle in the first place.

12.4.3.3 Cox, Ingersoll and Ross

Vasicek's model suffers from two main drawbacks. First, the short-term rate is normally distributed. This circumstance might be mitigated when σ is low, compared to \bar{r} , in which case the probability the short-term rate takes negative values can be small. At the same time, even a small probability of a negative interest rate might lead to severe mispricings when it comes to pricing interest rate derivatives, due to nonlinearities induced by optionality, as pointed out by Dybvig [cite reference]. Section 11.5.3 of the previous chapter displays numerical examples where small changes in assumptions can lead to quite substantial changes in the price of derivatives. The second drawback, related to the first, is that the short-term rate volatility is independent of the level of the short-term rate. It might be argued that short-term rates changes become more and more volatile as the level of the short-term rate increases, a phenomenon usually referred to as the *level-effect*.

The model proposed by Cox, Ingersoll and Ross (1985) (CIR, henceforth) addresses these two drawbacks at once, as it assumes that the short-term rate is solution to,

$$dr(\tau) = \kappa(\bar{r} - r(\tau))d\tau + \sigma\sqrt{r(\tau)}dW(\tau), \quad \tau \in (t, T].$$

The CIR model is also referred to as “square-root” process to emphasize that the diffusion function is proportional to the square-root of r . This feature makes the model address the level-effect phenomenon. The evidence about the level-effect is further discussed below (see Section 12.4.7). Moreover, this property prevents r from taking negative values. Intuitively, when r wanders just above zero, it is pulled back to the strictly positive region at a strength of the order $dr = \kappa\bar{r}d\tau$.⁸ The transition density of r is noncentral chi-square. The stationary density of r is a gamma distribution. The expected value is as in Vasicek.⁹ However, the variance is different, although its exact expression is really not important here.

CIR formulated a set of assumptions on the primitives of the economy (e.g., preferences) that led to a risk-premium function $\lambda = \ell\sqrt{r}$, where ℓ is a constant. By replacing this, $b(r) = \kappa(\bar{r} - r)$ and $a(r) = \sigma\sqrt{r}$ into the PDE (12.31), one gets (similarly as in the Vasicek model), that the bond price function takes the form in Eq. (12.38), but with functions A and B satisfying the following differential equations:

$$0 = A_1(\tau, T) - \kappa\bar{r}B(\tau, T) \quad \text{and} \quad 0 = -B_1(\tau, T) + (\kappa + \ell\sigma)B(\tau, T) + \frac{1}{2}\sigma^2B^2(\tau, T) - 1,$$

subject to the boundary conditions, $A(T, T) = 0$ and $B(T, T) = 0$.

⁸This is only intuition. The exact condition under which the zero boundary is unattainable by r is $\kappa\bar{r} > \frac{1}{2}\sigma^2$. See Karlin and Taylor (1981, vol II chapter 15) for a general analysis of attainability of boundaries for scalar diffusion processes.

⁹The expected value of linear mean-reverting processes is always as in Vasicek, independently of the functional form of the diffusion coefficient. This property follows by a direct application of a general result for diffusion processes given in Chapter 6 (Appendix A).

In their article, CIR also showed how to compute options on bonds. They even provided hints on how to “invert the term-structure,” a popular technique that we describe in detail in Section 12.4.6. For all these features, the CIR model and paper have been used in the industry for many years. And many of the more modern models are mere multidimensional extensions of the basic CIR model. (See Section 12.4.6).

12.4.3.4 Nonlinear drifts

Models that are analytically tractable are certainly quite valuable. Vasicek and CIR models do lead to closed-form solutions, because they have a *linear* drift, among other things. Is the empirical evidence consistent with linear mean-reversion of the short-term rate? This issue is subject to controversy. In the mid 1990s, three papers by Aït-Sahalia (1996), Conley et al. (1997) and Stanton (1997) produce evidence of nonlinear mean-reverting behavior. For example, Aït-Sahalia (1996) estimates a drift function of the following form:

$$b(r) = \beta_0 + \beta_1 r + \beta_2 r^2 + \beta_3 r^{-1}, \quad (12.42)$$

corresponding to a nonlinear diffusion function. Figure 12.3 reproduces this function using the parameter values in his Table 4, and relating to the sample period from 1983 to 1995. Similar results are reported in the other papers. To grasp the action the short-term rate dynamics are under, Figure 12.3 also depicts a linear drift, obtained with the parameter estimates of Aït-Sahalia (1996) (Table 4), and corresponding to a model with a CEV diffusion.

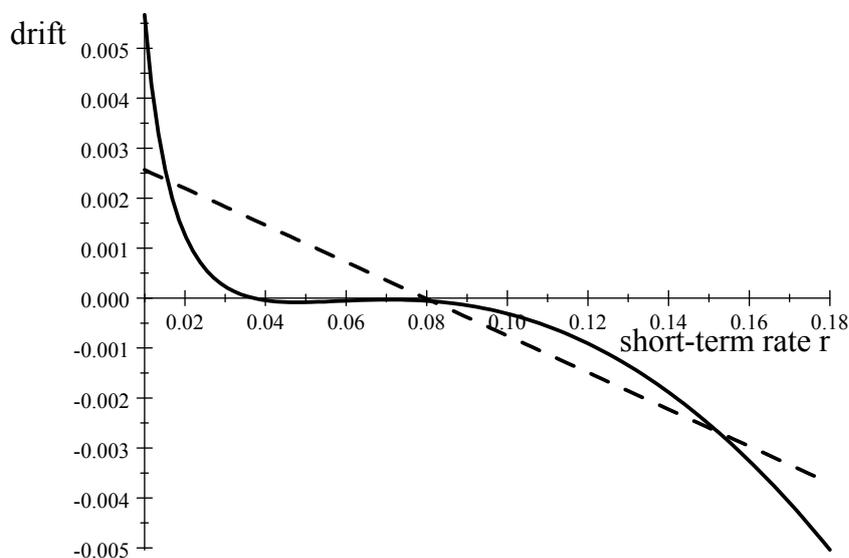


FIGURE 12.3. Nonlinear mean reversion? The solid line is the drift function in Eq. (12.42), estimated by Aït-Sahalia (1996), and relating to a parametric model with a nonlinear diffusion function. The dashed line is the estimated linear drift relating to a model with CEV diffusion.

The nonlinear drifts in Figure 12.3 might lead bond prices to exhibit unusual properties, though. As explained in Chapter 7 (Appendix 5), bond prices are *concave* in the short-term rate if the risk-neutralized drift function is sufficiently convex (Mele, 2003). While the results in Figure 12.3 relate to the *physical* drift functions, the point is nevertheless important as risk-premiums should look like quite unusual to destroy the nonlinearities of the short-term rate under the physical probability.

The compelling lesson from Figure 12.3 is that under the “nonlinear drift dynamics,” the short-term rate behaves in a way that can at least be *roughly* comparable with that it would behave under the “linear drift dynamics.” However, the behavior at the extremes is dramatically different. As the short-term rate moves to the extremes, it is pulled back to the “center” in a very abrupt way. At the moment, it is not clear whether these preliminary empirical results are reliable or not. New econometric techniques are currently being developed to address this and related issues.

One possibility is that such *single factor* models of the short-term rate are simply misspecified. For example, there is strong empirical evidence that the volatility of the short-term rate is time-varying, as we shall discuss in the next section. Moreover, the term-structure implications of a single factor model are counterfactual, since we know that a single factor cannot explain the entire variation of the yield curve, as explained in Section 12.3.4. We now describe more realistic models driven by more than one factor.

12.4.4 Multifactor models

The empirical evidence reviewed in Section 12.3.4 suggests that one-factor models cannot explain the entire variation of the term-structure of interest rates. Factor analysis suggests we need at least three factors. In this section, we succinctly review the advances made in the literature to address this important empirical issue.

12.4.4.1 Stochastic volatility

In the CIR model, the instantaneous short-term rate volatility is stochastic, as it depends on the level of the short-term rate, which is obviously stochastic. However, empirical evidence suggests that the short-term rate volatility depends on some additional factors. A natural extension of the CIR model is one where the instantaneous volatility of the short-term rate depends on (i) the level of the short-term rate, similarly as in the CIR model, and (ii) some additional random component. Such an additional random component is what we shall refer to as the “stochastic volatility” of the short-term rate. It is the term-structure counterpart to the stochastic volatility extension of the Black and Scholes (1973) model (see Chapter 10).

Fong and Vasicek (1991) write the first paper in which the volatility of the short-term rate is stochastic. They consider the following model:

$$\begin{aligned} dr(\tau) &= \kappa_r (\bar{r} - r(\tau)) d\tau + \sqrt{v(\tau)} r(t)^\gamma dW_1(\tau) \\ dv(\tau) &= \kappa_v (\bar{v} - v(\tau)) d\tau + \xi_v \sqrt{v(\tau)} dW_2(\tau) \end{aligned} \quad (12.43)$$

where κ_r , \bar{r} , κ_v , \bar{v} and ξ_v are constants, and $[W_1 \ W_2]$ is a vector Brownian motion. To obtain a closed-form solution, Fong and Vasicek set $\gamma = 0$. The authors also make assumptions about risk aversion corrections. Namely, they assume that the unit-risk-premia for the stochastic fluctuations of the short-term rate, λ_r , and the short-term rate volatility, λ_v , are both proportional to $\sqrt{v(\tau)}$, and then they find a closed-form solution for the bond price as of time t and maturing at time T , $P(r(t), v(t), T - t)$.

Longstaff and Schwartz (1992) propose another model of the short-term rate where the volatility of the short-term rate is stochastic. The remarkable feature of their model is that it is a general equilibrium model. Naturally, the Longstaff & Schwartz model predicts, as the Fong-Vasicek model, that the bond price is a function of both the short-term rate and its instantaneous volatility.

Note, then, the important feature of these models. The pricing function, $P(r(t), v(t), T-t)$ and, hence, the yield curve $R(r(t), v(t), T-t) \equiv -(T-t)^{-1} \ln P(r(t), v(t), T-t)$, depends on the level of the short-term rate, $r(t)$, and one additional factor, the instantaneous variance of the short-term rate, $v(t)$. Hence, these models predict that we now have two factors that help explain the term-structure of interest rates, $R(r(t), v(t), T-t)$.

What is the relation between the volatility of the short-term rate and the term-structure of interest rates? Does this volatility help “track” one of the factors driving the variations of the yield curve? To develop intuition, consider the following binomial example. In the next period, the short-term rate is either $r^+ = r + d$ or $r^+ = r - d$ with equal probability, where r is the current interest rate level and $d > 0$. The price of a two-period bond is $P(r, d) = m(r, d)/(1+r)$, where $m(r, d) = E[1/(1+r^+)]$ is the expected discount factor of the next period. By Jensen’s inequality, $m(r, d) > 1/(1+E[r^+]) = 1/(1+r) = m(r, 0)$. Therefore, two-period bond prices increase upon activation of randomness. More generally, two-period bond prices are always increasing in the “volatility” parameter d in this example, as illustrated by Figure 12.4. The intuition is that the bond price is inversely related to the short-term rate and it is convex in it. Therefore, an increase in the volatility of the interest rate may only increase the price, as the loss in value of the price in bad times, when the interest rate increases, is less than the gain in value in good times, when the interest rate decreases. This property relates to an insight of Jagannathan (1984) (p. 429-430) that in a two-period economy with identical initial underlying asset prices, a terminal underlying asset price \tilde{y} is a mean preserving spread of another terminal underlying asset price \tilde{x} , in the Rothschild and Stiglitz (1970) sense, if and only if the price of a call option on \tilde{y} is higher than the price of a call option on \tilde{x} . This is because if \tilde{y} is a mean preserving spread of \tilde{x} , then $E[f(\tilde{y})] > E[f(\tilde{x})]$ for f increasing and convex.¹⁰

These properties arise because the expected short-term rate is independent of d . In an alternative setting, say a multiplicative setting, where either $r^+ = r(1+d)$ or $r^+ = r/(1+d)$ with equal probability, bond prices are decreasing in volatility at short maturities and increasing in volatility at longer maturities, as originally pointed out by Litterman, Scheinkman and Weiss (1991). It’s because expected future interest rates increase over time at a strength positively related to d . That is, *the expected variation of the short-term rate is increasing in the volatility of the short-term rate, d* , a property that can be re-interpreted as one arising in an economy with risk-averse agents. At short maturity dates, such an effect dominates the convexity effect illustrated in Figure 12.4. At longer maturity dates, convexity effects dominate.

¹⁰In our case, let $\tilde{m}_d(i^+) = 1/(1+i^+)$ denote the random discount factor when $i^+ = i \mp d$. We have that $x \mapsto -\tilde{m}_d(x)$ is increasing and concave and, hence, $E[-\tilde{m}_{d'}(x)] < E[-\tilde{m}_d(x)] \Leftrightarrow d' < d$, as shown in Figure 12.3. In Jagannathan (1984), f is increasing and convex, and so we must have: $E[f(\tilde{y})] > E[f(\tilde{x})] \Leftrightarrow \tilde{y}$ is riskier than, or a mean preserving spread of \tilde{x} .

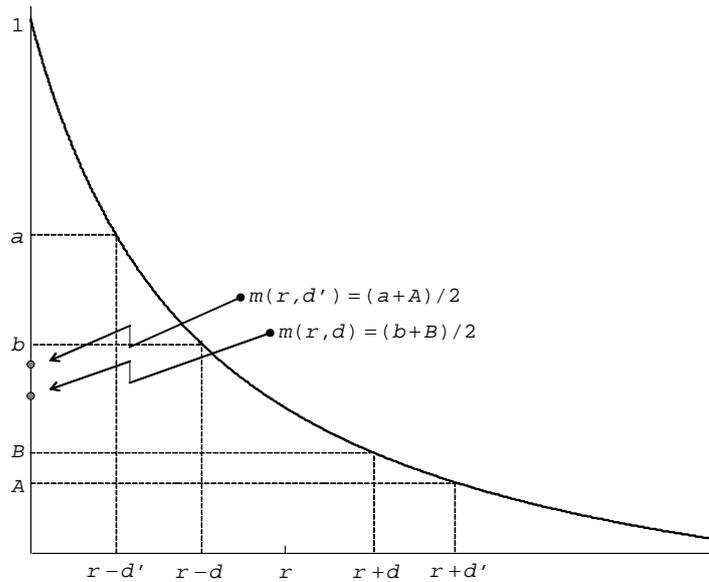


FIGURE 12.4. If the risk-neutralized interest rate of the next period is either $r^+ = r + d$ or $r^+ = r - d$ with equal probability, the random discount factor $1/(1 + r^+)$ is either B or b with equal probability. Hence $m(r, d) = E[1/(1 + r^+)]$ is the midpoint of \overline{bB} . Similarly, if volatility is $d' > d$, $m(r, d')$ is the midpoint of \overline{aA} . Since $\overline{ab} > \overline{BA}$, it follows that $m(r, d') > m(r, d)$. Therefore, the two-period bond price $P(r, d) = m(r, d)/(1 + r)$ satisfies: $P(r, d') > P(r, d)$ for $d' > d$.

To illustrate further, consider the basic Vasicek (1997) model. Naturally, volatility is constant in this model, but we can use this model to develop intuition stochastic volatility models, such those the Fong and Vasicek (1991) of Eqs. (12.43). For the Vasicek model, we have that, by Eq. (12.38),

$$\frac{\partial R(r(t), T - t)}{\partial \sigma} = -\frac{1}{T - t} \left[\sigma \int_t^T B^2(T - s) ds + \lambda \int_t^T B(T - s) ds \right]. \tag{12.44}$$

where $B(T - s)$ is as in Section 12.4.3.1. Eq. (12.44) shows that if $\lambda \geq 0$, the term-structure is decreasing in short-term rate volatility. That is, bond prices increase in σ , a conclusion that parallels that for options, where option prices are increasing in the volatility of the asset price. As explained in Chapter 10, this property arises through the optionality of the contract—say the convexity of a European call price with respect to the asset price.

Interesting properties arise in the empirically relevant case, $\lambda < 0$.¹¹ In this case, the sign of $\frac{\partial R(t, T)}{\partial \sigma}$ depends on both “convexity” and “slope” effects. “Convexity” effects, those relating to the second partial $\frac{\partial^2 P(r, T - t)}{\partial r^2} = P(r, T - t) B^2(T - t)$, arise through the term $\sigma \int_t^T B^2(T - s) ds$. “Slope” effects, those relating to $\frac{\partial P(r, T - t)}{\partial r} = P(r, T - t) B(T - t)$, arise, instead, through the term $\int_t^T B(T - s) ds$. If λ is negative, and sufficiently large in absolute value, slope effects dominate convexity effects, and the term-structure can actually increase in σ . For intermediate

¹¹In this simple model, the assumption that $\lambda < 0$ is reasonable, as we observe positive risk-premia more often than negative risk-premia. But in this same model, $u_r < 0$, which together with $\lambda < 0$, ensures that term-premia are positive.

values of λ , the term-structure can be both increasing and decreasing in σ . At short maturities, the convexity effects in Eq. (12.44) are typically dominated by slope effects, and the short-end of the term-structure can be increasing in σ . At longer maturity dates, however, convexity effects are more important and, sometimes, dominate slope effects.

More generally, changes in interest rate volatility are not mean-preserving spreads for the risk-neutral distribution, as Eq. (12.44) illustrates for the Vasicek model. In a world with complete markets, say Black-Scholes, the asset underlying derivative contracts is traded. In the case under study, the short-term rate is not a traded risk. Therefore, its risk-neutral drift depends on volatility through risk-adjustment—in Vasicek, for example, this dependence arises through the risk-premium parameter λ .

The previous reasoning, while relying on comparative statics for models with constant volatility, goes through even when volatility is random. Mele (2003) shows that in more complex stochastic volatility cases, provided the risk-premium required to bear the interest rate risk is negative, and sufficiently large in absolute value, slope effects dominate convexity effects at any finite maturity date, thus making bond prices decrease with volatility at any arbitrary maturity date. In general, we would expect that in bad times, i.e. when interest rate volatility is high, the risk-premium effects dominate the convexity effects, with the yield curve increasing following an increase in volatility. In good times, however, we would expect convexity effects to dominate, with the yield curve decreasing following an increase in volatility. For example, if risk-premiums are particularly sensitive to increases in volatility, we would expect that in good times, when volatility is small, convexity effects dominate and the yield curve lowers as volatility increases. In bad times, when volatility is large, risk-premium effects should dominate, instead, and the yield curve should increase following an increase in volatility. Consider, for example, the Vasicek model, and set the risk-premium equal to $\lambda = \frac{1}{3}\bar{\lambda}\sigma^3$. This functional form of the risk-premium ensures that the risk-premium is quite small when σ is small, although then it substantially increases in bad times, i.e. when σ is larger. With this risk-premium, Eq. (12.44) is replaced with:

$$\frac{\partial R(r(t), T-t)}{\partial \sigma} = -\frac{1}{T-t}\sigma \left[\int_t^T B^2(T-s) ds + \bar{\lambda}\sigma \int_t^T B(T-s) ds \right]. \quad (12.45)$$

That is, risk-premium effects become more and more relevant as σ increases. In fact, the previous equation reveals that we might even define a threshold value of σ such that convexity effects are exactly offset by risk-premium effects. Eq. (12.45) shows that for each time to maturity $T-t$, there exists a value of σ depending on $T-t$, say $\hat{\sigma}(T-t)$ such that the partial $\frac{\partial R(r(t), T-t)}{\partial \sigma} = 0$. We might go on and define an average value of $\hat{\sigma}(T-t)$, say $\hat{\sigma}_* \equiv \mathcal{T}^{-1} \int_0^{\mathcal{T}} \hat{\sigma}(u) du$, where \mathcal{T} denotes the highest time-to-maturity we want to consider. This threshold value, $\hat{\sigma}_*$, is the one that might lead to a definition of what are good or bad times—in terms of the term-structure implications of a volatility shock.

What are the implications of in terms of the factors reviewed in Section 12.3.4? Clearly, the very short-end of yield curve is not affected by movements of the volatility, as $\lim_{T \rightarrow t} R(r(t), v(t), T-t) = r(t)$, for all $v(t)$. Moreover, these models predict that $\lim_{T \rightarrow \infty} R(r(t), v(t), T-t) = \bar{R}$, where \bar{R} is a constant and, hence, independent of $v(t)$. Therefore, movements in the short-term volatility can only produce their effects on the middle of the yield curve. For example, if the risk-premium required to bear the interest rate risk is negative and sufficiently large, an upward movement in $v(t)$ can produce an effect on the yield curve qualitatively similar to that depicted in Figure 12.2 (“Curvature” panel), and would thus roughly mimic the “curvature” factor that we reviewed in Section 12.3.4.

12.4.4.2 Three-factor models

We need at least three factors to explain the entire variation in the yield-curve. A model where the interest rate volatility is stochastic may be far from being exhaustive in this respect. A natural extension is a model where the drift of the short-term rate contains some predictable component, $\bar{r}(\tau)$, which acts as a third factor, as in the following model:

$$\begin{aligned} dr(\tau) &= \kappa_r (\bar{r}(\tau) - r(\tau)) d\tau + \sqrt{v(\tau)} r(t)^\gamma dW_1(\tau) \\ dv(\tau) &= \kappa_v (\bar{v} - v(\tau)) d\tau + \xi_v \sqrt{v(\tau)} dW_2(\tau) \\ d\bar{r}(\tau) &= \kappa_{\bar{r}} (\bar{r} - \bar{r}(\tau)) d\tau + \xi_{\bar{r}} \sqrt{\bar{r}(\tau)} dW_3(\tau) \end{aligned} \quad (12.46)$$

where $\kappa_r, \gamma, \kappa_v, \bar{v}, \xi_v, \kappa_{\bar{r}}, \bar{r}$ and $\xi_{\bar{r}}$ are constants, and $[W_1 \ W_2 \ W_3]$ is vector Brownian motion.

Balduzzi et al. (1996) develop the first model for which the drift of the short-term rate changes stochastically, as in Eqs. (12.46). Dai and Singleton (2000) estimate a number of models that generalize that in Eqs. (12.46) (See Section 12.4.7 for details on the estimation strategy). The term-structure implications of these models can be understood very simply. First, under regularity conditions about the risk-premia, the yield curve is $R(r(t), \bar{r}(t), v(t), T-t) \equiv -(T-t)^{-1} \ln P(r(t), \bar{r}(t), v(t), T-t)$. Second, and intuitively, changes in the new factor $\bar{r}(t)$ should primarily affect the long-end of the yield curve. This is because empirically, the usual finding is that the short-term rate reverts relatively quickly to the long-term factor $\bar{r}(\tau)$ (i.e. κ_r is relatively large), where $\bar{r}(\tau)$ mean-reverts slowly (i.e. $\kappa_{\bar{r}}$ is relatively low). This mechanism makes the short-term rate quite persistent anyway. Ultimately, then, the slow mean-reversion of $\bar{r}(\tau)$ means that changes in $\bar{r}(\tau)$ last for the relevant part of the term-structure we are usually interested in (i.e. up to 30 years), despite the fact that $\lim_{T \rightarrow \infty} R(r(t), \bar{r}(t), v(t), T-t)$ is independent of the movements of the three factors $r(t)$, $\bar{r}(t)$ and $v(t)$.

However, it is difficult to see how to reconcile such a behavior of the long-end of the yield curve with the existence of any of the factors discussed in Section 12.3.4. First, the short-term rate cannot be taken as a “level factor,” since we know its effects die off relatively quickly. Instead, a *joint* change in both the short-term rate, $r(t)$, and the “long-term” rate, $\bar{r}(t)$, should be really needed to mimic the “Level” panel of Figure 12.2 in Section 12.3.4. However, this interpretation is at odds with the assumption that the factors discussed in Section 12.3.4 are uncorrelated! Moreover, and crucially, the empirical results in Dai and Singleton reveal that if any, $r(t)$ and $\bar{r}(t)$ are negatively correlated.

Finally, to emphasize how exacerbated these puzzles are, consider the effects of changes in the short-term rate $r(t)$. We know that the long-end of the term-structure is not affected by movements of the short-term rate. Hence, the short-term rate acts as a “steepness” factor, as in Figure 12.2 (“Slope” panel). However, this interpretation is restrictive, as factor analysis reveals that the short-end and the long-end of the yield curve move in opposite directions after a change in the steepness factor. Here, instead, a change in the short-term rate only modifies the short-end (and, perhaps, the middle) of the yield curve and, hence, does not produce any variation in the long-end curve.

12.4.4.3 Unspanned stochastic volatility

Unspanned stochastic volatility arises when

$$\frac{\partial}{\partial v} P(r(t), \bar{r}(t), v(t), T-t) = 0.$$

The hypothesis that fixed income markets have unspanned stochastic volatility has been put forward by Collin-Dufresne and Goldstein (2002). Mele (2003) provides conditions under which this occurs.

[In progress]

12.4.5 Affine and quadratic term-structure models

12.4.5.1 Affine

The Vasicek and CIR models predict that the bond price is exponential-affine in the short-term rate r . This property is the expression of a general phenomenon. Indeed, it is possible to show that bond prices are exponential-affine in r if, and only if, the functions b and a^2 are affine in r . Models that satisfy these conditions are known as *affine models*. More generally, these basic results extend to *multifactor* models, where bond prices are exponential-affine in the state variables.¹² In these models, the short-term rate is a function $r(y)$ such that

$$r(y) = r_0 + r_1 \cdot y,$$

where r_0 is a constant, r_1 is a vector, and y is a multidimensional diffusion, in \mathbb{R}^n , and is solution to.

$$dy(\tau) = \kappa(\mu - y(t)) dt + \Sigma V(y(\tau)) dW(\tau), \quad (12.47)$$

where W is a d -dimensional Brownian motion, Σ is a full rank $n \times d$ matrix, and V is a full rank $d \times d$ diagonal matrix with elements,

$$V(y)_{(ii)} = \sqrt{\alpha_i + \beta_i^\top y}, \quad i = 1, \dots, d, \quad (12.48)$$

for some scalars α_i and vectors β_i . Langetieg (1980) develops the first multifactor model of this kind, in which $\beta_i = 0$.

Next, Let $V^-(y)$ be a $d \times d$ diagonal matrix with elements

$$V^-(y)_{(ii)} = \begin{cases} \frac{1}{V(y)_{(ii)}} & \text{if } \Pr\{V(y(t))_{(ii)} > 0 \text{ all } t\} = 1 \\ 0 & \text{otherwise} \end{cases}$$

and set,

$$\Lambda(y) = V(y) \lambda_1 + V^-(y) \lambda_2 y, \quad (12.49)$$

for some d -dimensional vector λ_1 and some $d \times n$ matrix λ_2 . Duffie and Kan (1996) explained in a comprehensive way the benefit of this model. In their formulation $\lambda_2 = 0_{d \times n}$, and the bond price is exponential-affine in the state variables y . That is, the price of the zero has the following functional form,

$$P(y, T-t) = \exp[A(T-t) + B(T-t) \cdot y], \quad (12.50)$$

for some functions A and B of time to maturity, $T-t$ (B is vector-valued), such that $A(0) = 0$ and $B(0)_{(i)} = 0$.

The more general functional form for Λ in Eq. (12.49) has been suggested by Duffee (2002). Duffee noticed that in bond markets, risk-premiums, defined as $\Lambda(y) V(y) = V^2(y) \lambda_1 + \lambda_2 y$,

¹²More generally, we say that affine models are those that make the characteristic function exponential-affine in the state variables. In the case of the multifactor interest rate models of the previous section, this condition is equivalent to the condition that bond prices are exponential affine in the state variables.

are related not only to the *volatility* of fundamentals, but also to the *level* of the fundamentals, which justifies the inclusion of the additional term $\lambda_2 y$. In this case, bond prices still have an exponential affine form, just as in Eq. (12.50). When $\lambda_2 = 0_{d \times n}$, we say that the model is “completely affine,” and “essentially affine,” otherwise. The clear advantage of these affine models, then, is that they considerably simplify statistical inference, as explained in Section 12.4.7 below.

Ang and Piazzesi (2003) (AP, henceforth) and Hördahl, Tristani and Vestin (2006) (HTS, henceforth) introduce “no-arbitrage” regressions, to model the relations linking macroeconomic variables to the yield curve. In their models, the factors are taken to be a discrete-time version of Eq. (12.47), where some components of y are observable, and others are unobservable. The observables relate to macroeconomic factors such as inflation or industrial production. The authors, then, study how all these factors affect the yield curve, predicted by a pricing equation such as that in Eq. (12.50). While HTS have a structural model of the macroeconomy, AP have a reduced-form model.

Reduced-form model can be exposed to the critique that some of the parameters are not “variation-free.” [**Explain what variation-free parameters are, in mathematical statistics**] For example, in the simple Lucas economy of Part I, we know that the short-term rate is $r = \rho + \eta\mu + \frac{1}{2}\sigma^2\eta(1 - \eta)$, so by “tilting” η (risk-aversion), we should also have a change in the interest rate. This simple example shows that the parameters related to risk-aversion correction in Eq. (12.49) are not free to be “tilted,” in that tilting them has an effect on the parameters of the factor dynamics in Eq. (12.47). At the same time, reduced-form model offer a great deal of flexibility, as they do not restrict, so to speak, the model to track any market or economy such as the Lucas economy, say. Moreover, we can always find a theoretical market supporting the no-arb market underlying the reduced-form model. No-arb regressions such as those in AP give the data the power to say which parameter constellation make the model likely to perform, without imposing theoretical restrictions which the data might, then, be likely to reject. For example, the Lucas model, while clearly illustrates that some of the parameters are not variation-free, can be simply wrong, and might impose unreasonable restrictions on the data. For no-arb models, instead, cross-equations restrictions arise through the weaker requirement of absence of arbitrage opportunities.

12.4.5.2 Quadratic

Affine models are known to impose tight conditions on the structure of the volatility of the state variables. These restrictions arise to keep the square root in Eq. (12.48) real valued. But these constraints may hinder the actual performance of the models. There exists another class of models, known as quadratic models, that partially overcome these difficulties.

12.4.6 Short-term rates as jump-diffusion processes

Ahn and Thompson (1988) extend the CIR model to one where the short-term rate is a jump-diffusion process. In general, suppose that the short-term rate is a jump-diffusion process:

$$dr(\tau) = b^J(r(\tau))d\tau + a(r(\tau))dW(\tau) + \ell(r(\tau)) \cdot \mathcal{S} \cdot dZ(\tau),$$

where W and Z are under the risk-neutral probability, and b^J is, then, a jump-adjusted risk-neutral drift. The bond price $P(r, \tau, T)$ is solution to,

$$0 = \left(\frac{\partial}{\partial \tau} + L - r \right) P(r, \tau, T) + v^Q \int_{\text{supp}(\mathcal{S})} [P(r + \ell\mathcal{S}, \tau, T) - P(r, \tau, T)] p(d\mathcal{S}), \quad (12.51)$$

for all $(r, \tau) \in \mathbb{R}_{++} \times [t, T)$, and satisfies the boundary condition, $P(r, T, T) = 1 \forall r \in \mathbb{R}_{++}$. Eq. (12.51) follows because, as usual, $e^{-\int_t^\tau r(u)du} P(r, \tau, T)$ is a martingale under the risk-neutral probability. This model can also be extended to one where there are different quality, or types, of jumps, in which case Eq. (12.51) is:

$$0 = \left(\frac{\partial}{\partial \tau} + L - r \right) P(r, \tau, T) + \sum_{j=1}^N v_j^Q \int_{\text{supp}(\mathcal{S})} [P(r + \ell \mathcal{S}, \tau, T) - P(r, \tau, T)] p^j(d\mathcal{S}),$$

where N is the number of jump types. However, to simplify the exposition, we just set $N = 1$.

To identify risk-premiums related to jumps, we simply note that $v^Q = v \cdot \lambda^J$, where v is the intensity of the short-term rate jump under the *physical* distribution, and λ^J is the risk-premium demanded by agents to be compensated for the presence of jumps.

We assume default is an exogenously given rare event, driven by a Poisson process. Chapter 13 develops a comprehensive account of this approach, known as “reduced-form” approach—to be distinct from the “structural approach,” where the event of default is modeled in regard to the books of the issuer. It is instructive to anticipate some of the main features of this reduced-form approach. We follow the derivation of Mele (2003), relying on partial differential equations. Assume that the event of default at each instant of time is a Poisson process Z with intensity v , and that in the event of default at point τ , the bondholder receives a recovery payment $\bar{P}(\tau)$. This recovery payment can be a bounded deterministic function of time, or more generally a bounded process adapted to the short-term rate. Next, let $\hat{\tau}$ be the random default time, and let’s create an auxiliary state variable g with the following features:

$$g = \begin{cases} 0 & \text{if } t \leq \tau < \hat{\tau} \\ 1 & \text{otherwise} \end{cases}$$

Therefore, we have that under the risk-neutral probability,

$$\begin{cases} dr(\tau) = b(r(\tau))d\tau + a(r(\tau))dW(\tau) \\ dg(\tau) = \mathcal{S} \cdot dN(\tau), \text{ where } \mathcal{S} \equiv 1, \text{ with probability one} \end{cases} \quad (12.52)$$

Denote the rational bond price function with $P(r, g, \tau, T)$, $\tau \in [t, T]$, and assume that $\forall \tau \in [t, T]$ and $\forall v \in (0, \infty)$, $P(r, 1, \tau, T) = \bar{P}(\tau) < P(r, 0, \tau, T)$ and that $\bar{P}(\tau; v') \geq \bar{P}(\tau; v) \Leftrightarrow v' \geq v$ a.s. These assumptions, guarantee that default-free bond prices are higher than defaultable bond prices, as shown below. In the absence of arbitrage, the pre-default bond price $P(r, 0, \tau, T) = P^{\text{pre}}(r, \tau, T)$ satisfies:

$$\begin{aligned} 0 &= \left(\frac{\partial}{\partial \tau} + L - r \right) P(r, 0, \tau, T) + v(r) \cdot [P(r, 1, \tau, T) - P(r, 0, \tau, T)] \\ &= \left(\frac{\partial}{\partial \tau} + L - (r + v(r)) \right) P(r, 0, \tau, T) + v(r)\bar{P}(\tau), \end{aligned} \quad (12.53)$$

for all $\tau \in [t, T)$, and the boundary condition $P(r, 0, T, T) = 1$. The solution is, formally:

$$\begin{aligned} P^{\text{pre}}(x, t, T) &= \mathbb{E}_t^* \left[\exp \left(- \int_t^T (r(\tau) + v(r(\tau))) d\tau \right) \right] \\ &\quad + \mathbb{E}_t^* \left[\int_t^T \exp \left(- \int_t^\tau (r(u) + v(r(u))) du \right) \cdot v(r(\tau)) \bar{P}(\tau) d\tau \right], \end{aligned}$$

where \mathbb{E}_t^* [·] is the expectation taken with respect to only the first equation of system (12.52). This formulation coincides with that in Duffie and Singleton (1999, Eq. (10) p. 696), once we define a percentage loss process l in $[0, 1]$ such that $\bar{P} = (1 - l) \cdot P$. Indeed, inserting $\bar{P} = (1 - l) \cdot P$ into Eq. (12.53) leaves:

$$0 = \left(\frac{\partial}{\partial \tau} + L - (r + l(\tau)v(r)) \right) P(r, 0, \tau, T), \quad \forall (r, \tau) \in \mathbb{R}_{++} \times [t, T),$$

with the usual boundary condition, the solution of which is:

$$P^{\text{pre}}(x, t, T) = \mathbb{E}_t^* \left[\exp \left(- \int_t^T (r(\tau) + l(\tau) \cdot v(r(\tau))) d\tau \right) \right].$$

To validate the claim that P^{pre} is decreasing in v , consider two markets A and B , where the default intensities are v^A and v^B , and assume that the coefficients of L are independent of v^i . The pre-default bond price function in economy i is $P^i(r, \tau, T)$, $i = A, B$, and satisfies:

$$0 = \left(\frac{\partial}{\partial \tau} + L - r \right) P^i + v^i \cdot (\bar{P}^i - P^i), \quad i = A, B,$$

with the usual boundary condition. Assuming that $\bar{P}^A = \bar{P}^B$, subtracting these two equations, and rearranging terms, shows that the price difference $\Delta P(r, \tau, T) \equiv P^A(r, \tau, T) - P^B(r, \tau, T)$ satisfies,

$$0 = \left(\frac{\partial}{\partial \tau} + L - (r + v^A) \right) \Delta P(r, \tau, T) + (v^A - v^B) \cdot [\bar{P}^B(\tau) - P^B(r, \tau, T)],$$

with boundary condition, $\Delta P(r, T, T) = 0$ for all r . Because, clearly, $\bar{P}^B < P^B$, we have that $\Delta P(r, \tau, T) < 0$ whenever $v^A > v^B$, by an application of the maximum principle reviewed in Appendix 3 of Chapter 7.

12.4.7 Some stylized facts and estimation strategies

Does short-term rate volatility really increase with the level of the short-term rate? Consider the following simple model, aiming to address the issue of correlation between the short-term rate and its instantaneous volatility. Let $r(t)$ be the short-term rate process, solution to the following stochastic differential equation,

$$dr(t) = \kappa(\mu - r(t)) dt + \sqrt{v(t)} r(t)^\eta dW(t), \quad t \geq 0, \quad (12.54)$$

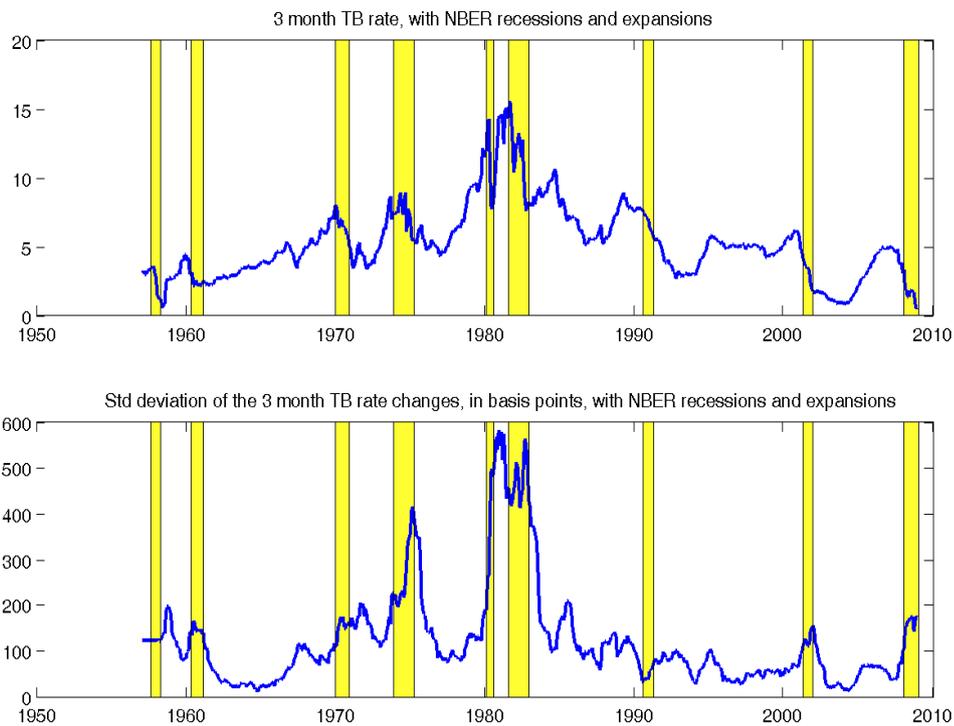
where $W(t)$ is a standard Brownian motion under the physical probability, and κ, μ and η are three positive constants. Suppose, also, that the instantaneous volatility process $\sqrt{v(t)} r(t)^\eta$ is such that $v(t)$ is solution to,

$$dv(t) = \beta(\alpha - v(t)) dt + \xi v(t)^\vartheta \left(\rho dW(t) + \sqrt{1 - \rho^2} dU(t) \right), \quad t \geq 0, \quad (12.55)$$

where $U(t)$ is another standard Brownian motion; β, α, ξ and ϑ are four positive constants, and ρ is a constant such that $|\rho| < 1$. This model generalizes the two-factor model discussed in Section 12.4.4.1, as it allows r and v to be instantaneously correlated.

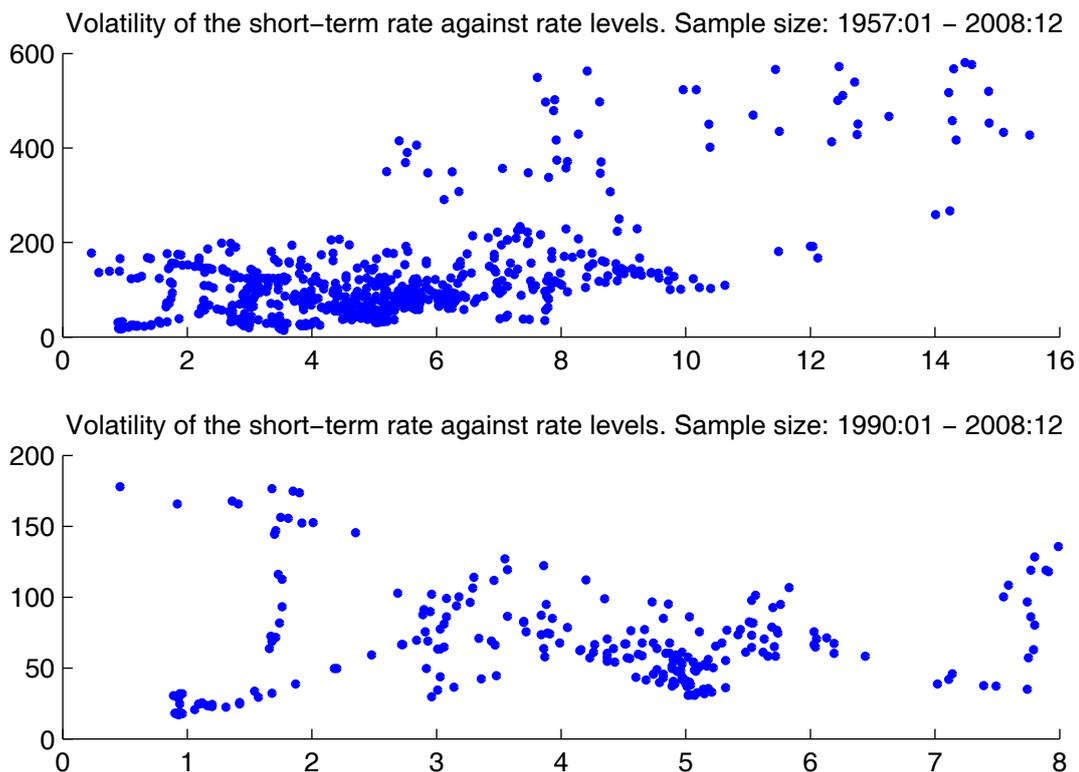
12.4.7.1 The level effect

Which empirical regularities would the short-term rate model in Eqs. (12.54)-(12.55) address? Which sign of the correlation coefficient ρ would be consistent with historical episodes such as the Monetary Experiment of the Federal Reserve System between October 1979 and October 1982? The following picture depicts the time series behavior of the nominal short-term rate, as measured by the three month TB rate, as well as the volatility of its changes, as measured through a formula similar to that in Section 7.2 of Chapter 7: $\text{Vol}_{r,t} \equiv 100^2 \sqrt{6\pi} \cdot \bar{\sigma}_{r,t}$, where $\bar{\sigma}_{r,t} \equiv \frac{1}{12} \sum_{i=1}^{12} |r_{t+1-i} - r_{t-i}|$, and r_t is the short-term rate as of month t . The multiplicative factor 100^2 arises because: (i) the short-rate is converted into percentage points, and (ii) volatility is converted into basis points, as in market conventions.



This picture depicts the time-series behavior of the 3 month TB rate (top panel, in percent) and its rolling, basis point volatility, $\text{Vol}_{r,t}$ (bottom panel), over the sampling period spanning 1957:01 through 2008:12.

The next picture plots a scatterplot of the short-term rate basis point volatility, $\text{Vol}_{r,t}$, against r_t , for two sampling periods.



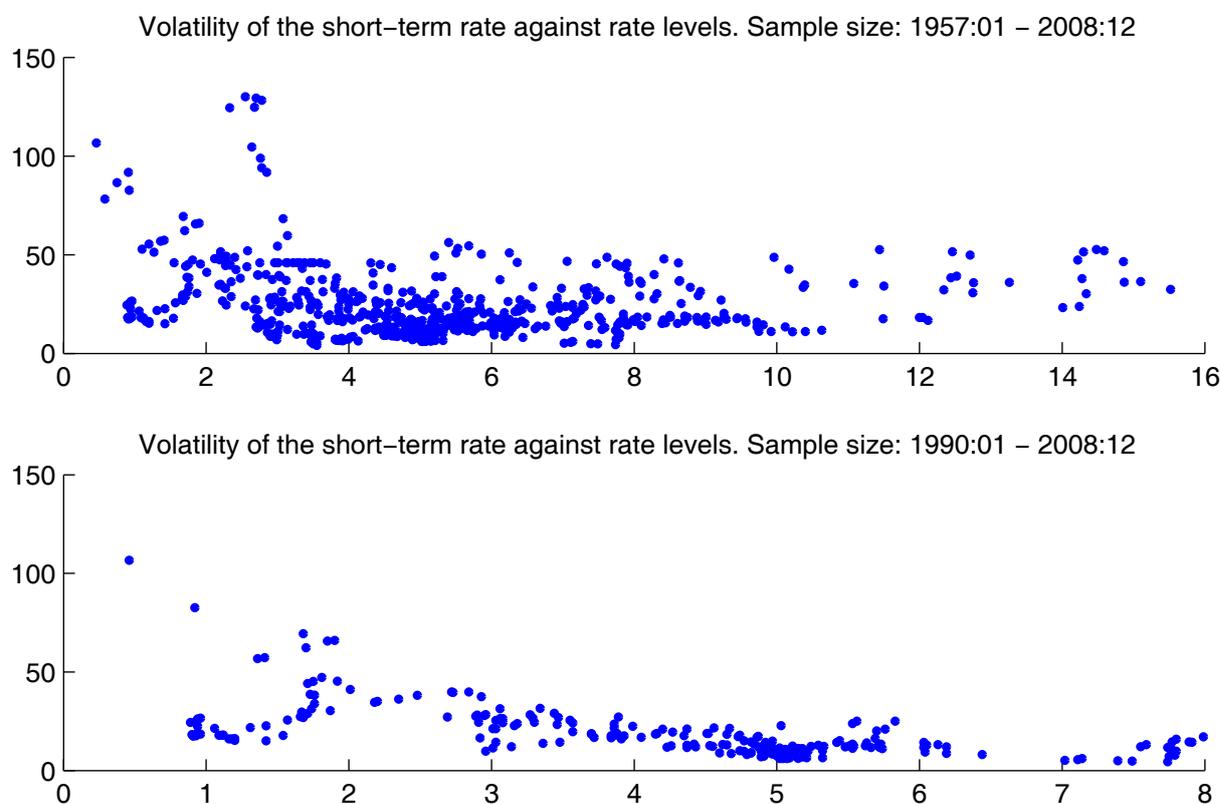
This picture is a scatterplot of the basis point volatility of the short-term rate, $\text{Vol}_{r,t}$ (on the vertical axis), against the level of the short-term rate (on the horizontal axis), in correspondence of the sampling period spanning 1957:01 through 2008:12 (top panel), and a more recent sample spanning 1990:01 through 2008:12 (bottom panel).

The short-term rate model in Eqs. (12.54)-(12.55) would then address two empirical regularities.

- (i) The volatility of the short-term rate is not constant over time. Rather, it seems to be driven by an additional source of randomness. All in all, the short-term process seems to be generated by the stochastic volatility model in Eqs. (12.54)-(12.55), in which the volatility component $v(t)$ is driven by a source of randomness only partially correlated with the source of randomness driving the short-term rate process itself.
- (ii) The “level effect,” i.e. the fact that at times, the volatility of the short-term rate is increasing in the level of the short-term rate. Perhaps, one explanation for the level effect is that when interest rates are high arise when liquidity is erratic, leading to an increase in the risk-premiums. But precisely because of erratic liquidity, interest rates are also very volatile in such periods. Eqs. (12.54)-(12.55) lead to a simple model capable to capture these effects through the two parameters, η and ρ . If $\eta > 0$, the instantaneous rate volatility increases with the level of the interest level. If the “correlation” coefficient $\rho > 0$, rate volatility is also partly related to sources of volatility not directly affected by the *level* of the interest rate.

The previous picture does actually cast doubts on the plausibility of a level effect, as it reveals that the volatility of the short-term rate is not necessarily increasing in the level of the short-term rate. Granted, during the Monetary Experiment, the FED target was essentially money supply, rather than interest rates. As a result, high volatility of money demand mechanically translated to high rate volatility, through market clearing. Moreover, the monetary base was kept deliberately low, as an attempt to fight against inflation. These facts lead to a period where both rate volatility and interest rates very high. Note that one additional reason for the high nominal rates at the time might link to a compensation for high *inflation volatility*—not only high inflation. The previous picture is also suggestive of a change in regime that possibly occurred over a more recent past. From 1990 on, rate volatility does not necessarily appear to positively link to rate levels, and there is evidence of the opposite.

The next picture suggests that “percentage” volatility, defined as $\text{Vol}_{r,t}^{\%} \equiv 100\sqrt{6\pi} \cdot \bar{\sigma}_{r,t}^{\%}$, where $\bar{\sigma}_{r,t}^{\%} \equiv \frac{1}{12} \sum_{i=1}^{12} \left| \ln \frac{r_{t+1-i}}{r_{t-i}} \right|$, is actually inversely related to the level rates, over the more recent sample periods too.



This picture is a scatterplot of the percentage volatility of the short-term rate $\text{Vol}_{r,t}^{\%}$ (on the vertical axis), against the level of the short-term rate (on the horizontal axis), in correspondence of the sampling period spanning 1957:01 through 2008:12 (top panel), and a more recent sample spanning 1990:01 through 2008:12 (bottom panel).

12.4.7.2 A simple case

Next, suppose we wish to estimate the parameter vector $\theta = [\kappa, \mu, \eta, \beta, \alpha, \xi, \vartheta, \rho]^{\top}$ of the model in Eqs. (12.54)-(12.55). Under which circumstances would Maximum Likelihood be a feasible estimation method?

The ML estimator would be feasible under two sets of conditions. *First*, the model in Eqs. (12.54)-(12.55) should not have stochastic volatility at all, viz, $\beta = \xi = 0$; in this case, the short-term rate would be solution to,

$$dr(t) = \kappa(\mu - r(t)) dt + \bar{\sigma} r(t)^\eta dW(t), \quad t \geq 0,$$

where $\bar{\sigma}$ is now a constant. *Second*, the value of the elasticity parameter η is important. If $\eta = 0$, the short-term rate process is the Gaussian one proposed by Vasicek (1977). If $\eta = \frac{1}{2}$, we obtain the square-root process of Cox, Ingersoll and Ross (1985). In the Vasicek case, the transition density of r is Gaussian, and in the CIR case, the transition density of r is a noncentral chi-square. So in both the Vasicek and CIR, we may write down the likelihood function of the diffusion process. Therefore, ML estimation is possible in these two cases. In more general cases, such those in the next section, one needs to go for simulation methods, such as those described in Chapter 5.

12.4.7.3 More general models

Estimating the model in Eqs. (12.54)-(12.55) is certainly instructive. Yet a more important question is to examine the term-structure implications of this model. More generally, how would the estimation procedure outlined in the previous subsection change if the task is to estimate a Markov model of the term-structure of interest rates? There are three steps.

Step 1

Collect data on the term structure of interest rates. We will need to use data on three maturities, say a time series of riskless 6 month, 5 year and 10 year yields.

Step 2

Let us consider the three-factor model in Eqs. (12.46) of Section 12.4.4.2, where the three Brownian motions W_i are now allowed to be correlated. The bond price predicted by this model is:

$$P^j(r(t), v(t), \bar{r}(t)) \equiv P(r(t), v(t), \bar{r}(t), N_j - t) = \mathbb{E} \left(e^{-\int_t^{N_j} r(s) ds} \middle| r(t), v(t), \bar{r}(t) \right), \quad (12.56)$$

where N_j is a sequence of expiration dates. Naturally, this price depends on the risk-aversion corrections needed to turn the dynamics the short-term rate in Eqs. (12.46) into the risk-neutral one. As discussed, one may impose analytically convenient conditions on the risk-adjustments, but we do not need to be more precise at this juncture. No matter the nature of the risk-adjustments, we have that they entail that Eq. (12.56) depends not only on the “physical” parameter vector $\boldsymbol{\theta} = [\kappa_r, \kappa_v, \kappa_{\bar{r}}, \gamma, \bar{v}, \xi_v, \bar{l}, \xi_{\bar{r}}, \rho]^\top$, where ρ is a vector containing all the correlation coefficients of W_i , but also on these very same risk-adjustment parameter vector, say $\boldsymbol{\lambda}$. Precisely, the Radon-Nikodym derivative of the risk-neutral probability with respect to the physical probability is $\exp \left(-\frac{1}{2} \int \|\boldsymbol{\Lambda}(t)\|^2 dt - \int \boldsymbol{\Lambda}(t) d\mathbf{Z}(t) \right)$, for some vector Brownian motion \mathbf{Z} , and $\boldsymbol{\Lambda}(t)$ is some process, assumed to take the form $\boldsymbol{\Lambda}(t) \equiv \boldsymbol{\Lambda}^m(r(t), v(t), \bar{r}(t); \boldsymbol{\lambda})$, for some vector-valued function $\boldsymbol{\Lambda}^m$ and some parameter vector $\boldsymbol{\lambda}$. The function $\boldsymbol{\Lambda}^m$ makes risk-adjustment corrections depend on the current value of the state vector $[r(t), v(t), \bar{r}(t)]$, which makes the model Markov, thereby simplifying statistical inference.

To summarize, the issue is now one where we need to estimate both the physical parameter vector $\boldsymbol{\theta}$ and the “risk-adjustment” parameter vector $\boldsymbol{\lambda}$. Next, we consider the yield curve in

correspondence of three maturities,

$$R^j(r(t), v(t), \bar{r}(t); \boldsymbol{\theta}, \boldsymbol{\lambda}) \equiv -\frac{1}{N_j} \ln P^j(r(t), v(t), \bar{r}(t)), \quad j = 1, 2, 3, \quad (12.57)$$

where the notation $R^j(r, v, \bar{r}; \boldsymbol{\theta}, \boldsymbol{\lambda})$ emphasizes that the theoretical yield curve depends on the parameter vector $(\boldsymbol{\theta}, \boldsymbol{\lambda})$. We can now use actual data, R_s^j say, and the model predictions about the data, R^j , create moment conditions, and proceed to estimate the parameter vector $(\boldsymbol{\theta}, \boldsymbol{\lambda})$ through some method of moments—provided of course the moments are enough to make $(\boldsymbol{\theta}, \boldsymbol{\lambda})$ identifiable. But there are two difficulties. The first is that the volatility process $v(t)$ and the long-term, moving value of the short-term rate, $\bar{r}(t)$, are not directly observable by the econometrician. We can use inference methods based on simulations to cope with this issue. Very simply, we simulate Eqs. (12.46), and apply moment conditions or auxiliary models to observable variables, as explained in Chapter 5. For example, we simulate Eqs. (12.46) for a given value of $(\boldsymbol{\theta}, \boldsymbol{\lambda})$. For each simulation, we compute a time series of interest rates R^j from Eq. (12.57). Then, we use these simulated data to create moment conditions or fit some auxiliary model to these artificial data that is as close as possible to the very same auxiliary model fit to real data. The parameter estimator, then, is the value of $(\boldsymbol{\theta}, \boldsymbol{\lambda})$ minimizing some norm of these moment conditions, obtained through the simulations, with any of the methods explained in Chapter 5. According to Theorem 5.4 in Chapter 5, fitting a sufficiently rich auxiliary model should result in a quite efficient estimator.

A second difficulty is that the bond pricing formula in Eq. (12.56) does not generally admit a closed-form, an issue we can address using affine models, as explained next.

Step 3

The use of *affine* models would considerably simplify the analysis. Affine models place restrictions on the data generating process in Eqs. (12.46) and in the risk-aversion corrections in Eq. (12.56), as originally illustrated by Dai and Singleton (2000), in such a way that the yield curve in Eq. (12.57) is,

$$R^j(r(t), v(t), \bar{r}(t); \boldsymbol{\theta}, \boldsymbol{\lambda}) = A(j; \boldsymbol{\theta}, \boldsymbol{\lambda}) + \mathbf{B}(j; \boldsymbol{\theta}, \boldsymbol{\lambda}) \cdot [r(t), v(t), \bar{r}(t)]^\top, \quad j = 1, 2, 3, \quad (12.58)$$

where $A(j; \boldsymbol{\theta}, \boldsymbol{\lambda})$ and $\mathbf{B}(j; \boldsymbol{\theta}, \boldsymbol{\lambda})$ are some functions of the maturity N_j (\mathbf{B} is vector valued), and generally depend on the parameter vector $(\boldsymbol{\theta}, \boldsymbol{\lambda})$. Once Eqs. (12.46) are simulated, the computation of a time series of yields R^j , then, straightforward, given Eq. (12.58).¹³

12.5 No-arbitrage models: early formulations

12.5.1 Fitting the yield-curve, perfectly

When it comes to price interest rate derivatives consistent with the price of already existing fixed income instruments, we do not really wish to *explain* the yield curve. Rather, as explained in Chapter 11, we wish to *take* it as given. To illustrate, consider a European option written

¹³Dai and Singleton (2000) implement this estimation strategy, although they make use of data on swap rates. The models they consider predict theoretical values for the swap rates, obtained through the formula in Eq. (12.97) of Section 12.7.5.4 below, where the bond prices in that formula are replaced by the pricing functions predicted by the models. Dai & Singleton consider three rates predicted by their models: two swap rates (with tenures of two and ten years), plus the six month Libor rate, $-\frac{1}{2} \ln P(t, t + \frac{1}{2})$, where P is the pricing function predicted by the models they consider.

on a bond. We may find it unsatisfactory to have a model that only “explains” the bond price. A model’s error on the bond price might generate a large error for the option price, due to the nonlinearities induced by the optionality. How can we trust an option pricing model, which is not even capable to pin down the value of the underlying asset? To begin address these points with a simple case, let again $P(r(\tau), \tau, S)$ be the price of a zero coupon bond maturing at some S . By the FTAP, the price, C^b say, of a European option written on this bond, struck at K and expiring at $T < S$, is:

$$C^b(r(t), t, T, S) = \mathbb{E}_t \left[e^{-\int_t^T r(\tau) d\tau} \cdot (P(r(T), T, S) - K)^+ \right].$$

For example, affine models predict that the bond price P is conditionally lognormally distributed, provided r is conditionally normally distributed, as in the case of Vasicek model. Insights from the Black and Scholes (1973) formula suggest, then, that in this case, the previous expectation is a nonlinear function of the *current* bond price $P(r(t), t, T)$. However, we cannot make use of the standard tools leading to the Black & Scholes formula. The main issue arising whilst evaluating fixed-income instruments such as simple option on a zero, is that their payoffs at expiry, $(P(r(T), T, S) - K)^+$ in our case, depends on the short-term rate at T , and yet the discounting factor, $e^{-\int_t^T r(\tau) d\tau}$, would also obviously depend on the realization of the short-term rate. However, the problem can be quite tractable, and addressed through the forward martingale probability introduced in Section 12.2. Let \mathbb{I}_{exe} be the indicator function which is always zero, or one when the option is exercised, i.e. when $P(r(T), T, S) \geq K$. We have:

$$\begin{aligned} C^b(r(t), t, T, S) &= \mathbb{E}_t \left[e^{-\int_t^T r(\tau) d\tau} P(r(T), T, S) \cdot \mathbb{I}_{\text{exe}} \right] - K \cdot \mathbb{E}_t \left[e^{-\int_t^T r(\tau) d\tau} \cdot \mathbb{I}_{\text{exe}} \right] \\ &= P(r(t), t, S) \cdot \mathbb{E}_t \left[\frac{e^{-\int_t^S r(\tau) d\tau}}{P(r(t), t, S)} \cdot \mathbb{I}_{\text{exe}} \right] - KP(r(t), t, T) \cdot \mathbb{E}_t \left[\frac{e^{-\int_t^T r(\tau) d\tau}}{P(r(t), t, T)} \cdot \mathbb{I}_{\text{exe}} \right] \\ &= P(r(t), t, S) \cdot \mathbb{E}_{Q_F^S} [\mathbb{I}_{\text{exe}}] - KP(r(t), t, T) \cdot \mathbb{E}_{Q_F^T} [\mathbb{I}_{\text{exe}}] \\ &= P(r(t), t, S) \cdot Q_F^S [P(r(T), T, S) \geq K] - KP(r(t), t, T) \cdot Q_F^T [P(r(T), T, S) \geq K], \quad (12.59) \end{aligned}$$

where the second equality follows by an argument nearly identical to that produced in Section 12.2.2.2 (see Footnote 1);¹⁴ Q_F^i ($i = T, S$) is the i -forward probability; and, finally, $\mathbb{E}_{Q_F^i} [\cdot]$ is the expectation taken under the i -forward martingale probability, as defined in Section 12.2.3.

Section 12.8 explains how the two probabilities in Eq. (12.59) are computed. The important issue, now, is to emphasize that the bond option price does depend on the *theoretical* bond prices $P(r(t), t, T)$ and $P(r(t), t, S)$, which, in turn, cannot equal the *current*, observed market prices. Theoretical prices are, after all, the output of a rational expectations model. This fact is obviously not a source of concern to those who wish to predict future term-structure movements with the help of a few, key state variables, as in the multifactor models discussed earlier.

¹⁴By the Law of Iterated Expectations,

$$\mathbb{E}_t \left[e^{-\int_t^T r(\tau) d\tau} P(r(T), T, S) \mathbb{I}_{\text{exe}} \right] = \mathbb{E}_t \left[e^{-\int_t^T r(\tau) d\tau} \mathbb{I}_{\text{exe}} \mathbb{E} \left(e^{-\int_T^S r(\tau) d\tau} \middle| F(T) \right) \right] = \mathbb{E}_t \left[e^{-\int_t^S r(\tau) d\tau} \mathbb{I}_{\text{exe}} \right].$$

However, a source of concern to sell-side practitioners might be that the option should be priced with a model that simultaneously matches the yield curve, at the time of evaluation. The aim of this section is to introduce a class of models that fit the yield curve without errors, which we call “perfectly fitting models.” These models are simply a more elaborated, continuous-time version of the no-arbitrage models introduced in Chapter 11. They predict that the price of any bond, say a bond expiring at some S , is, of course, random, at time $T < S$, but also exactly equal to the *current* market price, that of time t . Finally, and naturally, this price must be arbitrage-free. We now show that conditions can be met by augmenting the models seen in the previous sections with a set of “infinite dimensional parameters.” We begin with a discussion of two specific and old, and yet famous examples addressing these issues: the Ho and Lee (1986) model, and one generalization of it, introduced by Hull and White (1990). In Section 12.6, we move on towards a general model-building principle.

A final remark. In Section 12.8, we shall show that at least for the Vasicek model, Eq. (12.59) does not explicitly depend on r because it only “depends” on $P(r(t), t, T)$ and $P(r(t), t, S)$. So why do we look for perfectly fitting models in the first place? Wouldn’t it be enough, then, to just replace the theoretical prices $P(r(t), t, T)$ and $P(r(t), t, S)$ with the market values, say $P^s(t, T)$ and $P^s(t, S)$? This way, the model *is* perfectly fitting. Apart from being logically inconsistent (you would have a model predicting something generically different from prices), this way of proceeding also has practical drawbacks. Section 12.8 shows that option pricing formulae for European options, might well agree “in notation” with those relating to perfectly fitting models. However, Section 12.8.3 explains that as we move towards more complex interest rate derivatives products, such as options on *coupon* bonds and *swaption* contracts, the situation gets dramatically different. Finally, it can be the case that some maturity dates are actually not traded at some point in time. For example, it may happen that $P^s(t, T)$ is not observed and that we could still be interested in pricing more “exotic” or less liquid bonds or options on these bonds. An intuitive procedure to deal with this difficulty is to “interpolate” the traded maturities. In fact, the objective of perfectly fitting models is to allow for such an “interpolation” while preserving absence of arbitrage opportunities.

12.5.2 Ho & Lee

The original Ho and Lee (1986) model is in discrete-time and is analyzed in the context of Chapter 11, along with other models. The model below, represents the “diffusion limit” of the original Ho & Lee model, as put forward in Section 11.6.6 of Chapter 11:

$$dr(\tau) = \theta(\tau) d\tau + \sigma d\tilde{W}(\tau), \quad \tau \geq t, \quad (12.60)$$

where t is the time of evaluation, \tilde{W} is a Brownian motion under Q , σ is a constant, and $\theta(\tau)$ is an “infinite dimensional” parameter, which we need to pin down the initial, observed yield curve, as we now explain. The reason we refer to $\theta(\tau)$ as “infinite dimensional” parameter is that we assume $\theta(\tau)$ is a function of calendar time $\tau \geq t$. We assume this function is known at t . Clearly, Eq. (12.60) defines an affine model. Therefore, the bond price takes the following form,

$$P(r(\tau), \tau, T) = e^{A(\tau, T) - B(\tau, T) \cdot r(\tau)}, \quad (12.61)$$

for two functions A and B to be determined below. It is easy to show that,

$$A(\tau, T) = \int_{\tau}^T \theta(s) (s - T) ds + \frac{1}{6} \sigma^2 (T - \tau)^3, \quad B(\tau, T) = T - \tau.$$

Let $f_{\S}(t, \tau)$ denote the instantaneous, observed forward rate. By matching the instantaneous forward rate $f(\tau, T)$ predicted by the model to $f_{\S}(\tau, T)$ yields:

$$f_{\S}(t, T) = f(t, T) = -\frac{\partial \ln P(r(t), t, T)}{\partial T} = \int_t^T \theta(s) ds - \frac{1}{2} \sigma^2 (T - t)^2 + r(t). \quad (12.62)$$

Because $P(t, T) = \exp\left(-\int_t^T f(t, \tau) d\tau\right)$, the drift term $\theta(s)$ satisfying Eq. (12.62) guarantees an exact fit of the yield curve. By differentiating Eq. (12.62) with respect to T , leaves $\theta(T) = \frac{\partial}{\partial T} f_{\S}(t, T) + \sigma^2 (T - t)$, or:

$$\theta(\tau) = \frac{\partial}{\partial \tau} f_{\S}(t, \tau) + \sigma^2 (\tau - t). \quad (12.63)$$

To check that θ is indeed the solution we were looking for, we replace Eq. (12.63) into Eq. (12.62) and verify indeed that Eq. (12.62) holds as an identity. By Eqs. (12.60) and (12.63), the short-term rate is, then:

$$r(t) = f_{\S}(0, t) + \frac{1}{2} \sigma^2 t + \sigma \tilde{W}(t).$$

Moreover, by Eq. (12.62), and Eq. (12.60), the instantaneous forward rate satisfies,

$$d_{\tau} f(\tau, T) = -\theta(\tau) d\tau + \sigma^2 (T - \tau) d\tau + dr(\tau) = \sigma^2 (T - \tau) d\tau + \sigma d\tilde{W}(t).$$

These results are the continuous time counterparts to those introduced in Section 11.6.6 of the previous chapter. In Section 12.6, they will be shown to be a particular case of a general framework, known as the HJM.

12.5.3 Hull & White

Hull and White (1990) consider the following model:

$$dr(\tau) = \kappa \left(\frac{\theta(\tau)}{\kappa} - r(\tau) \right) d\tau + \sigma d\tilde{W}(\tau), \quad (12.64)$$

where \tilde{W} is a Q -Brownian motion, and κ, σ are constants. The model generalizes the Ho and Lee model (1986) in Eq. (12.60) and the Vasicek (1977) model in Eq. (12.34). In the original formulation of Hull and White, κ and σ are both time-varying, but the main points of this model can be learnt by working out this particularly simple case.

Eq. (12.64) also gives rise to an affine model. Therefore, the solution for the bond price is given by Eq. (12.61). It is easy to show that the functions A and B are given by

$$A(\tau, T) = \frac{1}{2} \sigma^2 \int_{\tau}^T B^2(s, T) ds - \int_{\tau}^T \theta(s) B(s, T) ds, \quad (12.65)$$

and

$$B(\tau, T) = \frac{1}{\kappa} [1 - e^{-\kappa(T-\tau)}]. \quad (12.66)$$

By reiterating the same reasoning produced to show (12.63), one shows that the solution for θ is:

$$\theta(\tau) = \frac{\partial}{\partial \tau} f_{\S}(t, \tau) + \kappa f_{\S}(t, \tau) + \frac{\sigma^2}{2\kappa} [1 - e^{-2\kappa(\tau-t)}]. \quad (12.67)$$

A proof of this result is in Appendix 5.

Why did we need to go for this more complex model? After all, the Ho & Lee model is already able to pin down the entire yield curve. The answer is that in practice, investment banks typically prices a large variety of derivatives. The yield curve is not the only thing to be exactly fit. Rather it is only the starting point. In general, the more flexible a given perfectly fitting model is, the more successful it is to price more complex derivatives.

12.6 The Heath-Jarrow-Morton framework

12.6.1 Framework

The bond price representation in Eq. (12.3),

$$P(\tau, T) = e^{-\int_{\tau}^T f(\tau, \ell) d\ell}, \quad \text{all } \tau \in [t, T], \quad (12.68)$$

underlies the modeling approach started by Heath, Jarrow and Morton (1992) (HJM, henceforth). Given Eq. (12.68), this approach takes as a primitive the stochastic evolution of the entire structure of forward rates, not only the special case of the short-term rate, $r(t) = \lim_{\ell \downarrow t} f(t, \ell) \equiv f(t, t)$. The goal is to start with Eq. (12.68) and take the initial, observed structure of forward rates $\{f(t, \ell)\}_{\ell \in [t, T]}$, as given, and find, then, no-arb, “cross-equation” restrictions on the stochastic behavior of $\{f(\tau, \ell)\}_{\tau \in (t, \ell]}$, for any $\ell \in [t, T]$.

By construction, the HJM approach *allows for a perfect fit of the initial term-structure*. This point can be illustrated quite simply, as the bond price $P(\tau, T)$ is,

$$\begin{aligned} P(\tau, T) &= e^{-\int_{\tau}^T f(\tau, \ell) d\ell} \\ &= \frac{P(t, T)}{P(t, \tau)} \cdot \frac{P(t, \tau)}{P(t, T)} e^{-\int_{\tau}^T f(\tau, \ell) d\ell} \\ &= \frac{P(t, T)}{P(t, \tau)} \cdot e^{-\int_t^{\tau} f(t, \ell) d\ell + \int_t^T f(t, \ell) d\ell - \int_{\tau}^T f(\tau, \ell) d\ell} \\ &= \frac{P(t, T)}{P(t, \tau)} \cdot e^{\int_{\tau}^T f(t, \ell) d\ell - \int_{\tau}^T f(\tau, \ell) d\ell} \\ &= \frac{P(t, T)}{P(t, \tau)} \cdot e^{-\int_{\tau}^T [f(\tau, \ell) - f(t, \ell)] d\ell}. \end{aligned}$$

The key point of the HJM methodology is to take the current forward rates structure $f(t, \ell)$ as given, i.e. perfectly fitted, and to model, then, the future forward rate movements,

$$f(\tau, \ell) - f(t, \ell).$$

Therefore, the HJM methodology takes the current term-structure as perfectly fitted, as we observe both $P(t, T)$ and $P(t, \tau)$. In contrast, the approach to interest rate modeling in Section 12.3, is to *model* the *current* bond price $P(t, T)$ through assumptions relating to developments in the short-term rate. For this reason, these models of the short-term rate do not fit the initial term structure. As explained in the previous chapter, and in the previous section, fitting the initial term-structure is, instead, critical, when it comes to price interest-rate derivatives.

12.6.2 The model

12.6.2.1 Primitives

Because the primitive is still a Brownian information structure, once we want to model future movements of $\{f(\tau, T)\}_{\tau \in [t, T]}$, we also have to accept that for every T , $\{f(\tau, T)\}_{\tau \in [t, T]}$ is $\mathcal{F}(\tau)$ -adapted. Thus, there exist functionals α and σ such that, for a given T ,

$$d_\tau f(\tau, T) = \alpha(\tau, T) d\tau + \sigma(\tau, T) dW(\tau), \quad \tau \in (t, T], \quad (12.69)$$

where $f(t, T)$ is given. The solution to Eq. (12.69) is:

$$f(\tau, T) = f(t, T) + \int_t^\tau \alpha(s, T) ds + \int_t^\tau \sigma(s, T) dW(s), \quad \tau \in (t, T]. \quad (12.70)$$

In other terms, W “doesn’t depend” on T . In some sense, however, we may also want to “index” W by T . The so-called *stochastic string models* are capable of doing that, and are discussed in Section 12.8.

12.6.2.2 No-arb restrictions

The next step is to derive restrictions on α that rule out arbitrage. Let $X(\tau) \equiv - \int_\tau^T f(\tau, \ell) d\ell$. We have

$$dX(\tau) = f(\tau, \tau) d\tau - \int_\tau^T (d_\tau f(\tau, \ell)) d\ell = [r(\tau) - \alpha^I(\tau, T)] d\tau - \sigma^I(\tau, T) dW(\tau),$$

where

$$\alpha^I(\tau, T) \equiv \int_\tau^T \alpha(\tau, \ell) d\ell, \quad \sigma^I(\tau, T) \equiv \int_\tau^T \sigma(\tau, \ell) d\ell.$$

By Eq. (12.68), $P = e^X$. By Itô’s lemma,

$$\frac{d_\tau P(\tau, T)}{P(\tau, T)} = \left[r(\tau) - \alpha^I(\tau, T) + \frac{1}{2} \|\sigma^I(\tau, T)\|^2 \right] d\tau - \sigma^I(\tau, T) dW(\tau).$$

By the FTAP, there are no arbitrage opportunities if and only if

$$\frac{d_\tau P(\tau, T)}{P(\tau, T)} = \left[r(\tau) - \alpha^I(\tau, T) + \frac{1}{2} \|\sigma^I(\tau, T)\|^2 + \sigma^I(\tau, T) \lambda(\tau) \right] d\tau - \sigma^I(\tau, T) d\tilde{W}(\tau),$$

where $\tilde{W}(\tau) = W(\tau) + \int_t^\tau \lambda(s) ds$ is a Q -Brownian motion, and λ satisfies:

$$\alpha^I(\tau, T) = \frac{1}{2} \|\sigma^I(\tau, T)\|^2 + \sigma^I(\tau, T) \lambda(\tau). \quad (12.71)$$

By differentiating the previous relation with respect to T gives us the arbitrage restriction that we were looking for:

$$\alpha(\tau, T) = \sigma(\tau, T) \int_\tau^T \sigma(\tau, \ell)^\top d\ell + \sigma(\tau, T) \lambda(\tau). \quad (12.72)$$

12.6.3 The dynamics of the short-term rate

By Eq. (12.70), the short-term rate satisfies:

$$r(\tau) \equiv f(\tau, \tau) = f(t, \tau) + \int_t^\tau \alpha(s, \tau) ds + \int_t^\tau \sigma(s, \tau) dW(s), \quad \tau \in (t, T]. \quad (12.73)$$

Differentiating with respect to τ yields

$$dr(\tau) = \left[f_2(t, \tau) + \sigma(\tau, \tau)\lambda(\tau) + \int_t^\tau \alpha_2(s, \tau) ds + \int_t^\tau \sigma_2(s, \tau) dW(s) \right] d\tau + \sigma(\tau, \tau) dW(\tau),$$

where

$$\alpha_2(s, \tau) = \sigma_2(s, \tau) \int_s^\tau \sigma(s, \ell)^\top d\ell + \sigma(s, \tau) \sigma(s, \tau)^\top + \sigma_2(s, \tau) \lambda(s).$$

As is clear, the short-term rate is in general non-Markov. However, the short-term rate can be “risk-neutralized,” and used to price exotics through simulations. A special case of Eq. (12.73) is the Ho and Lee model, where $\sigma(s, \tau) = \sigma$, a constant, such that, by Eq. (12.72), $\alpha(s, \tau) = \sigma^2(\tau - s) + \sigma\lambda(\tau)$.

12.6.4 Embedding

At first glance, it might be guessed that HJM models are quite distinct from the models of the short-term rate introduced in Section 12.4. However, there exist “embeddability” conditions turning HJM into short-term rate models, and viceversa, a property known as “universality” of HJM models.

12.6.4.1 Markovianity

One natural question to ask is whether there are conditions under which HJM-type models predict the short-term rate to be a Markov process. The question is natural insofar as it relates to the early literature surveyed in Section 12.4, where the whole yield curve is driven by a scalar Markov process: the short-term rate. The answer to this question is in the contribution of Carverhill (1994). Another important contribution in this area is due to Ritchken and Sankarasubramanian (1995), who studied conditions under which it is possible to enlarge the original state vector in such a manner that the resulting “augmented” state vector is Markov and at the same time, includes that short-term rate as a component. The resulting model quite resembles some of the short-term rate models surveyed in Section 12.4. In these models, the short-term rate is not Markov, yet it is part of a system that is Markov. Here we only consider the simple Markov scalar case.

Assume the forward-rate volatility is deterministic and takes the following form:

$$\sigma(t, T) = g_1(t)g_2(T) \quad \text{all } t, T. \quad (12.74)$$

By Eq. (12.73), r is then:

$$r(\tau) = f(t, \tau) + \int_t^\tau \alpha(s, \tau) ds + g_2(\tau) \cdot \int_t^\tau g_1(s) dW(s), \quad \tau \in (t, T],$$

such that

$$\begin{aligned}
dr(\tau) &= \left[f_2(t, \tau) + \sigma(\tau, \tau)\lambda(\tau) + \int_t^\tau \alpha_2(s, \tau)ds + g_2'(\tau) \int_t^\tau g_1(s)dW(s) \right] d\tau + \sigma(\tau, \tau)dW(\tau) \\
&= \left[f_2(t, \tau) + \sigma(\tau, \tau)\lambda(\tau) + \int_t^\tau \alpha_2(s, \tau)ds + \frac{g_2'(\tau)}{g_2(\tau)}g_2(\tau) \int_t^\tau g_1(s)dW(s) \right] d\tau + \sigma(\tau, \tau)dW(\tau) \\
&= \left[f_2(t, \tau) + \sigma(\tau, \tau)\lambda(\tau) + \int_t^\tau \alpha_2(s, \tau)ds + \frac{g_2'(\tau)}{g_2(\tau)} \left(r(\tau) - f(t, \tau) - \int_t^\tau \alpha(s, \tau)ds \right) \right] d\tau + \sigma(\tau, \tau)dW(\tau).
\end{aligned}$$

Done. This is Markov. Precisely, the condition in Eq. (12.74) ensures the HJM model predicts the short-term rate is Markov. Mean reversion, then, obtains assuming that $g_2'(T) < 0$ for all T . For example, take λ to be a constant, and:

$$g_1(t) = \sigma \cdot e^{\kappa t}, \quad \sigma > 0, \quad g_2(t) = e^{-\kappa t}, \quad \kappa \geq 0.$$

This is the Hull-White model discussed in Section 12.4, and of course, the Ho and Lee model obtains in the special case $\kappa = 0$.

12.6.4.2 Short-term rate reductions

We prove everything in the Markov case. Let the short-term rate be solution to:

$$dr(\tau) = \bar{b}(\tau, r(\tau))d\tau + a(\tau, r(\tau))d\tilde{W}(\tau),$$

where \tilde{W} is a Q -Brownian motion, and \bar{b} is some risk-neutralized drift function. The rational bond price function is $P(r(t), t, T)$, and the forward rate implied by the model is:

$$f(r(t), t, T) = -\frac{\partial}{\partial T} \ln P(r(t), t, T).$$

By Itô's lemma,

$$df = \left[\frac{\partial}{\partial t} f + \bar{b}f_r + \frac{1}{2}a^2 f_{rr} \right] d\tau + af_r d\tilde{W}.$$

But for $f(r, t, T)$ to be consistent with the solution to Eq. (12.70), it must be the case that

$$\begin{aligned}
\alpha(t, T) - \sigma(t, T)\lambda(t) &= \frac{\partial}{\partial t} f(r, t, T) + \bar{b}(t, r)f_r(r, t, T) + \frac{1}{2}a(t, r)^2 f_{rr}(r, t, T) \\
\sigma(t, T) &= a(t, r)f_r(t, r)
\end{aligned} \tag{12.75}$$

and

$$f(t, T) = f(r, t, T). \tag{12.76}$$

In particular, the last condition can only be satisfied if the short-term rate model under consideration is of the perfectly fitting type.

12.7 Stochastic string shocks models

The first papers are Kennedy (1994, 1997), Goldstein (2000) and Santa-Clara and Sornette (2001). Heaney and Cheng (1984) are also very useful to read.

12.7.1 Addressing stochastic singularity

Let $\sigma(\tau, T) = [\sigma_1(\tau, T), \dots, \sigma_N(\tau, T)]$ in Eq. (12.69). For any $T_1 < T_2$,

$$E[df(\tau, T_1)df(\tau, T_2)] = \sum_{i=1}^N \sigma_i(\tau, T_1) \sigma_i(\tau, T_2) d\tau,$$

and,

$$c(\tau, T_1, T_2) \equiv \text{corr}[df(\tau, T_1)df(\tau, T_2)] = \frac{\sum_{i=1}^N \sigma_i(\tau, T_1) \sigma_i(\tau, T_2)}{\|\sigma(\tau, T_1)\| \cdot \|\sigma(\tau, T_2)\|}. \quad (12.77)$$

By replacing this result into Eq. (12.72),

$$\begin{aligned} \alpha(\tau, T) &= \int_{\tau}^T \sigma(\tau, T) \cdot \sigma(\tau, \ell)^{\top} d\ell + \sigma(\tau, T)\lambda(\tau) \\ &= \int_{\tau}^T \|\sigma(\tau, \ell)\| \|\sigma(\tau, T)\| c(\tau, \ell, T) d\ell + \sigma(\tau, T)\lambda(\tau). \end{aligned}$$

One drawback of this model is that the correlation matrix of any $(N + M)$ -dimensional vector of forward rates is degenerate for $M \geq 1$. Stochastic string models overcome this difficulty by *modeling* in an independent way the correlation structure $c(\tau, \tau_1, \tau_2)$ for all τ_1 and τ_2 rather than *implying* it from a given N -factor model (as in Eq. (12.77)). In other terms, the HJM methodology uses functions σ_i to accommodate both volatility and correlation structure of forward rates. This is unlikely to be a good model in practice. As we will now see, stochastic string models have two separate functions with which to model volatility and correlation.

The starting point is a model where the forward rate is solution to,

$$d_{\tau}f(\tau, T) = \alpha(\tau, T) d\tau + \sigma(\tau, T) d_{\tau}Z(\tau, T),$$

where the *string* Z satisfies the following five properties:

- (i) For all τ , $Z(\tau, T)$ is continuous in T ;
- (ii) For all T , $Z(\tau, T)$ is continuous in τ ;
- (iii) $Z(\tau, T)$ is a τ -martingale and, hence, a local martingale i.e. $E[d_{\tau}Z(\tau, T)] = 0$;
- (iv) $\text{var}[d_{\tau}Z(\tau, T)] = d\tau$;
- (v) $\text{cov}[d_{\tau}Z(\tau, T_1) d_{\tau}Z(\tau, T_2)] = \psi(T_1, T_2)$ (say).

Properties (iii), (iv) and (v) make Z Markovian. The functional form for ψ is crucially important to guarantee this property. Given the previous properties, we can deduce a key property of the forward rates. We have,

$$\begin{aligned} \sqrt{\text{var}[df(\tau, T)]} &= \sigma(\tau, T) \\ c(\tau, T_1, T_2) &\equiv \text{corr}[df(\tau, T_1)df(\tau, T_2)] = \frac{\sigma(\tau, T_1) \sigma(\tau, T_2) \psi(T_1, T_2)}{\sigma(\tau, T_1) \sigma(\tau, T_2)} = \psi(T_1, T_2) \end{aligned}$$

As claimed before, we now have two separate functions with which to model volatility and correlation.

12.7.2 No-arbitrage restrictions

Similarly as in the HJM-Brownian case, let $X(\tau) \equiv -\int_{\tau}^T f(\tau, \ell) d\ell$. We have,

$$dX(\tau) = f(\tau, \tau) d\tau - \int_{\tau}^T d_{\tau}f(\tau, \ell) d\ell = [r(\tau) - \alpha^I(\tau, T)] d\tau - \int_{\tau}^T [\sigma(\tau, \ell) d_{\tau}Z(\tau, \ell)] d\ell,$$

where as usual, $\alpha^I(\tau, T) \equiv \int_{\tau}^T \alpha(\tau, \ell) d\ell$. But $P(\tau, T) = \exp(X(\tau))$. Therefore,

$$\begin{aligned} \frac{dP(\tau, T)}{P(\tau, T)} &= dX(\tau) + \frac{1}{2} \text{var}[dX(\tau)] \\ &= \left[r(\tau) - \alpha^I(\tau, T) + \frac{1}{2} \int_{\tau}^T \int_{\tau}^T \sigma(\tau, \ell_1) \sigma(\tau, \ell_2) \psi(\ell_1, \ell_2) d\ell_1 d\ell_2 \right] d\tau \\ &\quad - \int_{\tau}^T [\sigma(\tau, \ell) d_{\tau}Z(\tau, \ell)] d\ell. \end{aligned}$$

Next, suppose that the pricing kernel ξ satisfies:

$$\frac{d\xi(\tau)}{\xi(\tau)} = -r(\tau) d\tau - \int_{\mathbb{T}} \phi(\tau, T) d_{\tau}Z(\tau, T) dT,$$

where \mathbb{T} denotes the set of all “risks” spanned by the string Z , and ϕ is the corresponding family of “unit risk-premia.”

By absence of arbitrage opportunities,

$$0 = E[d(P\xi)] = E \left[P\xi \cdot \left(\text{drift} \left(\frac{dP}{P} \right) + \text{drift} \left(\frac{d\xi}{\xi} \right) + \text{cov} \left(\frac{dP}{P}, \frac{d\xi}{\xi} \right) \right) \right].$$

By exploiting the dynamics of P and ξ ,

$$\alpha^I(\tau, T) = \frac{1}{2} \int_{\tau}^T \int_{\tau}^T \sigma(\tau, \ell_1) \sigma(\tau, \ell_2) \psi(\ell_1, \ell_2) d\ell_1 d\ell_2 + \text{cov} \left(\frac{dP}{P}, \frac{d\xi}{\xi} \right),$$

where

$$\begin{aligned} \text{cov} \left(\frac{dP}{P}, \frac{d\xi}{\xi} \right) &= E \left[\int_{\mathbb{T}} \phi(\tau, S) d_{\tau}Z(\tau, S) dS \cdot \int_{\tau}^T \sigma(\tau, \ell) d_{\tau}Z(\tau, \ell) d\ell \right] \\ &= \int_{\tau}^T \int_{\mathbb{T}} \phi(\tau, S) \sigma(\tau, \ell) \psi(S, \ell) dS d\ell. \end{aligned}$$

By differentiating α^I with respect to T we obtain,

$$\alpha(\tau, T) = \int_{\tau}^T \sigma(\tau, \ell) \sigma(\tau, T) \psi(\ell, T) d\ell + \sigma(\tau, T) \int_{\mathbb{T}} \phi(\tau, S) \psi(S, T) dS. \quad (12.78)$$

A proof of Eq. (12.78) is in the Appendix.

12.8 Interest rate derivatives

12.8.1 Introduction

Options on bonds, caps and swaptions are the main interest rate derivatives traded in the market. The purpose of this section is to price these assets. In principle, the pricing problem could be solved very elegantly. Let w denote the value of any of such instrument, and π be the *instantaneous* payoff process paid by it. Consider any model of the short-term rate considered in Section 12.4. To simplify, assume that $d = 1$, and that all uncertainty is subsumed by the short-term rate process in Eq. (12.28). By the FTAP, w is then the solution to the following partial differential equation:

$$0 = \frac{\partial w}{\partial \tau} + \bar{b}w_r + \frac{1}{2}a^2w_{rr} + \pi - rw, \quad \text{for all } (r, \tau) \in \mathbb{R}_{++} \times [t, T] \quad (12.79)$$

subject to some appropriate boundary conditions. In the previous PDE, \bar{b} is some risk-neutralized drift function of the short-term rate. The additional π term arises because to the average instantaneous increase rate of the derivative, viz $\frac{\partial w}{\partial \tau} + \bar{b}w_r + \frac{1}{2}a^2w_{rr}$, one has to add its payoff π . The sum of these two terms must equal rw to avoid arbitrage opportunities. In many applications considered below, the payoff π can be *approximated* by a function of the short-term rate itself $\pi(r)$. However, such an approximation is at odds with standard practice. Market participants define the payoffs of interest-rate derivatives in terms of LIBOR discretely-compounded rates. Moreover, intermediate payments do not occur continuously, only discretely. The aim of this section is to present more models that are more realistic than those emanating from Eq. (12.79).

The next section introduces notation to cope expeditiously with the pricing of these interest rate derivatives. Section 12.8.3 shows how to price options within the Gaussian models discussed in Section 12.4. Section 12.8.4 provides precise definitions of the remaining most important fixed-income instruments: fixed coupon bonds, floating rate bonds, interest rate swaps, caps, floors and swaptions. It also provides exact solutions based on short-term rate models. Finally, Section 12.8.5 presents the “market model,” which is a HJM-style model intensively used by practitioners.

12.8.2 A put-call parity for fixed income markets

Consider the identity,

$$[K - P(T, S)]^+ \equiv [P(T, S) - K]^+ + K - P(T, S), \quad T \leq S.$$

Taking risk-neutral, discounted expectations of both sides of this equation leaves,

$$\begin{aligned} & \mathbb{E}_t \left[e^{-\int_t^T r(\tau) d\tau} (K - P(T, S))^+ \right] \\ &= \mathbb{E}_t \left[e^{-\int_t^T r(\tau) d\tau} (P(T, S) - K)^+ \right] + P(t, T)K - \mathbb{E}_t \left[e^{-\int_t^T r(\tau) d\tau} P(T, S) \right] \\ &= \mathbb{E}_t \left[e^{-\int_t^T r(\tau) d\tau} (P(T, S) - K)^+ \right] + P(t, T)K - P(t, S), \end{aligned}$$

where the last equality follows by the same argument leading to Eq. (12.59). Therefore, we have the put-call parity relation:

$$\text{Put}(t, T; P(t, S), K) = \text{Call}(t, T; P(t, S), K) + P(t, T)K - P(t, S), \quad (12.80)$$

where $\text{Put}(t, T; P(t, S), K)$ is the price of a European put written on a zero expiring at time S , expiring at time $T < S$, and struck at K , and $\text{Call}(\cdot)$ denotes the corresponding call price.

Likewise, let $B(t, S)$ be the price of a coupon bearing bond, such that the put-call parity for options written on it reads as:

$$\text{Put}^{\text{cb}}(t, T; P(t, S), K) = \text{Call}^{\text{cb}}(t, T; P(t, S), K) + P(t, T)K - B(t, S), \quad (12.81)$$

where straight forward notation.

12.8.3 European options on bonds

Let T be the expiration date of a European call option on a bond and $S > T$ be the expiration date of the bond. We consider a simple model of the short-term rate with $d = 1$, and a rational *bond pricing function* of the form $P(\tau) \equiv P(r, \tau, S)$. We also consider a rational *option price function* $C^b(\tau) \equiv C^b(r, \tau, T, S)$. By the FTAP, there are no arbitrage opportunities if and only if,

$$C^b(t) = \mathbb{E}_t \left[e^{-\int_t^T r(\tau) d\tau} (P(r(T), T, S) - K)^+ \right], \quad (12.82)$$

where K is the strike of the option. In terms of PDEs, C^b is solution to Eq. (12.79) with $\pi \equiv 0$ and boundary condition $C^b(r, T, T, S) = (P(r, T, S) - K)^+$, where $P(r, \tau, S)$ is also the solution to Eq. (12.79) with $\pi \equiv 0$, but with boundary condition $P(r, S, S) = 1$. In terms of PDEs, the situation seems hopeless. As we show below, the problem can considerably be simplified with the help of the T -forward martingale probability introduced in Section 12.2.3. In fact, we shall show that under the assumption that the short-term rate is a Gaussian process, Eq. (12.82) has a closed-form expression. We now present two models enabling this. The first one is that developed in a seminal paper by Jamshidian (1989), and the second one is, simply, its perfectly fitting extension.

12.8.3.1 Jamshidian & Vasicek

Suppose that the short-term rate is solution to the Vasicek model considered in Section 12.4 (see Eq. (12.34)):

$$dr(\tau) = \kappa(r^* - r(\tau)) d\tau + \sigma d\tilde{W},$$

where \tilde{W} is a Q -Brownian motion and $r^* \equiv \bar{r} - \frac{\lambda\sigma}{\kappa}$. As shown in Section 12.4, Eq. (12.38), the bond price is:

$$P(r(\tau), \tau, S) = e^{A(\tau, S) - B(\tau, S)r(\tau)},$$

for some function A , and for $B(t, T) = \frac{1}{\kappa} (1 - e^{-\kappa(T-t)})$ (see Eq. (12.66)).

In Section 12.4, Eq. (12.59), it was also shown that

$$\begin{aligned} & \mathbb{E}_t \left[e^{-\int_t^T r(\tau) d\tau} (P(r(T), T, S) - K)^+ \right] \\ &= P(r(t), t, S) \cdot Q_F^S [P(r(T), T, S) \geq K] - KP(r(t), t, T) \cdot Q_F^T [P(r(T), T, S) \geq K], \end{aligned} \quad (12.83)$$

where Q_F^T denotes the T -forward martingale probability introduced in Section 12.2.3.

In Appendix 8, we show that the two probabilities in Eq. (12.83) can be evaluated by the *changes of numéraire* described in Section 12.2.3, such that the solution for $P(r, T, S)$ is:

$$\begin{aligned} P(r, T, S) &= \frac{P(r, T, S)}{P(r, t, T)} e^{-\frac{1}{2}\sigma^2 \int_t^T [B(\tau, S) - B(\tau, T)]^2 d\tau - \sigma \int_t^T [B(\tau, S) - B(\tau, T)] dW^{Q_F^T}(\tau)} \quad \text{under } Q_F^T \\ P(r, T, S) &= \frac{P(r, T, S)}{P(r, t, T)} e^{\frac{1}{2}\sigma^2 \int_t^T [B(\tau, S) - B(\tau, T)]^2 d\tau - \sigma \int_t^T [B(\tau, S) - B(\tau, T)] dW^{Q_F^S}(\tau)} \quad \text{under } Q_F^S \end{aligned} \quad (12.84)$$

where $W^{Q_F^T}$ is a Brownian motion under the forward probability Q_F^T . Therefore, simple algebra now reveals that:

$$Q_F^S [P(T, S) \geq K] = \Phi(d_1), \quad Q_F^T [P(T, S) \geq K] = \Phi(d_1 - v), \quad d_1 = \frac{\ln \left[\frac{P(r(t), t, S)}{KP(r(t), t, T)} \right] + \frac{1}{2}v^2}{v},$$

where

$$v^2 = \sigma^2 \int_t^T [B(\tau, S) - B(\tau, T)]^2 d\tau = \sigma^2 \frac{1 - e^{-2\kappa(T-t)}}{2\kappa} B(T, S)^2. \quad (12.85)$$

12.8.3.2 Perfectly fitting extension

We now consider the perfectly fitting extension of the previous results. Namely, we consider the Hull and White (1990) model in Eq. (12.64) of Section 12.4, viz

$$dr(\tau) = (\theta(\tau) - \kappa r(\tau))d\tau + \sigma d\tilde{W}(\tau),$$

where $\theta(\tau)$ is the infinite dimensional parameter with which we “invert the term-structure.”

The solution to Eq. (12.82) is the same as in the previous section. However, in Section 12.8.3 we shall argue that the advantage of using such a perfectly fitting extension arises as soon as one is concerned with the evaluation of more complex options on fixed coupon bonds.

12.8.3.3 Bond price volatility and the persistence of the short-term rate

The implied vol on options on bonds is typically very large, in fact comparable to that on stocks. Why is it that this implied vol is so large, when in fact, the volatility of the short-term rate is one order of magnitude less than that on stock markets? The answer is that the short-term rate is very persistent, and it is “a risk for the long-run,” pretty much in the same spirit of the explanations attempting to explain the equity premium puzzle, reviewed in Chapter 8. To make this point precise, define, first, the *term-structure of volatility*. It is the function, $\tau \mapsto \text{Vol}(R(\tau))$, where $R(\tau)$ is the spot rate for the maturity τ , and $\text{Vol}(R(\tau))$ is the standard deviation of this spot-rate. By the definition of $R(\tau)$, the term-structure of volatility can also be written as the function

$$\tau \mapsto \text{Vol} \left(-\frac{1}{\tau} \ln P(\tau) \right),$$

where $P(\tau)$ is the price of a zero with maturity equal to τ . It is instructive to see what this volatility looks like, for a concrete model. Consider again the Vasicek model. This model assumes that the short-term rate is solution to,

$$dr_t = \kappa(\mu - r_t) dt + \sigma dW_t,$$

where W_t is a Brownian motion, and κ , μ and σ are three positive constants. By previous results given in this chapter, we know that for this model,

$$R(\tau) = \frac{A(\tau)}{\tau} + \frac{1}{\tau} B(\tau) r, \quad B(\tau) = \frac{1 - e^{-\kappa\tau}}{\kappa}.$$

for some function $A(\tau)$. Therefore, we have that,

$$\text{Vol}[R(\tau)] = \frac{1}{\tau} B(\tau) \text{Vol}_\infty(r), \quad (12.86)$$

where $\text{Vol}_\infty(r)$ is the “ergodic” volatility of the short-term rate, defined as, $\text{Vol}_\infty(r) = \sqrt{\sigma^2/2\kappa}$. For example, if $\kappa = 0.2$ and $\sigma = 0.03$, then $\text{Vol}_\infty(r) \approx 4.7\%$. Given the previous values for κ and σ , Figure 12.5 depicts the term-structure of volatility, i.e. Eq. (12.86).

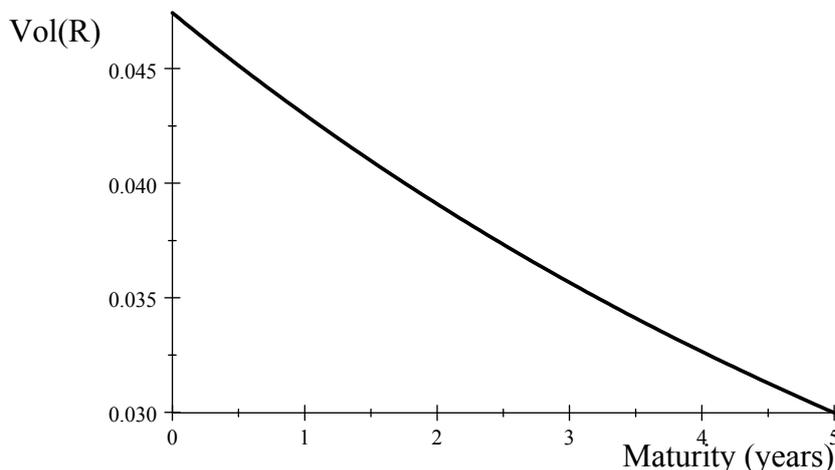


FIGURE 12.5. The term-structure of volatility predicted by the Vasicek model.

As we can see, the term-structure of volatility is decreasing in the maturity of the zero, and attains its maximum at $\text{Vol}_\infty(r) \approx 4.7\%$. It is natural, as the yield curve in this model flattens out, converging towards a constant long-term value, the asymptotic interest rate, as we say sometimes.

Despite this, the *volatility of bond returns* can be much higher, as we now illustrate. We need to figure out the dynamics of the bond price, for the Vasicek model. By Itô’s lemma,

$$\frac{dP(\tau)}{P(\tau)} = [\dots] dt + [-\sigma \cdot B(\tau)] dW_t$$

Therefore, the volatility of bond returns is,

$$\text{Vol}\left(\frac{dP}{P}\right) = \sigma B(\tau). \quad (12.87)$$

Compare Eq. (12.87) with Eq. (12.86). The main difference between the two equations is that the right hand side of Eq. (12.86) is divided by τ , which makes $\text{Vol}[R(\tau)]$ decreasing in τ . (Otherwise, $\text{Vol}_\infty(r)$ and σ have roughly the same order of magnitude.) The point is, indeed, that the yield, $R(\tau)$, is simply an *average return* which we obtain were we to decide not to sell the bond until its expiry. This average return is, of course, progressively less volatile as time to maturity gets large and it becomes a constant, eventually. The return $\frac{dP}{P}$ is, instead, measuring the capital gains we may obtain by trading the bond, and tends to be more and more volatile as time to maturity gets large. Indeed, even if σ is very small, the volatility of bond return, $\text{Vol}\left(\frac{dP}{P}\right)$, can be quite high. For example, if κ is close to zero, then, $\text{Vol}\left(\frac{dP}{P}\right) \approx \sigma \cdot \tau$, which is

15% for a 5Y zero. This fact is illustrated by Figure 12.6, which depicts Eq. (12.87), evaluated at the previous parameter values, $\kappa = 0.2$ and $\sigma = 0.03$.

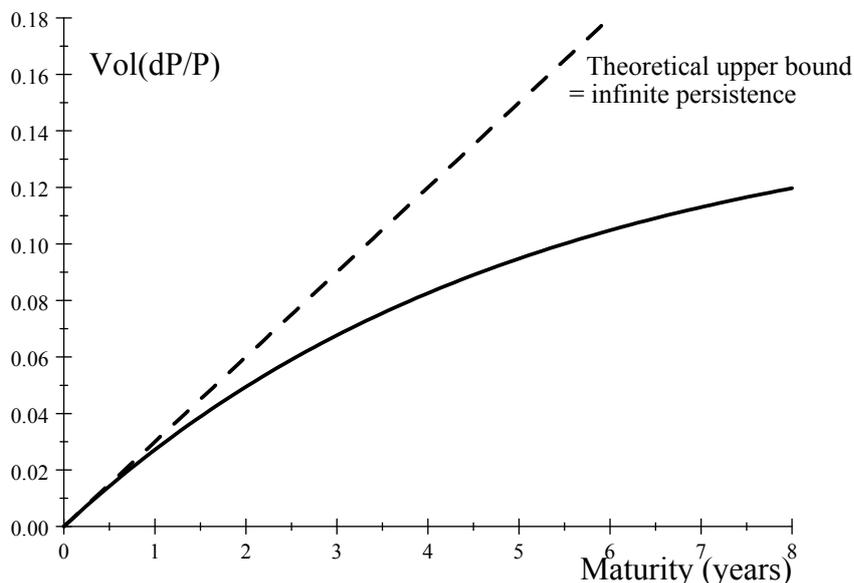


FIGURE 12.6. The dashed line depicts the bond return volatility, $\text{Vol}\left(\frac{dP}{P}\right)$, arising when the persistence parameter $\kappa = 0$, and the solid line is the bond return volatility for $\kappa = 0.2$.

The high persistence of the short-term rate, as measured by a low value of κ , makes long maturity bond returns quite volatile. Intuitively, this high persistence implies that a shock in the short-term rate has long lasting effects on the future path of the short-term rate. This makes the short-term rate very volatile in the long-run, which makes the value of long maturity zeros very volatile as a result. Intuitively, interest rates exhibit inertia: (i) it takes a number of shocks to move interest rates away from their equilibrium paths and so, short-term bonds are not volatile; and (ii) it takes time for interest rates to absorb shocks and so, medium/long-term bonds are volatile. For example, Figure 12.7 depicts the dynamics of the three month rate and those of the three months into five years forward swap rate, an interest rate that refers to relatively higher maturities, as explained in Section 12.8.5.4 below (see Eq. (12.97)). The forward swap rate is orders of magnitude more volatile than the short-term rate.

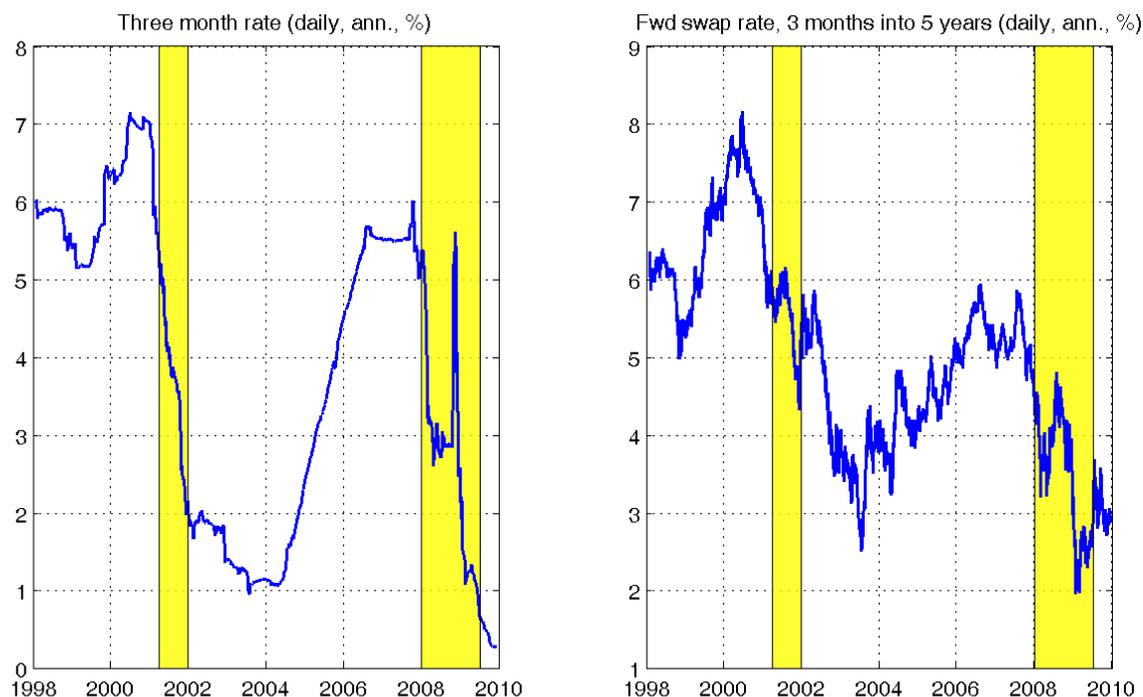


FIGURE 12.7. Interest rate volatility increases with maturity.

These facts are confirmed by the “implicit” (not implied) option-based volatility in Eq. (12.85),

$$v = \sqrt{T-t} \cdot \text{Vol}^O, \quad \text{Vol}^O \equiv \sigma \sqrt{\frac{1 - e^{-2\kappa(T-t)}}{2\kappa(T-t)} \frac{1 - e^{-\kappa(S-T)}}{\kappa}}.$$

As κ gets small, Vol^O tends to $\sigma \times (S - T)$, which increases with the bond’s time to maturity left at its expiration, $S - T$.

The previous reasoning does, of course, still hold in the more realistic case of a three-factor model, such as that in Eqs. (12.46). In that case, as explained, κ_r is large and $\kappa_{\bar{r}}$ is small: the short-term rate is quite persistent because it mean-reverts, quickly, to a persistent process, which we denoted as $\bar{r}(\tau)$. Naturally, in such a three-factor model, Eq. (12.87) does not hold anymore, as we should add two more volatility components, related to stochastic volatility, $v(\tau)$, and the persistent process $\bar{r}(\tau)$. However, the bond return volatility would be boosted by the high persistence of $\bar{r}(\tau)$.

12.8.4 Callable and puttable bonds

Callable bonds are assets that give the issuer the right to buy them back at certain times and predetermined prices. *Puttable bonds*, on the other hand, give the investor the right to sell them back to the issuer at a certain strike price. The previous chapter, Section 11.8.1, illustrates how to evaluate callable bonds, using binomial trees. In this section, we illustrate some useful properties of both callable and puttable bonds, with the help of a simple continuous-time model. For simplicity, we consider *non-defaultable*, and zero coupon, bonds.

Consider, first, callable bonds, and let K be the strike price of the callable bond maturing at time S , and suppose that the date of exercise, if any, is some future time $T < S$. In Section 11.8 of the previous chapter, the bond-issuer has the option to call the bond at any fixed date before the expiration, such that at each time τ , the value of the callable bond is $\min\{D_\tau, K\}$, where, as explained, D_τ is the time τ present value of the future expected discounted cash flows promised at time τ , by a callable bond with the same strike price K . When the option to exercise occurs at only one maturity date, at $T < S$, the callable bond is, instead, worth $\min\{K, P\}$, where P is the price of a *non-callable* bond. Indeed, if $K < P$, then, the issuer can buy its bonds back at K and re-issue the same bond at better market conditions, P . The difference, $P - K$, is just a net gain for the issuer. Therefore, the callable bond is worth just K when $K < P$. Instead, if $K > P$, the issuer does not have any incentives to exercise and, then, the value of the callable bond is just that of a non-callable bond. Therefore, the callable bond is worth P when $K > P$.

It easy to see that,

$$\min\{P, K\} = P - \max\{P - K, 0\}.$$

Therefore, we see that the price of a *callable* bond with maturity date S , equals the price of a *non-callable* bond with the same maturity date S , minus the value to call the bond, which equals the price of an hypothetical option on the *non-callable* bond, struck at K .

We can apply these insights to price a callable option in a concrete example. Consider, for example, the short-term rate in the Vasicek model. Then, if the short-term rate is $r(t)$ at time t , the value as of time t of the non-defaultable zero coupon bond maturing at time S , callable at time $T < S$, at a strike price equal to K , is,

$$P^{\text{callable}}(r(t), t, T, S) = P(r(t), t, S) - \text{Call}^b(r(t), t, T, S), \quad (12.88)$$

where $P(r(t), t, S)$ is the value of the *non-callable* zero maturing at time S , and $\text{Call}^b(r(t), t, T, S)$ is the value of a call option on the *non-callable* S -zero, maturing at time T and having a strike price equal to K .

Eq. (12.88) shows that the presence of the option to call the bond raises the cost of capital for the issuer.

In the context of the Vasicek model, the solution to $C^b(r(t), t, T, S)$ in Eq. (12.88) is given by the Jamshidian's (1989) formula in Eq. (12.83), which we now use below. Figure 12.8 depicts the behavior of the price of the callable bond in Eq. (12.88), $P^{\text{callable}}(r, 0, T, S)$, as a function of the short-term rate, r , when the exercise price, $K = 0.65$, and $S = 10$, $T = 0.5$, $\kappa = 0.2$, $\sigma = 0.03$, $\bar{\theta} = 0.06 * \kappa - \lambda$, where λ , the unit risk-premium, equals -1.7146×10^{-2} .¹⁵

¹⁵To evaluate Eq. (12.88), we make use of the closed-form solution for the bond price, given by $P(r, t, T) = e^{A(T-t) - B(T-t)r}$, where the functions A and B are given by $A(T-t) = -(T - \frac{1 - e^{-\kappa(T-t)}}{\kappa})\bar{r} - \frac{\sigma^2}{4\kappa^3}(1 - e^{-\kappa(T-t)})^2$, $\bar{r} = \frac{1}{\kappa}\bar{\theta} - \frac{1}{2}(\frac{\sigma}{\kappa})^2$ and $B(T-t) = \frac{1}{\kappa}[1 - e^{-\kappa(T-t)}]$.

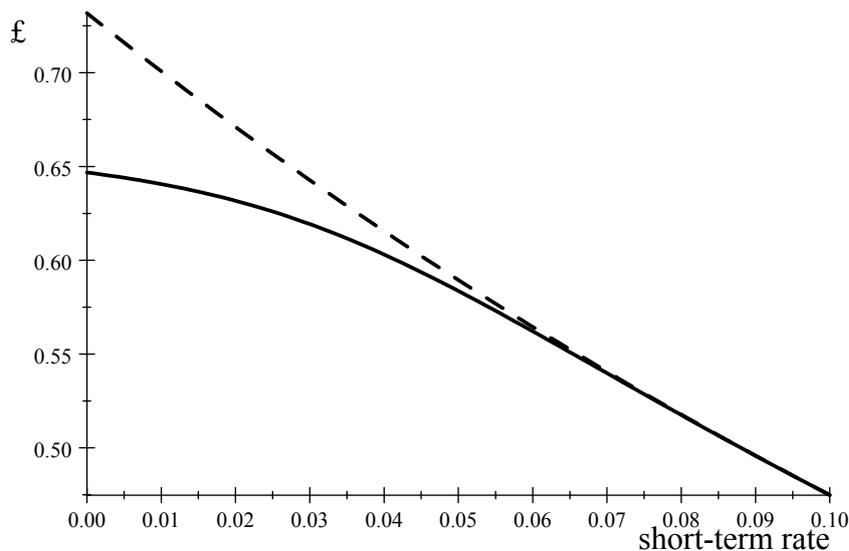


FIGURE 12.8. “Negative convexity.” Solid line: the price of a callable bond. Dashed line: the price of a non-callable bond. The price of a callable bond exhibits negative convexity with respect to the short-term rate.

As Figure 12.7 illustrates, the convexity of the non-callable bond price is destroyed by the convexity of the price of the option embedded in the callable bond. Intuitively, as the short-term rate gets small, callable and non-callable bond prices increase. However, the price of callable bonds increases less because as the short-term rate decreases, bond prices increase and then, the probability the issuer will exercise the option, at maturity, increases. This makes the risk-neutral distribution of the callable bond price markedly shifted towards the value of the strike price, $K = 0.65$, which entails a progressively lower decay rate for the bond price as the short-term rate gets small.

What is the duration of a callable bond? Naturally, a five year bond with fixed-coupons issued when interest rates are relatively high might resemble, so to speak, a three year conventional bond, as a likely decrease in the interest rates would lead the bond-issuer to redeem its debt at the strike price. To formalize this intuition, we can compute the stochastic duration of the callable bond predicted by this model, using Eq. (12.20). For the Vasicek model, we have that the semi-elasticity of the non-callable bond price with respect to r is $\Psi(r, T-t) = \frac{1}{\kappa} (1 - e^{-\kappa(T-t)})$, and its inverse with respect to time-to-maturity is given by:

$$\Psi^{-1}(x) = -\frac{1}{\kappa} \ln(1 - \kappa x).$$

Therefore, the stochastic duration for the callable bond predicted by the Vasicek model is, by Eq. (12.20):

$$D(r, S-t) = -\frac{1}{\kappa} \ln \left(1 + \kappa \frac{P_r(r, t, S) - C_r^b(r, t, T, S)}{P(r, t, S) - C^b(r, t, T, S)} \right),$$

where subscripts denote partial derivatives.

[In progress]

Next, we proceed with pricing the puttable bond. As explained in the previous chapter, Section 11.8, the payoff at the expiration of the bondholders right to tender the bonds is:

$$\max\{P, K\} = P + \max\{K - P, 0\},$$

where P is the price of a non-puttable bond. We can use, again, the Vasicek model to price the previous payoff. The price at t of a non-defaultable zero-coupon bond maturing at time S , puttable at time $T < S$, at a strike price equal to K , when the short-term rate is $r(t)$, is:

$$P^{\text{puttable}}(r(t), t, T, S) = P(r(t), t, S) + \text{Put}^b(r(t), t, T, S) = \text{Call}^b(r(t), t, T, S) + P(r(t), t, T)K,$$

where $P(r(t), t, S)$ is the value of the *non-puttable* zero maturing at time S ; $\text{Put}^b(r(t), t, T, S)$ is the value of a put option on the *non-puttable* zero maturing at S , maturing at T , struckable at K ; and the second equality follows by the put-call parity of Eq. (12.80), with $\text{Call}^b(r(t), t, T, S)$ defined as in Eq. (12.88).

[In progress]

12.8.5 Related fixed income products

12.8.5.1 Fixed coupon bonds

Given a set of dates $\{T_i\}_{i=0}^n$, a fixed coupon bond pays off a fixed *coupon* c_i at T_i , $i = 1, \dots, n$ and one unit of numéraire at time T_n . Ideally, one generic coupon at time T_i pays off for the time-interval $T_i - T_{i-1}$. It is assumed that the various coupons are known at time $t < T_0$. By the FTAP, the value of a fixed coupon bond is

$$P_{\text{fcb}}(t, T_n) = P(t, T_n) + \sum_{i=1}^n c_i P(t, T_i).$$

12.8.5.2 Floating rate bonds

A floating rate bond works the same as a fixed coupon bond, with the important exception that the coupon payments are defined as:

$$c_i = \delta_{i-1} L(T_{i-1}, T_i) = \frac{1}{P(T_{i-1}, T_i)} - 1, \quad (12.89)$$

where $\delta_i \equiv T_{i+1} - T_i$, and where the second equality is the definition of the simply-compounded LIBOR rates introduced in Section 12.2.2.1 (see Eq. (12.1)). By the FTAP, the price p_{frb} as of time t of a floating rate bond is:

$$\begin{aligned} p_{\text{frb}}(t) &= P(t, T_n) + \sum_{i=1}^n \mathbb{E}_t \left[e^{-\int_t^{T_i} r(\tau) d\tau} \delta_{i-1} L(T_{i-1}, T_i) \right] \\ &= P(t, T_n) + \sum_{i=1}^n \mathbb{E}_t \left[\frac{e^{-\int_t^{T_i} r(\tau) d\tau}}{P(T_{i-1}, T_i)} \right] - \sum_{i=1}^n P(t, T_i) \\ &= P(t, T_n) + \sum_{i=1}^n P(t, T_{i-1}) - \sum_{i=1}^n P(t, T_i) \\ &= P(t, T_0). \end{aligned}$$

where the second line follows from Eq. (12.89) and the third line from Eq. (12.7) given in Section 12.2. That is, a floating rate bond would quote at par at its first reset date, $p_{\text{frb}}(T_0) = P(T_0, T_0) = 100\%$.

We may arrive to the same property, once we consider a market where the floating rates continuously pay off the instantaneous short-term rate r . Let $T_0 = t$, and let p_{frb} is solution to the partial differential equation (12.79), with $\pi(r) = r$, and boundary condition $p_{\text{frb}}(T) = 1$. Then, it can be easily verified that $p_{\text{frb}} = 1$ is indeed solution to Eq. (12.79).

12.8.5.3 Options on fixed coupon bonds

We ignore issues relating to accruals of coupon payments, and assume the expiration date of these options occurs at any of the reset dates. Therefore, the payoff of an option maturing at T_0 on a fixed coupon bond paying off at dates T_1, \dots, T_n is:

$$[P_{\text{fcb}}(T_0, T_n) - K]^+ = \left[P(T_0, T_n) + \sum_{i=1}^n c_i P(T_0, T_i) - K \right]^+. \quad (12.90)$$

Evaluating the expectation of the payoff in Eq. (12.90) might seem challenging, but only apparently, once a number of assumptions are made. The challenge might arise, because the maximum between zero and a sum, such as that in Eq. (12.90) is, in general, and obviously, not the same as the sum of the maxima between zero and each element of the sum. Even a model where bond prices are log-normal, appears to be problematic, as we know that the sum of log-normals is not log-normal. However, this issue can be dealt with in an elegant manner. Suppose we wish to model the bond price $P(t, T)$ through any one of the models of the short-term rate reviewed in Section 12.4, and denote, as usual, the pricing function as $P(t, T) = P(r, t, T)$. Assume, further, that

$$\text{For all } t, T, \quad \frac{\partial P(r, t, T)}{\partial r} < 0, \quad (12.91)$$

and that

$$\text{For all } t, T, \quad \lim_{r \rightarrow 0} P(r, t, T) > K \quad \text{and} \quad \lim_{r \rightarrow \infty} P(r, t, T) = 0. \quad (12.92)$$

Under conditions (12.91) and (12.92), there is one and only one value of r , say $r^*(K)$, that solves the following equation:

$$P(r^*, T_0, T_n) + \sum_{i=1}^n c_i P(r^*(K), T_0, T_i) = K. \quad (12.93)$$

Then, the payoff in Eq. (12.90) can be written as:

$$\left[\sum_{i=1}^n \bar{c}_i P(r(T_0), T_0, T_i) - K \right]^+ = \left[\sum_{i=1}^n \bar{c}_i (P(r(T_0), T_0, T_i) - P(r^*(K), T_0, T_i)) \right]^+,$$

where $\bar{c}_i = c_i$, $i = 1, \dots, n-1$, and $\bar{c}_n = 1 + c_n$.

Next, note that by condition (12.91), the terms $P(r(T_0), T_0, T_i) - P(r^*(K), T_0, T_i)$ have all the same sign for all i .¹⁶ Therefore, the payoff in Eq. (12.90) is,

$$\left[\sum_{i=1}^n \bar{c}_i P(r(T_0), T_0, T_i) - K \right]^+ = \sum_{i=1}^n \bar{c}_i [P(r(T_0), T_0, T_i) - \mathcal{K}_i^*(K)]^+, \quad \mathcal{K}_i^*(K) \equiv P(r^*(K), T_0, T_i), \quad (12.94)$$

Each term of the sum in Eq. (12.94) can be evaluated as an option on a pure discount bond with strike price equal to $P(r^*(K), T_0, T_i)$, where the threshold $r^*(K)$ is found numerically. The device to reduce the problem of an option on a fixed coupon bond to a problem involving the sum of options on zero coupon bonds was invented by Jamshidian (1989).¹⁷

¹⁶Suppose that $P(r(T_0), T_0, T_1) > P(r^*, T_0, T_1)$. By Eq. (12.91), $r(T_0) < r^*$. Hence $P(r(T_0), T_0, T_2) > P(r^*, T_0, T_2)$, etc.

¹⁷The conditions in Eqs. (12.91) and (12.92) hold, within the Vasicek's model that Jamshidian considered in his paper. In fact, bond prices are always decreasing in the short-term rate in all one-factor stationary, Markov models of the short-term rate. Note that this monotonicity property is not a general property in multi-factor models (see Mele (2003)).

The price of the call on the fixed coupon bond is, therefore,

$$\overline{\text{Call}}(t, T_0; P_{\text{fcb}}(t, T_n), K, v) = \sum_{i=1}^n \bar{c}_i \cdot \text{Call}(t, T_0; P(t, T_i), \mathcal{K}_i^*(K), v_i), \quad (12.95)$$

where r^* solves Eq. (12.93), and,

$$\begin{aligned} \text{Call}(t, T_0; P_i, \mathcal{K}_i^*(K), v_i) &= P_i \Phi(d_{1,i}) - \mathcal{K}_i^*(K) P(t, T_0) \Phi(d_{1,i} - v_i), \\ d_{1,i} &= \frac{\ln \frac{P_i}{\mathcal{K}_i^*(K) P(t, T_0)} + \frac{1}{2} v_i^2}{v_i}, \quad v_i = \sigma \sqrt{\frac{1 - e^{-2\kappa(T_0 - t)}}{2\kappa}} B(T_0, T_i), \quad B(t, T) = \frac{1}{\kappa} (1 - e^{-\kappa(T-t)}). \end{aligned}$$

Finally, we determine the price of a put through the put-call parity in Eq. (12.81).

Why are perfectly fitting models so important, in practice? Suppose that in Eq. (12.93), the critical value r^* is computed by means of the Vasicek model. This assumption is attractive because it leads to evaluate the payoff in Eq. (12.94) through the Jamshidian's formula of Section 12.8.3. However, this way to proceed does not ensure that the yield curve is perfectly fitted. The natural alternative is to use the corresponding perfectly fitting extension, the Hull and White model in Section 12.5.3, i.e. Eq. (12.64), and use this price to calibrate r^* in Eq. (12.93). Note, now, the importance of a perfectly fitting model. As mentioned in Section 12.8.3.1, both Jamshidian and its perfectly fitting extension agree on the price of an option on a zero. However, Jamshidian and its perfectly fitting extension would assign different values to options on coupon bearing bonds, because they would lead to different values for r^* in Eq. (12.93) and, hence, different values for the fictitious strikes $P(r^*, T_0, T_i)$ in Eq. (12.94).

12.8.5.4 Interest rate swaps

A Savings and Loan (S&L, henceforth) is an institution that makes mortgage, car and personal loans to individual members, financed through savings. During the 1980s through the beginning of the 1990s, these forms of cooperative ventures entered into a deep and persistent crisis, leading to a painful Government bailout of about \$125b under George H.W. Bush administration. There are many causes of this crisis, but one of them was certainly the rise in short-term rates arising as a result of inflation and the attempts at fighting against it—the so-called Monetary Experiment mentioned in Section 12.4.7. But banking is risky precisely because it involves lending at horizons longer than those relating to borrowing, and S&L “banking” was not an exception to such *modus operandi*. Certainly, interest rate swaps could have helped coping with the inversion of the yield curve of the time.

An interest rate swap is simply an exchange of interest rate payments. Typically, one counterparty exchanges a fixed against a floating interest rate payment. The floating payment is typically a short-term interest rate. For example, the counterparty receiving a floating interest rate payment has “good” (or only) access to markets for “variable” interest rates, but wishes to pay fixed interest rates. Alternatively, the counterparty receiving a floating interest rate wants to hedge itself against changes in short-term rates, as it might have been the case for S&L institutions during the 1980s. The counterparty receiving a floating interest rate payment and paying a fixed interest rate K_{irs} has a payoff equal to,

$$\delta_{i-1} [L(T_{i-1}, T_i) - K_{\text{irs}}]$$

at time T_i , $i = 1, \dots, n$. Each of this payment is a FRA really, and can be evaluated as in Section 12.2. By convention, we say that the *swap payer* is the counterparty who pays the fixed

interest rate K_{irs} , and that the *swap receiver* is the counterparty receiving the fixed interest rate K_{irs} .

With a dedicated interest swap of this kind, a S&L institution would have locked-in the yield curve: at time t , the payoff for the financial institution is, in this stylized example, $\delta_{i-1} [L^{\text{long}}(T_{i-1}, T_i) - L(T_{i-1}, T_i)] + \delta_{i-1} [L(T_{i-1}, T_i) - K_{\text{irs}}] = \delta_{i-1} [L^{\text{long}}(T_{i-1}, T_i) - K_{\text{irs}}]$, where $L^{\text{long}}(T_{i-1}, T_i)$ is the interest rate gained over long-term assets. Naturally, if short-term interest rates had to go down, relative to K_{irs} , a S&L institution would not have benefited from the increased long-term/short-term spread, $\delta_{i-1} [L^{\text{long}}(T_{i-1}, T_i) - L(T_{i-1}, T_i)]$. But clearly insuring against yield curve inversions is the thing to do, if yield curve inversions lead to bankruptcy and bankruptcy is costly. We shall see, below, that other products exist, such as caps or swaptions, which ensure against the upside while at the same time freeing up the downside.

By the FTAP, the price as of time t of an interest rate swap *payer*, $p_{\text{irs}}(t)$, say, is:

$$p_{\text{irs}}(t) = \sum_{i=1}^n \mathbb{E}_t \left[e^{-\int_t^{T_i} r(\tau) d\tau} \delta_{i-1} (L(T_{i-1}, T_i) - K_{\text{irs}}) \right] = \sum_{i=1}^n \text{IRS}(t, T_{i-1}, T_i; K_{\text{irs}}), \quad (12.96)$$

where IRS is the value of a FRA and, by Eq. (12.8) in Section 12.2, is:

$$\text{IRS}(t, T_{i-1}, T_i; K_{\text{irs}}) = \delta_{i-1} [F(t, T_{i-1}, T_i) - K_{\text{irs}}] P(t, T_i).$$

The *forward swap rate* R_{swap} is the value of K_{irs} such that $p_{\text{irs}}(t) = 0$. Simple computations yield:

$$R_{\text{swap}}(t) = \frac{\sum_{i=1}^n \delta_{i-1} F(t, T_{i-1}, T_i) P(t, T_i)}{\sum_{i=1}^n \delta_{i-1} P(t, T_i)} = \frac{P(t, T_0) - P(t, T_n)}{\sum_{i=1}^n \delta_{i-1} P(t, T_i)}, \quad (12.97)$$

where the last equality holds by Eq. (12.3) in Section 12.2: $\delta_{i-1} F(t, T_{i-1}, T_i) P(t, T_i) = P(t, T_{i-1}) - P(t, T_i)$.¹⁸ This expression collapses to the par coupon rate derived in Section 11.2.2.2 of Chapter 11, once we set $t = T_0$. That is, the *spot* swap rate is a par yield.

By plugging the expression for the forward swap rate in Eq. (12.97) into Eq. (12.96), we obtain the following intuitive expression for the swap payer:

$$\begin{aligned} p_{\text{irs}}(t) &= \sum_{i=1}^n \delta_{i-1} F(t, T_{i-1}, T_i) P(t, T_i) - K_{\text{irs}} \sum_{i=1}^n \delta_{i-1} P(t, T_i) \\ &= \sum_{i=1}^n \delta_{i-1} P(t, T_i) (R_{\text{swap}}(t) - K_{\text{irs}}) \\ &\equiv \text{PVBP}_t(T_1, \dots, T_n) (R_{\text{swap}}(t) - K_{\text{irs}}), \end{aligned} \quad (12.98)$$

where $\text{PVBP}_T(T_1, \dots, T_n)$ is the so-called swap ‘‘Present Value of the Basis Point’’ (see, e.g., Brigo and Mercurio, 2006), i.e. the present value impact of one basis point move in the forward swap rate at T .

12.8.5.5 Caps & floors

A cap works as an interest rate swap, with the important exception that the exchange of interest rates payments takes place only if actual interest rates are higher than K . A cap protects against

¹⁸To cast this problem in terms of continuous time swap exchanges and, then, PDEs, we set $p_{\text{irs}}(T) \equiv 0$ as a boundary condition, and $\pi(r) = r - k$, where k plays the same role as K_{irs} above. Then, if the bond price $P(\tau)$ is solution to Eq. (12.79), the following function, $p_{\text{irs}}(\tau) = 1 - P(\tau) - k \int_{\tau}^T P(s) ds$, does also satisfy Eq. (12.79).

upward movements of the interest rates, freeing up the downside. By going long a cap, the S&L institution in the example of the previous section, then, would benefit from the downside in the short-term interest rates through a cap on them, literally. Precisely, a cap is made up of *caplets*. The payoff as of time T_i of a caplet is:

$$\delta_{i-1} [L(T_{i-1}, T_i) - K]^+, \quad i = 1, \dots, n.$$

Floors are defined in a similar way, with a single *floorlet* paying off,

$$\delta_{i-1} [K - L(T_{i-1}, T_i)]^+$$

at time T_i , $i = 1, \dots, n$.

We will only focus on caps. By the FTAP, the value p_{cap} of a cap as of time t is:

$$p_{\text{cap}}(t) = \sum_{i=1}^n \mathbb{E}_t \left[e^{-\int_t^{T_i} r(\tau) d\tau} \delta_{i-1} (L(T_{i-1}, T_i) - K)^+ \right]. \quad (12.99)$$

We can develop explicit solutions to this problem, relying upon models of the short-term rate. First, we use the standard definition of simply compounded rates given in Section 12.2 (see Eq. (12.1)), viz $\delta_{i-1} L(T_{i-1}, T_i) = \frac{1}{P(T_{i-1}, T_i)} - 1$, and rewrite the caplet payoff as follows:

$$(\delta_{i-1} L(T_{i-1}, T_i) - \delta_{i-1} K)^+ = \frac{1}{P(T_{i-1}, T_i)} (1 - (1 + \delta_{i-1} K) P(T_{i-1}, T_i))^+.$$

We have,

$$\begin{aligned} p_{\text{cap}}(t) &= \sum_{i=1}^n \mathbb{E}_t \left[\frac{e^{-\int_t^{T_i} r(\tau) d\tau}}{P(T_{i-1}, T_i)} (1 - (1 + \delta_{i-1} K) P(T_{i-1}, T_i))^+ \right] \\ &= \sum_{i=1}^n \mathbb{E}_t \left[e^{-\int_t^{T_{i-1}} r(\tau) d\tau} \frac{1}{\mathcal{K}_i} (\mathcal{K}_i - P(T_{i-1}, T_i))^+ \right], \quad \mathcal{K}_i = (1 + \delta_{i-1} K)^{-1}, \end{aligned} \quad (12.100)$$

where the last equality follows by a simple computation.¹⁹ For the models of Jamshidian or in Hull & White, bond prices are such that the cap price in Eq. (12.100) can be expressed in closed-form. Indeed, Eq. (12.100) makes clear a cap is a basket of puts on zero coupon bonds, with strikes \mathcal{K}_i . As such, it can be priced in closed form, using the models in Sections 12.7.4.1 and 12.7.4.2. We have:

$$p_{\text{cap}}(t) = \sum_{i=1}^n \frac{1}{\mathcal{K}_i} \text{Put}(t, T_{i-1}; P(t, T_i), \mathcal{K}_i, v), \quad (12.101)$$

¹⁹By the law of iterated expectations,

$$\begin{aligned} \mathbb{E}_t \left[\frac{e^{-\int_t^{T_i} r(\tau) d\tau}}{P(T_{i-1}, T_i)} [1 - \mathcal{K}_i P(T_{i-1}, T_i)]^+ \right] &= \mathbb{E}_t \left[\mathbb{E} \left[\frac{e^{-\int_t^{T_i} r(\tau) d\tau}}{P(T_{i-1}, T_i)} (1 - \mathcal{K}_i P(T_{i-1}, T_i))^+ \middle| \mathcal{F}(T_i) \right] \right] \\ &= \mathbb{E}_t \left[\mathbb{E} \left[e^{-\int_t^{T_i} r(\tau) d\tau} e^{\int_{T_{i-1}}^{T_i} r(\tau) d\tau} (1 - \mathcal{K}_i P(T_{i-1}, T_i))^+ \middle| \mathcal{F}(T_i) \right] \right] \\ &= \mathbb{E}_t \left[\mathbb{E} \left[e^{-\int_t^{T_{i-1}} r(\tau) d\tau} (1 - \mathcal{K}_i P(T_{i-1}, T_i))^+ \middle| \mathcal{F}(T_i) \right] \right] \\ &= \mathbb{E}_t \left[e^{-\int_t^{T_{i-1}} r(\tau) d\tau} (1 - \mathcal{K}_i P(T_{i-1}, T_i))^+ \right] \end{aligned}$$

where $\text{Put}(\cdot)$ satisfies the put-call parity in Eq. (12.80), and, by the pricing formulae in Section 12.8.4.1,

$$\begin{aligned} \text{Call}(t, T_{i-1}; P(t, T_i), \mathcal{K}_i, v) &= P(t, T_i) \Phi(d_{1,i}) - \mathcal{K}_i P(t, T_{i-1}) \Phi(d_{1,i} - v), \\ d_{1,i} &= \frac{\ln \frac{P(t, T_i)}{\mathcal{K}_i P(t, T_{i-1})} + \frac{1}{2} v^2}{v}, \quad v = \sigma \sqrt{\frac{1 - e^{-2\kappa(T_{i-1}-t)}}{2\kappa}} B(T_{i-1}, T_i), \quad B(t, T) = \frac{1}{\kappa} (1 - e^{-\kappa(T-t)}). \end{aligned} \quad (12.102)$$

Naturally, caps on interest rates, which are nothing but baskets of calls, are portfolios of puts on fixed coupon bonds, due to the inverse relation between prices and interest rates.²⁰

12.8.5.6 Swaptions

Let us proceed with the example of the S&L institution in the previous sections. The benefits for a S&L institution long of caps is to be protected against upward movements in the short-term rates while ensuring the downside is freed up. These benefits arise, so to speak, period per period in that, a cap is a basket of options with different maturities. A swaption works differently, in that the optionality kicks in “all together.” Suppose at time t , the S&L institution is still concerned about future inversions of the yield curve and, therefore, anticipates it might need to go for going long a swap payer at some future date. At the same time, the institution might fear that in the future, swap rates will be lower relative to some reference strike. Swaptions allow to free up such a downside risk, in that they simply are options to enter a swap contract on a future date. Let the maturity date of this option be T_0 . Then, at time T_0 , the payoff for a *payer* swaption is the maximum between zero and the value of a *payer* interest rate swap at T_0 , $p_{\text{irs}}(T_0)$, viz

$$(p_{\text{irs}}(T_0))^+ = \left[\sum_{i=1}^n \text{IRS}(T_0, T_{i-1}, T_i; K_{\text{irs}}) \right]^+ = \left[\sum_{i=1}^n \delta_{i-1} (F(T_0, T_{i-1}, T_i) - K_{\text{irs}}) P(T_0, T_i) \right]^+. \quad (12.103)$$

By the FTAP, the value of the payer swaption at time t is:

$$\begin{aligned} p_{\text{swaption}}(t) &= \mathbb{E}_t \left[e^{-\int_t^{T_0} r(\tau) d\tau} \left(\sum_{i=1}^n \delta_{i-1} (F(T_0, T_{i-1}, T_i) - K_{\text{irs}}) P(T_0, T_i) \right)^+ \right] \\ &= \mathbb{E}_t \left[e^{-\int_t^{T_0} r(\tau) d\tau} \left(1 - P(T_0, T_n) - K_{\text{irs}} \sum_{i=1}^n \delta_{i-1} P(T_0, T_i) \right)^+ \right], \end{aligned} \quad (12.104)$$

where we used the relation $\delta_{i-1} F(T_0, T_{i-1}, T_i) = \frac{P(T_0, T_{i-1})}{P(T_0, T_i)} - 1$.

Eq. (12.104) is the expression for the price of a *put* option on a *fixed coupon bond* struck at one. Therefore, we can price this contract in closed-form, through the models in Section 12.8.4.1 and 12.8.4.2, similarly to that we did in the previous section for caps pricing. We have:

$$p_{\text{swaption}}(t) = \overline{\text{Put}}(t, T_0; P_{\text{fcb}}(t, T_n), 1, v),$$

²⁰We might also price caps and floors through the partial differential equation (12.79), after setting $\pi(r) = (r - k)^+$ (caps) and $\pi(r) = (k - r)^+$ (floors), for some strike k . However, this type of contracts, where payoffs are paid continuously in time, is highly stylized, and does not exist in the markets.

where $\overline{\text{Put}}(\cdot)$ satisfies the put-call parity in Eq. (12.80). By the pricing formulae in Section 12.8.4.1,

$$\overline{\text{Call}}(t, T_0; P_{\text{fcb}}(t, T_n), 1, v) = K_{\text{irs}} \sum_{i=1}^n \delta_{i-1} \text{Call}(t, T_0; P(t, T_i), P_i^*, v) + \text{Call}(t, T_0; P(t, T_n), P_n^*, v),$$

where $\text{Call}(t, T_0; P(t, T_i), P_i^*, v)$ is as in Eq. (12.102), with $P_i^* = P(r^*, T_0, T_i)$, and r^* solution to Eq. (12.93) for $K = 1$.

12.8.6 Market models

12.8.6.1 Models and market practice

As illustrated in the previous sections, models of the short-term rate can be used to obtain closed-form solutions of virtually every important interest rate derivative product. The typical examples are the Vasicek model and its perfectly fitting extension. Yet practitioners evaluate caps through the Black's (1976) formula. The assumption underlying the market practice is that the simply-compounded forward rate is lognormally distributed. As it turns out, the analytically tractable (Gaussian) short-term rate models are *not* consistent with this assumption. Clearly, the (Gaussian) Vasicek model does not predict that the simply-compounded forward rates are Geometric Brownian motions.²¹

Is it possible to address these issues through a non-Markovian HJM? The answer is in the affirmative, although some qualifications are necessary. A practical difficulty with HJM is that *instantaneous* forward rates are not observed, which at a first sight seems to be an hindrance to realistic pricing of caps and swaptions, a so important portion of the interest rate derivative markets. This point has been addressed by Brace, Gatarek and Musiela (1997), Jamshidian (1997) and Miltersen, Sandmann and Sondermann (1997), who observed that the HJM framework can be somehow “forced” to produce models ready to be used consistently with the market practice. The key feature of these models is the emphasis on the dynamics of the *simply-compounded* forward rates. One additional, and technical, assumption is that these simply-compounded forward rates are lognormal under the risk-neutral probability Q . That is, given a non-decreasing sequence of reset times $\{T_i\}_{i=0,1,\dots}$, each simply-compounded rate, F_i , is solution to the following stochastic differential equation:²²

$$\frac{dF_i(\tau)}{F_i(\tau)} = m_i(\tau)d\tau + \gamma_i(\tau)d\tilde{W}(\tau), \quad \tau \in [t, T_i], \quad i = 0, \dots, n-1, \quad (12.105)$$

where to simplify notation, we have set, $F_i(\tau) \equiv F(\tau, T_i, T_{i+1})$, and m_i and γ_i are some deterministic functions of time (γ_i is vector valued). On a mathematical point of view, that assumption that F_i follows Eq. (12.105) is innocuous.²³

As we shall show, this simple framework can be used to use the simple Black's (1976) formula to price caps and floors. However, we need to emphasize that there is nothing wrong with the

²¹Indeed, $1 + \delta_i F_i(\tau) = \frac{P(\tau, T_i)}{P(\tau, T_{i+1})} = \exp[\Delta A_i(\tau) - \Delta B_i(\tau)r(\tau)]$, where $\Delta A_i(\tau) = A(\tau, T_i) - A(\tau, T_{i+1})$, and $\Delta B_i(\tau) = B(\tau, T_i) - B(\tau, T_{i+1})$. Hence, $F_i(\tau)$ is not a Geometric Brownian motion, despite the fact that the short-term rate r is Gaussian and, hence, the bond price is log-normal. Black '76 can not be applied in this context.

²²Brace, Gatarek and Musiela (1997) formulate a model in terms of the spot simply-compounded LIBOR interest rates. Because $F_i(T_i) = L(T_i)$, the two derivations are essentially the same.

²³It is well-known that lognormal *instantaneous* forward rates create mathematical problems to the money market account (see, for example, Sandmann and Sondermann (1997) for a succinct overview on how this problem is easily handled with *simply-compounded* forward rates).

short-term rate models analyzed in previous sections. The real advance of the so-called market model is to give a rigorous foundation to the standard market practice to price caps and floors by means of the Black's (1976) formula.

12.8.6.2 Simply-compounded forward rate dynamics, and no-arb restrictions

By the definition of the simply-compounded forward rates in Eq. (12.3),

$$\ln \left[\frac{P(\tau, T_i)}{P(\tau, T_{i+1})} \right] = \ln [1 + \delta_i F_i(\tau)]. \quad (12.106)$$

The logic we follow, now, is the same as that underlying the HJM representation of Section 12.5. We wish to express the volatility of bond prices in terms of the volatility of forward rates. To achieve this task, we first assume that bond prices are driven by Brownian motions and expand the l.h.s. of Eq. (12.106) (step 1). Then, we expand the r.h.s. of Eq. (12.106) (step 2). Finally, we identify the two diffusion terms derived from the previous two steps (step 3).

Step 1: Let $P_i \equiv P(\tau, T_i)$, and assume that under the risk-neutral probability Q , P_i is solution to:

$$\frac{dP_i}{P_i} = r d\tau + \sigma_{bi} d\tilde{W}.$$

In terms of the HJM framework in Section 12.5,

$$\sigma_{bi}(\tau) = -\sigma^I(\tau, T_i) = -\int_{\tau}^{T_i} \sigma(\tau, \ell) d\ell, \quad (12.107)$$

where $\sigma(\tau, \ell)$ is the instantaneous volatility of the instantaneous ℓ -forward rate as of time τ . By Itô's lemma,

$$d \ln \left[\frac{P(\tau, T_i)}{P(\tau, T_{i+1})} \right] = -\frac{1}{2} [\|\sigma_{bi}\|^2 - \|\sigma_{b,i+1}\|^2] d\tau + (\sigma_{bi} - \sigma_{b,i+1}) d\tilde{W}. \quad (12.108)$$

Step 2: Applying Itô's lemma to $\ln [1 + \delta_i F_i(\tau)]$, and using Eq. (12.105), yields:

$$\begin{aligned} d \ln [1 + \delta_i F_i(\tau)] &= \frac{\delta_i}{1 + \delta_i F_i} dF_i - \frac{1}{2} \frac{\delta_i^2}{(1 + \delta_i F_i)^2} (dF_i)^2 \\ &= \left[\frac{\delta_i m_i F_i}{1 + \delta_i F_i} - \frac{1}{2} \frac{\delta_i^2 F_i^2 \|\gamma_i\|^2}{(1 + \delta_i F_i)^2} \right] d\tau + \frac{\delta_i F_i}{1 + \delta_i F_i} \gamma_i d\tilde{W}. \end{aligned} \quad (12.109)$$

Step 3: By Eq. (12.106), the diffusion terms in Eqs. (12.108) and (12.109) have to be the same. Therefore,

$$\sigma_{bi}(\tau) - \sigma_{b,i+1}(\tau) = \frac{\delta_i F_i(\tau)}{1 + \delta_i F_i(\tau)} \gamma_i(\tau), \quad \tau \in [t, T_i].$$

By summing over i , we get the following no-arbitrage restriction applying to the volatility of the bond prices:

$$\sigma_{bi}(\tau) - \sigma_{b,0}(\tau) = -\sum_{j=0}^{i-1} \frac{\delta_j F_j(\tau)}{1 + \delta_j F_j(\tau)} \gamma_j(\tau). \quad (12.110)$$

As is clear, Eq. (12.110) is merely a restriction to the general HJM framework. In other words, assume the instantaneous forward rates are as in Eq. (12.69) of Section 12.5. As we demonstrated in Section 12.5, then, the bond prices volatility is given by Eq. (12.107). But if we also assume that simply-compounded forward rates are solution to Eq. (12.105), then, the bond prices volatility is also equal to Eq. (12.110). Comparing Eq. (12.107) with Eq. (12.110) produces,

$$\int_{T_0}^{T_i} \sigma(\tau, \ell) d\ell = \sum_{j=0}^{i-1} \frac{\delta_j F_j(\tau)}{1 + \delta_j F_j(\tau)} \gamma_j(\tau).$$

The practical interest to restrict the forward-rate volatility dynamics in this way lies in the possibility to obtain closed-form solutions for some of the interest rates derivatives surveyed in Section 12.8.3.

12.8.6.3 Pricing formulae

Caps & floors

We provide analytical results for the price of caps only. We have:

$$\begin{aligned} p_{\text{cap}}(t) &= \sum_{i=1}^n \mathbb{E}_t \left[e^{-\int_t^{T_i} r(\tau) d\tau} \delta_{i-1} (L(T_{i-1}, T_i) - K)^+ \right] \\ &= \sum_{i=1}^n \mathbb{E}_t \left[e^{-\int_t^{T_i} r(\tau) d\tau} \delta_{i-1} (F(T_{i-1}, T_{i-1}, T_i) - K)^+ \right] \\ &= \sum_{i=1}^n \delta_{i-1} P(t, T_i) \cdot \mathbb{E}_{Q_F^{T_i}} [F(T_{i-1}, T_{i-1}, T_i) - K]^+, \end{aligned} \quad (12.111)$$

where $\mathbb{E}_{Q_F^{T_i}}[\cdot]$ denotes, as usual, the expectation taken under the T_i -forward martingale probability $Q_F^{T_i}$; the first equality is Eq. (12.99); and the second equality has been obtained through the usual change of probability technique introduced Section 12.3.4.

The key point is that

$$F_{i-1}(\tau) \equiv F_{i-1}(\tau, T_{i-1}, T_i), \quad \tau \in [t, T_{i-1}], \text{ is a martingale under } Q_F^{T_i}.$$

A proof of this statement was given in Section 12.2. By Eq. (12.105), this means that $F_{i-1}(\tau)$ is solution to:

$$\frac{dF_{i-1}(\tau)}{F_{i-1}(\tau)} = \gamma_{i-1}(\tau) dW^{Q_F^{T_i}}(\tau), \quad \tau \in [t, T_{i-1}], \quad i = 1, \dots, n,$$

under $Q_F^{T_i}$. Therefore, the cap price in Eq. (12.111) reduces to the Black's (1976) formula discussed in Chapter 10 (see Section 10.4.4 and Appendix 2 to Chapter 10), once we assume γ is deterministic:

$$\mathbb{E}_{Q_F^{T_i}} [F(T_{i-1}, T_{i-1}, T_i) - K]^+ = F_{i-1}(t) \Phi(d_{1,i-1}) - K \Phi(d_{1,i-1} - s_i), \quad (12.112)$$

where

$$d_{1,i-1} = \frac{\ln \frac{F_{i-1}(t)}{K} + \frac{1}{2} s_i^2}{s_i}, \quad s_i^2 = \int_t^{T_{i-1}} \gamma_{i-1}^2(\tau) d\tau.$$

Swaptions

By Eq. (12.98), the payoff of a *payer* swaption expiring at time T_0 is:

$$[p_{\text{irs}}(T_0)]^+ = \text{PVBP}_{T_0}(T_1, \dots, T_n) (R_{\text{swap}}(T_0) - K_{\text{irs}})^+, \quad \text{PVBP}_{T_0}(T_1, \dots, T_n) = \sum_{i=1}^n \delta_{i-1} P(T_0, T_i).$$

Therefore, by the FTAP, and a change of measure,

$$\begin{aligned} p_{\text{swaption}}(t) &= \mathbb{E}_t \left[e^{-\int_t^{T_0} r(\tau) d\tau} \text{PVBP}_{T_0}(T_1, \dots, T_n) (R_{\text{swap}}(T_0) - K_{\text{irs}})^+ \right] \\ &= \text{PVBP}_t(T_1, \dots, T_n) \cdot \mathbb{E}_{Q_{\text{swap}}} (R_{\text{swap}}(T_0) - K_{\text{irs}})^+, \end{aligned} \quad (12.113)$$

where $\mathbb{E}_{Q_{\text{swap}}}$ denotes the expectation taken under the so-called *forward swap probability*, defined by:

$$\left. \frac{dQ_{\text{swap}}}{dQ} \right|_{\mathbb{F}_{T_0}} = e^{-\int_t^{T_0} r(\tau) d\tau} \frac{\text{PVBP}_{T_0}(T_1, \dots, T_n)}{\text{PVBP}_t(T_1, \dots, T_n)}.$$

It is easy to see that $\mathbb{E} \left(\left. \frac{dQ_{\text{swap}}}{dQ} \right|_{\mathbb{F}_{T_0}} \right) = 1$, by using the definition of $\text{PVBP}_{T_0}(T_1, \dots, T_n)$, and the pricing equation, $P(t, T_i) = \mathbb{E}_t \left[e^{-\int_t^{T_0} r(\tau) d\tau} P(T_0, T_i) \right]$. The key point underlying this change of measure is that the forward swap rate R_{swap} is a Q_{swap} -martingale,²⁴ and clearly, positive. Therefore, it must satisfy:

$$\frac{dR_{\text{swap}}(\tau)}{R_{\text{swap}}(\tau)} = \gamma_{\text{swap}}(\tau) dW_{\text{swap}}(\tau), \quad \tau \in [t, T_0], \quad (12.114)$$

where W_{swap} is a Q_{swap} -Brownian motion, and $\gamma_{\text{swap}}(\tau)$ is adapted.

If the volatility $\gamma_{\text{swap}}(\tau)$ in Eq. (12.114) is deterministic, we can use Black 76 to price the payer swaption in Eq. (12.113) in closed-form. We have:

$$p_{\text{swaption}}(t) = \text{PVBP}_t(T_1, \dots, T_n) \cdot \text{Black76}(R_{\text{swap}}(t); T_0, K_{\text{irs}}, \sqrt{\bar{V}}), \quad (12.115)$$

where $\text{Black76}(\cdot)$ is given by Black's (1976) formula:

$$\begin{aligned} \text{Black76}(R_{\text{swap}}(t); T_0, K_{\text{irs}}, \sqrt{\bar{V}}) &= R_{\text{swap}}(t) \Phi(d_t) - K_{\text{irs}} \Phi(d_t - \sqrt{\bar{V}}), \\ d_t &= \frac{\ln \frac{R_{\text{swap}}(t)}{K_{\text{irs}}} + \frac{1}{2} \bar{V}}{\sqrt{\bar{V}}}, \quad \bar{V} = \int_t^{T_0} \gamma_{\text{swap}}^2(\tau) d\tau. \end{aligned}$$

Inconsistencies

If the forward rate is solution to Eq. (12.105), γ_{swap} cannot be deterministic. Unfortunately, if forward swap rates are lognormal, then, Eq. (12.105) does not hold. Therefore, we may use Black's formula to price either caps or swaptions, not both. This might limit the importance of market models. A couple of tricks that seem to work in practice. The best known is based

²⁴By Eq. (12.97), and one change of measure,

$$\mathbb{E}_{Q_{\text{swap}}} [R_{\text{swap}}(\tau)] = \mathbb{E}_{Q_{\text{swap}}} \left[\frac{P(\tau, T_0) - P(\tau, T_n)}{\text{PVBP}_\tau(T_1, \dots, T_n)} \right] = \mathbb{E}_t \left[\frac{e^{-\int_t^\tau r(\tau) d\tau} (P(\tau, T_0) - P(\tau, T_n))}{\text{PVBP}_t(T_1, \dots, T_n)} \right] = \frac{P(t, T_0) - P(t, T_n)}{\text{PVBP}_t(T_1, \dots, T_n)} = R_{\text{swap}}(t).$$

on a suggestion by Rebonato (1998), to replace the true pricing problem with an approximating pricing problem where γ_{swap} is deterministic. That works in practice, but in a world with stochastic volatility, we should expect that trick to generate unstable things in periods experiencing highly volatile volatility. See, also, Rebonato (1999) for an essay on related issues. The next section suggests to use numerical approximation based on Montecarlo techniques.

12.8.6.4 Numerical approximations

Suppose forward rates are lognormal. Then, we can price caps using Black's formula. As for swaptions, Montecarlo integration should be implemented as follows. By a change of measure,

$$\begin{aligned} p_{\text{swaption}}(t) &= \mathbb{E}_t \left[e^{-\int_t^{T_0} r(\tau) d\tau} \left(\sum_{i=1}^n \delta_{i-1} (F(T_0, T_{i-1}, T_i) - K) P(T_0, T_i) \right)^+ \right] \\ &= P(t, T_0) \mathbb{E}_{Q_F^{T_0}} \left[\sum_{i=1}^n \delta_{i-1} (F(T_0, T_{i-1}, T_i) - K) P(T_0, T_i) \right]^+, \end{aligned}$$

where $F(T_0, T_{i-1}, T_i)$, $i = 1, \dots, n$, can be simulated under $Q_F^{T_0}$.

Details are as follows. We know that

$$\frac{dF_{i-1}(\tau)}{F_{i-1}(\tau)} = \gamma_{i-1}(\tau) dW^{Q_F^{T_i}}(\tau). \quad (12.116)$$

By results in Appendix 3, we also know that:

$$\begin{aligned} dW^{Q_F^{T_i}}(\tau) &= dW^{Q_F^{T_0}}(\tau) - [\sigma_{bi}(\tau) - \sigma_{b0}(\tau)] d\tau \\ &= dW^{Q_F^{T_0}}(\tau) + \sum_{j=0}^{i-1} \frac{\delta_j F_j(\tau)}{1 + \delta_j F_j(\tau)} \gamma_j(\tau) d\tau, \end{aligned}$$

where the second line follows from Eq. (12.110) in the main text. Replacing this into Eq. (12.116) leaves:

$$\frac{dF_{i-1}(\tau)}{F_{i-1}(\tau)} = \gamma_{i-1}(\tau) \sum_{j=0}^{i-1} \frac{\delta_j F_j(\tau)}{1 + \delta_j F_j(\tau)} \gamma_j(\tau) d\tau + \gamma_{i-1}(\tau) dW^{Q_F^{T_0}}(\tau), \quad i = 1, \dots, n.$$

These can easily be simulated with the methods described in any standard textbook of this kind, such as that of Kloeden and Platen (1992).

12.8.6.5 Volatility surfaces

Caps & floors

The market practice quotes volatility surfaces by relying on the models of this section, rather than those of Sections 12.7.4.1-12.7.4.2. In the models of Sections 12.7.4.1-12.7.4.2, volatility surfaces might be produced, but only indirectly, after calibration of the two parameters κ and σ , as Eq. (12.101) indicates. It is easier, however, to provide volatility surfaces in the first place, through the models of this section. Quite simply, practitioners use Eq. (12.112) and quote volatilities such that the market price of a cap equals to the value predicted by Eq. (12.112) using the desired implied volatility s_i . In Eq. (12.112),

$$s_i = \sqrt{T_{i-1} - t} \cdot \gamma(i),$$

for some $\gamma(i)$, although, then, practitioners simply quote the value of $\hat{\gamma}_n$ that satisfies:

$$\hat{\gamma}_n : p_{\text{cap}}^{\$}(t; n) = \sum_{i=1}^n \delta_{i-1} P(t, T_i) \cdot \text{Black76}(F_{i-1}(t); K, \hat{s}_{i,n}),$$

where $p_{\text{cap}}^{\$}(t; n)$ is the market price of the cap, and:

$$\begin{aligned} \text{Black76}(F_{i-1}(t); K, \hat{s}_{i,n}) &= F_{i-1}(t) \Phi(d_{1,i-1}^n) - K \Phi(d_{1,i-1}^n - \hat{s}_{i,n}), \\ d_{1,i-1}^n &= \frac{\ln \frac{F_{i-1}(t)}{K} + \frac{1}{2} \hat{s}_{i,n}^2}{\hat{s}_{i,n}}, \quad \hat{s}_{i,n} = \sqrt{T_{i-1} - t} \cdot \hat{\gamma}_n \end{aligned}$$

Given n , we can bootstrap $\hat{\gamma}(i)$, i.e. we can recursively solve for $\hat{\gamma}(i)$, as follows:

$$0 = \sum_{i=1}^n \delta_{i-1} P(t, T_i) \cdot [\text{Black76}(F_{i-1}(t); K, \hat{s}_{i,n}) - \text{Black76}(F_{i-1}(t); K, \hat{s}_i)], \quad n = 1, \dots, N,$$

where N is the latest available maturity, and $\hat{s}_i = \sqrt{T_{i-1} - t} \cdot \hat{\gamma}(i)$. The values of $\hat{\gamma}(i)$ constitute what is known as the term structure of caps volatilities.

Swaptions

As for swaptions, the situation is much simpler. The market practice is to quote swaptions through standard implied vols, i.e. those vols IV_t such that, once inserted into Eq. (12.115), delivers the swaption market price:

$$p_{\text{swaption}}(t) = \text{PVB}_t(T_1, \dots, T_n) \cdot \text{Black76}(R_{\text{swap}}(t); T_0, K_{\text{irs}}, IV_t).$$

12.9 Appendix 1: The FTAP for bond prices

Suppose there exist m pure discount bond prices $\{\{P_i \equiv P(\tau, T_i)\}_{i=1}^m\}_{\tau \in [t, T]}$ satisfying:

$$\frac{dP_i}{P_i} = \mu_{bi} \cdot d\tau + \sigma_{bi} \cdot dW, \quad i = 1, \dots, m, \quad (12A.1)$$

where W is a Brownian motion in \mathbb{R}^d , and μ_{bi} and σ_{bi} are progressively $\mathcal{F}(\tau)$ -measurable functions guaranteeing the existence of a strong solution to the previous system (σ_{bi} is vector-valued). The value process V of a self-financing portfolio in these m bonds and a money market technology satisfies:

$$dV = \left(\pi^\top (\mu_b - \mathbf{1}_m r) + rV \right) d\tau + \pi^\top \sigma_b dW,$$

where π is some portfolio, $\mathbf{1}_m$ is a m -dimensional vector of ones, and

$$\mu_b = [\mu_{b1}, \dots, \mu_{bm}]^\top, \quad \sigma_b = [\sigma_{b1}, \dots, \sigma_{bm}]^\top.$$

Next, suppose that there exists a portfolio $\underline{\pi}$ such that $\underline{\pi}^\top \sigma_b = 0$. This is an arbitrage opportunity if there exist events for which at some time, $\mu_b - \mathbf{1}_m r \neq 0$ (use $\underline{\pi}$ when $\mu_b - \mathbf{1}_m r > 0$, and $-\underline{\pi}$ when $\mu_b - \mathbf{1}_m r < 0$: the drift of V will then be appreciating at a deterministic rate that is strictly greater than r). Therefore, arbitrage opportunities are ruled out if:

$$\pi^\top (\mu_b - \mathbf{1}_m r) = 0 \text{ whenever } \pi^\top \sigma_b = 0.$$

In other terms, arbitrage opportunities are ruled out when every vector in the null space of σ_b is orthogonal to $\mu_b - \mathbf{1}_m r$, or when there exists a λ taking values in \mathbb{R}^d satisfying some basic integrability conditions, and such that

$$\mu_b - \mathbf{1}_m r = \sigma_b \lambda$$

or,

$$\mu_{bi} - r = \sigma_{bi} \lambda, \quad i = 1, \dots, m. \quad (12A.2)$$

In this case,

$$\frac{dP_i}{P_i} = (r + \sigma_{bi} \lambda) \cdot d\tau + \sigma_{bi} \cdot dW, \quad i = 1, \dots, m.$$

Now define $\tilde{W} = W + \int \lambda d\tau$, $\frac{dQ}{dP} = \exp(-\int_t^T \lambda^\top dW - \frac{1}{2} \int_t^T \|\lambda\|^2 d\tau)$. The Q -martingale property of the “normalized” bond price processes now easily follows by Girsanov’s theorem. Indeed, define for a generic i , $P(\tau, T) \equiv P(\tau, T_i) \equiv P_i$, and:

$$g(\tau) \equiv e^{-\int_t^\tau r(u) du} \cdot P(\tau, T), \quad \tau \in [t, T].$$

By Girsanov’s theorem, and an application of Itô’s lemma,

$$\frac{dg}{g} = \sigma_{bi} \cdot d\tilde{W}, \quad \text{under } Q.$$

Therefore, for all $\tau \in [t, T]$, $g(\tau) = \mathbb{E}_t [g(T)]$, implying that:

$$g(\tau) \equiv e^{-\int_t^\tau r(u) du} \cdot P(\tau, T) = \mathbb{E}_t [g(T)] = \mathbb{E}_t [e^{-\int_t^T r(u) du} \cdot \underbrace{P(T, T)}_{=1}] = \mathbb{E}_t \left[e^{-\int_t^T r(u) du} \right],$$

or

$$P(\tau, T) = e^{\int_t^\tau r(u) du} \cdot \mathbb{E}_t \left[e^{-\int_t^T r(u) du} \right] = \mathbb{E}_t \left[e^{-\int_\tau^T r(u) du} \right], \quad \text{all } \tau \in [t, T],$$

which is Eq. (12.2).

Notice that no assumption has been made on m . The previous result holds for all m , be they less or greater than d . Suppose, for example, that there are no other traded assets in the economy. Then, if $m < d$, there exists an infinite number of risk-neutral probabilities Q . If $m = d$, there exists one and only one risk-neutral probability Q . If $m > d$, there exists one and only one risk-neutral probability but then, the various bond prices have to satisfy some basic no-arbitrage restrictions. As an example, take $m = 2$ and $d = 1$. Eq. (12A.2) then becomes

$$\frac{\mu_{b1} - r}{\sigma_{b1}} = \lambda = \frac{\mu_{b2} - r}{\sigma_{b2}}.$$

In other terms, the Sharpe ratio of any two bonds must be identical. Relation (12A.2) will be used several times in this chapter.

- In Section 12.4, the primitive of the economy is the short-term rate, solution of a multidimensional diffusion process, and μ_{bi} and σ_{bi} will be derived via Itô's lemma.
- In Section 12.5, μ_{bi} and σ_{bi} are restricted by a model for the forward rates.

12.10 Appendix 2: Certainty equivalent interpretation of forward prices

Multiply both sides of the bond pricing equation (12.2) by the amount $S(T)$:

$$P(t, T) \cdot S(T) = \mathbb{E}_t \left[e^{-\int_t^T r(\tau) d\tau} \right] \cdot S(T).$$

Suppose momentarily that $S(T)$ is known at T . In this case, we have:

$$P(t, T) \cdot S(T) = \mathbb{E}_t \left[e^{-\int_t^T r(\tau) d\tau} \cdot S(T) \right].$$

But in the applications we have in mind, $S(T)$ is random. Define then its certainty equivalent by the number $\overline{S(T)}$ that solves:

$$P(t, T) \cdot \overline{S(T)} = \mathbb{E}_t \left[e^{-\int_t^T r(\tau) d\tau} \cdot S(T) \right],$$

or

$$\overline{S(T)} = \mathbb{E}_t [\eta_T(T) \cdot S(T)], \quad (12A.3)$$

where $\eta_T(T)$ has been defined in (12.15).

Comparing Eq. (12A.3) with Eq. (12.14) reveals that forward prices can be interpreted in terms of the previously defined certainty equivalent.

12.11 Appendix 3: Additional results on T -forward martingale probabilities

Eq. (12.15) defines $\eta_T(T)$ as:

$$\eta_T(T) = \frac{e^{-\int_t^T r(\tau)d\tau} \cdot 1}{\mathbb{E}_t \left[e^{-\int_t^T r(\tau)d\tau} \right]}$$

More generally, we can define a *density process* as:

$$\eta_T(\tau) \equiv \frac{e^{-\int_t^\tau r(u)du} \cdot P(\tau, T)}{\mathbb{E}_t \left[e^{-\int_t^T r(\tau)d\tau} \right]}, \quad \tau \in [t, T].$$

By the FTAP, $\{\exp(-\int_t^\tau r(u)du) \cdot P(\tau, T)\}_{\tau \in [t, T]}$ is a Q -martingale (see Appendix 1 to this chapter). Therefore, $\mathbb{E} \left[\frac{dQ_F^T}{dQ} \middle| \mathcal{F}_\tau \right] = \mathbb{E}[\eta_T(T) | \mathcal{F}_\tau] = \eta_T(\tau)$ all $\tau \in [t, T]$, and in particular, $\eta_T(t) = 1$. We now show that this works. And at the same time, we show this by deriving a representation of $\eta_T(\tau)$ that can be used to find “forward premia.”

We begin with the dynamic representation (12A.1) given for a generic bond price # i , $P(\tau, T) \equiv P(\tau, T_i) \equiv P_i$:

$$\frac{dP}{P} = \mu \cdot d\tau + \sigma \cdot dW,$$

where we have defined $\mu \equiv \mu_{bi}$ and $\sigma \equiv \sigma_{bi}$.

Under the risk-neutral probability Q ,

$$\frac{dP}{P} = r \cdot d\tau + \sigma \cdot d\tilde{W},$$

where $\tilde{W} = W + \int \lambda$ is a Q -Brownian motion.

By Itô's lemma,

$$\frac{d\eta_T(\tau)}{\eta_T(\tau)} = -[-\sigma(\tau, T)] \cdot d\tilde{W}(\tau), \quad \eta_T(t) = 1.$$

The solution is:

$$\eta_T(\tau) = \exp \left[-\frac{1}{2} \int_t^\tau \|\sigma(u, T)\|^2 du - \int_t^\tau (-\sigma(u, T)) \cdot d\tilde{W}(u) \right].$$

Under the usual integrability conditions, we can now use the Girsanov's theorem and conclude that

$$W^{Q_F^T}(\tau) \equiv \tilde{W}(\tau) + \int_t^\tau \left(-\sigma(u, T)^\top \right) du \tag{12A.4}$$

is a Brownian motion under the T -forward martingale probability Q_F^T .

Finally, note that for all integers i and non decreasing sequences of dates $\{T_i\}_{i=0,1,\dots}$,

$$W^{Q_F^{T_i}}(\tau) = \tilde{W}(\tau) + \int_t^\tau \left(-\sigma(u, T_i)^\top \right) du, \quad i = 0, 1, \dots.$$

Therefore,

$$W^{Q_F^{T_i}}(\tau) = W^{Q_F^{T_{i-1}}}(\tau) - \int_t^\tau \left[\sigma(u, T_i)^\top - \sigma(u, T_{i-1})^\top \right] du, \quad i = 1, 2, \dots, \tag{12A.5}$$

is a Brownian motion under the T_i -forward martingale probability $Q_F^{T_i}$. Eqs. (12A.5) and (12A.4) are used in Section 12.8 on interest rate derivatives.

12.12 Appendix 4: Principal components analysis

Principal component analysis transforms the original data into a set of uncorrelated variables, the principal components, with variances arranged in descending order. Consider the following program,

$$\max_{C_1} [\text{var}(Y_1)] \quad \text{s.t.} \quad C_1^\top C_1 = 1,$$

where $\text{var}(Y_1) = C_1^\top \Sigma C_1$, and the constraint is an identification constraint. The first order conditions lead to,

$$(\Sigma - \lambda I) C_1 = 0,$$

where λ is a Lagrange multiplier. The previous condition tells us that λ must be one eigenvalue of the matrix Σ , and that C_1 must be the corresponding eigenvector. Moreover, we have $\text{var}(Y_1) = C_1^\top \Sigma C_1 = \lambda$ which is clearly maximized by the largest eigenvalue. Suppose that the eigenvalues of Σ are distinct, and let us arrange them in descending order, i.e. $\lambda_1 > \dots > \lambda_p$. Then,

$$\text{var}(Y_1) = \lambda_1.$$

Therefore, the first principal component is $Y_1 = C_1^\top (R - \bar{R})$, where C_1 is the eigenvector corresponding to the largest eigenvalue, λ_1 .

Next, consider the second principal component. The program is, now,

$$\max_{C_2} [\text{var}(Y_2)] \quad \text{s.t.} \quad C_2^\top C_2 = 1 \quad \text{and} \quad C_2^\top C_1 = 0,$$

where $\text{var}(Y_2) = C_2^\top \Sigma C_2$. The first constraint, $C_2^\top C_2 = 1$, is the usual identification constraint. The second constraint, $C_2^\top C_1 = 0$, is needed to ensure that Y_1 and Y_2 are orthogonal, i.e. $E(Y_1 Y_2) = 0$. The first order conditions for this problem are,

$$0 = \Sigma C_2 - \lambda C_2 - \nu C_1$$

where λ is the Lagrange multiplier associated with the first constraint, and ν is the Lagrange multiplier associated with the second constraint. By pre-multiplying the first order conditions by C_1^\top ,

$$0 = C_1^\top \Sigma C_2 - \nu,$$

where we have used the two constraints $C_1^\top C_2 = 0$ and $C_1^\top C_1 = 1$. Post-multiplying the previous expression by C_1^\top , one obtains, $0 = C_1^\top \Sigma C_2 C_1^\top - \nu C_1^\top = -\nu C_1^\top$, where the last equality follows by $C_1^\top C_2 = 0$. Hence, $\nu = 0$. So the first order conditions can be rewritten as,

$$(\Sigma - \lambda I) C_2 = 0.$$

The solution is now λ_2 , and C_2 is the eigenvector corresponding to λ_2 . (Indeed, this time we cannot choose λ_1 as this choice would imply that $Y_2 = C_1^\top (R - \bar{R})$, implying that $E(Y_1 Y_2) \neq 0$.) It follows that $\text{var}(Y_2) = \lambda_2$.

In general, we have,

$$\text{var}(Y_i) = \lambda_i, \quad i = 1, \dots, p.$$

Let Λ be the diagonal matrix with the eigenvalues λ_i on the diagonal. By the spectral decomposition of Σ , $\Sigma = C \Lambda C^\top$, and by the orthonormality of C , $C^\top C = I$, we have that $C^\top \Sigma C = \Lambda$ and, hence,

$$\sum_{i=1}^p \text{var}(R_i) = \text{Tr}(\Sigma) = \text{Tr}(\Sigma C C^\top) = \text{Tr}(C^\top \Sigma C) = \text{Tr}(\Lambda).$$

Hence, Eq. (12.26) follows.

12.13 Appendix 5: A few analytics for the Hull and White model

As in the Ho and Lee model, the instantaneous forward rate $f(\tau, T)$ predicted by the Hull and White model is as in Eq. (12.62), where functions A_2 and B_2 can be easily computed from Eqs. (12.65) and (12.66) as:

$$A_2(\tau, T) = \sigma^2 \int_{\tau}^T B(s, T)B_2(s, T)ds - \int_{\tau}^T \theta(s)B_2(s, T)ds, \quad B_2(\tau, T) = e^{-\kappa(T-\tau)}.$$

Therefore, the instantaneous forward rate $f(\tau, T)$ predicted by the Hull and White model is obtained by replacing the previous equations in Eq. (12.62). The result is then equated to the observed forward rate $f_{\S}(t, \tau)$ so as to obtain:

$$f_{\S}(t, \tau) = -\frac{\sigma^2}{2\kappa^2} \left[1 - e^{-\kappa(\tau-t)}\right]^2 + \int_t^{\tau} \theta(s)e^{-\kappa(\tau-s)}ds + e^{-\kappa(\tau-t)}r(t).$$

By differentiating the previous equation with respect to τ , and rearranging terms,

$$\begin{aligned} \theta(\tau) &= \frac{\partial}{\partial \tau} f_{\S}(t, \tau) + \frac{\sigma^2}{\kappa} \left(1 - e^{-\kappa(\tau-t)}\right) e^{-\kappa(\tau-t)} + \kappa \left[\int_t^{\tau} \theta(s)e^{-\kappa(\tau-s)}ds + e^{-\kappa(\tau-t)}r(t) \right] \\ &= \frac{\partial}{\partial \tau} f_{\S}(t, \tau) + \frac{\sigma^2}{\kappa} \left(1 - e^{-\kappa(\tau-t)}\right) e^{-\kappa(\tau-t)} + \kappa \left[f_{\S}(t, \tau) + \frac{\sigma^2}{2\kappa^2} \left(1 - e^{-\kappa(\tau-t)}\right)^2 \right], \end{aligned}$$

which reduces to Eq. (12.67) after using simple algebra.

12.14 Appendix 6: Expectation theory and embedding in selected models

A. Expectation theory

Suppose that

$$\sigma(\cdot, \cdot) = \sigma \quad \text{and} \quad \lambda(\cdot) = \lambda, \quad (12A.6)$$

where σ and λ are constants. We derive the dynamics of r and compare them with f to deduce something about the expectation theory. We have:

$$r(\tau) = f(t, \tau) + \int_t^\tau \alpha(s, \tau) ds + \sigma (W(\tau) - W(t)),$$

where

$$\alpha(\tau, T) = \sigma(\tau, T) \int_\tau^T \sigma(\tau, \ell) d\ell + \sigma(\tau, T)\lambda(\tau) = \sigma^2(T - \tau) + \sigma\lambda.$$

Hence,

$$\int_t^\tau \alpha(s, \tau) ds = \frac{1}{2}\sigma^2(\tau - t) + \sigma\lambda(\tau - t).$$

Finally,

$$r(\tau) = f(t, \tau) + \frac{1}{2}\sigma^2(\tau - t) + \sigma\lambda(\tau - t) + \sigma(W(\tau) - W(t)),$$

and since $E(W(\tau)|\mathcal{F}(t)) = W(t)$,

$$E[r(\tau)|\mathcal{F}(t)] = f(t, \tau) + \frac{1}{2}\sigma^2(\tau - t) + \sigma\lambda(\tau - t).$$

Even with $\lambda < 0$, this model is *not* able to *always* generate $E[r(\tau)|\mathcal{F}(t)] < f(t, \tau)$. As shown in the following exercise, this is due to the nonstationary nature of the volatility function. Indeed, suppose, next, that instead of Eq. (12A.6), we have that

$$\sigma(t, T) = \sigma \cdot \exp(-\gamma(T - t)) \quad \text{and} \quad \lambda(\cdot) = \lambda,$$

where σ, γ and λ are constants. In this case, we have:

$$r(\tau) = f(t, \tau) + \int_t^\tau \alpha(s, \tau) ds + \sigma \int_t^\tau e^{-\gamma(\tau-s)} \cdot dW(s),$$

where

$$\alpha(s, \tau) = \sigma^2 e^{-\gamma(\tau-s)} \int_s^\tau e^{-\gamma(\ell-s)} d\ell + \sigma\lambda e^{-\gamma(\tau-s)} = \frac{\sigma^2}{\gamma} \left[e^{-\gamma(\tau-s)} - e^{-2\gamma(\tau-s)} \right] + \sigma\lambda e^{-\gamma(\tau-s)}.$$

Finally,

$$E[r(\tau)|\mathcal{F}(t)] = f(t, \tau) + \int_t^\tau \alpha(s, \tau) ds = f(t, \tau) + \frac{\sigma}{\gamma} \left(1 - e^{-\gamma(\tau-t)} \right) \left[\frac{\sigma}{2\gamma} \left(1 - e^{-\gamma(\tau-t)} \right) + \lambda \right].$$

Therefore, it is sufficient to have a risk-premium such that $-\lambda > \frac{\sigma}{2\gamma}$, to generate the prediction that:

$$E[r(\tau)|\mathcal{F}(t)] < f(t, \tau) \quad \text{for any } \tau.$$

In other words, $\lambda < 0$ is a necessary condition, not sufficient. Notice that when $\lambda = 0$, it always holds that $E(r(\tau)|\mathcal{F}(t)) > f(t, \tau)$.

B. Embedding

We now embed the Ho and Lee model in Section 12.6.2 in the HJM format. In the Ho and Lee model,

$$dr(\tau) = \theta(\tau)d\tau + \sigma d\tilde{W}(\tau),$$

where \tilde{W} is a Q -Brownian motion. By Eq. (12.62) in Section 12.5,

$$f(r, t, T) = -A_2(t, T) + B_2(t, T)r,$$

where $A_2(t, T) = \int_t^T \theta(s)ds - \frac{1}{2}\sigma^2(T-t)$ and $B_2(t, T) = 1$. Therefore, by Eqs. (12.75),

$$\sigma(t, T) = B_2(t, T) \cdot \sigma = \sigma,$$

$$\alpha(t, T) - \sigma(t, T)\lambda(t) = -A_{12}(t, T) + B_{12}(t, T)r + B_2(t, T)\theta(t) = \sigma^2(T-t).$$

Next, we embed the Vasicek model in Section 12.5 in the HJM format. The Vasicek model is:

$$dr(\tau) = (\theta - \kappa r(\tau))d\tau + \sigma d\tilde{W}(\tau),$$

where \tilde{W} is a Q -Brownian motion. By results in Section 12.4,

$$f(r, t, T) = -A_2(t, T) + B_2(t, T)r,$$

where $-A_2(t, T) = -\sigma^2 \int_t^T B(s, T)B_2(s, T)ds + \theta \int_t^T B_2(s, T)ds$, $B_2(t, T) = e^{-\kappa(T-t)}$ and $B(t, T) = \frac{1}{\kappa} [1 - e^{-\kappa(T-t)}]$. By Eqs. (12.75),

$$\sigma(t, T) = \sigma \cdot B_2(t, T) = \sigma \cdot e^{-\kappa(T-t)};$$

$$\alpha(t, T) - \sigma(t, T)\lambda(t) = -A_{12}(t, T) + B_{12}(t, T)r + (\theta - \kappa r)B_2(t, T) = \frac{\sigma^2}{\kappa} [1 - e^{-\kappa(T-t)}] e^{-\kappa(T-t)}.$$

Naturally, this model can never be embedded within a HJM model because it is not of the perfectly fitting type. In practice, condition (12.76) can never hold in the simple Vasicek model. However, the model is embeddable once θ is turned into an infinite dimensional parameter *à la* Hull and White (see Section 12.4).

12.15 Appendix 7: Additional results on string models

Here we prove Eq. (12.78). We have, $\alpha^I(\tau, T) = \frac{1}{2} \int_{\tau}^T g(\tau, T, \ell_2) d\ell_2 + \text{cov}(\frac{dP}{P}, \frac{d\xi}{\xi})$, where

$$g(\tau, T, \ell_2) \equiv \int_{\tau}^T \sigma(\tau, \ell_1) \sigma(\tau, \ell_2) \psi(\ell_1, \ell_2) d\ell_1.$$

Differentiation of the *cov* term is straight forward. Moreover,

$$\begin{aligned} \frac{\partial}{\partial T} \int_{\tau}^T g(\tau, T, \ell_2) d\ell_2 &= g(\tau, T, T) + \int_{\tau}^T \frac{\partial g(\tau, T, \ell_2)}{\partial T} d\ell_2 \\ &= \sigma(\tau, T) \left[\int_{\tau}^T \sigma(\tau, x) [\psi(x, T) + \psi(T, x)] dx \right] \\ &= 2\sigma(\tau, T) \left[\int_{\tau}^T \sigma(\tau, x) \psi(x, T) dx \right]. \end{aligned}$$

12.16 Appendix 8: Changes of numéraire and Jamshidian's (1989) formula

Consider the following *change-of-numéraire* arithmetics. Let

$$\frac{dX_j}{X_j} = \mu_j d\tau + \sigma_j dW, \quad j \in \{A, B\}.$$

We have:

$$\frac{d(X_A/X_B)}{X_A/X_B} = (\mu_A - \mu_B + \sigma_B^2 - \sigma_A \sigma_B) d\tau + (\sigma_A - \sigma_B) dW. \quad (12A.7)$$

Next, we apply this result to the process $y(\tau, S) \equiv \frac{P(\tau, S)}{P(\tau, T)}$, under Q_F^S as well as under Q_F^T . We aim to derive the solution of $y(\tau, S)$ at T , viz

$$y(T, S) \equiv \frac{P(T, S)}{P(T, T)} = P(T, S) \text{ under } Q_F^S \text{ as well as under } Q_F^T,$$

which would allow us to calculate the two probabilities in Eq. (12.83).

By Itô's lemma, the PDE (12.37) and the fact that $P_\tau = -BP$,

$$\frac{dP(\tau, x)}{P(\tau, x)} = r d\tau - \sigma B(\tau, x) d\tilde{W}(\tau), \quad x \geq T.$$

By applying Eq. (12A.7) to $y(\tau, S)$,

$$\frac{dy(\tau, S)}{y(\tau, S)} = \sigma^2 [B^2(\tau, T) - B(\tau, T)B(\tau, S)] d\tau - \sigma [B(\tau, S) - B(\tau, T)] d\tilde{W}(\tau). \quad (12A.8)$$

All we need to do now is to change measure with the tools of Appendix 3. We have that:

$$dW^{Q_F^x}(\tau) = d\tilde{W}(\tau) + \sigma B(\tau, x) d\tau$$

is a Brownian motion under the x -forward martingale probability. Replace then $W^{Q_F^x}$ into Eq. (12A.8), then integrate, and obtain:

$$\begin{aligned} \frac{y(T, S)}{y(t, S)} &= P(T, S) \frac{P(t, T)}{P(t, S)} = e^{-\frac{1}{2}\sigma^2 \int_t^T [B(\tau, S) - B(\tau, T)]^2 d\tau - \sigma \int_t^T [B(\tau, S) - B(\tau, T)] dW^{Q_F^T}(\tau)}, \\ \frac{y(T, S)}{y(t, S)} &= P(T, S) \frac{P(t, T)}{P(t, S)} = e^{\frac{1}{2}\sigma^2 \int_t^T [B(\tau, S) - B(\tau, T)]^2 d\tau - \sigma \int_t^T [B(\tau, S) - B(\tau, T)] dW^{Q_F^S}(\tau)}. \end{aligned}$$

Rearranging terms gives Eqs. (12.84) in the main text.

References

- Ait-Sahalia, Y. (1996): "Testing Continuous-Time Models of the Spot Interest Rate." *Review of Financial Studies* 9, 385-426.
- Ahn, C.-M. and H.E. Thompson (1988): "Jump-Diffusion Processes and the Term Structure of Interest Rates." *Journal of Finance* 43, 155-174.
- Ang, A. and M. Piazzesi (2003): "A No-Arbitrage Vector Autoregression of Term Structure Dynamics with Macroeconomic and Latent Variables." *Journal of Monetary Economics* 50, 745-787.
- Balduzzi, P., S. R. Das, S. Foresi and R. K. Sundaram (1996): "A Simple Approach to Three Factor Affine Term Structure Models." *Journal of Fixed Income* 6, 43-53.
- Black, F. (1976): "The Pricing of Commodity Contracts." *Journal of Financial Economics* 3, 167-179.
- Black, F. and M. Scholes (1973): "The Pricing of Options and Corporate Liabilities." *Journal of Political Economy* 81, 637-659.
- Brace, A., D. Gatarek and M. Musiela (1997): "The Market Model of Interest Rate Dynamics." *Mathematical Finance* 7, 127-155.
- Brigo, D. and F. Mercurio (2006): *Interest Rate Models—Theory and Practice, with Smile, Inflation and Credit*. Springer Verlag (2nd Edition).
- Brunnermeier, M. (2009): "Deciphering the Liquidity and Credit Crunch 2007-08." *Journal of Economic Perspectives* 23, 77-100.
- Carverhill, A. (1994): "When is the Short-Rate Markovian?" *Mathematical Finance* 4, 305-312.
- Cochrane, J. H. and M. Piazzesi (2005): "Bond Risk Premia." *American Economic Review* 95, 138-160.
- Collin-Dufresne, P. and R. S. Goldstein (2002): "Do Bonds Span the Fixed-Income Markets? Theory and Evidence for Unspanned Stochastic Volatility." *Journal of Finance* 57, 1685-1729.
- Conley, T. G., L. P. Hansen, E. G. J. Luttmer and J. A. Scheinkman (1997): "Short-Term Interest Rates as Subordinated Diffusions." *Review of Financial Studies* 10, 525-577.
- Cox, J. C., J. E. Ingersoll and S. A. Ross (1979): "Duration and the Measurement of Basis Risk." *Journal of Business* 52, 51-61.
- Cox, J. C., J. E. Ingersoll and S. A. Ross (1985): "A Theory of the Term Structure of Interest Rates." *Econometrica* 53, 385-407.
- Dai, Q. and K. J. Singleton (2000): "Specification Analysis of Affine Term Structure Models." *Journal of Finance* 55, 1943-1978.
- Duffie, D. and R. Kan (1996): "A Yield-Factor Model of Interest Rates." *Mathematical Finance* 6, 379-406.

- Duffie, D. and K. J. Singleton (1999): "Modeling Term Structures of Defaultable Bonds." *Review of Financial Studies* 12, 687-720.
- Estrella, A. and G. Hardouvelis (1991): "The Term Structure as a Predictor of Real Economic Activity." *Journal of Finance* 46, 555-76.
- Fama, E. F. and R. R. Bliss (1987): "The Information in Long-Maturity Forward Rates." *American Economic Review* 77, 680-692.
- Fong, H. G. and O. A. Vasicek (1991): "Fixed Income Volatility Management." *The Journal of Portfolio Management* (Summer), 41-46.
- Geman, H. (1989): "The Importance of the Forward Neutral Probability in a Stochastic Approach to Interest Rates." Unpublished working paper, ESSEC.
- Geman H., N. El Karoui and J. C. Rochet (1995): "Changes of Numeraire, Changes of Probability Measures and Pricing of Options." *Journal of Applied Probability* 32, 443-458.
- Goldstein, R. S. (2000): "The Term Structure of Interest Rates as a Random Field." *Review of Financial Studies* 13, 365-384.
- Harvey, C. R. (1991): "The Term Structure and World Economic Growth." *Journal of Fixed Income* 1, 4-17.
- Harvey, C. R. (1991): "The Term Structure Forecasts Economic Growth." *Financial Analysts Journal* May/June 6-8.
- Heaney, W. J. and P. L. Cheng (1984): "Continuous Maturity Diversification of Default-Free Bond Portfolios and a Generalization of Efficient Diversification." *Journal of Finance* 39, 1101-1117.
- Heath, D., R. Jarrow and A. Morton (1992): "Bond Pricing and the Term-Structure of Interest Rates: a New Methodology for Contingent Claim Valuation." *Econometrica* 60, 77-105.
- Ho, T. S. Y. and S.-B. Lee (1986): "Term Structure Movements and the Pricing of Interest Rate Contingent Claims." *Journal of Finance* 41, 1011-1029.
- Hördahl, P., O. Tristani and D. Vestin (2006): "A Joint Econometric Model of Macroeconomic and Term Structure Dynamics." *Journal of Econometrics* 131, 405-444.
- Hull, J. C. (2003): *Options, Futures, and Other Derivatives*. Prentice Hall. 5th edition (International Edition).
- Hull, J. C. and A. White (1990): "Pricing Interest Rate Derivative Securities." *Review of Financial Studies* 3, 573-592.
- Jagannathan, R. (1984): "Call Options and the Risk of Underlying Securities." *Journal of Financial Economics* 13, 425-434.
- Jamshidian, F. (1989): "An Exact Bond Option Pricing Formula." *Journal of Finance* 44, 205-209.

- Jamshidian, F. (1997): "Libor and Swap Market Models and Measures." *Finance and Stochastics* 1, 293-330.
- Jöreskog, K. G. (1967): "Some Contributions to Maximum Likelihood Factor Analysis." *Psychometrika* 32, 443-482.
- Karlin, S. and H. M. Taylor (1981): *A Second Course in Stochastic Processes*. San Diego: Academic Press.
- Kennedy, D. P. (1994): "The Term Structure of Interest Rates as a Gaussian Random Field." *Mathematical Finance* 4, 247-258.
- Kennedy, D. P. (1997): "Characterizing Gaussian Models of the Term Structure of Interest Rates." *Mathematical Finance* 7, 107-118.
- Kessel, R. A. (1965): "The Cyclical Behavior of the Term Structure of Interest Rates." National Bureau of Economic Research Occasional Paper No. 91.
- Kloeden, P. and E. Platen (1992): *Numeric Solutions of Stochastic Differential Equations*. Berlin: Springer Verlag.
- Knez, P. J., R. Litterman and J. Scheinkman (1994): "Explorations into Factors Explaining Money Market Returns." *Journal of Finance* 49, 1861-1882.
- Lamberton, D. and B. Lapeyre (1997): *Introduction au Calcul Stochastique Appliqué à la Finance*. Paris: Ellipses.
- Langtieg, T. (1980): "A Multivariate Model of the Term Structure of Interest Rates." *Journal of Finance* 35, 71-97.
- Laurent, R. D. (1988): "An Interest Rate-Based Indicator of Monetary Policy." *Federal Reserve Bank of Chicago Economic Perspectives* 12, 3-14.
- Laurent, R. D. (1989): "Testing the Spread." *Federal Reserve Bank of Chicago Economic Perspectives* 13, 22-34.
- Litterman, R. and J. Scheinkman (1991): "Common Factors Affecting Bond Returns." *Journal of Fixed Income* 1, 54-61.
- Litterman, R., J. Scheinkman, and L. Weiss (1991): "Volatility and the Yield Curve." *Journal of Fixed Income* 1, 49-53.
- Longstaff, F. A. and E. S. Schwartz (1992): "Interest Rate Volatility and the Term Structure: A Two-Factor General Equilibrium Model." *Journal of Finance* 47, 1259-1282.
- Mele, A. (2003): "Fundamental Properties of Bond Prices in Models of the Short-Term Rate." *Review of Financial Studies* 16, 679-716.
- Mele, A. and F. Fornari (2000): *Stochastic Volatility in Financial Markets: Crossing the Bridge to Continuous Time*. Boston: Kluwer Academic Publishers.
- Merton, R. C. (1973): "Theory of Rational Option Pricing." *Bell Journal of Economics and Management Science* 4, 141-183.

- Miltersen, K., K. Sandmann and D. Sondermann (1997): "Closed Form Solutions for Term Structure Derivatives with Lognormal Interest Rate." *Journal of Finance* 52, 409-430.
- Rebonato, R. (1998): *Interest Rate Option Models*. Wiley.
- Rebonato, R. (1999): *Volatility and Correlation*. Wiley.
- Ritchken, P. and L. Sankarasubramanian (1995): "Volatility Structure of Forward Rates and the Dynamics of the Term Structure." *Mathematical Finance* 5, 55-72.
- Rothschild, M. and J. E. Stiglitz (1970): "Increasing Risk: I. A Definition." *Journal of Economic Theory* 2, 225-243.
- Sandmann, K. and D. Sondermann (1997): "A Note on the Stability of Lognormal Interest Rate Models and the Pricing of Eurodollar Futures." *Mathematical Finance* 7, 119-125.
- Santa-Clara, P. and D. Sornette (2001): "The Dynamics of the Forward Interest Rate Curve with Stochastic String Shocks." *Review of Financial Studies* 14, 149-185.
- Stanton, R. (1997): "A Nonparametric Model of Term Structure Dynamics and the Market Price of Interest Rate Risk." *Journal of Finance* 52, 1973-2002.
- Stock, J. H. and M. W. Watson (1989): "New Indexes of Coincident and Leading Economic Indicators." In: Blanchard, O. J. and S. Fischer (Eds.): *NBER Macroeconomics Annual 1989*, MIT Press, 352-394.
- Stock, J. H. and M. W. Watson (2003): "Forecasting Output and Inflation: The Role of Asset Prices," *Journal of Economic Literature* 41, 788-829.
- Vasicek, O. (1977): "An Equilibrium Characterization of the Term Structure." *Journal of Financial Economics* 5, 177-188.
- Veronesi, P. (2010): *Fixed Income Securities: Valuation, Risk and Risk Management*. John Wiley and Sons.

13

Risky debt and credit derivatives

13.1 Introduction

13.2 The classics: Modigliani-Miller irrelevance results

Modigliani and Miller (1958) consider an economy where firms can be sorted by the expected returns of their shares, according to the sector, or class, they belong to. Let π be the constant, expected profit paid off by the each firm within sector k , and E_U be the price of an unlevered firm's share. Under standard conditions, we have that $E_U = \sum_{t=1}^{\infty} (1 + \rho_k)^{-t} \pi$, where ρ_k is the risk-adjusted discount rate prevailing in sector k , such that the return on equity (ROE) for the unlevered firm is,

$$\rho_k = \frac{\pi}{E_U},$$

a constant for all the unlevered firms belonging to sector k . Naturally, the value of the firm is equal to the value of equity, $V_U = E_U$, say. Next, consider a levered firm operating within the k -th sector. This firm issues debt with nominal value equal to D such that its value denoted as V_L , equals the sum of equity and debt, $V_L = E_L + D$. In the absence of any market frictions, we have the following irrelevance result:

THEOREM 13.1 (Modigliani & Miller theorem). *In the absence of arbitrage and frictions, the market value of any firm is independent of its capital structure and is given by discounting its expected profits at the discount rate appropriate to its class: $V_j = \frac{\pi}{\rho_k}$, for any firm $j \in \{U, L\}$ in class k .*

In other words, the return on investment (ROI), defined as $\rho_k = \frac{\pi}{V_j}$, is the same for two firms that earn the same expected profit π , regardless of the capital structure. Naturally, the ROE and ROI are the same for the unlevered firm.

The proof of Theorem 13.1 can proceed by applying the modern tools reviewed in Chapter 2 through 4. For sake of completeness, we use the original Modigliani and Miller arguments, which are very simple. Consider two firms: a first, unlevered and a second, levered. They both earn the same expected profit, π . Suppose to purchase the shares of the unlevered firm and

borrow the same amount of money issued by the levered firm. In the absence of arbitrage or any frictions, the value of this portfolio should equal the value of the levered firm, which is possible as soon as the value of the levered and the unlevered firm are the same.

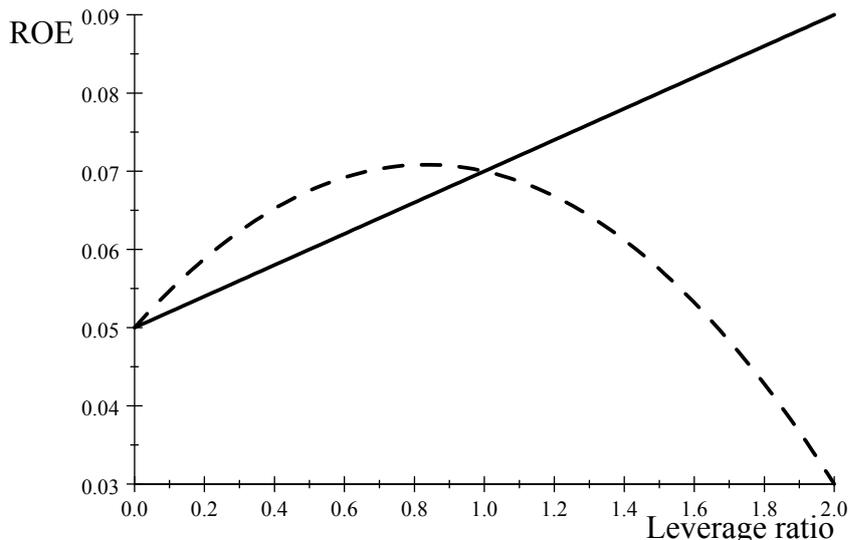
Mathematically, given an arbitrary $\alpha \in (0, 1)$, we do the following trade: (i) we buy $N_U = \alpha \frac{E_L + D}{E_U} = \alpha \frac{V_L}{V_U}$ of the unlevered firm; (ii) we sell $N_L = \alpha$ shares of the levered firm. These two trades make the balance of the position worth $-N_U E_U + N_L E_L = -\alpha D$, and so (iii) we borrow αD at the interest rate r , to make this initial position worthless. This portfolio yields: (i) $+N_U \pi$, due to the purchase of the shares of the unlevered firm, (ii) $-N_L (\pi - rD)$, due to the sale of the shares of the levered firm, which of course has to pay interests on its debt, and (iii) $-r\alpha D$, arising to honour the debt we are making to build up the worthless portfolio. Summing up, the profits are $N_U \pi - N_L (\pi - rD) - r\alpha D = \alpha \left(\frac{V_L}{V_U} - 1 \right) \pi$. If $V_L > V_U$, we have an arbitrage opportunity as we may make money out of a worthless portfolio, and if $V_L < V_U$, we have an arbitrage as well, as we could reverse the positions of the worthless portfolio. So we need to have that $V_L = V_U = E_U = \frac{\pi}{\rho_k}$.

[As mentioned, Theorem 13.1 can be proved through the modern tools in Chapters 2 through 4]

We have: $\pi = \text{ROI} \cdot V$. Therefore,

$$\text{ROE} = \frac{\pi - rD}{E} = \frac{\text{ROI} \cdot (E + D) - rD}{E} = \text{ROI} + (\text{ROI} - r) \frac{D}{E}.$$

If the financial conditions of the firm do not affect the interest rate on debt, the ROE is increasing in the leverage ratio, $\frac{D}{E}$, provided $\text{ROI} > r$. This situation arises when the arbitrage arguments underlying Theorem 13.1 assume no-arbitrage trades can be implemented with a cost of borrowing money equal to that of the firm. In the presence of market frictions such as asymmetric information between borrowers and lenders, this needs not to be the case. For example, debt markets might be concerned about the size of the leverage ratio. Assume, for example, that $r = f(\ell)$, where $\ell = \frac{D}{E}$, and in particular that $f(\ell) = 0.03\ell$. Then, we have that: $\text{ROE} = \text{ROI} + (\text{ROI} - 0.03\ell)\ell$. The picture below depicts the behavior of ROE as a function of ℓ , assuming that $\text{ROI} = 5\%$ and that the risk-free rate in case of no such frictions is $r = 3\%$.



The solid line depicts the ROE for a firm sustaining a cost of debt independent of the leverage ratio, with $ROI = 5\%$ and $r = 2\%$. The dashed line is the ROE for a firm that has a cost of debt increasing in the leverage ratio ℓ , $r(\ell) = 0.03\ell$.

Consider the firm with cost of capital depending on the current leverage ratio, ℓ . For a low level of ℓ , the ROE increases with ℓ , so as to magnify the difference $ROI - 0.03\ell$ through the multiplying effect $(ROI - 0.03\ell) \ell$. However, for higher leverage ratios, the difference $ROI - 0.03\ell$ becomes thinner and thinner, and an increase in ℓ then leads to marginally lower ROE. In this example, there is an interior value for the leverage ratio that maximizes the ROE, which is, approximately, $\ell = 0.83$.

13.3 Conceptual approaches to valuation of defaultable securities

13.3.1 Firm's value, or structural, approaches

Relies on the *structure* of the firm. Shares and bonds as derivatives written upon the firm's assets value. We begin reviewing the Merton's model, which stems from a Modigliani-Miller world, where, as we know, the value of the firm is not affected by the leverage. Note since the beginning the limitation of such a model: it's a model with which we evaluate debt, assuming a world where leverage does not even affect the value of the firm that is issuing the debt!

Consider the following stylized balance sheet.

Stylized balance sheet	
Assets (A)	Equity (E) (Shares)
	Debt (D) (Bonds)

Therefore, we have the accounting identity: $Assets = Equity + Debt$, or

$$A = E + D.$$

When debt expires, debtholders receive the minimum between the nominal value of debt and the value of the assets the firm can liquidate to honour debt. Debtholders are senior claimants. Equity holders are juniors, i.e. they are residual claimants to the firm's assets.

We use these basic insights and illustrate the first approach to the modeling of the risk-structure of interest rates—the Merton-KMV approach. In this approach, equity is the same as a European call option written on the firm's assets, with expiration equal to the debt expiration, and strike equal to the nominal value of debt. The current value of debt equals the value of the assets minus the value of equity, i.e. the value of a risk-free discount bond minus the value of a put option on the firm with strike price equal to the nominal value of debt, as shown by Eq. (13.3) below.

Merton (1974) uses the Black and Scholes (1973) formula to derive the price of debt. The main assumption underlying this model is that the assets of the firm can be traded, and that their value A_t satisfies,¹

$$\frac{dA_t}{A_t} = rdt + \sigma d\tilde{W}_t, \quad (13.1)$$

where \tilde{W}_t is a Brownian motion under the risk-neutral probability, σ is the instantaneous standard deviation, and r is the short-term rate on riskless bonds.

Let N be the nominal value of debt, T be time of expiration of debt; D_t the debt value as of at time $t \leq T$. As argued earlier, shareholders are long a European call option, and the bondholders are residual claimants. Mathematically,

$$D_T = \begin{cases} A_T, & \text{if the firm defaults, i.e. } A_T < N \\ N, & \text{if the firm is solvent, i.e. } A_T \geq N \end{cases}$$

We can decompose the assets value at time T , into the sum of the value of equity and the value of debt, at time T ,

$$D_T = \min \{A_T, N\} = A_T - \max \{A_T - N, 0\}. \quad (13.2)$$

\equiv Equity at T

Note, also, that,

$$D_T = \min \{A_T, N\} = N - \max \{N - A_T, 0\}. \quad (13.3)$$

\equiv Put on the firm

That is, credit risk raises the cost of capital.

¹Eq. (13.1) could be generalized to one in which $dA_t = (rA_t - \delta_t) dt + \sigma A_t d\tilde{W}_t$, where δ_t is the instantaneous cash flow to the firm. This would make the firm value equal to $A_0 = \mathbb{E}(\int_0^\infty e^{-rt} \delta_t dt)$. For example, one could take δ_t to be a geometric Brownian motion with parameters g and σ , in which case $A_t = (r - g)^{-1} \delta_t$, forever, but we're just ignoring this complication.

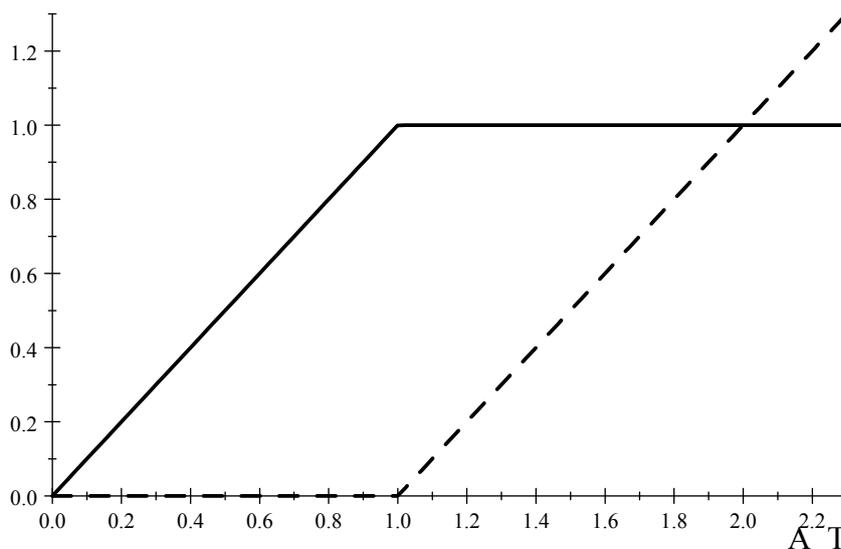


FIGURE 13.1. Dashed line: the value of equity at the debt maturity, T , $\max\{A_T - N, 0\}$, plotted as a function of the value of assets, A_T . Solid line: the value of debt at maturity, $\min\{A_T, N\}$ as a function of A_T . Nominal value of debt is fixed to $N = 1$.

A word on convexity, and risk-taking behavior. Convexity: Managers have incentives to invest in risky assets, as the terminal payoff to them is increasing in the assets volatility, σ . Concavity: The value of debt, instead, is decreasing in the assets volatility, as we shall show in detail in the next section.

13.3.1.1 Merton

The *current* value of the bonds equals the current value of the assets, A_0 , minus the current value of equity. The *current* value of equity can be obtained through the Black & Scholes formula, as equity is a European call option on the firm, struck at N . By Eq. (13.2), and standard risk-neutral evaluation, the *current* value of debt, D_0 , is,

$$D_0 = A_0 \Phi(-d_1) + N e^{-rT} \Phi(d_1 - \sigma\sqrt{T}), \quad d_1 = \frac{\ln(A_0/N) + (r + \frac{1}{2}\sigma^2)T}{\sigma\sqrt{T}}, \quad (13.4)$$

where $\Phi(\cdot)$ denotes the distribution function of a standard normal variable.²

²For the details, note that $D_0 = e^{-rT} \mathbb{E}[D_T | A_0]$ and, then, by Eq. (13.2),

$$D_0 = e^{-rT} \mathbb{E}(A_T | A_0) - e^{-rT} \mathbb{E}[\max\{A_T - N, 0\} | A_0] = A_0 - [A_0 \Phi(d_1) - N e^{-rT} \Phi(d_1 - \sigma\sqrt{T})],$$

where the last equality follows by the Black & Scholes formula. Eq. (13.4) follows after rearranging terms in the previous equation.

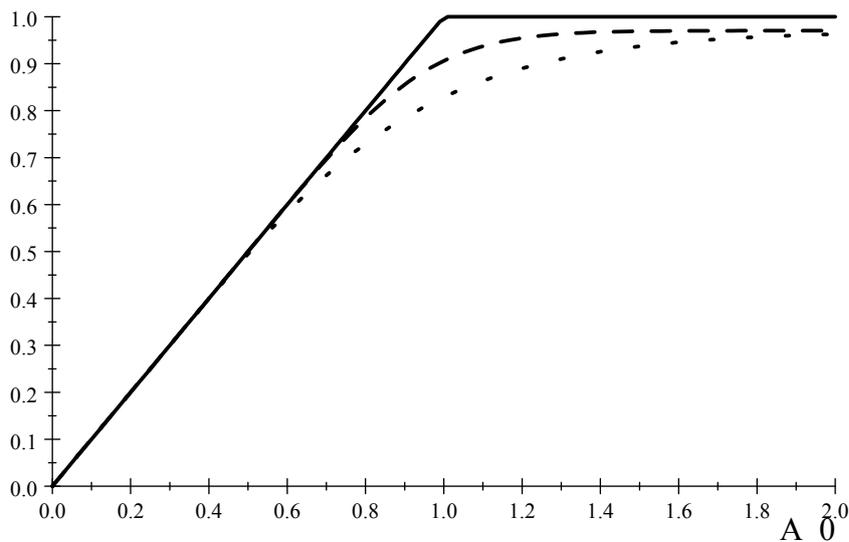


FIGURE 13.2. Solid line: the no-arbitrage bound, $\min\{A_0, N\}$, depicted as a function of A_0 , when the nominal value of debt is fixed to $N = 1$. Dashed line: the bond value predicted by the Merton's model when $T = 1$, $r = 3\%$ and $\sigma = 20\%$, annualized. Dotted line: same as the dashed line, but with a larger assets volatility, $\sigma = 40\%$.

Bond prices are decreasing in the assets volatility as bad outcomes are exaggerated on the downside, due to the concavity properties depicted in Figure 13.1. Note that the property that the term structure of interest rates increases with the volatility of the fundamentals is quite sharp here. In Chapter 12 (Section 12.3.4.1), it was argued that the relation between the yield curve and the volatility of the fundamentals (i.e. the volatility of the short-term rate) was quite complex, as it depends on which of two effects dominate—a “convexity” and a “risk-premium” effect. In bad times, it should be the risk-premium effect to dominate, thereby leading to a positive link between the volatility of the fundamentals and the yield curve. In good times, a convexity effect would lead the yield curve to be negatively related to the volatility of the fundamentals. The prediction that we have for risky debt in this section is, instead, neat: the term structure of interest rates or, as we shall say soon, the “risk-structure of interest rates,” or the “term-spread,” always increases with the volatility of the fundamentals—a prediction that relies on a channel, which is completely distinct from the risk-premium channel discussed in Chapter 12.

The risk-structure of interest rates is obtained with the standard formula for continuously compounded interest rates as,

$$R = -\frac{1}{T} \ln \left(\frac{D_0}{N} \right) = r + s_0,$$

where the term-spread is

$$s_0 = -\frac{1}{T} \ln \left(\frac{A_0}{N} e^{rT} \Phi(-d_1) + \Phi(d_1 - \sigma\sqrt{T}) \right). \quad (13.5)$$

Figure 13.3 depicts the spread predicted by this model. Credit spreads shrink to zero as time-to-maturity becomes smaller and smaller. This property of the model stands in sharp

contrast with the empirical behavior of credit spreads, which are high even for short-maturity bonds. This property arises because the model is driven by Brownian motions, which have continuous sample paths, such that given an assets value $A > N$, the probability of bankruptcy, arising when A hits N from above, approaches zero very fast as time-to-maturity goes to zero. Because credit spreads reflect default probabilities, as explained in detail below (see Eq. (13.9)), credit spreads shrink to zero quickly as time-to-maturity approaches zero.

Naturally, one might end up with credit spreads sufficiently high at short maturities, by assuming the assets value is sufficiently small. For example, in Figure 13.3.1, credit spreads are “high” at short maturities, when $A = 1.1$. However, even with $A = 1.1$, credit spreads are still zero at very short maturities. More fundamentally, requiring such a small value for A is problematic. Firms with such a low assets value would command a much higher spread than that in Figure 13.3.1. All in all, the Brownian motion model in this section lacks some source of risk driving the behavior of short-term spreads. In Section 13.3.2, we will show that this issue can be addressed assuming that firm’s default can be triggered by “jumps.”

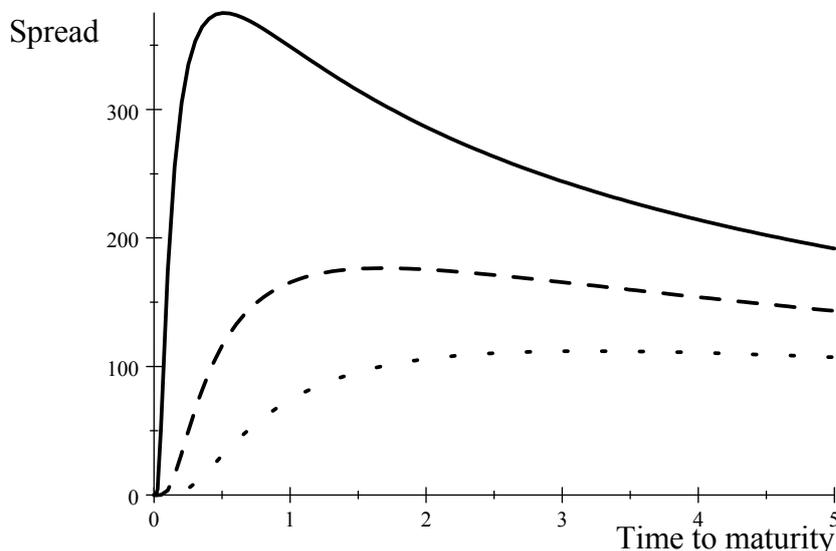


FIGURE 13.3.1. The term structure of spreads, s_0 , in basis points, predicted by Merton’s model, obtained with initial assets values $A_0 = 1.1$ (solid line), $A_0 = 1.2$ (dashed line), and $A_0 = 1.3$ (dotted line). The short-term rate, $r = 3\%$, and assets volatility is $\sigma = 0.20$. Nominal debt $N = 1$.

Naturally, the term-structure of credit spreads has a rather different shape, when the current assets value is below N , as depicted in Figure 13.3.2. In this case, the probability the firm defaults is close to one when time to maturity is close to zero, such that the spreads would then be arbitrarily large as we get closer and closer to maturity. For visualization purposes, Figure 13.3.2 is truncated so as to only include values of the spreads for maturities higher than one year.

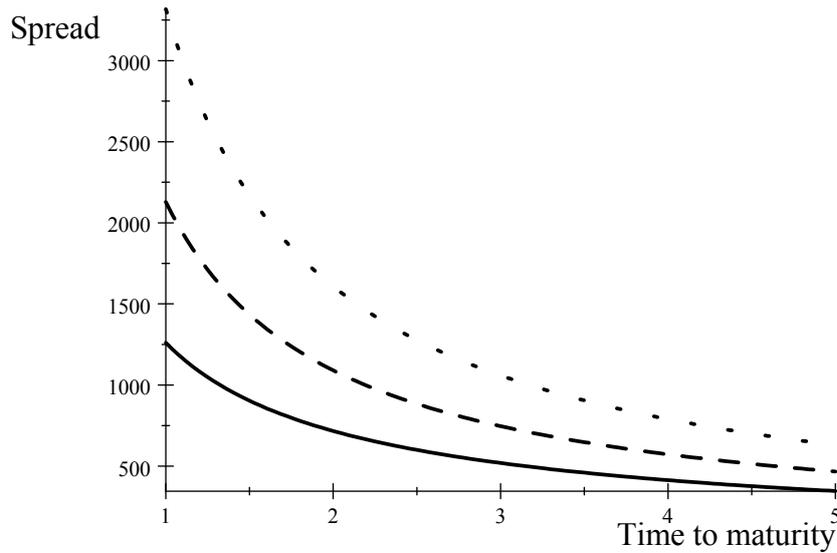


FIGURE 13.3.2. The term structure of spreads, s_0 , in basis points, predicted by Merton's model, obtained with initial assets values $A_0 = 0.9$ (solid line), $A_0 = 0.8$ (dashed line), and $A_0 = 0.7$ (dotted line). The short-term rate, $r = 3\%$, and assets volatility is $\sigma = 0.20$. Nominal debt $N = 1$.

What is the asymptotic behavior of the spread predicted by this model? If $r > \frac{1}{2}\sigma^2$, then, as $T \rightarrow \infty$, the probability of survival for the firm, which we shall show to be $\Phi(d_1 - \sigma\sqrt{T})$ below (see Eq. (13.7)), approaches one, $\Phi(d_1 - \sigma\sqrt{T}) \rightarrow 1$. That is, the assets value is expected to grow so large that default will never occur, such that the bond becomes riskless and $s_0 \rightarrow -\frac{1}{T} \ln \Phi(d_1 - \sigma\sqrt{T}) \rightarrow 0$. Intuitively, when $r > \frac{1}{2}\sigma^2$, the assets volatility is so small, that the exponential trend for A_t will make it unlikely that the assets value will fall below the constant value N . In other words, the Merton's model predicts that in the long-run, things can only go well for the firm, a view quite opposite to that leading to positive spreads for long maturities. Intensity models, such as those analyzed in Section 13.3.2, help mitigate this issue.

We can introduce a useful summary statistics for credit quality: *distance-to-default* (under \mathbb{Q}). We can use the previous model to estimate the likelihood of default for a given firm. First, we develop Eq. (13.2),

$$D_T = \min\{A_T, N\} = A_T \cdot \mathbb{I}_{\{A_T < N\}} + N \cdot \mathbb{I}_{\{A_T \geq N\}},$$

where $\mathbb{I}_{\{\mathcal{E}\}}$ is the indicator function, i.e. $\mathbb{I}_{\{\mathcal{E}\}} = 1$ if the event \mathcal{E} is true and $\mathbb{I}_{\{\mathcal{E}\}} = 0$ if the event \mathcal{E} is false. Second, we have,

$$\begin{aligned} D_0 &= e^{-rT} \mathbb{E}(D_T) \\ &= e^{-rT} [\mathbb{E}(A_T \cdot \mathbb{I}_{\{A_T < N\}}) + N \cdot \mathbb{E}(\mathbb{I}_{\{A_T \geq N\}})] \\ &= e^{-rT} [\mathbb{E}(A_T | \text{Default}) \mathbb{Q}(\text{Default}) + N \cdot \mathbb{Q}(\text{Survival})], \end{aligned} \quad (13.6)$$

where $\mathbb{E}(A_T | \text{Default})$ is the expected assets value given the event of default, $\mathbb{Q}(\text{Default})$ is the probability of default, and $\mathbb{Q}(\text{Survival}) = 1 - \mathbb{Q}(\text{Default})$ is the probability the firms does

not default. The last equality follows by the Law of Iterated Expectations, $\mathbb{E}(A_T \cdot \mathbb{I}_{\{A_T < N\}}) = \mathbb{E}(\mathbb{E}(A_T \cdot \mathbb{I}_{\{A_T < N\}} | \mathbb{I}_{\{A_T < N\}})) = \mathbb{E}(\mathbb{I}_{\{A_T < N\}} \cdot \mathbb{E}(A_T | \mathbb{I}_{\{A_T < N\}})) = \mathbb{E}(\mathbb{I}_{\{A_T < N\}} \cdot \mathbb{E}(A_T | \text{Default}))$. Comparing Eq. (13.6) and Eq. (13.4) reveals that for the Merton's model,

$$\mathbb{Q}(\text{Survival}) = \Phi(d_2), \quad (13.7)$$

where for obvious reasons, the quantity,

$$d_2 = \frac{\ln(A_0/N) + (r - \frac{1}{2}\sigma^2)T}{\sigma\sqrt{T}}, \quad (13.8)$$

is called *distance-to-default*. Distance to default is a useful summary statistics, as it is a very intuitive measure of how far the firm is from defaulting: the higher d_2 , the higher the (risk-adjusted) probability of surviving, $\Phi(d_2)$. The larger the current assets value A_0 is, the less likely it is the firm will default at T .

By Eq. (13.1), we have that $\mathbb{E}(\ln A_T | A_0) = \ln A_0 + (r - \frac{1}{2}\sigma^2)T$, so Eq. (13.8) tells us that distance-to-default is simply the difference $\mathbb{E}(\ln A_T | A_0) - \ln N$, normalized by the standard deviation of the assets over the life of debt. Some, then, might prefer to use the slightly different formula,

$$\text{Distance-to-default} = \frac{\text{Mkt value of Assets} - \text{Default value}}{\text{Mkt value of Assets} * \text{Assets volatility}}.$$

How does the probability of survival for a given firm relate to debt maturity or assets volatility? In Figure 13.4, the probability of survival decreases with, (i) debt maturity and (ii) assets volatility.

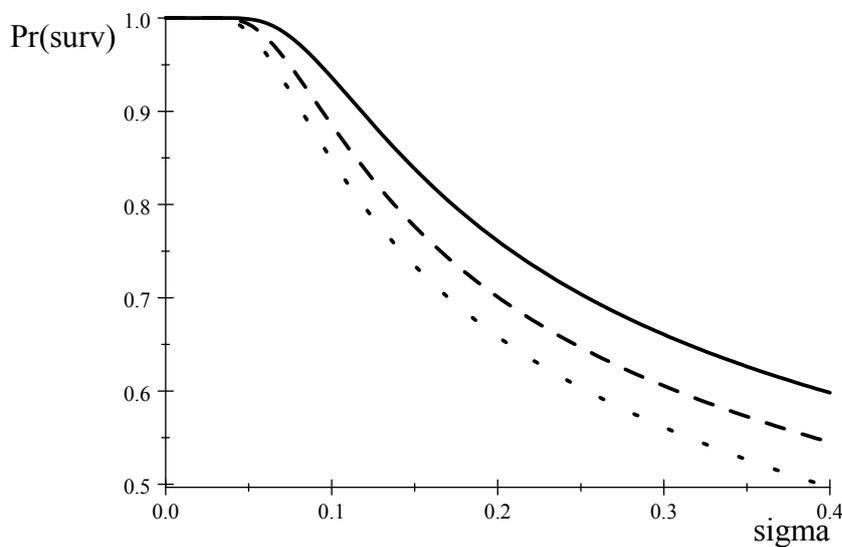
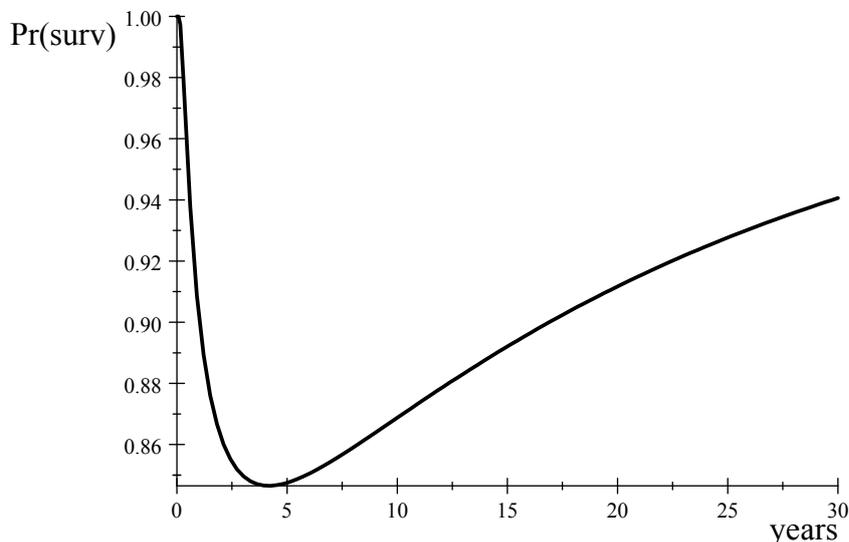


FIGURE 13.4. Probability of survival for a given firm predicted by the Merton's model, $\Phi(d_2)$, depicted as a function of the assets volatility, σ . Assets value is fixed at $A_0 = 1.1$, and plotted are survival probabilities for bonds maturing at $T = 0.5$ years (solid line), $T = 1$ year (dashed line) and $T = 2$ years (dotted line). The short-term rate, $r = 3\%$. Nominal debt $N = 1$.

Property (i) is not a general property, though. For example, we already pointed out that for large T , the probability of survival is close to one as soon $r > \frac{1}{2}\sigma^2$, a condition ensuring the assets value grows so large to ensure default becomes unlikely, eventually. The next picture shows that for a fixed σ , such that $r > \frac{1}{2}\sigma^2$, the probability of survival is non-monotonic in T .



Probability of survival for a given firm predicted by the Merton's model, $\Phi(d_2)$, depicted as a function of time-to-maturity, when assets value is fixed at $A_0 = 1.1$, and assets volatility is $\sigma = 0.10$. The short-term rate, $r = 3\%$. Nominal debt $N = 1$.

When T is small, the first term of d_2 in Eq. (13.8), and the distance to default increases with maturity, whereas for T large, the second term of d_2 dominates, and distance to default becomes eventually large. Non-monotonicities arise even at finite maturities, once we consider low values of A_0 , in which case the relation between maturity and probability of survival can be increasing or decreasing, according to the values of σ , as shown in Figure 13.5. Intuitively, when $A_0 \approx N$, the probability of survival is:

$$\mathbb{Q}(\text{Survival}) = \Phi(d_2), \text{ with } d_2 = \frac{\ln\left(\frac{A_0}{N}\right) + \left(r - \frac{1}{2}\sigma^2\right)T}{\sigma\sqrt{T}} \approx \frac{\left(r - \frac{1}{2}\sigma^2\right)\sqrt{T}}{\sigma},$$

such that the survival probability decreases in T for large σ although then it increases in T for small σ . The intuition underlying this property is that for large σ , the probability the assets value will end up below N from $A_0 \approx N$ can only increase with time to maturity, T . Analytically, $\mathbb{E}(\ln A_T | A_0) = \ln A_0 + \left(r - \frac{1}{2}\sigma^2\right)T \approx \ln N + \left(r - \frac{1}{2}\sigma^2\right)T$, such that the probability the assets value will be below N does indeed increase with σ .

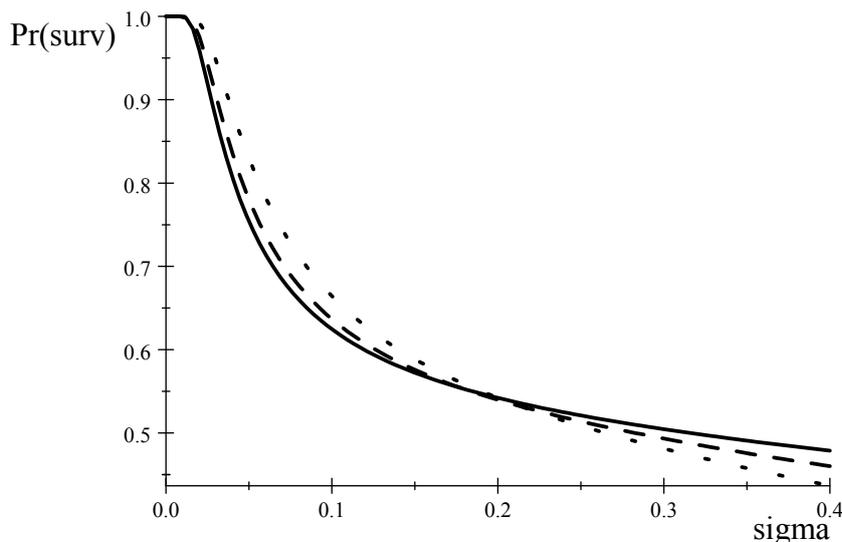


FIGURE 13.5. Probability of survival for a given firm predicted by the Merton's model, $\Phi(d_2)$, depicted as a function of the assets volatility, σ . Assets value is fixed at $A_0 = 1.01$, and plotted are survival probabilities for bonds maturing at $T = 0.5$ years (solid line), $T = 1$ year (dashed line) and $T = 2$ years (dotted line). The short-term rate, $r = 3\%$. Nominal debt $N = 1$.

A final useful concept is that of Loss-given-default, under \mathbb{Q} . Comparing Eq. (13.6) with Eq. (13.4) reveals another property of the Merton's model,

$$\mathbb{E}(A_T | \text{Default}) = \frac{A_0 e^{rT} \Phi(-d_1)}{\mathbb{Q}(\text{Default})} = A_0 e^{rT} \frac{\Phi(-d_1)}{\Phi(-d_2)} = \mathbb{E}(A_T) \frac{\Phi(-d_1)}{\Phi(-d_2)} \leq \mathbb{E}(A_T).$$

Recovery rates are defined as the fraction of the bond value the bondholders expect to obtain in the event of default, at maturity:

$$\text{Recovery Rate} = \frac{\mathbb{E}(A_T | \text{Default})}{N} = \frac{A_0 e^{rT} \Phi(-d_1)}{N \Phi(-d_2)}.$$

Loss-given-default is defined as the fraction of the bond value the bondholders expect to lose in the event of default, at maturity:

$$\text{Loss-given-default} = 1 - \text{Recovery Rate}.$$

Finally, by Eq. (13.5), we can write,

$$\begin{aligned} s_0 &= -\frac{1}{T} \ln \left(\frac{A_0}{N} e^{rT} \Phi(-d_1) + \Phi(d_2) \right) \\ &= -\frac{1}{T} \ln (\text{Recovery Rate} \cdot \mathbb{Q}(\text{Default}) + \mathbb{Q}(\text{Survival})) \\ &\approx \frac{1}{T} [\text{Loss-given-default} \cdot \mathbb{Q}(\text{Default})]. \end{aligned} \tag{13.9}$$

This is actually a general formula, which goes through beyond the Merton's model). It can easily be obtained through Eq. (13.6).

An important note. Previously, we defined survival probabilities, distance-to-default, and loss-given-default, under the risk-adjusted probability \mathbb{Q} . To calculate the same objects under the true probability \mathbb{P} , we replace r with the asset growth rate under the physical probability, μ , in the formulae for the survival probabilities, $\Phi(d_2)$, distance-to-default, d_2 , and loss-given-default.

However, it is hard to estimate μ for many single names. Moody's KMV EDFTM are based on dynamic structural models like these, although the details are not publicly known. Finally, we could use historical data about *default frequencies* to estimate the probability that a given single name within a certain industry will default. These frequencies are based on samples of firms that have defaulted in the past, with similar characteristics to those of the firm under evaluation (in terms, for instance, of distance-to-default).

How to estimate A_t and σ ? One algorithm is to start with some σ equal to the volatility of equity returns, say $\sigma^{(0)}$, and use Merton's formula for equity, to extract $A_t^{(0)}$ for each date $t \in \{1, \dots, T\}$, where T is the sample size. Then, use $A_t^{(0)}$ to compute the standard deviation of $\ln(A_t^{(0)}/A_{t-1}^{(0)})$. This gives say $\sigma^{(1)}$, which can be used as the new input to the Merton's formula to extract say $A_t^{(1)}$. We obtain a sequence of $(A_t^{(i)})_{t=1}^T$ and $\sigma^{(i)}$, and we stop for i sufficiently large, according to some criterion.

13.3.1.2 One example

Assume the assets value of a given firm is $A_0 = 110$, and that the instantaneous volatility of the assets value growth is $\sigma = 30\%$, annualized. The safe interest rate is $r = 2\%$, annualized, and the expected growth rate of the assets value is $\mu = 5\%$, annualized. The firm has outstanding debt with nominal value $N = 100$, which expires in two years.

First, we compute the distance-to-default implied by the Merton's model, which is,

$$\text{D-t-D} = \frac{\ln\left(\frac{A_0}{N}\right) + \left(r - \frac{1}{2}\sigma^2\right)T}{\sigma\sqrt{T}} = \frac{\ln(1.1) + \left(0.02 - \frac{1}{2}0.3^2\right)2}{0.3\sqrt{2}} = 0.10680.$$

Accordingly, the probability of default is,

$$1 - \Phi(0.10680) = 1 - 0.54253 = 0.45747.$$

We can compute the same probability, under the physical measure, by simply replacing $r = 2\%$ with $\mu = 5\%$, in the formula for D-t-D. We have,

$$\text{D-t-D}_{\text{physical}} = \frac{\ln\left(\frac{A_0}{N}\right) + \left(\mu - \frac{1}{2}\sigma^2\right)T}{\sigma\sqrt{T}} = \frac{\ln(1.1) + \left(0.05 - \frac{1}{2}0.3^2\right)2}{0.3\sqrt{2}} = 0.24822.$$

Therefore, the probability of default under the physical distribution is,

$$1 - \Phi_{\text{physical}}(0.24822) = 1 - 0.59802 = 0.40198.$$

It is, of course, lower under the physical probability than under the risk-neutral probability, due to the larger asset growth rate, $\mu > r$.

Finally, we can compute the spread on this bond, which is given by:

$$\text{Spread} = -\frac{1}{T} \ln \left(\frac{A_0}{N} e^{rT} \Phi(-d_1) + \Phi(d_2) \right),$$

where $d_2 = D-t-D$, and $d_1 = d_2 + \sigma\sqrt{T}$. So we have,

$$\begin{aligned}\text{Spread} &= -\frac{1}{2} \ln \left(1.1e^{0.02*2} \Phi \left(- \left(0.10680 + 0.30 * \sqrt{2} \right) \right) + \Phi (0.10680) \right) \\ &= -\frac{1}{2} \ln \left(1.1e^{0.02*2} * 0.29769 + 0.54253 \right) \\ &= 6.20\%.\end{aligned}$$

13.3.1.3 Stocks and bonds

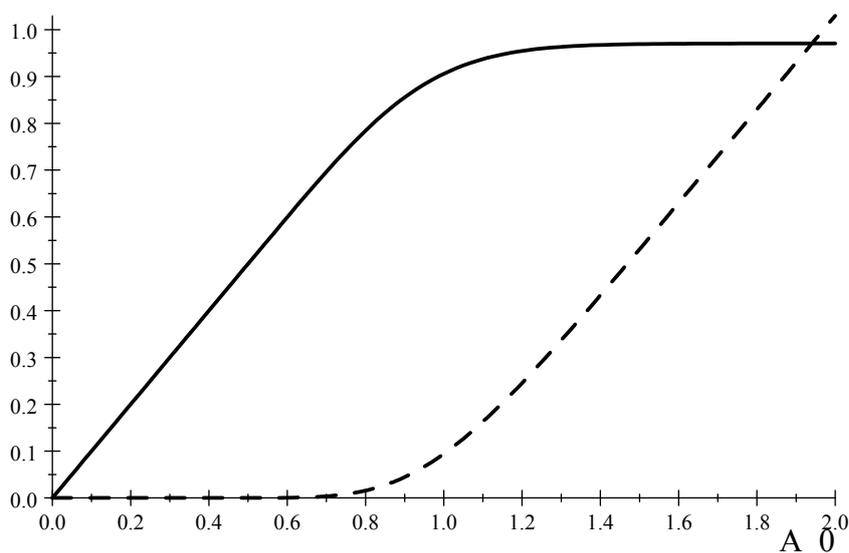
What happens to the price of a stock and a bond, after an adverse shock hits the fundamentals, A ? Granted, they both fall. But which of the two prices will drop more? It is an interesting question. For example, when the rating agency Standard & Poor's downgraded the US debt from AAA to AA+ for the first time in its history on the 5th of August 2011, the US equity market sunk nearly 7% on the first trading day following the announcement. Somehow paradoxically, Treasuries rallied on the very same day, a phenomenon many commentators described as a flight-to-quality response to a quite unique event in financial history.

Yet the flight-to-quality theory would also have its detractors. Indeed, during those days, doubts were arising as to whether the downgrade decision would really warrant a revolutionary view of the global financial system. For example, it might a reasonable assumption that a portion of the bad signals about US debt were related to a political gaming around the increase in the debt ceiling that was likely occurring before the 5th of August—a gaming that might have led to delinquencies. (In the US, the Congress has to approve an increase in debt capacity.) This gaming was arguably very transitory. Moreover, AA+ debt is still accepted as a collateral in many transactions, and is considered to be a high grade debt in most, if not all, investment mandates. Moreover, an issue at the time was to speculate whether the Standard & Poor's decision would also be followed by the other rating agencies, Moody's and Fitch, which might have made the 5th of August decision more solidly grounded, so to speak. (At the time of writing, Moody's and Fitch have still not downgraded US debt.) Also, the Standard & Poor's decision was not totally unexpected. Rumours about a possible downgrade in US debt would start circulate even many months earlier indeed—not to mention during the days preceding the agency's final announcement. Finally, even disregarding these arguments, how could we really say US debt was a safe-heaven asset over a period when the US debt/GDP ratio was flying at about 100% anyway?

Whilst these facts help mitigate the flight-to-quality theory, there remain issues: after all, if the Standard & Poor's decision is not to be taken as such a bad and unexpected piece of news, what else might explain the subsequent spectacular drop in the equity market? In fact, around the 5th of August, the US were flooded with substantially bad news about the fundamentals of the economy, with many leading indicators reaching levels historically consistent with recessions, making it likely that the US economy would spiral towards a recession—the second recession possibly occurring in less than four years, just after the subprime crisis related recession, discussed in Section 13.5. As we know from discussions in Chapters 12 and 13, recession fears translate into an expectation future short rates will lower or, at least, an expectation future short rates would not raise, as a result of an attempt of the FED to stimulate growth. This explanation helps rationalize the rally in Treasuries around the 5th of August and, clearly, the drop in equity markets. (Clearly, a drop in the expected short rates likely boosts medium- and long-term Treasuries, but is not guarantee of any future recovery whatsoever!) In this context, the Standard & Poor's downgrade might likely have come as yet-another negative signal during a particularly severe period, contributing to further compromise the general investment

climate, also affected by adverse developments in the debt crisis in Europe. Therefore, while a rally in Treasuries might have stemmed from a flight-to-quality effect, related to an uncertainty premium about the occurrence of possible disorderly tail events, another component of this rally might have related to market expectations about the FED policy response to business cycle developments.

There is a third channel that potentially helps explain the equity crash and the rally in US debt, following the Standard & Poor's downgrade decision. After all, debt is less risky than stocks (due to subordination), and bad news about the fundamentals should affect stocks more than bonds. The next picture depicts the price functions of a bond and a stock as predicted by the Merton's model. Naturally, the Merton's model is a caricature of the events we are discussing in this section, which relate to sovereign debt, not firm's debt! At the same time, this simple model helps shed light into these events. It predicts that bond prices do not move too much when the probability of default is small (i.e. when A_0 is large enough), which might roughly correspond to the situation where an agency announces a name to be downgraded from AAA to AA+. Instead, stocks prices fall, and substantially, due to convexity, following an increase in the probability of default (which occurs when A_0 falls in the Merton's model).



The solid line depicts the value of the bond and the dashed line depicts that of the stock, as predicted by the Merton's model, when the nominal value of debt is $N = 1$, and $T = 1$, $r = 3\%$ and $\sigma = 20\%$, annualized.

Whilst we do not have a model that explains the joint behavior of stock and sovereign bond prices, the Merton's model makes a sharp prediction about the behavior of stock and bond prices for a firm. It is an open issue whether these predictions would also apply to market developments related to sovereigns, although there are no apparent reasons they would not. Naturally, the informal arguments of this section do not aim to rule out a role, over the events around the 5th of August 2011, for flight-to-quality effects and market expectations about the business cycle and the FED. Rather, they put forward one additional, and perhaps complementary hypothesis: even absent any flight-to-quality effects or business cycle developments, bond prices should not substantially fall, once the probability of default for a name still remains very small.

13.3.1.4 First passage

The timing of default can be triggered by some exogenously specified events. For example, default occurs if the value of the assets hits some exogenously lower bound even before the expiration of debt. These models are known as “first passage” models, because they rely on mathematical techniques that solve for the probability the first time the assets value hit some exogenous “barrier,” as in Black and Cox (1976).

13.3.1.5 Strategic defaulting

The timing of default can be endogenous. Managers choose the defaulting barrier (i.e. the assets value that triggers bankruptcy) so as to maximize the equity value. Naturally, strategic defaulting works if the assumptions underlying the Modigliani-Miller theorem do not hold, such that an optimal level of leverage would arise. The mechanism is the following: on the one hand, debt is a tax-shielding device. On the other hand, issuing too much debt increases the likelihood of default, which triggers bankruptcy costs. The first effect raises the value of the firm while the second, decreases the value of the firm. Equity holders choose the value of the asset that triggers bankruptcy to maximize the value of equity. Leland (1994): Long-term debt. Leland and Toft (1996): Extension to finite maturity debt. Anderson and Sundaresan (1996): Debt re-negotiation.

The Leland’s model considers liquidation of the firm as a strategic choice of the equity holders. In fact, the US bankruptcy code includes both a liquidation process (Chapter 7) and a reorganization process (Chapter 11), but Leland’s model only considers firm’s liquidation at bankruptcy. Broadie, Chernov and Sundaresan (2007) generalize this setting to one where the firm may choose to default through a reorganization process, in which case no equity is issued to honour debt services, i.e. coupon payments, as it is instead the case in Leland, as we now explain.

The terms leading to the strategic defaulting in Leland’s model are as follows. First, the value of the assets, A_t , is solution, as usual, to Eq. (13.1). Second, debt is infinitely lived in that, it pays off an instantaneous coupon equal to Cdt , forever, conditionally upon survival; in the absence of default risk, the value of debt would simply equal C/r . Third, tax benefits are assumed to be proportional to the coupon, τCdt . Fourth, there are bankruptcy costs: if the firm defaults at $A = A^B$, recovery is $(1 - \alpha) A^B$. Equity holders choose A^B . Naturally, $A^B < A_0$.

The value of debt is a function of the assets value, A , say $D(A)$. Moreover, the firm finances the net cost of the coupon C by issuing additional equity, until the equity value is zero, i.e. until $A = A^B$, as seen below. Therefore, under the risk-neutral probability, the value of debt satisfies:

$$\underbrace{\frac{d}{dT} \mathbb{E}[D(A_T) | A_0] \Big|_{T=t}}_{=\text{Expected capital gains}} + \underbrace{C}_{=\text{coupon}} = rD(A_t).$$

By Itô’s lemma, this is an ordinary differential equation, subject to the following boundary conditions. First, at bankruptcy, $D(A^B) = (1 - \alpha) A^B$. Second, for large A , debt is substantially riskless, i.e. $\lim_{A \rightarrow \infty} D(A) = \frac{C}{r}$.

The solution to this is,

$$D(A) = (1 - p_B(A)) \frac{C}{r} + p_B(A) [(1 - \alpha) A^B], \quad (13.10)$$

where

$$p_B(A) \equiv \left(\frac{A^B}{A} \right)^{\frac{2r}{\sigma^2}}. \quad (13.11)$$

Note, we may interpret $p_B(A)$ as the present value of £1, contingent on future bankruptcy, as further explained in Appendix 1. Accordingly, $(1 - p_B(A))/r$ is the expected present value of the coupon payments up to bankruptcy.

The total benefits arising from tax shielding are,

$$TB(A) = (1 - p_B(A)) \tau \frac{C}{r}.$$

and the present value of bankruptcy costs is,

$$BC(A) = p_B(A) \alpha A^B.$$

We have,

$$\begin{aligned} \text{Equity} + \text{Debt} &= \text{Value of the firm} \\ &= \text{Value of Assets } (= A) + TB(A) - BC(A). \end{aligned}$$

Summing up,

$$E(A) \equiv \text{Equity} = A - (1 - p_B(A)) (1 - \tau) \frac{C}{r} - p_B(A) A^B.$$

Equity equals (i) the value of the assets, A ; minus (ii) the present value of debt contingent on no-bankruptcy, net of tax benefits, $(1 - p_B(A)) (1 - \tau) \frac{C}{r}$; minus (iii) the present value of debt contingent on bankruptcy, and net of bankruptcy costs, $p_B(A) A^B$. The second term decreases with the default boundary, A^B or, equivalently, $p_B(A)$. The third term, instead, increases with A^B . So the time equityholders wait before declaring bankruptcy, which is inversely related to A^B , affects in opposite ways these two terms. Equityholders choose A^B to maximize the value of equity. Their solution is a default boundary, A^B , such that the value of equity does not change for small changes in the value of the assets A around A^B , or $A^B : E'(A)|_{A=A^B} = 0$, a smooth pasting condition. The result is,

$$A^B = (1 - \tau) \frac{C}{r + \frac{1}{2}\sigma^2}.$$

Similarly as in the American option case, the value of the option to wait increases with uncertainty, σ^2 . Finally, and consistent with the real option theory, it is easy to check that this solution for A^B does maximize the value $E(A; A^B) \equiv E(A)$ in that $A^B : 0 = \partial E(A; A^B) / \partial A^B = 0$.

How is it that tax shielding does not seem to affect the *existence* of a solution to this problem? That is, the default boundary, A^B , still exists, even with $\tau = 0$. This issue is easily resolved. If $\tau = 0$, there are no reasons to issue debt in the first place, as no tax benefits would flow to the firm! In fact, when $\tau > 0$, there is a value of leverage that maximizes the value of the firm, according to simulations reported in Leland (1994).

13.3.1.6 Pros and cons of structural approaches to risky debt assessment

Pros. First, they allow to think about more complicated structures or instruments easily (e.g., convertibles, as we see in the next section). Second, they lead to simple yet consistent relations between different securities issued by the same name. Structural approaches were very useful for theoretical research during the 1990s.

Cons. The firm's assets value and assets volatility are not observed. Must rely on calibration/estimation methods. Bond prices generated by the model \neq market prices. These models are a bit difficult to use in practice, for trading or hedging purposes, as we know that in this case we need theoretical prices that exactly match market prices. Finally, how do we go for sovereign issuers?

Most important. Structural models predict unrealistically low short-term spreads: see, e.g., Figure 13.3. The intuition is that diffusion processes are smooth: the probability of default tends to zero as time to maturity approaches zero, because default cannot just jump in an unexpected way. This is not what we exactly observe. Jumps seem to be a more realistic device to modeling spreads, and will be introduced in Section 13.3.3.

13.3.2 An application of the structural approach: the pricing of convertible bonds

Convertible bonds offer bondholders the option to convert their bonds into shares of the firm. Chapter 11 (Section 11.8.2) provides an introductory discussion of these convertible bonds, and a numerical example of how to price them through a binomial tree. This section analyzes within the context of a continuous time model. We assume the option to convert can be exercised at any time up to maturity. By definition, the face value of the convertible is,

$$\text{Face value} = \text{£1} \equiv \text{CR} \times \text{CP}, \quad (13.12)$$

where CR is the *conversion ratio*, i.e. the number of shares this face value converts into, and CP is the *conversion price*, i.e. the stock price implicitly defined by Eq. (13.12).

Typically, the bond is any like other fixed income instrument, with coupon payments, callable features, credit risk, etc. Callable features are almost invariably embedded into this type of contracts. The parity, or *conversion value*, is the value of the bond if the bondholders decide to convert. It is defined as,

$$\text{CV} = \text{CR} \times S,$$

where S is the price of the common share. Not only is the convertible bond price affected by interest rates, credit risk, timing risk, etc. This price is also affected by the movements of the underlying stock price. This is quite natural as there is a positive probability that the bond will “become” a share in the future. To emphasize this fact, we also say that convertible bonds are *hybrid* instruments. The embedded option offers the bondholders the possibility to obtain equity returns (not just bond returns) in good times, while offering protection against the downside. As mentioned in Chapter 11 (Section 11.8), convertible bonds are usually callable as well. In this case, bond-holders are usually given the right to convert the bonds, once they are called. The rationale behind callability is to induce the bondholder to convert the bond earlier.

Useful to trade volatility. Simplest example of convertible arbitrage is going long a convertible and shorting a Treasury, which is the same as going long an option on the firm. Useful when there are no available options on the firm to trade, and/or when these are very illiquid.

Pricing convertible bonds is a topic that has been intensively studied, theoretically. Ingersoll (1977) provides the first theoretical insights into the pricing of convertible and callable bonds.

Let us define the *dilution factor*, denoted as γ , as the fraction of common equity that would be held by the convertible bondholders if the entire issue was converted. If there are n^{out} shares outstanding, and the convertible bond can be exchanged for n shares, then, in aggregate,

$$\gamma = \frac{n}{n + n^{\text{out}}}.$$

Let the market value of the firm be equal to the value of its assets, A , which we assume follows a Geometric Brownian motion, as in Eq. (13.1). Let $B^{\text{conv}}(A, \tau; N)$ be the aggregate value of the convertible bond with time to maturity τ and balloon payment N . To simplify the presentation, we do not consider callability issues. However, we shall provide some intuition about this issue later. Let us assume that the stocks and the convertible bonds are the only two claims in the capital structure of the firm. Since, after conversion, only the stocks will remain, then, the post-conversion value of the convertible bonds is simply the conversion value of the convertible, i.e. γA . Moreover, we have, for any $\tau \geq 0$,

$$\gamma A \leq B^{\text{conv}}(A, \tau; N) \leq A. \quad (13.13)$$

The first inequality in (13.13) is simple to understand. Indeed, suppose that $B^{\text{conv}}(A, \tau; N) < \gamma A$. Then, we can purchase the convertibles, convert them into shares and, finally, sell the shares for γA . The second inequality follows by limited liability equity holders, and the Modigliani-Miller theorem.

At maturity, we have that,

$$B^{\text{conv}}(A, 0; N) = \min\{A, \max\{N, \gamma A\}\}. \quad (13.14)$$

Indeed, $\bar{B} \equiv \max\{N, \gamma A\}$ is the value of the convertible, in case of no-default. Then, $\min\{A, \bar{B}\}$ is what the firm will pay, to the bondholders: A in case of default, and \bar{B} in case of no-default.

We can re-express the terminal payoff in Eq. (13.14) in a manner that allows a better understanding of the issues underlying the exercise of the convertibles. In particular, we have that,

$$B^{\text{conv}}(A, 0; N) = \min\{A, \max\{N, \gamma A\}\} = \max\{\gamma A, \min\{A, N\}\}. \quad (13.15)$$

Indeed, let $\hat{B} \equiv \min\{A, N\}$, which is what the firm is ready to pay, to the bondholders, if the bondholders do not exercise the option to convert. Then, $\max\{\gamma A, \hat{B}\}$ is obviously the payoff profile to the bondholders.

The terminal payoff in Eq. (13.15) illustrates very clearly that convertible bonds embed an option to convert - on top of the plain vanilla non-convertible bond. Intuitively, at maturity, a non-convertible bond is worth $\min\{A, N\}$, and the option to convert is either worthless (in case of non conversion) or $\gamma A - N$ (in case of conversion), i.e. it is $\max\{\gamma A - N, 0\}$. This intuition is confirmed, mathematically, as we have that:

$$\max\{\gamma A, \min\{A, N\}\} = \min\{A, N\} + \max\{\gamma A - N, 0\}.$$

Therefore, the value of the terminal payoff is, by Eq. (13.15),

$$B^{\text{conv}}(A, 0; N) = \min\{A, N\} + \gamma \max\{A - N/\gamma, 0\}. \quad (13.16)$$

It is possible to show that it is never optimal to exercise the option to convert before maturity. Therefore, to price the convertible bond, we only need to be concerned with the risk-neutral evaluation of the terminal payoff in Eq. (13.16).

Eq. (13.16) shows that the current value of the convertible bond is the sum of the value of a “straight” bond plus the value of γ options on the firm with strike price equal to N/γ . Accordingly, let $B(A, \tau; N)$ and $W(A, \tau; N/\gamma)$ be the prices of the straight bond and the option on the firm. We have,

$$B^{\text{conv}}(A, \tau; N) = B(A, \tau; N) + \gamma W(A, \tau; N/\gamma). \quad (13.17)$$

We may use the Merton’s (1974) model to find the price of the straight bond, $B(A, \tau; N)$. By the results in Section 13.2, it is:

$$B(A, \tau; N) = A\Phi(-d_1) + Ne^{-r\tau}\Phi(d_1 - \sigma\sqrt{\tau}), \quad d_1 = \frac{\ln(A/N) + (r + \frac{1}{2}\sigma^2)\tau}{\sigma\sqrt{\tau}}, \quad (13.18)$$

where σ is the instantaneous volatility of the assets, r is the (constant) instantaneous short-term rate, and Φ is the cumulative distribution of a standard normal. Similarly, we may use the Black-Scholes formula to compute the function W :

$$W(A, \tau; N/\gamma) = A\Phi(d_1) - \frac{N}{\gamma}e^{-r\tau}\Phi(d_1 - \sigma\sqrt{\tau}). \quad (13.19)$$

Eq. (13.18) reveals the intuitive property that as A gets large, $B(A, \tau; N) \approx Ne^{-r\tau}$: the probability of default gets extremely tiny as the value of the assets gets large. Moreover, the Black-Scholes formula, Eq. (13.19), suggests that $W(A, \tau; N/\gamma) \approx A - e^{-r\tau}N/\gamma$ as A gets large. Therefore, by Eq. (13.17), we have that, for large A , $B^{\text{conv}}(A, \tau; N) \approx Ne^{-r\tau} + \gamma(A - e^{-r\tau}N/\gamma) = \gamma A$. Eq. (13.18) also shows that for small values of A , $B^{\text{conv}}(A, \tau; N) \approx 0$. To sum-up, the value of the convertible bond is less than the value of the firm, A , and larger than the conversion value, γA . Moreover, it approaches γA , as the value of the firm gets large. Figure 13.6 depicts the price of the convertible bond as function of the value of the firm, as predicted by Eq. (13.17), for a particular example. It is possible to show that the value of a callable convertible bond is between the value of the straight and that of the convertible.

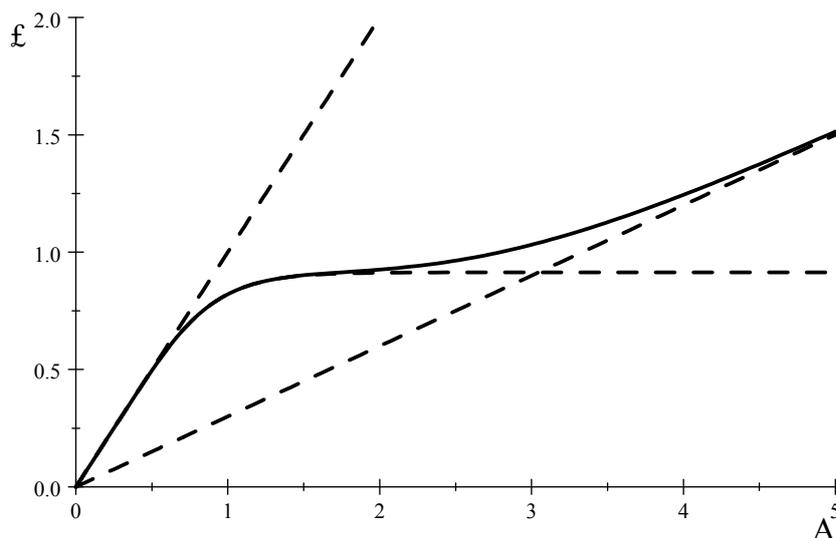


FIGURE 13.6. The value of convertible and straight bonds as a function of the current assets value, A , when the short-term rate $r = 3\%$, the assets volatility $\sigma = 0.20$, time to maturity $\tau = 3$ years, the dilution factor $\gamma = 10\%$, and nominal debt $N = 1$. The solid

line depicts the value of the convertible bond. The dashed straight line starting from the origins, and flattening out to the constant $Ne^{-r\tau} = 0.91393$, is the value of the straight bond. The two dashed straight lines starting from the origins are the no-arbitrage bounds γA and A in Eq. (13.13).

13.3.3 Reduced form approaches: rare events, or intensity, models

Default often displays a few striking features. It arrives unexpectedly, it is rare, and causes discontinuous price changes. The structural models in the previous section do not accommodate for these features because diffusion processes are continuous, as explained in Chapter 4. As a result, passage times are known, “locally,” so to speak. This feature is responsible of the low short-term spreads these models predict.

13.3.3.1 Poisson-driven defaults

As an alternative to Brownian motions, we can model defaults, by assuming their arrival is a Poisson process, of the kind introduced in Chapter 4. Suppose to “count” the number of times some event happens. Denote with N_t the corresponding “counting process,” as in Figure 13.7.

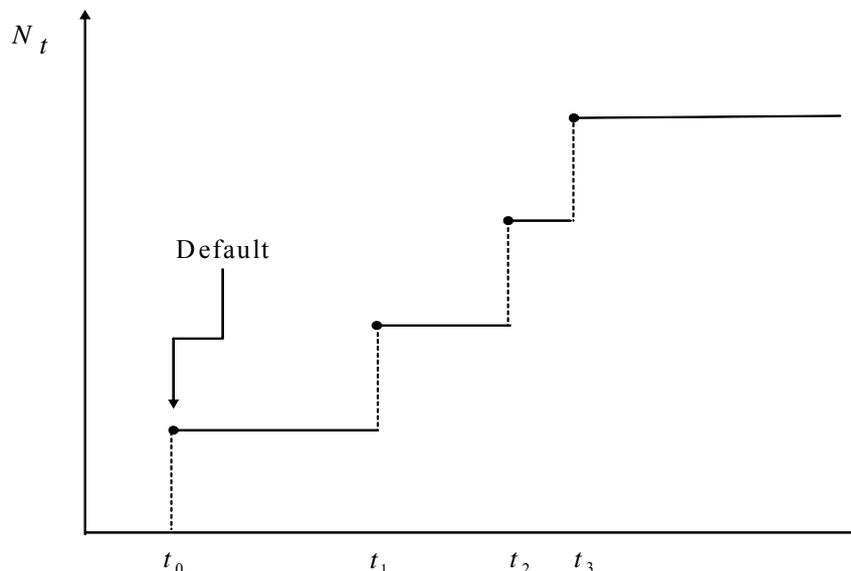


FIGURE 13.7.

Default time is simply defined as the first time N_t “jumps,” e.g. t_0 in Figure 13.7. So assume we chop a given interval $[0, T]$ in n pieces, and consider each resulting interval $\Delta t = \frac{T}{n}$. Assume that the jump probability over each of these small intervals of time Δt is proportional to Δt , with proportionality factor equal to λ ,

$$p \equiv \Pr \{ \text{One jump over } \Delta t \} = \lambda \Delta t. \tag{13.20}$$

Assume the number of jumps over the n intervals follows a binomial distribution:

$$\Pr \{ k \text{ jumps over } [0, T] \} = \binom{n}{k} p^k (1 - p)^{n-k}, \quad \text{where } p = \lambda \frac{T}{n}.$$

For n large, or, equivalently, for small intervals Δt ,

$$\Pr \{k \text{ jumps over } [0, T]\} \approx \frac{(\lambda T)^k}{k!} \left(1 - \frac{\lambda T}{n}\right)^{n-k} \approx \frac{(\lambda T)^k}{k!} e^{-\lambda T}.$$

We rely on these basic computations and give a few basic properties of default. We have,

$$\begin{aligned} \Pr \{\text{Survival}\} &= \Pr \{0 \text{ jumps over } [0, T]\} = e^{-\lambda T} \\ \Pr \{\text{Default}\} &= \Pr \{\text{at least one jump}\} = 1 - \Pr \{\text{Survival}\} = 1 - e^{-\lambda T} \\ \Pr \{\text{Default occurs at some } t\} &= \lambda e^{-\lambda t} dt \end{aligned}$$

Note that the expected time to default equals λ^{-1} .

We can use these probabilities to value debt subject to default risk. Consider Eq. (13.6):

$$D_0 = e^{-rT} \underbrace{[\text{Rec} \cdot \mathbb{Q}(\text{Default}) + N \cdot \mathbb{Q}(\text{Survival})]}_{\equiv B_0},$$

where Rec is the expected recovery value of the asset. Using the probabilities predicted by the Poisson model, we obtain:

$$B_0 = \text{Rec} \cdot (1 - e^{-\lambda T}) + N \cdot e^{-\lambda T}. \quad (13.21)$$

Appendix 2 supplies an alternative derivation of Eq. (13.21).

13.3.3.2 Predicted spreads

The implications for the spreads corresponding to small maturities T can be easily seen after some approximations,

$$\text{Spread} = -\frac{1}{T} \ln \left(\frac{B_0}{N} \right) \approx -\frac{1}{T} \left(\frac{B_0}{N} - 1 \right) = \frac{1}{T} (1 - e^{-\lambda T}) \cdot \text{Loss-given-default}.$$

Note, for T small, and in contrast to the structural models reviewed in Section 13.3.1, the spread is not zero. Rather, it is given by the expected default loss per period, defined as the instantaneous probability of default times loss-given-default,

$$\text{Short-Term Spread} = \lambda \cdot \text{Loss-given-default}.$$

Therefore, models with jumps have the potential to explain the empirical behavior of credit spreads at short maturities discussed in Section 13.3.1. As explained, structural models, which are typically driven by Brownian motions, cannot lead to positive spreads at very short maturities, as they imply that the probability of default decays quickly as time-to-maturity goes to zero. Instead, in models with jumps, there is always a possibility of “sudden death” for the firm: at any instant of time, and even when the debt is about to expire, default can occur with positive probability, and this fact is reflected by positive short-term spreads. A theoretical model of Duffie and Lando (2001) shows how a structural model of the firm can lead to positive short-term spreads, once we assume incomplete information and learning about the assets value. In their model, learning takes place with some delay, which leaves investors concerned about what they really know about the firm’s assets value. It is this concern that leads to positive credit spreads in their model, to an extent comparable to that generated by a jump process.

Figure 13.8.1 depicts the behavior of the spread predicted by the model at all maturities, given by,

$$\text{Spread} = -\frac{1}{T} \ln \left(\frac{B_0}{N} \right) = -\frac{1}{T} \ln \left(\frac{\text{Rec}}{N} \cdot (1 - e^{-\lambda T}) + e^{-\lambda T} \right).$$

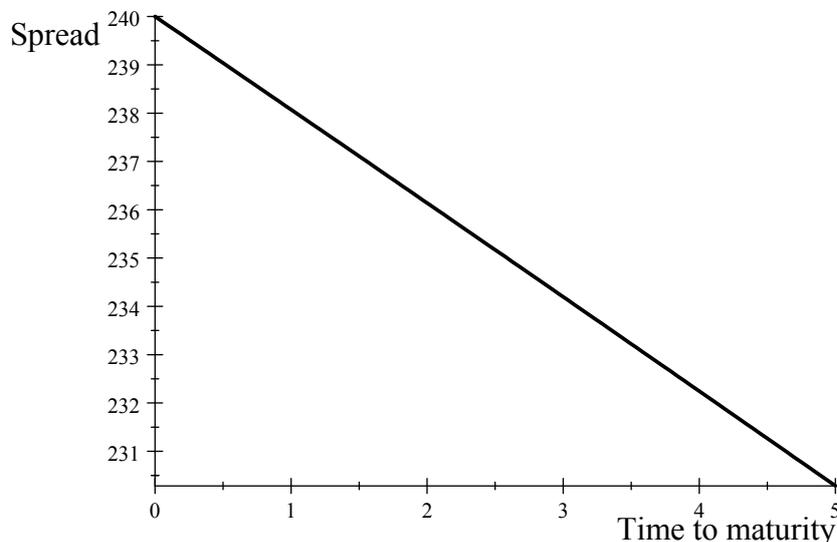


FIGURE 13.8.1. The term structure of bond spreads (in basis points) implied by an intensity model, with recovery rate equal to 40% and intensity equal to $\lambda = 0.04$, implying an expected time-to-default equal to $\lambda^{-1} = 25$ years.

In this example, spreads are decreasing in time-to-maturity. Eventually, as time to maturity gets large, the bond becomes, so to speak, certain to default, with the unusual feature to deliver, for sure, some recovery rate at some point—the bond is certain to deliver the recovery rate. Indeed, in Appendix 2, we show that if the recovery value of the bond is not constant, but shrinks exponentially to zero as $\text{Rec}_T \equiv R \cdot e^{-\kappa T}$, for two constants R and κ , then, asymptotically, the spread is:

$$\lim_{T \rightarrow \infty} s(T) = \begin{cases} \lambda, & \text{if } \kappa \geq \lambda \\ \kappa, & \text{if } \kappa \leq \lambda \end{cases} \quad (13.22)$$

The interpretation of κ is not discounting. Rather, we might refer to κ as a “recovery dissipation rate” due to unfolding of time. That is, as time unfolds, there might only occur bad events leading the recovery rate to deteriorate. Eq. (13.22) shows that if this dissipation rate is sufficiently large, term spreads can be increasing, as we discuss more comprehensively in a moment. An instance leading to such an expected recovery rate is one where the recovery value of the bond equals R , if the firm defaults at any time T , and provided an hidden risk does not materialize, namely the risk that the firm will not distribute any recovery value at all, in case of bankruptcy. If this risk is independent of bankruptcy, and Poisson, with instantaneous risk-neutral probability κ , the expected recovery is precisely $\text{Rec}_T = R \cdot e^{-\kappa T}$. This is indeed a quite simple way to model stochastic recovery rates.

Figure 13.8.2 plots the term-structure of spreads predicted by this model, obtained with the same parameter values used to produce the spreads in Figure 13.8.1, and utilizing three values for the dissipation rate: $\kappa = 0.05, 0.03$ and 0.013 . Naturally, instantaneous spreads (i.e. those corresponding to time to maturity equal to zero) are $(1 - R) \cdot \lambda = 240$ (in basis points) in

all cases. When $\kappa > \lambda$, large maturity spreads are always higher than short, by Eq. (13.22). In this particular example, they equal 400 basis points, i.e. the default intensity, λ . When $\kappa < \lambda$, instead, spreads for large maturities can be either higher or lower than short, according to whether κ is higher or lower than $(1 - R) \cdot \lambda$. When $\kappa = 0.03$, they are higher and when $\kappa = 0.013$, they are lower. In fact, when $\kappa = 0.013$, the term structure of the spreads is even hump-shaped, although this feature is not clearly visible from the picture. As is clear, this very simple model predicts features of both short-term and long-term spreads that the Merton's model in Section 13.3.1.1 cannot, realistically.

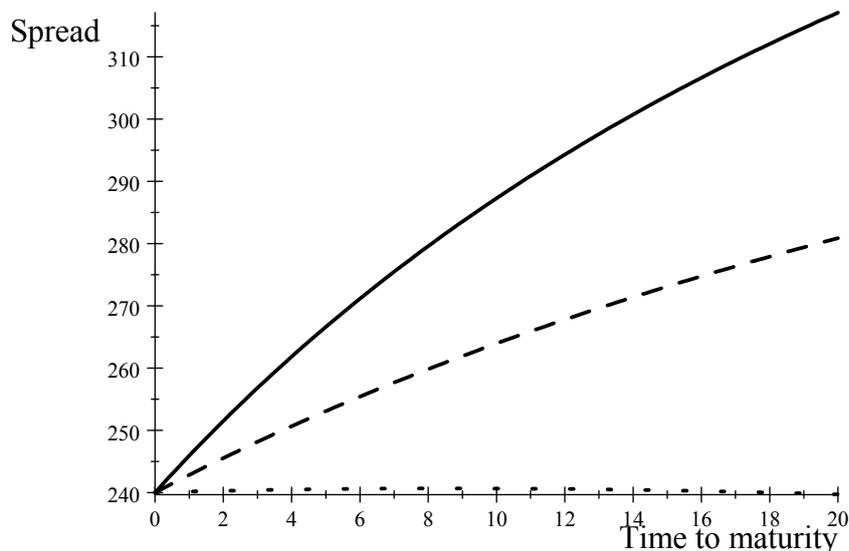


FIGURE 13.8.2. The term structure of bond spreads (in basis points) implied by an intensity model with recovery rate equal to $0.40e^{-\kappa\tau}$, where τ is time to maturity and κ is the recovery dissipation rate, taken to equal $\kappa = 0.05$ (solid line), $\kappa = 0.03$ (dashed line), and $\kappa = 0.013$ (dotted line). The instantaneous probability of default is taken to equal to $\lambda = 0.04$, implying an expected time-to-default equal to $\lambda^{-1} = 25$ years.

This behavior of term spreads is one we can interpret as follows. Suppose we are in good times, when λ is small relative to κ . We are in good times precisely because we expect things would change adversely in the future, captured by a large value of κ —even larger than λ . In this case, the term structure of spreads is increasing. Instead, in bad times, when λ is large compared to κ , we might expect future times to improve, which we might model with small values of κ —even smaller than λ . We see, from Figure 13.8.2, that spreads for large maturity become smaller than those we have in good times. Naturally, we would expect that in bad times, spreads should increase for any maturity, although this property is not captured by the numerical examples in Figure 13.8.2, where we fix $\lambda = 0.04$. Rather, the point of this exercise, is to show that the *slope* of the term structure of the spreads lowers as we enter bad times, when we only consider changes in the dissipation rate, κ . Allowing for a countercyclical λ would reinforce the conclusions of this exercise. Finally, these conclusions rely on comparative statics, although in Section 13.5.5.5, we shall show they still hold in a truly dynamic context, where the intensity λ follows a mean-reverting continuous-time model.

13.3.3.3 One example

Naturally, the intensity, λ , is the risk-neutral instantaneous probability of default, not the physical probability of default, λ^* say. The ratio λ/λ^* is generally larger than one. Its inverse, λ^*/λ , is an indicator of the risk-appetite in the credit market. Similarly, loss-given-default is an expectation under the risk-neutral probability, and should contain useful indications about market participants risk appetite.

Assume that under the risk-neutral probability, the instantaneous intensity of default for a given firm is $\lambda = 4\%$, annualized, and that under the physical probability, the instantaneous probability of default for the same firm is $\lambda^* = 2\%$, annualized. From here, we can compute the probability of survival of the firm within 5 years, under both probabilities. They are:

$$e^{-5\lambda} = e^{-5*0.04} = 0.81873, \quad e^{-5\lambda^*} = e^{-5*0.02} = 0.90484.$$

Naturally, the probability of survival is lower under the risk-neutral measure.

Next, assume that the spread on a 5 year bond with face value $N = 1$, equals 3%. What is the implied expected recovery rate from this spread? We have,

$$D_0 = e^{-rT} [\text{Rec} \cdot \mathbb{Q}(\text{Default}) + N \cdot \mathbb{Q}(\text{Survival})] = e^{-rT} [\text{Rec} * (1 - 0.81873) + 1 * 0.81873].$$

The spread is,

$$s_0 = 3\% = -\frac{1}{5} \ln \left(\frac{\text{Rec} * (1 - 0.81873) + 1 * 0.81873}{1} \right).$$

Solving for Rec, gives, Rec = 23.16%.

13.3.4 Ratings

In practice, corporate debt is rated by rating agencies, such as Moodys and Standard and Poors. Depending on the rating, corporate debt may be either investment grade or non-investment grade (“junk”). Moodys ratings range from Aaa to C. Standard and Poor’s range from Aaa to D. One can compute the probability of “migrations” based on past experience \rightarrow Transition probabilities. Consider, for example, the following table:

		One year rating transition probabilities (%), S&P's 1981-1991							
		To							
		AAA	AA	A	BBB	BB	B	CCC	D
From	AAA	89.1	9.63	0.78	0.19	0.3	0	0	0
	AA	0.86	90.1	7.47	0.99	0.29	0.29	0	0
	A	0.09	2.91	88.94	6.49	1.01	0.45	0	0.09
	BBB	0.06	0.43	6.56	84.27	6.44	1.6	0.18	0.45
	BB	0.04	0.22	0.79	7.19	77.64	10.43	1.27	2.41
	B	0	0.19	0.31	0.66	5.17	82.46	4.35	6.85
	CCC	0	0	1.16	1.16	2.03	7.54	64.93	23.19
	D	0	0	0	0	0	0	0	100

TABLE 13.1

13.3.4.1 Foundations

A natural approach, then, is to assess credit risk by making reference to probabilities of default built up on transition probabilities like those in Table 13.1.

Such an approach, also known as a migration approach, is somewhat less drastic than that based on rare events, and hopefully more realistic. However, it is also technically more complex than the intensity approach of the previous section. We provide the most foundational issues of this approach, leaving some details in the Appendix.

At time t , there exists several rating classes, Z say, denoted as Rat_t ,

$$\text{Rat}_t \in \{1, 2, \dots, Z\}.$$

Transition probabilities of rating from time t to time T are,

$$P(T - t)_{ij} \equiv \Pr(\text{Rat}_T = j | \text{Rat}_t = i), \quad i, j \leq Z.$$

We can build a Markov chain from here, by assuming that $P(T - t)_{ij}$ only depends on $T - t$. Finally, we must have that,

$$P(T - t)_{ij} \geq 0 \quad \text{and} \quad \sum_{j=1}^Z P(T - t)_{ij} = 1.$$

For example, the probability of transition from rating $\text{Rat}_t = i$ to rating $\text{Rat}_{t+1} = j$ in one year is, $P(1)_{ij}$. Table 13.1 contains one possible example of $P(1)_{ij}$. The probability of transition from rating $\text{Rat}_t = i$ to rating $\text{Rat}_{t+2} = j$ in two years is $P(2)_{ij}$, and is obtained as follows,

$$P(2)_{ij} = \sum_{k=1}^Z \underbrace{P(1)_{ik}}_{\text{Pr(transition from } i \text{ to } k \text{ in one year)}} \cdot \underbrace{P(1)_{kj}}_{\text{Pr(transition from } k \text{ to } j \text{ in one further year)}}$$

More generally, we have, $P(T) = P(1)^T$, where $P(T)$ is the matrix with elements $\{P(T)_{ij}\}$. For example, the probability transition matrix P in Table 13.1 is,

$$P = \begin{bmatrix} 89.1 & 9.63 & 0.78 & 0.19 & 0.3 & 0 & 0 & 0 \\ 0.86 & 90.1 & 7.47 & 0.99 & 0.29 & 0.29 & 0 & 0 \\ 0.09 & 2.91 & 88.94 & 6.49 & 1.01 & 0.45 & 0 & 0.09 \\ 0.06 & 0.43 & 6.56 & 84.27 & 6.44 & 1.6 & 0.18 & 0.45 \\ 0.04 & 0.22 & 0.79 & 7.19 & 77.64 & 10.43 & 1.27 & 2.41 \\ 0 & 0.19 & 0.31 & 0.66 & 5.17 & 82.46 & 4.35 & 6.85 \\ 0 & 0 & 1.16 & 1.16 & 2.03 & 7.54 & 64.93 & 23.19 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 100 \end{bmatrix}$$

The 15 year transition matrix is:

$$P(15) \approx \begin{bmatrix} 20.01 & 35.82 & 23.91 & 9.92 & 4.05 & 3.06 & 0.43 & 2.66 \\ 3.38 & 30.28 & 32.71 & 15.91 & 6.38 & 5.11 & 0.77 & 5.34 \\ 1.17 & 13.12 & 34.21 & 21.93 & 9.69 & 8.01 & 1.29 & 10.33 \\ 0.64 & 6.76 & 22.21 & 22.40 & 12.42 & 11.93 & 2.09 & 21.39 \\ 0.33 & 3.22 & 10.71 & 13.616 & 11.36 & 14.68 & 2.78 & 43.16 \\ 0.14 & 1.65 & 5.01 & 6.75 & 7.48 & 13.17 & 2.64 & 63.04 \\ 0 & 1.08 & 3.54 & 3.90 & 3.51 & 5.60 & 1.22 & 81.02 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 100 \end{bmatrix}$$

13.3.4.2 Evaluation

The previous probabilities, $\{P(T)_{ij}\}$, are meant to be taken under the *physical* world, not the *risk-neutral*. They can be used for risk-management purposes, but certainly not for pricing. Indeed, historical default rates are too low to explain the price of defaultable securities. A natural explanation relies on the presence of risk-premia. To use migration data for pricing, it is vital to implement a number of steps.

First, clean up the data — smoothing. For example, it might well be that downgrades from class i to class $i + 2$ are more frequent than downgrades from class i to class $i + 1$. Moreover, remove zero entries: although some rating event did not happen in the past, they might well occur in the future. Second, add positive risk-premia to the previous smoothed data so as to obtain realistic asset prices.

As regards pricing, according to the migration model, there are Z classes of assets. Each single asset may migrate from one class to another. Because evaluation is a dynamic business, we cannot evaluate defaultable securities within a given asset class without simultaneously evaluate the defaultable securities in the remaining classes. For example, there could be a chance that a given asset will “mutate” into a different one in the next year. Given this, the price of this asset, today, must reflect the price of the asset in the other classes where it can possibly migrate. Hence, we must simultaneously solve for all the asset prices in all the rating classes. This approach, developed by Jarrow, Lando and Turnbull (1997), is quite complex and is given a succinct account in the Appendix.

Consider the simplest case, which arises when the expected recovery rate is zero. In this case, by Eq. (13.6),

$$\frac{D_{0,i}}{N} = e^{-rT} (1 - Q_i(T - t)),$$

where $Q_i(T - t)$ is the *risk-neutral* probability the firm defaults, by time T , given it belongs to rating i at time T .

More generally, by Eq. (13.6),

$$\frac{D_{0,i}}{N} = e^{-rT} \left[\frac{\text{Rec}}{N} Q_i(T - t) + (1 - Q_i(T - t)) \right].$$

The risk neutral probabilities, $Q_i(T - t)$, must be found using migration frequencies such as those in Table 13.1, which we must “clean up” and correct with appropriate risk-premia, as discussed.

13.3.4.3 One example

Consider the following transition matrix:

		To		
		A	B	Def
From	A	0.9	0.07	0.03
	B	0.15	0.75	0.10
	Def	0	0	1

where Def denotes the state of default. What is the probability that a name A will remain name A in two years? What is the probability that a name A will default in two years?

Consider the following two year transition matrix:

$$\mathbb{Q}(2) = \underbrace{\begin{bmatrix} 0.90 & 0.07 & 0.03 \\ 0.15 & 0.75 & 0.10 \\ 0 & 0 & 1 \end{bmatrix}}_{\equiv \mathbb{Q}(1)} \cdot \underbrace{\begin{bmatrix} 0.90 & 0.07 & 0.03 \\ 0.15 & 0.75 & 0.10 \\ 0 & 0 & 1 \end{bmatrix}}_{\equiv \mathbb{Q}(1)},$$

such that:

$$\begin{aligned} \Pr \{A \text{ is A in 2 years}\} &= \underbrace{0.90 * 0.90}_{A \rightarrow A \rightarrow A} + \underbrace{(0.07) * (0.15)}_{A \rightarrow B \rightarrow A} + \underbrace{0.03 * 0}_{A \rightarrow Def \rightarrow A} \\ &= 0.8205, \end{aligned}$$

and

$$\begin{aligned} \Pr \{A \text{ defaults in 2 years}\} &= \underbrace{0.90 * 0.03}_{A \rightarrow A \rightarrow Def} + \underbrace{(0.07) * (0.10)}_{A \rightarrow B \rightarrow Def} + \underbrace{0.03 * 1}_{A \rightarrow Def \rightarrow Def} \\ &= 0.064. \end{aligned}$$

In general, we have that:

$$\mathbb{Q}(2)_{ij} = \sum_{k=1}^3 \mathbb{Q}(1)_{ik} \cdot \mathbb{Q}(1)_{kj},$$

and for any T ,

$$\mathbb{Q}(T) = \mathbb{Q}(1)^T = \begin{bmatrix} 0.90 & 0.07 & 0.03 \\ 0.15 & 0.75 & 0.10 \\ 0 & 0 & 1 \end{bmatrix}^T.$$

Next, consider the following transition matrix, under the risk-neutral probability:

		To		
		A	B	Def
From	A	0.80	0.20	0
	B	0.15	0.75	0.10
	Def	0	0	1

From here, we may easily compute, again, the (risk-neutral) probability A will default in two years, and the probability B will default in two years. We have,

$$\mathbb{Q}(2) = \underbrace{\begin{bmatrix} 0.80 & 0.20 & 0 \\ 0.15 & 0.75 & 0.10 \\ 0 & 0 & 1 \end{bmatrix}}_{\equiv \mathbb{Q}(1)} \cdot \underbrace{\begin{bmatrix} 0.80 & 0.20 & 0 \\ 0.15 & 0.75 & 0.10 \\ 0 & 0 & 1 \end{bmatrix}}_{\equiv \mathbb{Q}(1)},$$

such that:

$$\begin{aligned} \Pr \{A \text{ defaults in 2 years}\} &= \mathbb{Q}(2)_{13} \\ &= \underbrace{0.80 * 0}_{A \rightarrow A \rightarrow Def} + \underbrace{(0.20) * (0.10)}_{A \rightarrow B \rightarrow Def} + \underbrace{0 * 1}_{A \rightarrow Def \rightarrow Def} \\ &= 0.02. \end{aligned}$$

(multiply first row by the third column), and,

$$\begin{aligned} \Pr \{B \text{ defaults in 2 years}\} &= \mathbb{Q}(2)_{23} \\ &= \underbrace{0.15 * 0}_{B \rightarrow A \rightarrow Def} + \underbrace{(0.75) * (0.10)}_{B \rightarrow B \rightarrow Def} + \underbrace{0.10 * 1}_{B \rightarrow Def \rightarrow Def} \\ &= 0.175. \end{aligned}$$

(multiply second row by the third column).

Finally, suppose that the bonds issued by both A and B mature in two years. Furthermore, assume that if these two bonds default, they pay off the same recovery rate, equal to 30%, and only at the end of the second period. From here, we can compute the credit spreads for the two bonds. We have,

$$\begin{aligned} e^{rT} * \text{Price}_A &= (0.30) * (0.02) + (1 - 0.02) = 0.986 \\ \Rightarrow \text{Spread}_A &= -\frac{1}{2} \ln(0.986) = 7.0495 \times 10^{-3}. \end{aligned}$$

and,

$$\begin{aligned} e^{rT} * \text{Price}_B &= (0.30) * (0.175) + (1 - 0.175) = 0.8775 \\ \Rightarrow \text{Spread}_B &= -\frac{1}{2} \ln(0.8775) = 6.5339 \times 10^{-2}. \end{aligned}$$

13.4 Credit derivatives, and structured products based thereon

13.4.1 A brief history of credit risk and financial innovation

Credit risk is a pervasive feature of many securities, and the previous section has highlighted the main conceptual approaches to it, as well as a few applications such as the pricing of convertible bonds. But credit risk can be transferred, or hedged, through dedicated derivatives, similarly as we do when we transfer and hedge market risk. This section is an introduction to these credit derivatives, and some basic structured products stemming from those. It is instructive to have a perspective on their origins, which is a history about financial innovation really. The starting point of this story is perhaps the birth of the first interest rates derivatives, created around the mid 80s, and grown to an extent that in the late '80s already, this business proliferated and became fairly complex. Note that financial innovation is relatively easy to imitate, which led banks to become increasingly creative, so as to exploit their initial competitive advantage as longer as possible. During the early 1990s, just after the 1991 recession, interest rates were quite low, and the volatility of capital markets extraordinarily low. Derivatives, then, could be used as devices to boost investors' returns. JPMorgan introduced new products such as LIBOR squared, inverse floaters, power options, etc. But after the 1994 financial turmoil, the interest rate climate suddenly changed while some of these derivative products were producing large losses, which triggered a call for regulation by public opinion and certain policy makers. At the same time, the International Swaps and Derivatives Association would argue that more regulation would destroy markets creativity.

These regulatory pressures vanished by the mid 1990s, and markets started to innovate again, with a general consensus arising that derivative risks could be assessed, and controlled, through market discipline, rather than regulation. Swap markets recovered. They did do slowly though,

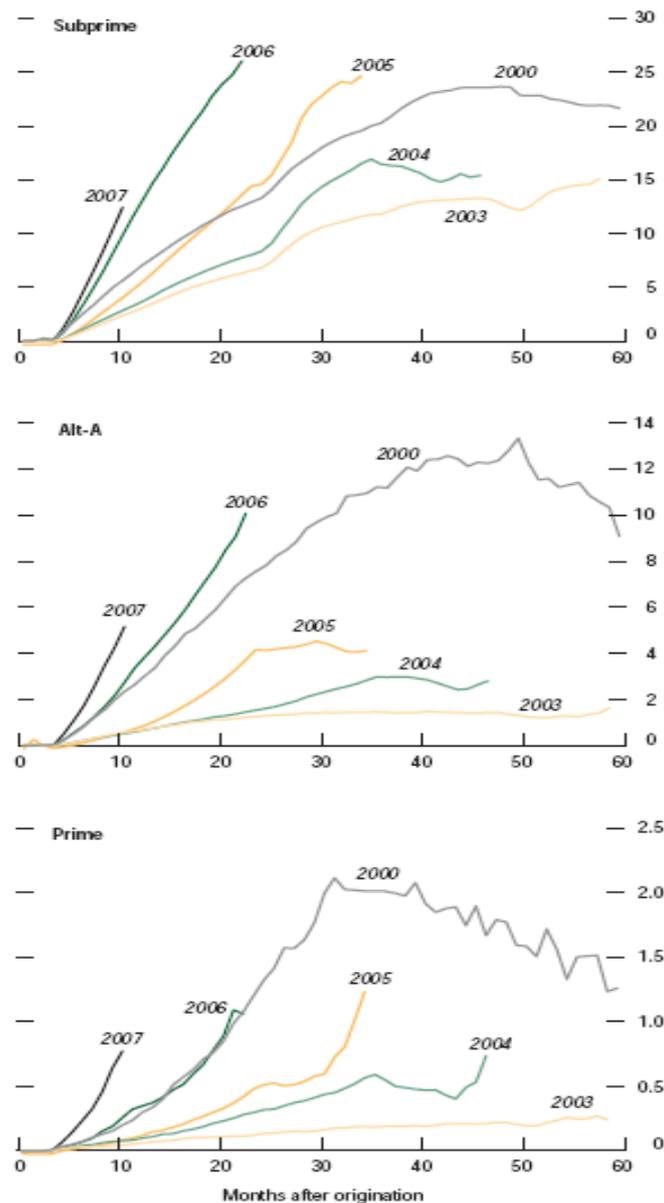
as these products were already in the end of the innovation cycle. They had been massively imitated to the extent they had become a mass-product so to speak, with profit margins having consequently been eroded. The markets were ready for a new major innovation wave as a result. The next innovation had naturally to do with credit risk. Global players such as JPMorgan, Credit Suisse, Bankers Trust soon realized that borrower defaulting was a source of substantial risk, which could be conveniently reallocated through the use of dedicated derivatives. Credit risks could be transferred in pretty much the same way as market risks can be transferred through the underwriting of options written on stocks, or on interest rates. JPMorgan had additional motivations to innovate, as its books contained vast pools of loans, which could be used as practical material to experiment with. Importantly, these loans required too many reserves and were consequently expensive.

The main idea was to repackage the loans into derivatives and/or proceed with *securitization*. It was a natural idea, because securitization is simply a process by which some illiquid assets are gathered into a common pool, which backs the issuance of new securities aimed to display an enhanced liquidity obtained through packaging, credit and liquidity enhancements. Two leading examples of this process include the securitization of mortgages and receivables. Financial institutions find the securitization process attractive, as they can carve out certain items in their balance sheet, thus boosting their return on investments or simply because by securitizing assets, less capital is needed to meet capital requirements standards. For example, the accounts receivables of a corporation may be used to back the issue of commercial paper known as *asset-backed commercial paper*. A well-functioning securitization system is a way (not the only way) to trade and transfer credit risk.

Global players would then repackage loans into derivatives, in a way that *default risk* and/or *part of the securitized loans*, or both, could be sold to outside investors. In a sense, credit derivatives were also a regulatory mitigation device, partly useful as a response to regulation. The idea was simple: to turn loans into derivatives that could be sold, and create new insurance products, such as credit default swaps. At the very beginning, derivatives were just designed to have single loans as underlying. Afterwards, the idea emerged to create structures organized in derivatives bundles, with cash flow indexed to baskets of loans—the ancestors to collateralized debt obligations. For example, JPMorgan created “Bistro” (Broad Index Secured Trust Offering), a structure relying on a variety of assets, ranging from corporate debt to student loans. ABN - Amro created similar structures, “Heineken” and “Amstel.” But then, competition increased and profit margins fell again, leading to an appetite for new innovation.

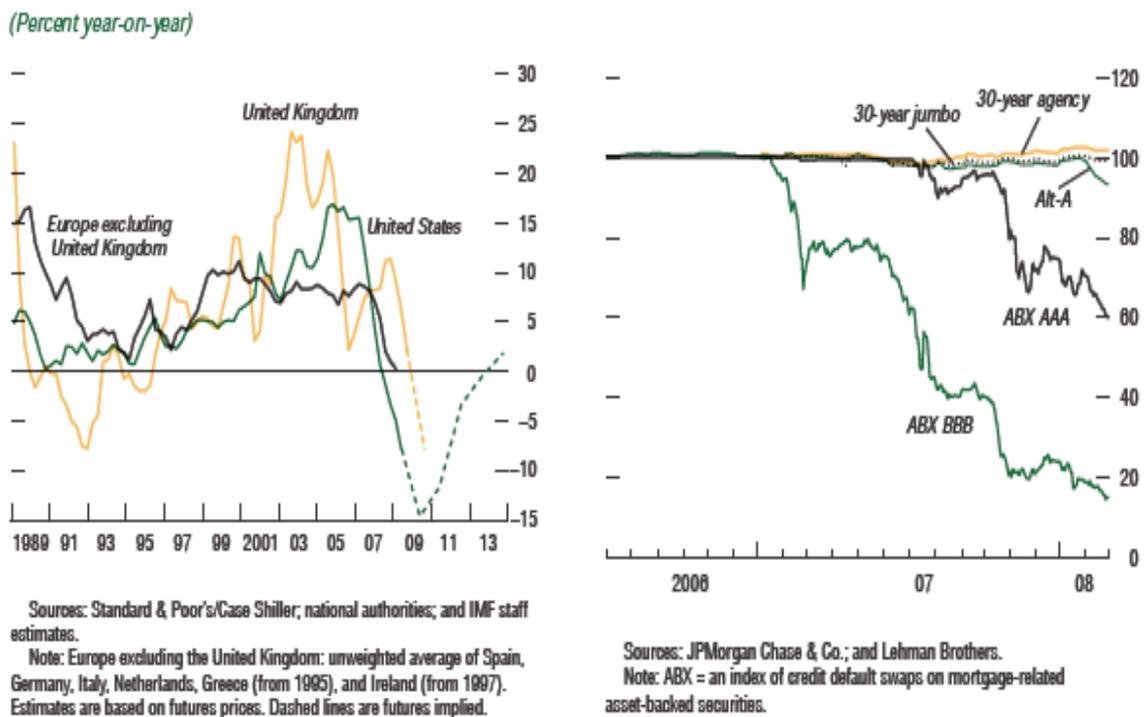
The response to increased competition was the creation of structured products with riskier assets. For example, during the mid 1990s, derivatives teams begun to interact with teams managing loans extended to borrowers with poor credit history—subprime mortgages. Subprime loans would begin to be securitized and structured into CDOs. JPMorgan was not the leader in the creation of these products, compared to other institutions such as Merrill Lynch or UBS. The subprime turmoil arose out of mechanics that are by now well-understood. First, there was a boom, sustained by (i) low interest rates and house price appreciation and (ii) a business model changing from *buy to hold* to *originate and distribute*, as explained. After the boom, the burst, caused by increasing interest rates and falling housing prices. Evaluation models, if any, relied on the assumption delinquencies would remain the same, and small risk-aversion adjustments to the calculation of expected actuarial losses were made (if any). The picture below shows this wasn’t true and that in fact, the pieces of information emanating from those simple pictures could have helped predict the crisis. Finally, correlation issues were simply ignored or, at best, badly calibrated. We shall see the importance of badly calibrated models later in this section.

Section 13.4.7 provides a more systematic analysis of these issues, but it is instructive to discuss since now, some of the causes leading to the burst and the 2007 crisis. One of them is certainly related to “model misspecification,” or an inappropriate rating “mapping” system, by which rating agencies tended to rate structured products relying on MBS, by using the rating system they had in place for corporations. A second cause is that these products were channeled through off-balance-sheet vehicles, a sort of “shadow banking system,” escaping the attention of the official financial community. The resulting substantial uncertainty about the identity of holders of these assets was a reason magnifying the crisis, with a sharp liquidity dry-up, and then a credit crunch, followed by a drop in the real economic activity, which further fed the credit crunch, etc.



Mortgage Delinquencies by Vintage Year (60+ day delinquencies, in percent of balance).

Source: IMF, Global Financial Stability Report, April 2008.



Left hand side panel: U.S. and European House Price Changes. Right hand side panel: U.S. Mortgage-Related Securities Prices. Source: IMF, Global Financial Stability Report, April 2008.

13.4.2 Options and spreads

13.4.2.1 Total Return Swaps

In a total return swap (TRS, henceforth), one party, who owns some asset, that underlying the TRS, receives from the counterparty payments based on a mutually agreed rate, either fixed or variable, and makes payments to the counterparty based on the *return* of the underlying asset, which includes both the income it generates and any capital gains. The underlying asset can be a loan, a bond, an equity index, or a basket of assets. The interest payments are typically based on the LIBOR plus a spread. Consider the following example. Party A receives LIBOR + fixed spread equal to 3%. Party B receives the total return of the S&P 500 on a principal amount of £1 million. If the LIBOR is 7% and the S&P 500 is up by 12%, A pays B 12% and B pays A 7% + 3%. By netting, A pays B £20,000, i.e. £1 million \times (12% - 10%).

While TRS are usually categorized as credit derivatives, they combine both market risk and credit risk. The main benefit from going long a TRS is that the party with the asset on the balance sheet buys protection against loss in value. The main benefit from shorting a TRS is that it allows the counterparty to receive the payoffs of the underlying without necessarily having to put this underlying in the balance sheet. Hedge funds find it quite convenient to short a TRS, as this allows them to have views with limited collateral upfront. The market for TRS is over-the-counter and market participants include institutions only.

13.4.2.2 Spread Options

Spread Options (SO, henceforth) are options written on the difference between two indexes. For example, let $S_1(T)$ and $S_2(T)$ be the prices of two assets at time T . The payoff promised by a SO entered at some time $t < T$, might be $\max\{S_1(T) - S_2(T) - K, 0\}$, where K is the strike of the SO. A SO can be written on the spread between two rates of returns too. Importantly, a SO can be written on the spread between the yield of a corporate bond and the yield of a Treasury bond. Examples include: (i) NOB spread (notes - bonds), which are spreads between maturities; (ii) Spreads between quality levels, such as the TED spread (treasury bills – Eurodollars); (iii) MOB spreads, i.e. the difference between municipal bonds and treasury bonds. More generally, the definition of a SO has now been extended to include payoffs written as a *linear combination* of indexes, interest rates and yields.

13.4.2.3 Credit spread options

Credit spread options (CSO, henceforth) are options where the payoff is the difference between (i) the spread between two reference securities (say Italian Government bonds and US Government bonds having the same maturity, or the spread between some stock return in excess of the LIBOR, or two credit instruments), and (ii) a given strike spread, for a certain maturity date. It may be an American or European option. So CSOs allow to hedge against, or take specific views about, changes in credit spreads. For example, an investor, while bullish on Italian bonds, might hedge against the uncertain outcome of a political election, which could trigger a widening of short-term spreads of Italians versus US. The investor, then, might go long a CSO, with time to maturity around the days of the political election, where the underlying are the Italian and Government bonds expiring in ten years, say. A possible payoff to the CSO holder might be proportional to, $(ITA/US - K)$, where ITA/US is the ten year Italian-US spread in three months, and K is the strike spread.

13.4.3 Credit Default Swaps

13.4.3.1 Single name swaps

TRS provide protection against a general loss in assets value, which could be triggered by both market or credit risk, although it is obviously more often market risk than credit to kick in. Credit Default Swaps (CDS, henceforth) differ from TRS insofar as they provide protection against a credit event.

The premium, assumed to be paid quarterly, on a CDS contract agreed at time t , is obtained by equating the expected discounted value of the protection (the *floating protection leg*), to the expected discounted value of all the premiums paid over the life of the contract (the *fixed premium leg*), i.e. at dates $t_i : t < t_1 < t_2 < \dots < t_{4M}$, where $t_i = t + \frac{i}{4}$, and M is the number of years the CDS refers to. The discounted expected floating protection leg is:

$$\text{Protection}_t = \sum_{i=1}^{4M} e^{-r(t_i-t)} \text{LGD}(t_i) \Pr\{\text{Default} \in (t_{i-1}, t_i)\},$$

and the discounted expected fixed leg is:

$$\text{Premium}_t = \sum_{i=1}^{4M} e^{-r(t_i-t)} \text{CDS}_t(M) \Pr\{\text{Survival at } t_i\},$$

where r is the (constant) risk free rate, $\text{CDS}_t(M)$ is the premium paid every quarter, prevailing at time t , and $\text{LGD}(t_i)$ is the Loss-Given-Default at time t_i , which for simplicity is assumed to be constant, i.e. known at time t .

Equating Premium_t and Protection_t , and solving for $\text{CDS}_t(M)$, leaves:

$$\text{CDS}_t(M) = \frac{\sum_{i=1}^{4M} e^{-r(t_i-t)} \text{LGD}(t_i) \Pr\{\text{Default} \in (t_{i-1}, t_i)\}}{\sum_{i=1}^{4M} e^{-r(t_i-t)} \Pr\{\text{Survival at } t_i\}}. \quad (13.23)$$

It is a forward premium, as long as $t < t_0$. We assume that if the obligor defaults prior to the start date, t_0 , the contract is terminated.

At first glance, the previous derivation might look like “actuarial,” although it is not, actually. The reason is that the probabilities in Eq. (13.23) are *risk-neutral* probabilities. As such, they are, obviously, the same as those we use to price the bonds underlying the CDS contract. Therefore, there are no-arb relations that link bond prices to CDS premiums, which shall be emphasized later on (see Section 13.4.5.4). This point illustrates in a remarkable way one key difference between finance and insurance. Even if in insurance, one may end up pricing some products through risk-adjusted probabilities, finance is where we typically end up having many more traded risks than in insurance, and these risks are tightly related through no-arb restrictions.

Eq. (13.23) is a general formula we can use, once we have a model determining the risk-neutral probability of default. In this chapter, we implement Eq. (13.23) through a reduced-form approach, which will allow us to find the quarterly premium (or spread) $\text{CDS}_t(M)$ quite easily, as follows.

We have, denoting again with λ the instantaneous probability of default, that $\Pr\{\text{Survival at } t_i\} = e^{-\lambda(t_i-t)}$, and that $\Pr\{\text{Default at any } z \in (t_{i-1}, t_i)\} = e^{-\lambda(t_{i-1}-t)} - e^{-\lambda(t_i-t)}$. Intuitively, if the name survives at t_i (event E_i), it must necessarily have survived at t_{i-1} (event E_{i-1}), but the converse is not true: $E_i \subset E_{i-1}$, and the complement of E_i to E_{i-1} is nothing but the event of default between t_{i-1} and t_i .³ Substituting the previous probabilities into Eq. (13.23), we find that:

$$\text{CDS}_t(M) = \frac{\sum_{i=1}^{4M} e^{-r(t_i-t)} \text{LGD}(t_i) (e^{-\lambda(t_{i-1}-t)} - e^{-\lambda(t_i-t)})}{\sum_{i=1}^{4M} e^{-(r+\lambda)(t_i-t)}}. \quad (13.24)$$

For example, if $\text{LGD}(t_i)$ is constant and equal to LGD for each t_i , then, for $\Delta t = t_i - t_{i-1} = \frac{1}{4}$,

$$\text{CDS}_t(M) \approx \lambda \cdot \text{LGD} \cdot \Delta t \equiv (\text{expected losses per unit of time}) \cdot \Delta t, \quad (13.25)$$

where the approximation is obtained by making $e^{-\lambda(t_{i-1}-t)} - e^{-\lambda(t_i-t)} \approx \lambda e^{-\lambda(t_i-t)} \Delta t$. Naturally, λ is the *risk-neutral* instantaneous probability of default for the security.

Note that in this simple model, the CDS premiums for a fixed maturity are constant over time, as a result of the assumption that the intensity of default, λ and the short-term rate, r , are constant. Moreover, note that Eq. (13.25) shows that the CDS premium is approximately the same as the instantaneous spread of a defaultable bonds, as explained in Section 13.2. This property is to be expected, so to speak, as a purchase of a defaultable bond and protection on it is nothing but a synthetic default-free bond. Therefore, there must be a no-arbitrage relation between CDS spreads and defaultable bond spreads, as we anticipated earlier. However, in

³Mathematically, we have that $\Pr\{\text{Default at any } z \in (t_{i-1}, t_i)\} = \int_{t_{i-1}}^{t_i} \Pr\{\text{Default at } z\} dz$, where $\Pr\{\text{Default at } z\} = \lambda e^{-\lambda(z-t)} dz$.

general, Eq. (13.25) does not hold, as the assumptions made to achieve it (λ is constant, LGD is constant, r is constant, etc.) are quite unrealistic. On the contrary, we often observe CDS spreads curves that increase with maturity, as we shall explain in more analytical detail in Section 13.4.5.4. Indeed, we may take interesting views. For example, buying CDS for two years and sell CDS for three years is a view that default will not occur between the second and the third year from now.

13.4.3.2 Marking to market

Suppose a party goes long a CDS, meaning that at time t , he commits to a swap agreement whereby it pays $\text{CDS}_t(M)$ at time t_i , if the name survives by time t_i , and receives $\text{LGD}(t_i)$, if default occurs in the time interval $[t_{i-1}, t_i]$, for $4M$ time intervals. Each swap payoff—the “CDS-let” so to speak—is:

$$\text{cds}_t(t_i) \equiv \text{LGD}(t_i) \cdot \mathbb{I}_{\{\text{Default} \in (t_{i-1}, t_i)\}} - \text{CDS}_t(M) \cdot \mathbb{I}_{\{\text{Survival at } t_i\}}, \quad (13.26)$$

such that

$$\text{CDS}_t(M) : 0 = \sum_{i=1}^{4M} e^{-r(t_i-t)} \mathbb{E}_t[\text{cds}_t(t_i)],$$

where \mathbb{E}_t denotes the expectation conditional upon the information set at time t , taken under the risk-neutral probability. The solution to this equation is just that in Eq. (13.24).

What happens to the value of this contract at some subsequent time $\tau \in (t, t_0)$? The marking to market value of the CDS is the present value of the risk-neutral expectation of the single swaps payments $\text{cds}_t(t_i)$ in Eq. (13.26), consistently with the explanations in Section 10.4.6 of Chapter 10. So the marking to market value of the CDS at τ is,

$$\begin{aligned} \text{MtM}_\tau(M) &\equiv \sum_{i=1}^{4M} e^{-r(t_i-\tau)} \mathbb{E}_\tau[\text{cds}(t_i)] \\ &= \sum_{i=1}^{4M} [e^{-r(t_i-\tau)} \text{LGD}(t_i) (e^{-\lambda(t_{i-1}-\tau)} - e^{-\lambda(t_i-\tau)}) - \text{CDS}_t(M) e^{-(r+\lambda)(t_i-\tau)}] \\ &= [\text{CDS}_\tau(M) - \text{CDS}_t(M)] \sum_{i=1}^{4M} e^{-(r+\lambda)(t_i-\tau)}, \end{aligned}$$

where the last line follows by the definition of $\text{CDS}_\tau(M)$, i.e. by setting $t \equiv \tau$ in Eq. (13.24).

Note that in this model, marking-to-market is deterministic, because CDS premiums for a fixed maturity are constant over time, due to the fact that both λ and r are constant.

13.4.3.3 CDS on indexes, and options based thereon

A CDS index is a basket of credit entities in which the protection buyer pays the same premium on all the names in the index, until a fixed expiration date. Credit events are typically bound to bankruptcy or delinquencies. After a credit event, the entity is removed from the index and the contract goes through with a reduced notional amount, until expiration, as explained in more detail below. While CDS on single names are over-the-counter, CDS indexes are standardized and can give rise to relatively more liquid markets, as historical data on bid-ask spreads show. In fact, it can be cheaper to hedge a portfolio of CDS or bonds with a CDS index than it would be to buy many CDS to achieve a similar effect. There exist two main indices: (i) CDX index,

which contains North American and Emerging Market companies; and (ii) iTraxx index, which contains companies from the rest of the world.

Credit default swaptions are options to enter a CDS—typically a CDS index. Consider, first, swaptions on single names. A payer swaption gives the right to buy protection at some future date at some CDS fixed strike, and a receiver swaption gives the right to sell protection. If default of the name occurs prior to the swaption maturity, the contract is terminated. Note that evaluating credit default swaptions is trivial in the pricing context of Section 13.5.3.1, because CDS premiums move deterministically over time when the intensity of default and the short-term rate are both constant. Section 13.5.3.9 hinges upon a continuous-time model of stochastic intensity rates, and supplies an evaluation framework for these products.

Credit default index swaptions work differently. Firstly, as noted, at inception, a credit default index swap (CDIS) is referenced to a number of fixed companies chosen by a market maker, each carrying a given weight. Secondly, buyers of CDIS are typically those who provide protection to market makers: they stand ready to pay a predetermined loss-given-default for any default that occurs before maturity, which is constant and identical for all reference entities in the index. In exchange, the market makers pay the CDIS buyer a periodic fixed premium—the credit default index spread. After a default takes place, the nominal value of the CDIS is reduced by one, and no replacement of the defaulted firm would take place, as further explained in Section 13.5.3.9.

13.4.3.4 Disentangling default probability from risk-aversion

The following picture, taken from Fender and Hördahl (2007), illustrates the behavior of the credit market risk appetite before the 2007 credit market turmoil.

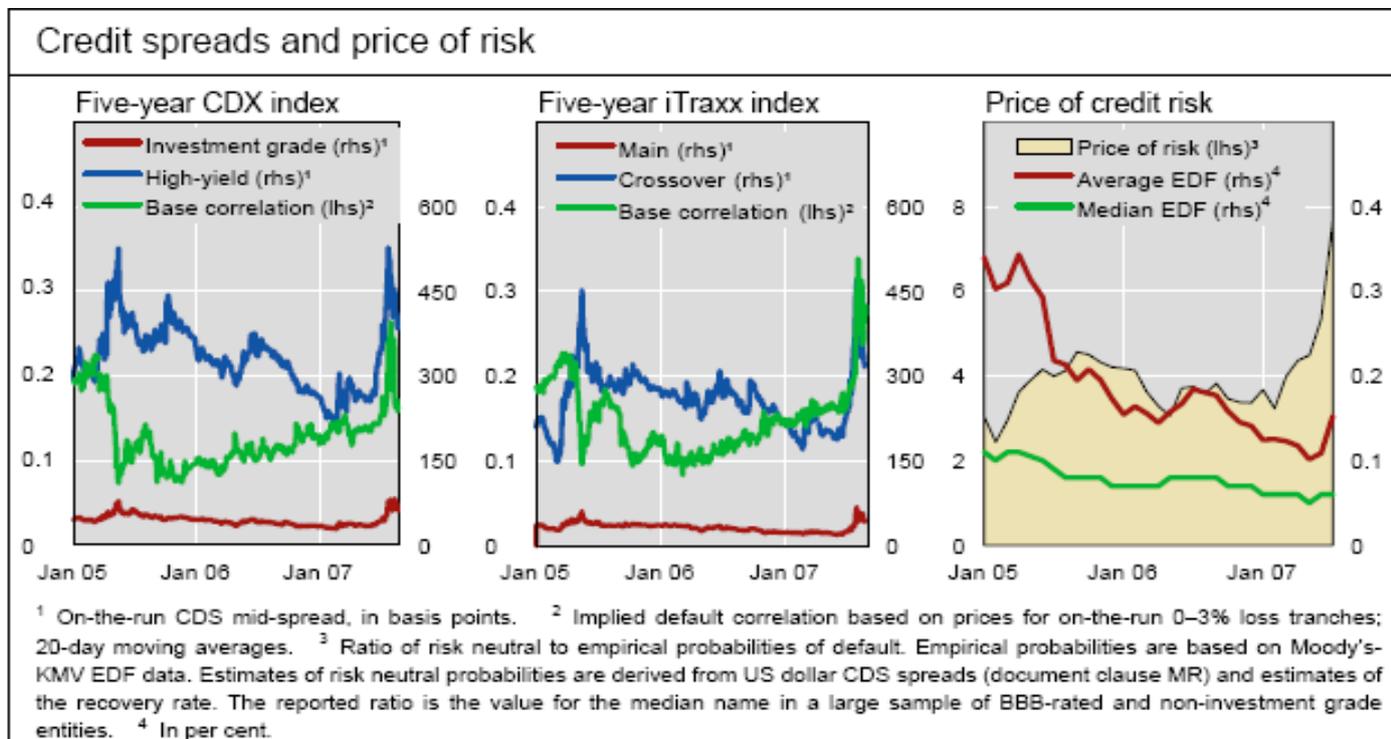


FIGURE 13.9. Antonio Mele does not claim any copyright on this picture, which is taken from Fender and Hördahl (2007). The picture has been put here for illustrative purposes only, and permission to the authors shall be duly asked before the book will be published.

How did the authors estimate the price of risk? Consider the expected losses under the actuarial, or physical probability for a given security. The counterpart to Eq. (13.25), under the physical probability, is:

$$\text{Expected Losses}^P \equiv \lambda^P \cdot \text{LGD} \cdot \Delta t,$$

where λ^P is the *physical* instantaneous probability of default for a given security. Assume that LGD is constant, to simplify. If investors require compensation for default events, the actuarial losses should be less than the CDS spread, i.e. $\text{Expected Losses}^P < \text{CDS}$, or,

$$\lambda > \lambda^P.$$

The risk-premium is defined as the difference between the actuarial losses, Expected Losses^P , and the CDS premium,

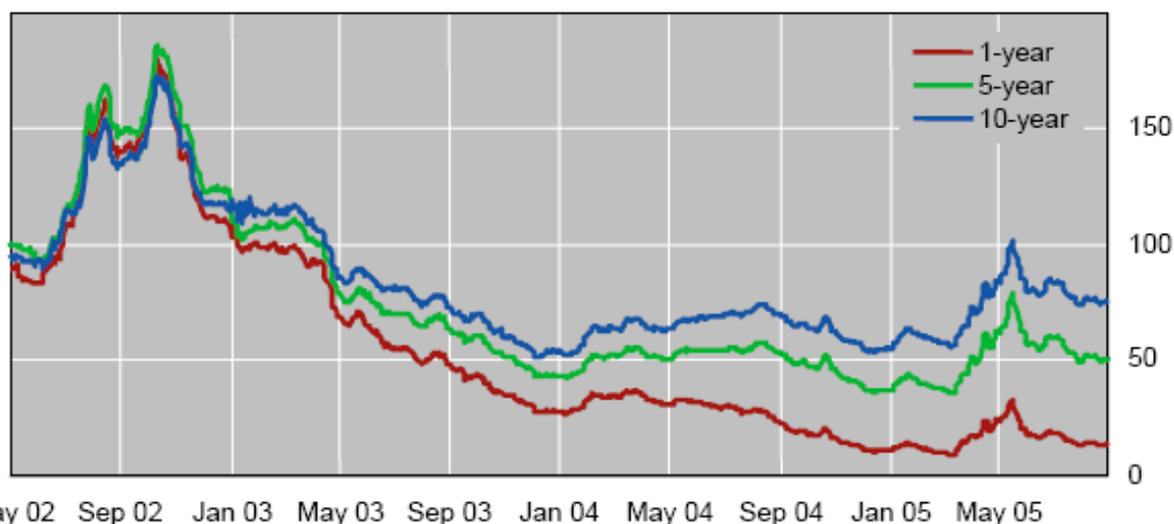
$$\text{Risk-Premium} = (\lambda - \lambda^P) \cdot \text{LGD} \cdot \Delta t.$$

The price of risk in Figure 13.9 is defined as the ratio of the CDS spread over Expected Losses^P ,

$$\text{Price-of-Risk} = \frac{\lambda}{\lambda^P}.$$

Early references to estimation methods are Duffie et al. (2005) and Amato (2005). Typically, Expected Losses^P are proxied by Moody's KMV's Expected Default Frequencies (EDFsTM), obtained through fully specified structural models for credit risk. The next pictures are taken from Amato (2005). As we can see, during the 2003-2005 period, credit spreads were so low, and this in turn gave incentives to CDO issuers to look for illiquid and relatively more complex assets to put as collateral, which led to the issuance of CDO relying on ABS such as MBS, or CDO², explained below.

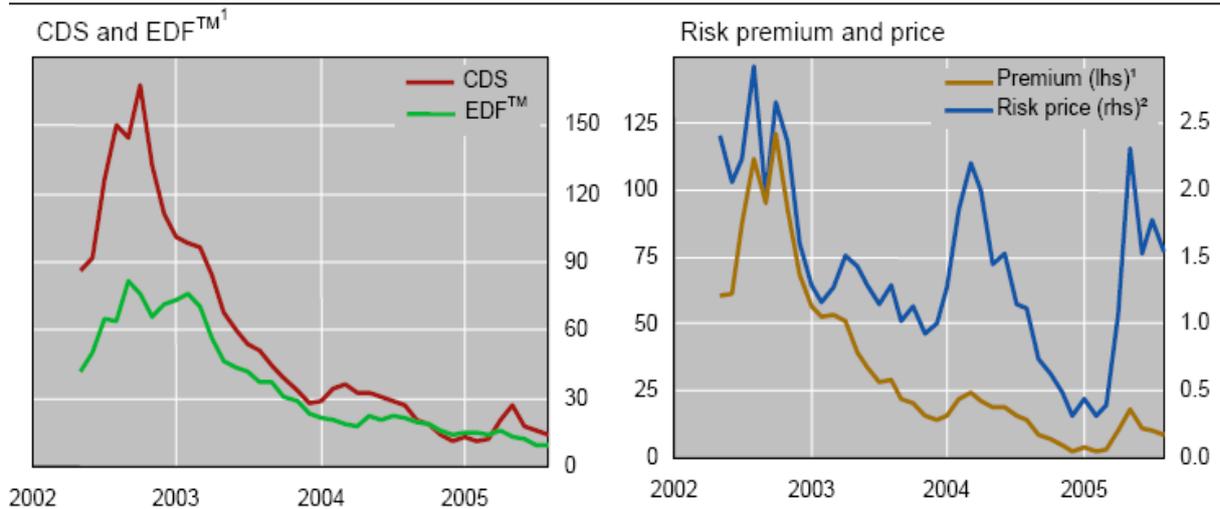
CDS spreads¹



¹ Based on the aggregate index, in basis points.

FIGURE 13.10. Antonio Mele does not claim any copyright on this picture, which is taken from Amato (2005). The picture has been put here for illustrative purposes only, and permission to the author shall be duly asked before the book will be published.

CDS risk premium and price of default risk



¹ One-year horizon, in basis points. ² One-year price of default risk.
 FIGURE 13.11. Antonio Mele does not claim any copyright on this picture, which is taken from Amato (2005). The picture has been put here for illustrative purposes only, and permission to the author shall be duly asked before the book will be published.

The following picture illustrates the behavior of CDS indexes during approximately 20 years before the 2007-2009 credit market turmoil.

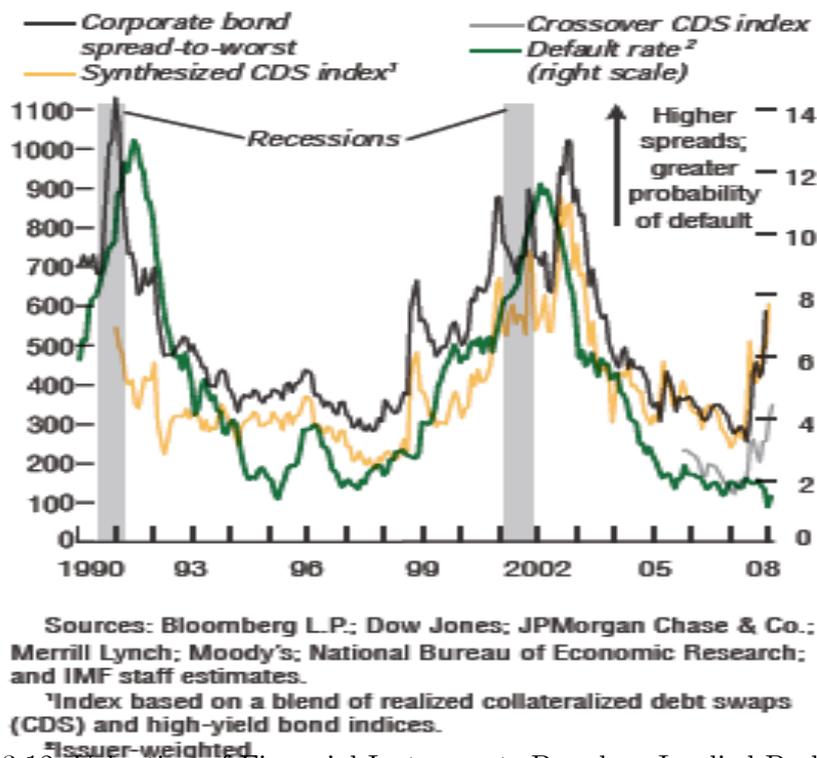


FIGURE 13.12. Valuation of Financial Instruments Based on Implied Probability of Default. Antonio Mele does not claim any copyright on this picture, which is taken from IMF (2008). The picture has been put here for illustrative purposes only, and permission to the authors shall be duly asked before the book will be published.

13.4.3.5 Continuous time

We may relax the assumption the instantaneous intensity of default, λ , is constant. This intensity is defined under the risk-neutral probability and can change either because the intensity of default under the physical probability changes or because risk-appetite changes, or both. We examine the asset pricing implications of time-varying intensities, by exploring how probabilities of survival change in a simple setting, where we do not single out the reasons leading to variations in λ .

First, we assume the instantaneous probability of default can only change discretely, giving rise to random intensities λ_t , meaning that λ_t is the intensity of default in the time interval $[t - 1, t]$. Let \mathcal{F}_t be the information set as of time t . We assume that λ_t is \mathcal{F}_t -measurable. What is the probability of survival of any given name in this case? We have, by Bayes's theorem,

$$\Pr \{ \text{Surv at } t | \text{Surv at } t - 1 \} = \frac{\Pr \{ \text{Surv at } t \}}{\Pr \{ \text{Surv at } t - 1 \}}. \tag{13.27}$$

By a repeated use of Eq. (13.27),

$$\begin{aligned} \Pr \{ \text{Surv at } t \} &= \Pr \{ \text{Surv at } t | \text{Surv at } t - 1 \} \Pr \{ \text{Surv at } t - 1 \} \\ &= \dots \\ &= \prod_{n=1}^t \Pr \{ \text{Surv at } n | \text{Surv at } n - 1 \}. \end{aligned} \tag{13.28}$$

So we are left with finding $\Pr \{ \text{Surv at } n | \text{Surv at } n - 1 \}$. Consider the following arguments. If λ_n was not random and fixed at some $\bar{\lambda}_n$, then, $\Pr \{ \text{Surv at } n | \text{Surv at } n - 1 \} = e^{-\bar{\lambda}_n}$. When λ_n is random, $e^{-\lambda_n}$ is the probability of survival, conditioned upon some particular value the intensity could possibly take. Heuristically, then, $\Pr \{ \text{Surv at } n | \text{Surv at } n - 1 \} = \sum_{s \in \mathcal{S}} e^{-\lambda_n(s)} \Pr \{ s \}$, where $\lambda_n(s)$ is, so to speak, the value λ_n would take in state s , $\Pr \{ s \}$ is the likelihood that state s occurs and, finally, \mathcal{S} is the set of all possible states, as illustrated by Figure 13.3.

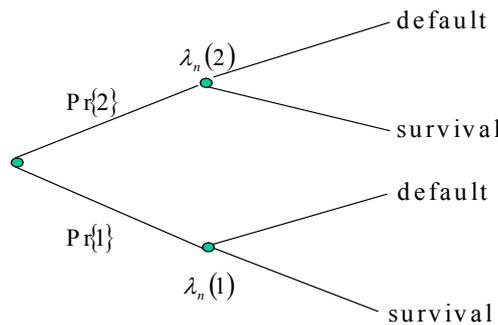


FIGURE 13.13. This picture illustrates the determination of the probability of survival in the case of random default intensities going over one period and two states. At the beginning of period n , nature draws the event defining the intensity of the default, which is either $\lambda_n(1)$ with probability $\Pr \{ 1 \}$, or $\lambda_n(2)$ with probability $\Pr \{ 2 \} = 1 - \Pr \{ 1 \}$. Then, the two paths leading to survival have probability of occurrence equal to $\Pr \{ 1 \} e^{-\lambda_n(1)}$ and $\Pr \{ 2 \} e^{-\lambda_n(2)}$, such that the total probability of survival equals $\Pr \{ 1 \} e^{-\lambda_n(1)} + \Pr \{ 2 \} e^{-\lambda_n(2)}$.

Therefore, $\Pr\{\text{Surv at } n | \text{Surv at } n-1\} = \mathbb{E}[e^{-\lambda_n} | \mathcal{F}_{n-1}]$, where \mathbb{E} denotes the expectation taken under the risk-neutral probability. Inserting this result into Eq. (13.28), and using the Law of Iterated Expectations, leaves:

$$\Pr\{\text{Surv at } t\} = \mathbb{E}\left[e^{-\sum_{n=1}^t \lambda_n}\right].$$

Under regularity conditions, we can easily extend the previous result to a continuous time setting. For example, we may assume that the risk-neutral default intensity, $\lambda(t)$, is solution to:

$$d\lambda(t) = \phi(\bar{\lambda} - \lambda(t))dt + \sigma\sqrt{\lambda(t)}dW(t), \quad \lambda(0) = \lambda. \quad (13.29)$$

where W is a standard Brownian motion under the risk-neutral probability, and ϕ , $\bar{\lambda}$ and σ are three positive constants. This is the same as the Cox, Ingersoll and Ross (1985) (CIR) model of the short-term rate reviewed in Chapter 12. Therefore, under the parameter restrictions in Chapter 12, $\lambda(t)$ is always positive, and

$$P_{\text{surv}}(\lambda, t) \equiv \Pr\{\text{Surv at } t\} = \mathbb{E}\left[e^{-\int_0^t \lambda(s)ds}\right]. \quad (13.30)$$

Eq. (13.30) is, formally, the same as the Feynman-Kac representation of a solution to a PDE, solved by a bond price in the CIR model. In other words, the model for the survival probability in Eqs. (13.29)-(13.30) has the same mathematical structure as that leading to the price of a bond in the CIR model. Therefore, a closed-form solution is available for $P_{\text{surv}}(\lambda, t)$. It is given by:

$$P_{\text{surv}}(\lambda, N) = \Phi(N)e^{-B(N)\lambda},$$

$$\Phi(N) = \left(\frac{2\gamma e^{\frac{1}{2}(\phi+\gamma)N}}{(\phi+\gamma)(e^{\gamma N}-1) + 2\gamma}\right)^{\frac{2\phi\bar{\lambda}}{\sigma^2}}, \quad B(N) = \frac{2(e^{\gamma N}-1)}{(\phi+\gamma)(e^{\gamma N}-1) + 2\gamma}, \quad \gamma = \sqrt{\phi^2 + 2\sigma^2}. \quad (13.31)$$

More generally, we can build up a whole family of models with a closed-form solution, the affine class reviewed in Chapter 12, by assuming that:

$$\lambda(t) = \lambda_0 + \lambda_1 \cdot y(t), \quad (13.32)$$

where λ_0 is a constant, λ_1 is a vector of constants, and y is a multivariate jump-diffusion process, with drift and diffusion terms as in Section 12.4.6 of Chapter 12. This model is interesting, as we can judiciously choose the components of $y(t)$ which we suppose may affect the default intensity. For example, some of them could be unobservable, and others could be observable, and relate, say, to the business cycle or even the structure of the firm.

Given any solution for the survival probability predicted by any of these affine models when $y(0) = y$, $P_{\text{surv}}(y, t)$ say, we can easily compute

$$\Pr\{\text{Default} \in (t_{i-1}, t_i)\} = P_{\text{surv}}(y, t_{i-1}) - P_{\text{surv}}(y, t_i). \quad (13.33)$$

We can look at the bond spreads and the CDS spreads implied by this modeling choice. In Appendix 3, we show the price of a defaultable pure discount bond expiring in N years is:

$$P(y, N) = e^{-rN}P_{\text{surv}}(y, N) + \int_0^N e^{-rt} \Pr\{\text{Default} \in dt\} \text{Rec}(t) dt, \quad (13.34)$$

where $\text{Rec}(t)$ denotes the recovery value in case of default, supposed to be known. This evaluation result is, naturally, consistent with a similar derivation provided in Section 12.4.7 of Chapter 12, although in this chapter we are emphasizing more “survival arguments.”

As for the forward CDS spreads, we have, by Eq. (13.23),

$$\text{CDS}_t(N) = \frac{\sum_{i=1}^{4N} e^{-r(t_i-t)} \text{LGD}(t_i) [P_{\text{surv}}(y, t_{i-1}) - P_{\text{surv}}(y, t_i)]}{\sum_{i=1}^{4N} e^{-r(t_i-t)} P_{\text{surv}}(y, t_i)},$$

where N is, again, the number of years the CDS refers to, and $t_i = t + \frac{i}{4}$.

Assume the short-term rate, r , is zero, and that loss-given-default is constant and equal to LGD . Then, as shown in Appendix 3, the price of a defaultable pure discount bond, $P(\lambda, N)$, and the CDS premium, obtained from the forward once we set $t = 0$, $\text{CDS}_0(N)$, are given by:

$$P(\lambda, N) = 1 - \text{LGD} \cdot (1 - P_{\text{surv}}(\lambda, N)), \quad \text{CDS}_0(N) = \text{LGD} \cdot \frac{1 - P_{\text{surv}}(\lambda, N)}{\sum_{i=1}^{4N} P_{\text{surv}}(\lambda, t_i)}. \quad (13.35)$$

Figure 13.14 depicts the bonds spread, $\frac{1}{N} \ln P(\lambda, N)$, and the annualized credit default spreads, $4 \times \text{CDS}_0(N)$, when the parameters in Eq. (13.29) are $\phi = 0.25$, $\bar{\lambda} = 0.04$ and $\sigma = \sqrt{\bar{\lambda}}$, with loss-given-default $\text{LGD} = 0.60$, and two values of the current intensity: $\lambda = \bar{\lambda} = 0.04$, and $\lambda = 0.02$. Assuming that LGD is constant is not plausible, empirically. Instead, we know LGD moves countercyclically for most names, although it does not exhibit strong business cycle features, for sovereigns. For sovereigns, the size of the country and debt distribution seem to be by far more important.

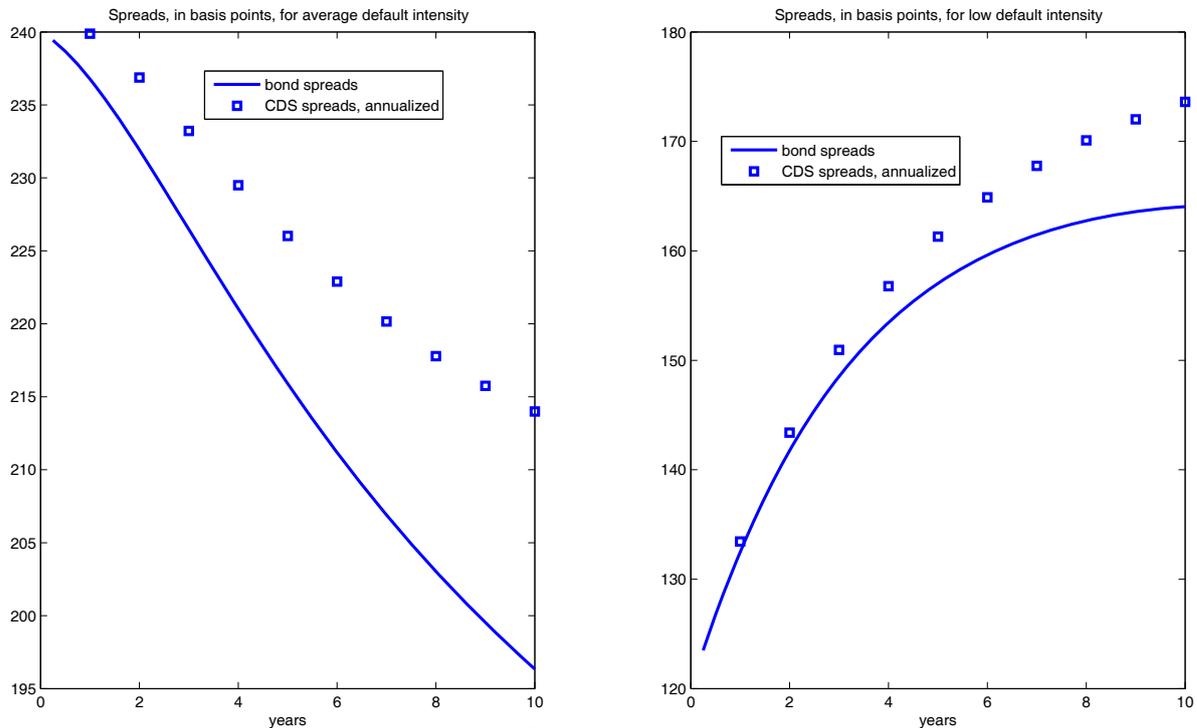


FIGURE 13.14. Spreads on bonds and CDS predicted by the affine model in Eq. (13.29). The left panel depicts the spreads when the current default intensity equals the long-run mean, $\lambda = \bar{\lambda} = 0.04$. The right panel depicts the spreads in good times, i.e., when the current intensity of default takes a low value, $\lambda = 0.02$. In each case the recovery rate equals 40%.

The mechanism is that given the mean-reverting behavior of λ , good times are likely followed by bad, and so when $\lambda = 0.02$, we expect default rates to rise in the future. As a consequence, spreads are increasing in maturity. Moreover, we easily see that bond spreads are approximately equal to CDS spreads at short maturities. At longer maturities, the two spreads diverge, with CDS spreads, $4 \times \text{CDS}_0(N)$, dominating bonds spreads, $-\frac{1}{N} \ln P(\lambda, N)$. Moreover, we have that the two curves are decreasing in time to maturity even when the current value of the intensity equals the long-run one, $\bar{\lambda}$.

Where do these two properties originate from? The first follows because we have, approximately, that:

$$\begin{aligned} \frac{-1}{N} \ln P(\lambda, N) &= \frac{-1}{N} \ln [1 - \text{LGD} \cdot (1 - P_{\text{surv}}(\lambda, N))] \\ &\approx \text{LGD} \cdot \frac{1 - P_{\text{surv}}(\lambda, N)}{N} \\ &\leq \text{LGD} \cdot \frac{1 - P_{\text{surv}}(\lambda, N)}{\frac{1}{4} \sum_{i=1}^{4N} P_{\text{surv}}(\lambda, t_i)} \\ &= 4 \times \text{CDS}_0(N). \end{aligned}$$

The second property is a convexity effect. We tackle this issue with arguments similar to those made while studying a related topic in Chapter 12, Section 12.3.4. For the bond spreads, since $\mathbb{E}(\lambda(s)) = \bar{\lambda} + e^{-\phi s}(\lambda - \bar{\lambda})$, we have, approximately,

$$\begin{aligned} \frac{-1}{N} \ln P(\lambda, N) &= \frac{-1}{N} \ln [1 - \text{LGD} \cdot (1 - P_{\text{surv}}(\lambda, N))] \\ &= \frac{-1}{N} \ln \left[1 - \text{LGD} \cdot \left(1 - \mathbb{E} \left(e^{-\int_0^N \lambda(s) ds} \right) \right) \right] \\ &\leq \frac{-1}{N} \ln \left[1 - \text{LGD} \cdot \left(1 - e^{-\int_0^N \mathbb{E}(\lambda(s)) ds} \right) \right] \\ &\approx \text{LGD} \cdot \frac{1 - e^{-\int_0^N \mathbb{E}(\lambda(s)) ds}}{N} \\ &= \text{LGD} \cdot \frac{1 - e^{-\bar{\lambda}N - (\lambda - \bar{\lambda}) \frac{1 - e^{-\phi N}}{\phi}}}{N}, \end{aligned}$$

so that even if $\lambda = \bar{\lambda}$, then, bond spreads are bounded away by a decreasing function (in N). Of course, it doesn't necessarily mean that bond spreads have to be decreasing as well, but that bounding function helps this happening. As for the CDS spreads, we have, approximately:

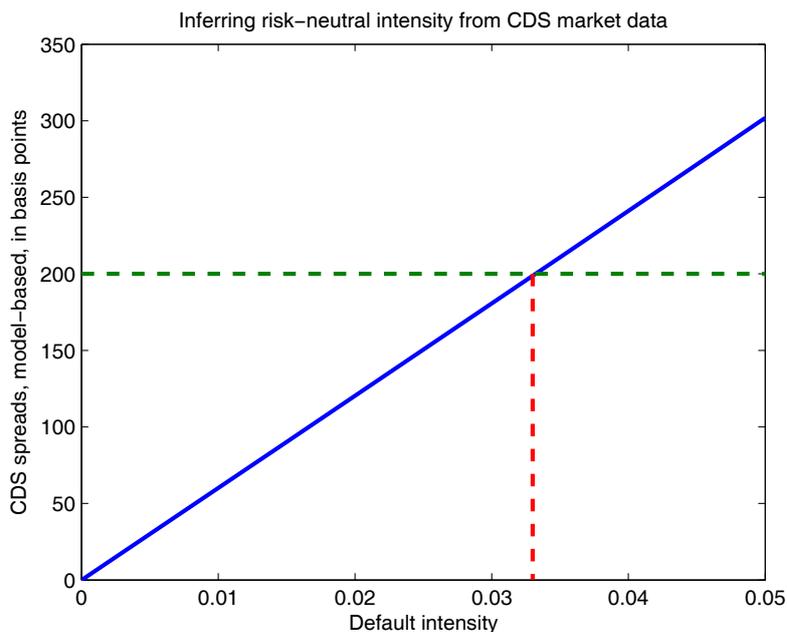
$$4 \times \text{CDS}_0(N) = \text{LGD} \cdot \frac{1 - P_{\text{surv}}(\lambda, N)}{\frac{1}{4} \sum_{i=1}^{4N} P_{\text{surv}}(\lambda, t_i)} \leq \text{LGD} \cdot \frac{1 - P_{\text{surv}}(\lambda, N)}{N \cdot P_{\text{surv}}(\lambda, N)} \approx -\frac{1}{N} \ln P_{\text{surv}}(\lambda, N),$$

such that for $\lambda = \bar{\lambda}$, $\text{CDS}_0(N)$ is bounded away by a decreasing function (in N), for the same arguments made as regards the bond spreads, $-\frac{1}{N} \ln P(\lambda, N)$.

13.4.3.6 A trading strategy

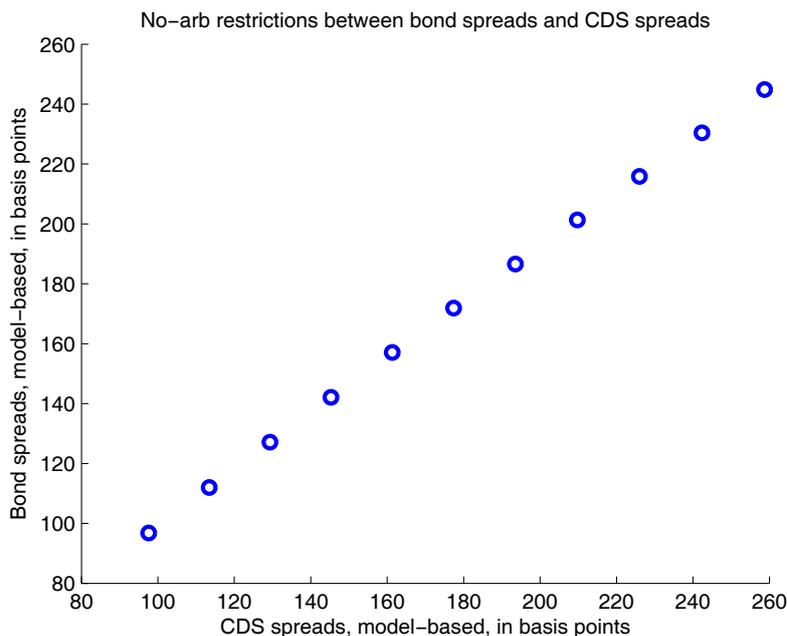
Bond prices and CDS spreads are driven by the same state variable, the default intensity, and so they are restricted to lie on some space, to be consistent with no-arbitrage. To illustrate,

consider, first, the simple case where the default intensity is constant, such that CDS spreads are given by Eq. (13.24). Given this model, we can look at the market data for CDS spreads, and infer the risk-neutral intensity, as in the picture below.



In this picture, the CDS spreads predicted by Eq. (13.24) are depicted as a function of the risk-neutral intensity, λ , assuming $N = 5$ years, $\text{LGD} = 0.60$ and the short-term rate r is zero. For example, if we had to observe a CDS equal to 200 basis points, we would infer a value of λ approximately equal to $\hat{\lambda} = 0.033$. The key point is this very same $\hat{\lambda}$ should be pricing the zero as well, such that for $N = 5$, $P(N) = 1 - \text{LGD} \cdot (1 - e^{-\hat{\lambda}N}) = 0.90874$, and so we might go long (short) the zero if its market price is lower (higher) than 0.90874 and short (go long) the CDS. While this example relies on a constant default intensity, we can use the same strategy in the more general case where default intensities are stochastic. In this case, bond prices and CDS spreads should also be restricted, by no-arbitrage. The picture below shows the restrictions between bond spreads and CDS spreads, obtained with the same parameter values as those used to produce Figure 13.14, and values of current default intensities ranging from

nearly zero to up to 0.05.



13.4.3.7 Hazard rates

In a pricing context, the relevant probabilities of survival are obviously conditioned upon the time of evaluation, time 0 say. For example, the probability of default in Eq. (13.33) is only conditioned to the information we have at time zero. More generally, the probability of defaulting in the time interval (t_{i-1}, t_i) , conditional upon survival at time $t < t_{i-1}$, is:

$$\Pr\{\text{Default} \in (t_{i-1}, t_i) | \text{Survival at } t\} = \frac{P_{\text{surv}}(y, t_{i-1}) - P_{\text{surv}}(y, t_i)}{P_{\text{surv}}(y, t)}. \quad (13.36)$$

For example, for $t = t_{i-1}$, and (t_{i-1}, t_i) small, and λ deterministic, a simple approximation to this conditional probability can be,

$$\begin{aligned} \Pr\{\text{Default} \in (t_{i-1}, t_i) | \text{Survival at } t\} &\approx \frac{\frac{\partial}{\partial t} P_{\text{surv}}(y, t)}{P_{\text{surv}}(y, t)} (t_i - t_{i-1}) \\ &\equiv \frac{p_{\text{default}}(y, t)}{1 - P_{\text{default}}(y, t)} (t_i - t_{i-1}) \\ &= \lambda(t) (t_i - t_{i-1}), \end{aligned}$$

with straight forward notation. The previous expressions are known as *hazard rates*. They coincide with $\lambda(t) dt$, when $\lambda(t)$ is deterministic. If $\lambda(t)$ is not deterministic, simple computations lead to:

$$\Pr\{\text{Default} \in (t, t + dt) | \text{Survival at } t\} = \mathbb{E}_{Q_\lambda} [\lambda(t)] dt, \quad (13.37)$$

where Q_λ is a new probability, with Radon-Nikodym derivative given by:

$$\frac{dQ_\lambda}{dQ} \Big|_{\mathcal{F}_t} = \frac{e^{-\int_0^t \lambda(s) ds}}{P_{\text{surv}}(\lambda, t)}. \quad (13.38)$$

Accordingly, under Q_λ , the state variables in Eq. (13.32) follow a diffusion process, with a drift process tilted, due to this change of measure. For example, in the simple setting of Eq. (13.29), we have that, for a fixed t ,

$$\begin{aligned} d\lambda(s) &= (\mathcal{B}_0 - \mathcal{B}_1^t(s) \lambda(s)) ds + \sigma \sqrt{\lambda(s)} dW_\lambda(s), \quad s \in (0, t], \quad \lambda(0) = \lambda, \\ \mathcal{B}_0 &= \phi \bar{\lambda}, \quad \mathcal{B}_1^t(s) = \phi + B(t-s) \sigma^2, \quad B(\cdot) \text{ as in Eq. (13.31)}, \end{aligned} \quad (13.39)$$

where W_λ is a Brownian motion under Q_λ . Therefore, by Eq. (13.37), and computations,

$$\Pr\{\text{Default} \in (t, t+dt) | \text{Survival at } t\} = \left[\frac{\lambda}{G(t)} + \mathcal{B}_0 \int_0^t \frac{G(s)}{G(t)} ds \right] dt, \quad G(x) \equiv e^{\int_0^x \mathcal{B}_1^t(u) du}.$$

Appendix 5 provides a proof of these results, which to the best of our knowledge, are developed here for the first time.

13.4.3.8 Extracting probabilities of default from market data

Market data obviously convey information about probabilities of default, which might be extracted from these data, under a number of assumptions. To illustrate this possibility in a simple case, assume that the recovery rate is zero, and that the short-term rate and the instantaneous probability of default are both continuous time Markov and independent of each other. Then, the price of a defaultable zero is: $P_{\text{def}}(\lambda, N) = P(N) \cdot P_{\text{surv}}(\lambda, N)$, where $P_{\text{def}}(\lambda, N)$ is the price of a defaultable zero and $P(N)$ is the price of a non-defaultable zero. Therefore, we can read the *risk-neutral* probability of survival from the defaultable/non-defaultable bond price ratio:

$$P_{\text{surv}}(\lambda, N) = \frac{P_{\text{def}}(\lambda, N)}{P(N)}. \quad (13.40)$$

Naturally, surviving until some time N_2 means having survived until some time $N_1 < N_2$ and having survived from N_1 to N_2 . Therefore, $P_{\text{surv}}(\lambda, N_2) = P_{\text{surv}}(\lambda, N_1) \cdot P_{\text{surv}}(\lambda, N_1, N_2)$, where $P_{\text{surv}}(\lambda, N_1, N_2)$ is the risk-neutral probability of survival between N_1 and N_2 . Using Eq. (13.40), then, we can extract this probability, as follows:

$$P_{\text{surv}}(\lambda, N_1, N_2) = \frac{P_{\text{def}}(\lambda, N_2) P(N_1)}{P_{\text{def}}(\lambda, N_1) P(N_2)}.$$

The previous example relies on the simplifying assumption of a zero recovery rate, but it can be generalized to the case where the recovery rate is nonzero. But in this case, bond prices would convey information about both probabilities of default and recovery rates, an identification issue to be dealt with.

13.4.3.9 Pricing credit default swaptions

Swaptions on single names

The model of stochastic intensity rates allows us to think about the pricing of the credit default swaptions briefly mentioned in Section 13.5.3.3. We now actually assume that both intensity rates and the short-term rate are stochastic: we take the short-term rate r to be a diffusion process, and default arriving as a Cox process with intensity λ adapted to r . Eq. (13.23) needs to be modified to accommodate for the presence of a stochastic short-term rate. We also generalize the pricing framework to one, where the value of a default swap is not necessarily zero. Consider

the following definition of a *default swap*, as a contract whereby a party stands ready to pay his counterparty a loss determined by a credit event for a given flows of premiums, which we refer to as CDS premiums. We assume that loss-given-default is constant and equal to LGD. Denote the CDS premium agreed at time t with $\overline{\text{CDS}}_t(M)$, $t \leq t_0$. It is a forward default CDS premium really, as in Section 13.4.3.1. Therefore, we shall still assume that the contract is terminated should the underlying obligor default prior to the start date, t_0 .

In a such a forward default swap, the protection buyer is in fact committing to a swap agreement whereby it pays $\overline{\text{CDS}}_t(M)$ at time t_i , if the name survives by time t_i , and receives LGD, if default occurs in the time interval $[t_{i-1}, t_i]$, for $4M$ time intervals. Each swap payoff is:

$$\text{cds}_t(t_i) \equiv \text{LGD} \cdot \mathbb{I}_{\{\text{Default} \in (t_{i-1}, t_i)\}} - \overline{\text{CDS}}_t(M) \cdot \mathbb{I}_{\{\text{Survival at } t_i\}}, \quad (13.41)$$

where $\mathbb{I}_{\{\cdot\}}$ is the indicator function.

The value of the default swap agreed at time t is,

$$\text{DSw}_t = \sum_{i=1}^{4M} \mathbb{E}_t \left[e^{-\int_t^{t_i} r(\tau) d\tau} \text{cds}_t(t_i) \right] = \text{LGD} \cdot V_{0t} - \overline{\text{CDS}}_t(M) \cdot V_{1t}, \quad (13.42)$$

where \mathbb{E}_t denotes the risk-neutral expectation, taken conditional upon the information set at time t , and,

$$V_{0t} \equiv \sum_{i=1}^{4M} \mathbb{E}_t \left[e^{-\int_t^{t_i} r(\tau) d\tau} \cdot \mathbb{I}_{\{\text{Default} \in (t_{i-1}, t_i)\}} \right], \quad V_{1t} \equiv \sum_{i=1}^{4M} \mathbb{E}_t \left[e^{-\int_t^{t_i} r(\tau) d\tau} \cdot \mathbb{I}_{\{\text{Survival at } t_i\}} \right].$$

The interpretation of V_{0t} is that of the value of one dollar paid off the first time after default, provided default occurs prior to the maturity of the default swap, $4M$. Instead, the interpretation of V_{1t} is that of the value of an annuity of one dollar paid at the dates t_1, t_2, \dots, t_{4M} , until default or maturity of the default swap, whichever occurs first. In other words, V_{1t} is the value of a basket of defaultable bonds with zero recovery value—a *defaultable present value of the basis point*.

The forward default spread is the value of $\overline{\text{CDS}}_\tau(M)$ such that $\text{DSw}_\tau = 0$, and is given by:

$$\text{CDS}_\tau(M) \equiv \text{LGD} \cdot \frac{V_{0\tau}}{V_{1\tau}}, \quad (13.43)$$

such that we can express the value of the default swap at time τ , agreed at time $t \leq \tau$, and denoted as $\text{DSw}_{t,\tau}$, as the product of the annuity factor $V_{1\tau}$ times the difference between the forward default spread and any fixed default spread $\overline{\text{CDS}}_t(M)$,

$$\text{DSw}_{t,\tau} = \text{LGD} \cdot V_{0\tau} - \overline{\text{CDS}}_t(M) \cdot V_{1\tau} = V_{1\tau} \cdot (\text{CDS}_\tau(M) - \overline{\text{CDS}}_t(M)). \quad (13.44)$$

Note that the derivation leading to Eq. (13.44) generalizes that underlying the marks-to-market updates in Section 13.5.3.2.

Next, note that for any $\mathbb{F}(T)$ -measurable random variable $X(T)$ adapted to $(r(\tau))_{\tau \in [t, T]}$ and satisfying enough regularity conditions, we have that for fixed T , and by the Law of Iterated

Expectations,

$$\begin{aligned}
& \mathbb{E} \left[e^{-\int_t^T r(\tau) d\tau} \cdot X(T) \cdot \mathbb{I}_{\{\text{Survival at } T\}} \right] \\
&= \mathbb{E} \left[\mathbb{E} \left(e^{-\int_t^T r(\tau) d\tau} \cdot X(T) \cdot \mathbb{I}_{\{\text{Survival at } T\}} \middle| \mathbb{F}^r(T) \right) \right] \\
&= \mathbb{E} \left[e^{-\int_t^T r(\tau) d\tau} \cdot X(T) \cdot \mathbb{E} \left(\mathbb{I}_{\{\text{Survival at } T\}} \middle| \mathbb{F}^r(T) \right) \right] \\
&= \mathbb{E} \left[e^{-\int_t^T (r(\tau) + \lambda(\tau)) d\tau} \cdot X(T) \right], \tag{13.45}
\end{aligned}$$

where $\mathbb{F}^r(t)$ is the information set including the path of the short-term rate only.

Define the probability Q_{sc} through the Radon-Nikodym derivative:

$$\frac{dQ_{\text{sc}}}{dQ} \Bigg|_{\mathbb{F}_T^r} = e^{-\int_t^T (r(\tau) + \lambda(\tau)) d\tau} \frac{V_{1T}}{V_{1t}}, \tag{13.46}$$

where Q is the risk-neutral probability. It is easy to see that Q_{sc} does indeed integrate to one, using results in Schönbucher (2003, Chapter 7), Lando (2004, Chapter 5) and Chapter 12 of these Lectures. Following Schönbucher (2003, p. 180), we refer to Q_{sc} as the *survival contingent measure*.

We can also show that for any $T \leq t_0$,

$$V_{0t} = \mathbb{E} \left[e^{-\int_t^T (r(\tau) + \lambda(\tau)) d\tau} V_{0T} \right]. \tag{13.47}$$

Indeed, V_{0t} is the value of a basket of securities paying off contingent upon default not having occurred prior to time T , with $T \leq t_0$, in which case the value drops to zero. Following derivations in Section 12.3.6 of Chapter 12 of these Lectures, we have that $LV_{0t} - rV_0 + \lambda(0 - V_0) = 0$, where L is the infinitesimal generator for diffusions, whence Eq. (13.47).

Therefore, given the definition of Q_{sc} in Eq. (13.46) and the martingale property of V_{0t} in Eq. (13.47), we have that the forward default spread in Eq. (13.43) is a martingale under the survival contingent measure:

$$\begin{aligned}
& \mathbb{E}^{\text{sc}} [\text{CDS}_T(M)] \\
&= \mathbb{E} \left[e^{-\int_t^T (r(\tau) + \lambda(\tau)) d\tau} \frac{V_{1T}}{V_{1t}} \text{CDS}_T(M) \right] = \mathbb{E} \left[e^{-\int_t^T (r(\tau) + \lambda(\tau)) d\tau} \frac{V_{0T}}{V_{1t}} \text{LGD} \right] = \text{CDS}_t(M),
\end{aligned}$$

where \mathbb{E}^{sc} denotes the expectation taken under the survival contingent measure.

We can evaluate default swaptions by relying on the survival contingent measure. By Eq. (13.44), the payoff of a swaption payer is,

$$\text{DSw}_{t,T} = \mathbb{I}_{\{\text{Survival at } T\}} \cdot V_{1T} \cdot (\text{CDS}_T(M) - K)^+, \quad T = t_0,$$

for a strike K . We have, using the property in Eq. (13.45),

$$\begin{aligned}
& \mathbb{E} \left[e^{-\int_t^T r(\tau) d\tau} \text{DSw}_{t,T} \right] \\
&= \mathbb{E} \left[e^{-\int_t^T (r(\tau) + \lambda(\tau)) d\tau} V_{1T} \cdot (\text{CDS}_T(M) - K)^+ \right] = V_{1t} \cdot \mathbb{E}^{\text{sc}} [(\text{CDS}_T(M) - K)^+].
\end{aligned}$$

We know that $\text{CDS}_t(M)$ is a martingale under the survival contingent measure. Let $W^{\text{sc}}(\tau)$ be a Brownian motion under Q^{sc} . Assume that

$$\frac{d\text{CDS}_\tau(M)}{\text{CDS}_\tau(M)} = \sigma dW^{\text{sc}}(\tau), \quad \tau \in [t, T],$$

where σ is the volatility parameter, a constant. We can apply Black (1976) to obtain evaluation formulae in this environment.

Swaptions on CDS indexes

Consider, first, a *CDS index*, as succinctly described in Section 13.5.3.3. Let n be the initial number of names in the index decided at time t , each having a notional value equal to $\frac{1}{n}$, the same loss-given-default LGD and the same default intensity λ . Denote with $\mathcal{D}(t_{i-1}, t_i)$ the number of names having defaulted over the time interval (t_{i-1}, t_i) ,

$$\mathcal{D}(t_{i-1}, t_i) \equiv \sum_{j=1}^n \mathbb{I}_{\{\text{Def}_j \in (t_{i-1}, t_i)\}},$$

where $\mathbb{I}_{\{\text{Def}_j \in (t_{i-1}, t_i)\}}$ is the indicator of the event that the j -th name defaults over the time interval (t_{i-1}, t_i) . Define the following swap payoff, occurring at time t_i , and generalizing that in Eq. (13.41) holding for single names,

$$\text{cdx}_t(t_i) \equiv \text{LGD} \frac{1}{n} \mathcal{D}(t_{i-1}, t_i) - \overline{\text{CDX}}_t(M) \left(1 - \frac{1}{n} \sum_{h=0}^i \mathcal{D}(t_{h-1}, t_h) \right), \quad (13.48)$$

where $\mathcal{D}(t_{-1}, t_0)$ denotes the number of defaults occurred over the time interval (t, t_0) . The first term of $\text{cdx}_t(t_i)$ is the loss in the index occurring at time t_i , paid off by the protection seller, whereas the second term is the protection premium, which equals the constant premium $\overline{\text{CDX}}_t(M)$ times the outstanding notional.⁴

The value of the protection leg minus that of the premium leg over the life of the index is obtained as:

$$\text{DSX}_t = \mathbb{E}_t \left[\sum_{i=1}^{4M} e^{-\int_t^{t_i} r(\tau) d\tau} \text{cdx}_t(t_i) \right] = \text{LGD} \cdot V_{0t}^x - \overline{\text{CDX}}_t(M) \cdot V_{1t}^x, \quad (13.49)$$

where,

$$V_{0t}^x \equiv \frac{1}{n} \sum_{j=1}^n \sum_{i=1}^{4M} \mathbb{E}_t \left[e^{-\int_t^{t_i} r(\tau) d\tau} \cdot \mathbb{I}_{\{\text{Def}_j \in (t_{i-1}, t_i)\}} \right],$$

and

$$V_{1t}^x \equiv \frac{1}{n} \sum_{j=1}^n \sum_{i=1}^{4M} \mathbb{E}_t \left[e^{-\int_t^{t_i} r(\tau) d\tau} \left(1 - \sum_{h=0}^i \mathbb{I}_{\{\text{Def}_j \in (t_{h-1}, t_h)\}} \right) \right] = \frac{1}{n} \sum_{j=1}^n \sum_{i=1}^{4M} \mathbb{E}_t \left[e^{-\int_t^{t_i} r(\tau) d\tau} \mathbb{I}_{\{\text{Surv}_j \text{ at } t_i\}} \right], \quad (13.50)$$

⁴According to standard market practice, the loss in the index would actually occur as soon obligors default—without any need to wait until the end of any of the time intervals t_i . However, we cast the discussion in terms of a different timing convention, as this makes the nature of the “swap transaction” in Eq. (13.48) quite transparent.

where $\mathbb{I}_{\{\text{Surv}_j \text{ at } t_i\}}$ is the indicator of the event that the j -th name has survived by time t_i . Because the names have the same credit quality, the value of the previous baskets of assets reduce to, $V_{0t}^x = V_{0t}$ and $V_{1t}^x = V_{1t}$, where V_{0t} and V_{1t} are the values of securities indexed on default events of an hypothetical representative firm, as in Eq. (13.42).

A CDS index at the time of origination $t \equiv t_0$ is, then, simply, the value of $\overline{\text{CDX}}_{t_0}(M)$ in Eq. (13.49), which makes $\text{DSX}_{t_0} = 0$, viz

$$\overline{\text{CDX}}_{t_0}(M) = \text{LGD} \cdot \frac{V_{0,t_0}}{V_{1,t_0}}.$$

Next, consider a *forward starting credit default index*, which is an index starting at time t_0 , as before, but decided at some point prior to t_0 , say at time t . Clearly, the value of the protection leg minus that of the premium leg over the life of the index is the same as that in Eq. (13.49), for a generic $t \leq t_0$. Moreover, in Appendix 6, we show that for any time $\tau \in (t, t_0)$,

$$\text{DSX}_\tau = \sum_{i=1}^{4M} \mathbb{E}_\tau \left[e^{-\int_\tau^{t_i} r(\tau) d\tau} \text{cdx}_t(t_i) \right] = N(\tau) \cdot (\text{LGD} \cdot V_{0\tau} - \overline{\text{CDX}}_t(M) \cdot V_{1\tau}), \quad (13.51)$$

where $N(\tau)$ denotes the outstanding notional,

$$N(\tau) = \frac{1}{n} \sum_{j=1}^n \mathbb{I}_{\{\text{Surv}_j \text{ at } \tau\}}, \quad N(t) \equiv 1. \quad (13.52)$$

Finally, an *index default swaption payer* with strike K , gives the holder the option to enter a CDS index as a protection buyer with an index strike spread equal to K . Upon exercise, the protection buyer would also receive a *front-end protection*, defined as the losses occurring from the option origination to the exercise date. Let t be the option origination and $T = t_0$ the maturity of the swaption. The front end protection is, $F_T = \text{LGD} \frac{1}{n} \mathcal{D}(t, T)$, where $\mathcal{D}(t, T)$ is the number of defaults occurred over the time interval (t, T) . In Appendix 6, we show that the value of the front-end protection is,

$$V_\tau^F = \mathbb{E}_\tau \left[e^{-\int_\tau^T r(u) du} F_T \right] = \text{LGD} \left(\frac{1}{n} \mathcal{D}(t, \tau) P(\tau, T) + N(\tau) (P(\tau, T) - P_{\text{def}}(\tau, T)) \right), \quad (13.53)$$

where $P(t, T)$ and $P_{\text{def}}(t, T)$ denote the price of a non-defaultable and a defaultable zero expiring at time T , with zero recovery value and default intensity equal to that of the representative firm, λ . The underlying of a default swaption payer equals $\text{DSX}_T + F_T$. Accordingly, we can define the *loss-adjusted forward default swap index*, as $\text{DSX}_\tau^L \equiv \text{DSX}_\tau + V_\tau^F$, and find the value of $\overline{\text{CDS}}_\tau(M)$ such that $\text{DSX}_\tau^L = 0$, denoted as $\text{CDX}_\tau(M)$, which is,

$$\text{CDX}_\tau(M) = \text{LGD} \frac{V_{0\tau}}{V_{1\tau}} + \frac{V_\tau^F}{N(\tau) V_{1\tau}}, \quad (13.54)$$

such that,

$$\text{DSX}_\tau^L = N(\tau) V_{1\tau} (\text{CDX}_\tau(M) - \overline{\text{CDX}}_t(M)).$$

We wish to use $N(\tau) V_{1\tau}$ as a numéraire such that $\text{CDX}_t(M)$ is a martingale under a suitable probability,⁵ similarly as for the probability Q_{sc} in Eq. (13.46). Define the probability \hat{Q}_{sc}

⁵A technical issue with the definition of $\text{CDX}_t(N)$ in Eq. (13.54) relates to a “denominator problem”—the possibility of a total collapse of the index, $N(\tau) = 0$. The occurrence of such an event has been taken into account by Rutkowski and Armstrong (2009) and Morini and Brigo (2011).

through the Radon-Nikodym derivative:

$$\left. \frac{d\hat{Q}_{sc}}{dQ} \right|_{\mathbb{F}_T} = e^{-\int_{\tau}^T r(u)du} \frac{N(T) V_{1T}}{N(\tau) V_{1\tau}}. \quad (13.55)$$

The probability \hat{Q}_{sc} is the index counterpart to Q_{sc} in Eq. (13.46). For simplicity, we shall keep on referring to \hat{Q}_{sc} as the “survival contingent measure.” Appendix 6 contains a proof that \hat{Q}_{sc} does indeed integrate to one. It also shows that $CDX_t(M)$ in Eq. (13.54) is a martingale under \hat{Q}_{sc} . Therefore, the price of a swaption payer with strike K is, for any $\tau \in [t, T]$,

$$SW_{\tau}^p(K, T) \equiv \mathbb{E}_{\tau} \left[e^{-\int_{\tau}^T r(u)du} N(T) V_{1T} (CDX_T(M) - K)^+ \right] = N(\tau) V_{1\tau} \cdot \hat{\mathbb{E}}_{\tau}^{sc} [(CDX_T(M) - K)^+],$$

where $\hat{\mathbb{E}}_{\tau}^{sc}[\cdot]$ denotes the time τ conditional expectation under the survival contingent measure \hat{Q}_{sc} in Eq. (13.55). We know $CDX_{\tau}(M)$ is a martingale under \hat{Q}_{sc} . We can use Black (1976) to evaluate the previous expression, once we assume that under \hat{Q}_{sc} , $CDX_{\tau}(M)$ is a geometric Brownian motion with constant volatility.

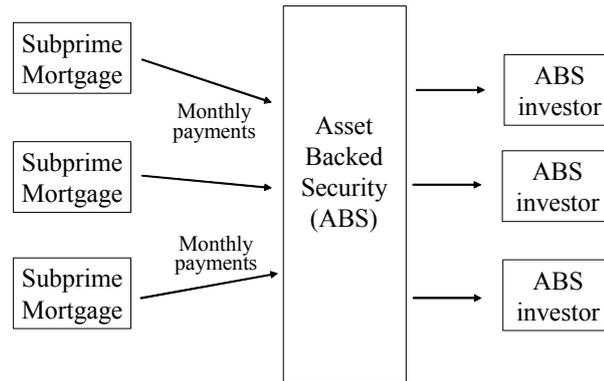
13.4.4 Collateralized Debt Obligations (CDOs)

13.4.4.1 A crash description of securitization

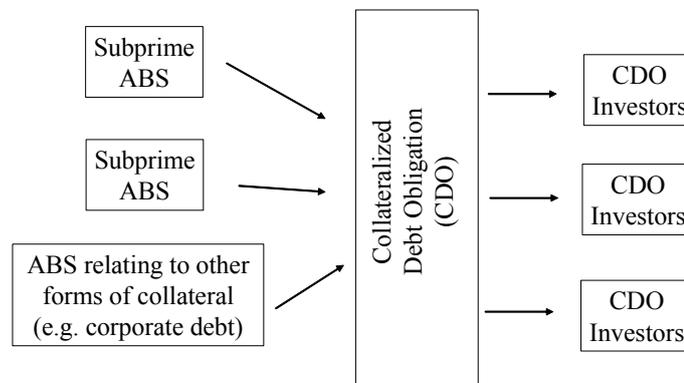
Historically, an important input to the process of securitization came from the financial innovation put forward by the US government during the 1980s. Up to the 1970s, the financial system used to live in a “buy to hold” world, whereby banks making loans to businesses or individuals would typically hold the loans in their portfolios. During the 1970s, another trend began, where the Government National Mortgage Association (“GNMA” or “Ginnie Mae”) would buy the mortgages from banks so as to incentivize these banks to extend more loans, thereby making houses accessible to families. The second step would then be for GNMA to sell securities based on the cash flows generated by these mortgages. Securitization would then begin to take on a higher level when the Federal National Mortgage Association (“Fannie Mae”) and the Federal Home Loan Mortgage Corporation (“Freddie Mac”) would securitize the assets through “tranching.” Once the tranching model was initially developed, investment banks applied this same idea to other kinds of assets, such as corporate bonds, student loans, small business loans, automobile loans, etc.

How does tranching work? What is a CDO? CDOs are securitized shares in pools of assets. Collateral assets include loans or debt instruments. A CDO may be a collateralized loan obligation (CLO) or collateralized bond obligation (CBO) according to whether it relies only on loans or bonds, respectively. CDO investors bear the credit risk of the collateral. Multiple tranches of securities are issued by the CDO, offering investors various maturity and credit risk characteristics. Tranches are categorized as senior, mezzanine, and subordinated, or junior, or equity, according to their degree of credit risk. If there are defaults or the CDO’s collateral otherwise underperforms, scheduled payments to senior tranches take precedence over those of mezzanine tranches, and scheduled payments to mezzanine tranches take precedence over those to junior tranches. Typically, senior tranches are rated, with ratings of A to AAA. Mezzanine are also rated, typically with ratings of B to BBB. In principle, these ratings should reflect both the credit quality of the collateral and the protection a given tranche is given by the tranches subordinating to it. CDOs are part of a more general securitization process, and can also include mortgages, as in the stylized example below.

- (i) In a first step, subprime mortgages are securitized, as illustrated below:



- (ii) In a second step, a CDO is created, out of the securitized subprime mortgages and additional Asset Backed Securities (ABS):



- (iii) In a third, and final step, the structuring process involves creating seniority rules.

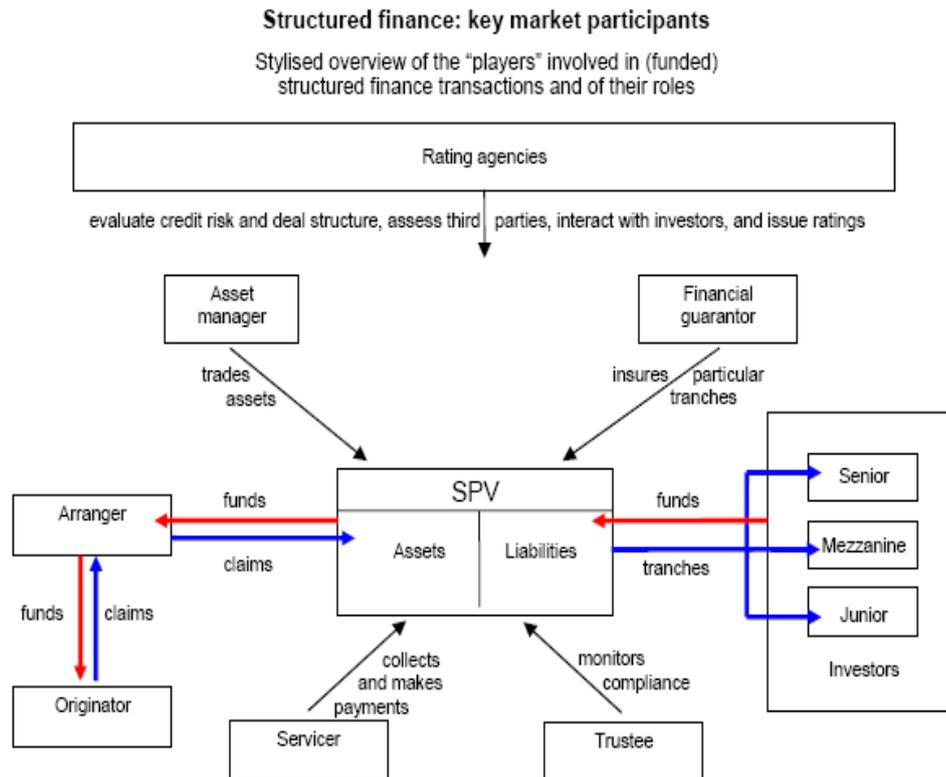
Investors in CDOs *senior tranches* include banks and pension funds, which might benefit from the expertise of the asset managers, and the risk-return profiles difficult to find in the market. Investors in *junior tranches* are hedge funds searching for highly risky investment opportunities that at the same time, are quite rewarding and certainly unavailable in the market. Additional investors in junior tranches were dedicated off-balance-sheets entities such as “SIV,” “conduits,” and “SIV-lites,” which will be reviewed in Section 13.4.7.

Underwriters of CDOs are investment banks, typically. They work closely with the asset manager and create the “right” debt/equity ratio and perform collateral quality tests. They liaise with law firms and create the special purpose vehicle (possibly in some tax heaven system) that will purchase the assets and issue the tranches, price the various tranches, and obviously find the investors. Fees to underwriters are very generous, due to the complexity of the CDOs. According to Thomson Financial, top underwriters in 2006 were: Bear Sterns, Merrill Lynch, Wachovia, Citigroup, Deutsche Bank, and Bank of America Securities.

Involved in the structuring process are also (i) trustee and collateral administrator, who distribute noteholder reports, check compliance and execute priority of payments; (ii) accountants, who perform due diligence on the CDOs collateral pool, verifying for example credit ratings for each asset; and (iii) rating agencies, which we shall discuss in the next subsection.

The economics behind structured finance is quite interesting. An originator may have private information about the quality of certain assets and/or a comparative advantage in evaluating

these assets relative to other market participants. If the originator wishes to sell some of its assets, an adverse selection problem will arise: because investors do not know the true quality of the assets, they will demand a premium to purchase them or even worse, a market might fail to arise. Structured finance helps originators mitigate this problem. First, by pooling the assets, diversification benefits can be achieved. Second, tranching allows relatively poorly informed investors to access senior tranches, and be relatively protected from default. In the process, the originator or arranger may retain subordinated exposure to alleviate investors' concerns about incentive compatibility. The following scheme summarizes the structuring process.



Source: Committee on the Global Financial System: "The role of ratings in structured finance: issues and implications," January 2005.

13.4.4.2 The role of rating agencies

Structured finance has always been a "rated" market. Issuers of structured instruments had a natural appetite for a rating to occur at a scale comparable to that available for bonds. The main reason was this would facilitate the sale of these products to investors bound by ratings-based constraints defined by their investment mandates.

However, the involvement of rating agencies into the delivery of their opinion about credit risk differs from that related to traditional bonds. As regards traditional instruments, rating agencies simply aim to assess the risk of default as given, which they take as given. As regards structured finance transactions, rating agencies play a much more ex-ante, reverse engineering role. A tranche rating reflects a view about both the credit risk of the asset pool and the extent of credit support to be provided. These two elements are organized to reverse engineer the tranche rating targeted by the deal's arrangers. Deal origination thus involves rating agencies in the structuring process.

13.4.4.3 Types of CDOs

In practice, CDOs are considerably more complex than the stylized examples outlined earlier. We have a number of cases. We say that a CDO is *static*, if it holds the same set of assets. Instead, a CDO is *managed*, if the asset manager is allowed to change the composition of assets. If the claims to the CDO arise from the cash flows originated by the assets, we have a *cash-flow CDO*. If the claims to the CDO arise from the cash flows originated by the assets and/or active asset management, we have a *market-value CDO*. CDOs can also be created to carve out balance sheets, in which case we have *balance-sheet CDOs*. Moreover, and interestingly, CDOs can be created (i) to achieve investment grade bonds through a pool of noninvestment grade bonds, and (ii) to create riskier securities than those in the asset pool. In these cases, we have *arbitrage CDOs*. Naturally, “arbitrage” CDOs do not give rise to any arbitrage opportunity. These instruments merely “reshuffle” risk and returns of the assets in the pool, as we shall see in the next section. Arbitrage CDOs differ from balance sheet CDOs, because issuers of arbitrage CDOs do not necessarily hold the underlying collateral in advance, which is obviously the case for issuers of balance-sheet CDOs. Therefore, the assets to be put into the an arbitrage CDO pool have to be reasonably liquid.

Furthermore, we have *synthetic CDOs*, which are exposed to a pool of assets that are not strictly owned or in the asset pool, typically through CDS underwriting. Like a cash-flow CDO, the vehicle receives payments (the premium), which is then transferred to the tranche holders. Naturally, there can be default events, which are also passed through to the investors, according to the prespecified seniority rules. A synthetic CDO is *funded*, if the relevant tranche holders are to pay for in the case of a credit event related to the assets the CDO is exposed to. Typically, some funding is made available at the very time of investment. At maturity, the investor receives a payoff equal to the funding minus the realized losses. Junior tranches are typically funded, and senior are typically not. However, senior tranches investors might have to make payments in the unlikely event losses had ever to erode their tranches.

Finally, we have *hybrid CDOs*, which are partly cash-flow CDOs and partly synthetic CDOs. In a *single-tranche CDO*, the entire CDO is structured to accommodate the specific needs of a small group of investors, with some remaining tranche held by the dealer. And we have CDO^2 , where a large portion of the assets in the pool are tranches from other CDOs; or more generally, CDO^n .

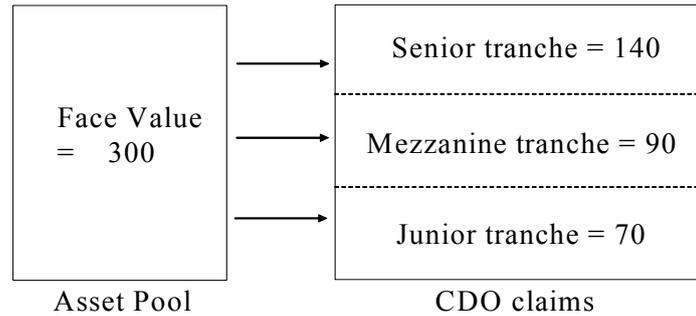
13.4.4.4 Pricing

CDOs repackage cash flows from a set of assets. We provide simple examples to show how to price this repackaging process. We begin with a simple example, taken from McDonald (2006, p. 583), which we further elaborate. Suppose we have three one-year bonds with face value = 100. For each of these bonds, the *risk-neutral* probabilities of default equal 10% and the recovery rates are 40. The safe interest rate for one year is 6%. So each bond price equals,

$$b = e^{-0.06} \cdot \left(\underbrace{0.10}_{\equiv \text{Def. Prob}} \cdot 40 + \underbrace{0.90}_{\equiv \text{Surv. Prob}} \cdot 100 \right) = 88.526.$$

The yield is, naturally, $-\ln \frac{b}{100} = 12.19\%$.

A CDO can restructure the payments promised by the three bonds in a way that transforms the riskiness and attractiveness of the initial assets. Consider the following example:



In this example, each tranche receives the minimum between (i) the nominal value claimed by the tranche and (ii) what is left available to the tranche after having satisfied the other tranches by order of seniority.

Let N_i be the nominal values claimed by the tranches, so that $N_1 = 140$, $N_2 = 90$ and $N_3 = 70$. Let $\tilde{\pi}$ be the realized payoff of the asset pool, defined as,

$$\tilde{\pi} = \text{No. of Defaults} \cdot 40 + \underbrace{(3 - \text{No. of Defaults}) \cdot 100}_{\equiv \text{No. of surviving bonds}}.$$

Naturally, $\tilde{\pi}$ is random because the number of defaults is random. At the expiration,

- (i) the senior tranches receives the minimum between N_1 and $\tilde{\pi}$. For example, if only one bond defaults, $\tilde{\pi} = 240$, and the senior tranche receives 140. If, however, three bonds default, $\tilde{\pi} = 120$, which is less than the senior tranche nominal value, and the senior tranche then receives 120. So a quite severe loss is needed to erode the senior tranche claims.
- (ii) The mezzanine tranche receives the minimum between N_2 and the “left-over” from the senior tranche.
- (iii) Finally, at the expiration, the junior tranche receives the minimum between N_3 and the “left-over” from the senior and mezzanine tranches.

More generally, tranche no. i receives,

$$\pi_i = \min \{ \text{Left-over from previous tranches up to tranche } i - 1, N_i \},$$

where

$$\text{Left-over from previous tranches up to tranche } i - 1 = \max \left\{ \tilde{\pi} - \sum_{k=1}^{i-1} \pi_k, 0 \right\}.$$

Synthetically,

$$\pi_i = \min \left\{ \max \left\{ \tilde{\pi} - \sum_{k=1}^{i-1} \pi_k, 0 \right\}, N_i \right\}.$$

All we need, now, is to model the risk-neutral probability of default for each firm. Initially, we assume the default events are independent across firms. Assume binomial distribution,

$$\Pr(\text{No. of Defaults} = k) = \binom{3}{k} p^k (1 - p)^{n-k}, \quad p = 10\%, \quad k \in \{1, 2, 3\}.$$

We can then derive the following payoff structure

Payoffs to CDO tranches, and prices: with independent defaults						
Defaults	Pr(Defaults)	π : pool payoff ⁽¹⁾	π_1 : Senior	π_2 : Mezzanine	π_3 : Junior	
0	0.729	300	140	90	70	
1	0.243	240	140	90	10	
2	0.027	180	140	40	0	
3	0.001	120	120	0	0	
			Price	131.8281994	83.40266709	50.34673197
			Yield	0.060142867	0.076129382	0.329561531

⁽¹⁾ π : pool payoff = Def*40+(3-Def)*100
 $N_1 = 140$
 $N_2 = 90$
 $N_3 = 70$

The price of each tranche is computed as the tranche payoff, averaged across states, discounted at the safe interest rate. For example, the price of the mezzanine tranche is,

$$\text{Price Mezzanine} = e^{-0.06} (0.729 * 90 + 0.243 * 90 + 0.027 * 40 + 0.001 * 0) = 83.403.$$

Its yield is, $\text{Yield Mezzanine} = -\ln \frac{83.403}{90} = 7.61\%$. Naturally, the sum of the three bond prices, $88.526 \times 3 = 265.58$, is equal to the total value of the three tranches, $131.828 + 83.403 + 50.347 = 265.58$. As anticipated, a CDO is a mere re-packaging device. It doesn't add or destroy value. It merely redistributes risks (and returns).

The assumption defaults among names are uncorrelated is unrealistic, as argued in Section 13.5.4. We now remove this assumption. First, what happens in the special case where default events are *perfectly correlated*? In this case, either the three firms all default (with probability 0.10) or none defaults (with probability 0.90), and we have the situation summarized by the table below.

Payoffs to CDO tranches, and prices: with perfectly correlated defaults						
Defaults	Pr(Defaults)	π : pool payoff ⁽¹⁾	π_1 : Senior	π_2 : Mezzanine	π_3 : Junior	
0	0.9	300	140	90	70	
1	0	NA	NA	NA	NA	
2	0	NA	NA	NA	NA	
3	0.1	120	120	0	0	
			Price	129.9635056	76.28292722	59.33116562
			Yield	0.074388737	0.165360516	0.165360516

⁽¹⁾ π : pool payoff = Def*40+(3-Def)*100
 $N_1 = 140$
 $N_2 = 90$
 $N_3 = 70$

Note that mezzanine and junior tranches now yield the same, because they each pay off either their nominal value or zero in exactly the same states of nature. In other words, default clustering implies that good times are really good, in that the probability to have no defaults is now 90%, much higher than the 72.9% arising when the correlation of defaults is zero.

The previous cases (with independent or perfectly correlated defaults) are extreme. What happens when defaults are only imperfectly correlated? In this case, the pricing of tranches is more complex, and requires a model of default correlations. We use the so-called Gaussian copulae, reviewed in Appendixes 7 and 8, and simulations. Figure 13.14 illustrates how the

yield on each tranche changes as a result of a change in the default correlation underlying the assets in the CDO.

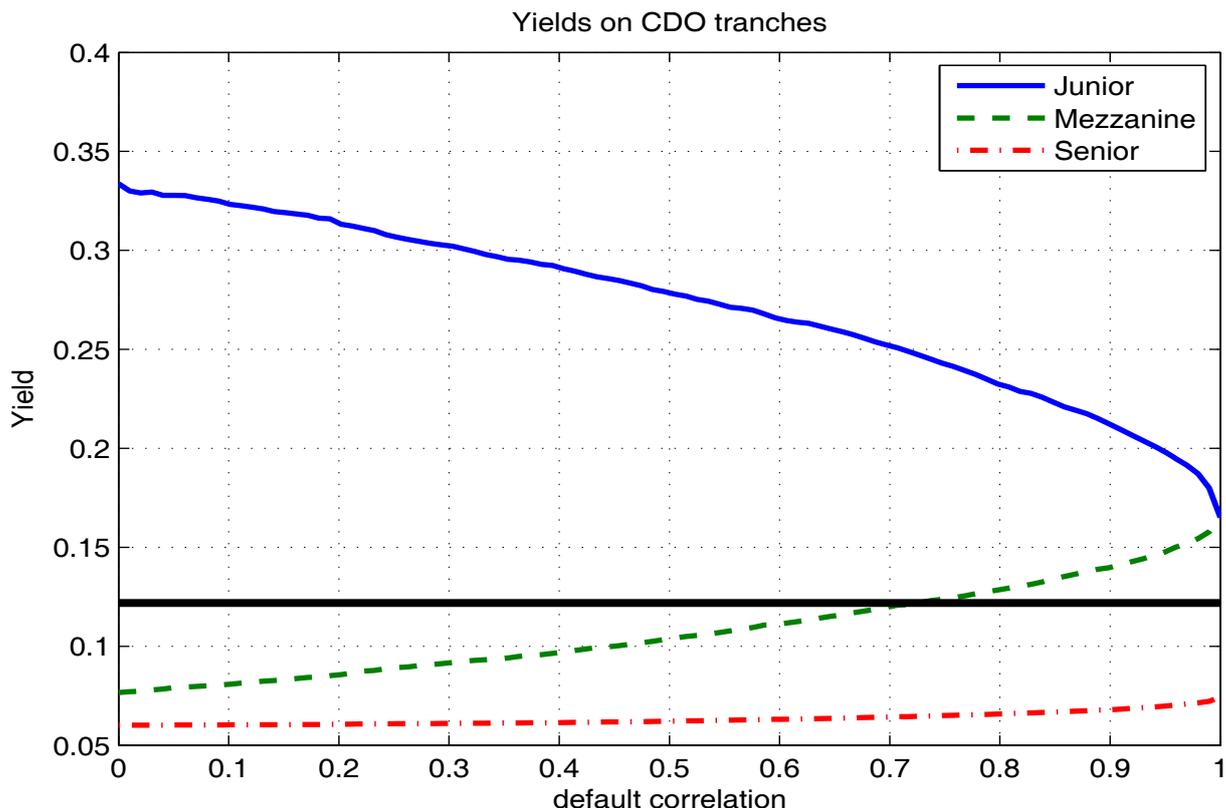


FIGURE 13.15. Yields on the three CDO tranches, as functions of the default correlation among the assets in the structure, with probability of default for each name $p = 20\%$. The thick, horizontal, line is the yield on each securitized asset.

“Arbitrage” CDOs

Figure 13.15 illustrates how arbitrage CDOs work. The CDO has three assets yielding the same, 12.19% (the horizontal line in the picture). However, by restructuring the asset base through a CDO, we can create claims (Senior and Mezzanine tranches) that yield less than 12.19%, as they are considerably less risky than the asset base. Such an excess return, $(12.19\% - \text{Yield}_{\text{tranche}})$, with $\text{Yield}_{\text{tranche}} \in \{\text{Senior}, \text{Mezzanine}\}$, is “made available” to the Junior tranche/equity holders—once management fees and expenses are accounted for. Note that such a redistribution of risk works quite effectively as soon as the default correlation is relatively low. As the default correlation in the asset base increases, the situation may change dramatically, with the mezzanine tranche becoming more risky and, then, yielding a higher expected return. Finally, Figure 13.16 depicts the output of a comparative statics where we increase p from 10% to 20%. The yields are obviously higher for each tranche, and the three assets now yield 18.78%, reflecting the higher marginal probability of default for each of the securities in the pool, p .

Correlation assumptions

In Figures 13.15 and 13.16, the yield on the junior tranche decreases with default correlation. This happens because we are assuming that the probability of default is fixed at $p = 10\%$ for each default correlation ρ (say). As ρ increases, the probability of clustering events increases,

which makes the Senior and Mezzanine tranches relatively less valuable and, correspondingly, the Junior tranches more valuable. A more appropriate model is one in which p increases as ρ increases, to capture the fact that in bad times, both default correlation and probability of defaults increase as these two things are intimately connected—by, e.g., some common business cycle factors.

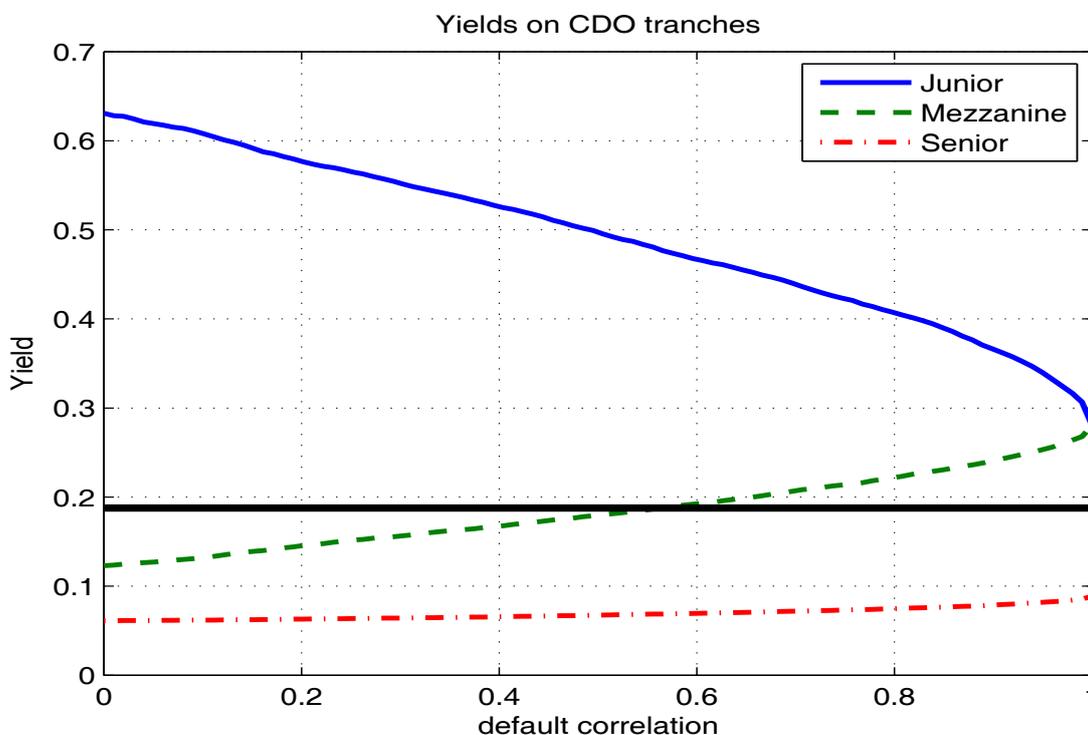


FIGURE 13.16. Yields on the three CDO tranches, as functions of the default correlation among the assets in the structure, with probability of default for each name $p = 20\%$. The thick, horizontal, line is the yield on each securitized asset.

Addressing the correlation assumption

We relax the assumption that the probability of default, p , and the default correlation, ρ are independent. We assume that ρ and p are tied up through the following relation, $\rho = 3.8116 * \ln(p + 1)$, and let p vary from 0.10 to 0.30, such that ρ varies from 0.3633 to 1. The situation now changes, dramatically. Figure 13.17 depicts the results, which show how modeling might substantially affect effective pricing. First, and naturally, the yield on each securitized asset is increasing in ρ because ρ is also increasing in the probability of default. Second, the Junior tranche has a yield that increases over a wide spectrum of values for the default correlation, ρ . Note that the Junior tranche bends back to lower values as the default correlation is close to one, reflecting the fact that default clustering makes this tranche quite valuable in good times, as explained earlier.

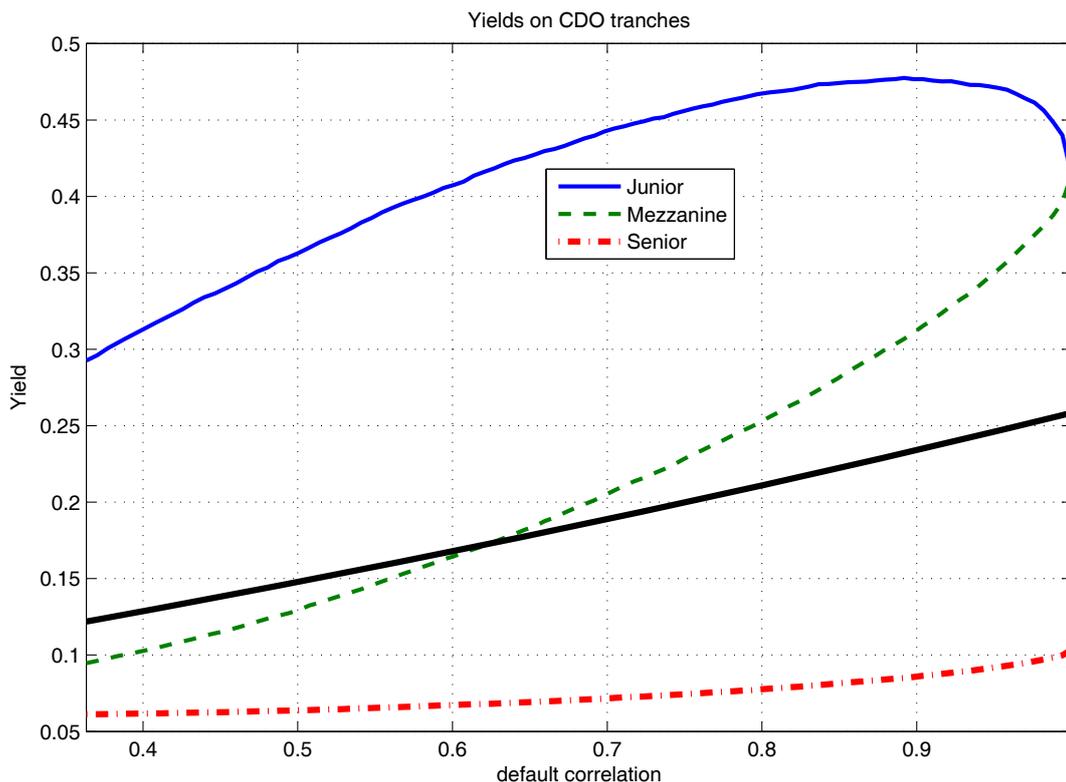


FIGURE 13.17. Yields on the three CDO tranches, as functions of the default correlation among the assets in the structure, with probability of default and default correlation related by $\rho = 3.8116 * \ln(p + 1)$, $p \in [0.10, 0.30]$. The thick curve line depicts the yield on each securitized asset.

13.4.4.5 Nth to default

In this contract, the owner of the 1st to default bears the risk of the first default that occurs in the asset pool:

$$\text{Payoff} = \Pr(\text{No. of Defaults} \geq 1) * 40 + \Pr(\text{No. of Defaults} < 1) * 100.$$

Likewise, the owner of the 2nd to default bears the risk of the second default that occurs in the asset pool:

$$\text{Payoff} = \Pr(\text{No. of Defaults} \geq 2) * 40 + \Pr(\text{No. of Defaults} < 2) * 100.$$

Finally, the owner of the 3rd to default bears the risk of the third default that occurs in the asset pool:

$$\text{Payoff} = \Pr(\text{No. of Defaults} = 3) * 40 + \Pr(\text{No. of Defaults} < 3) * 100.$$

Let us assume that default correlation is zero for simplicity. We have previously computed the previous probabilities as:

$$\begin{aligned} \Pr(\text{No. of Defaults} \geq 1) &= 0.243 + 0.027 + 0.001 = 0.271 \\ \Pr(\text{No. of Defaults} \geq 2) &= 0.027 + 0.001 = 0.028 \\ \Pr(\text{No. of Defaults} = 3) &= 0.001 \end{aligned}$$

Thus, we have the following prices,

$$\begin{aligned}\text{Price}_{1^{st}\text{-to-default}} &= e^{-0.06} * [0.271 * 40 + (1 - 0.271) * 100] = 78.863 \\ \text{Price}_{2^{nd}\text{-to-default}} &= e^{-0.06} * [0.028 * 40 + (1 - 0.028) * 100] = 92.594 \\ \text{Price}_{3^{rd}\text{-to-default}} &= e^{-0.06} * [0.001 * 40 + (1 - 0.001) * 100] = 94.120\end{aligned}$$

From here, we can compute the yields as follows, $\text{Yield}_{1^{st}\text{-to-def}} = -\ln(78.863/100) = 23.74\%$, $\text{Yield}_{2^{nd}\text{-to-def}} = -\ln(92.594/100) = 7.69\%$, and $\text{Yield}_{3^{rd}\text{-to-def}} = -\ln(94.120/100) = 6.06\%$.

13.4.4.6 One numerical example of a stylized structured product

A. Defaultable bonds

Suppose we observe the following risk-structure of spreads, related to two bonds maturing in two years:

$$\text{Spread}_A(2 \text{ years}) = 1.5\%, \quad \text{Spread}_B(2 \text{ years}) = 2.5\%,$$

where A and B denote the rating classes the bond issuers belong to. Assume that the one-year transition rating matrix, defined under the risk-neutral probability, is:

		To		
		A	B	Def
From	A	0.7	0.3	0
	B	0.3	0.5	0.2
	Def	0	0	1

where “Def” denotes default. We assume that in the event of default, the recovery value of the bond is paid off at the end of the second period. We want to determine the expected recovery rates for the two bonds, and which expected recovery rate is the largest. We have:

$$e^{rT} \frac{D_{0,i}}{N} = \left[\frac{\text{Rec}_i}{N} \mathbb{Q}_i(2) + (1 - \mathbb{Q}_i(2)) \right], \quad i \in \{A, B\}.$$

Therefore,

$$\text{Spread}_A(2 \text{ years}) = 1.5\% = -\frac{1}{2} \ln \left[\frac{\text{Rec}_A}{N} \mathbb{Q}_A(2) + (1 - \mathbb{Q}_A(2)) \right] \quad (13.56)$$

$$\text{Spread}_B(2 \text{ years}) = 2.5\% = -\frac{1}{2} \ln \left[\frac{\text{Rec}_B}{N} \mathbb{Q}_B(2) + (1 - \mathbb{Q}_B(2)) \right] \quad (13.57)$$

We have to find $\mathbb{Q}_A(2)$ and $\mathbb{Q}_B(2)$. The transition matrix for two years is,

$$\mathbb{Q}(2) = \begin{bmatrix} 0.7 & 0.3 & 0 \\ 0.3 & 0.5 & 0.2 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0.7 & 0.3 & 0 \\ 0.3 & 0.5 & 0.2 \\ 0 & 0 & 1 \end{bmatrix},$$

such that,

$$\begin{aligned}\text{Pr}\{\text{A defaults in 2 years}\} &= \mathbb{Q}_A(2) \\ &= \underbrace{0.70 * 0}_{A \rightarrow A \rightarrow Def} + \underbrace{0.30 * 0.20}_{A \rightarrow B \rightarrow Def} + \underbrace{0 * 1}_{A \rightarrow Def \rightarrow Def} \\ &= 0.06\end{aligned}$$

$$\begin{aligned}
\Pr \{B \text{ defaults in 2 years}\} &= \mathbb{Q}_B(2) \\
&= \underbrace{0.20 * 1}_{B \rightarrow Def \rightarrow Def} + \underbrace{0.50 * 0.20}_{B \rightarrow B \rightarrow Def} + \underbrace{0.30 * 0}_{B \rightarrow A \rightarrow Def} \\
&= 0.20 + 0.10 = 0.30.
\end{aligned}$$

Hence, using Eqs. (13.56)-(13.57), we have

$$\begin{aligned}
\text{Spread}_A(2 \text{ years}) &= 1.5\% = -\frac{1}{2} \ln \left[\frac{\text{Rec}_A}{N} 0.06 + (1 - 0.06) \right] \\
\text{Spread}_B(2 \text{ years}) &= 2.5\% = -\frac{1}{2} \ln \left[\frac{\text{Rec}_B}{N} 0.30 + (1 - 0.30) \right]
\end{aligned}$$

Solving, yields,

$$\frac{\text{Rec}_A}{N} = 50.7\%, \quad \frac{\text{Rec}_B}{N} = 83.7\%.$$

The expected recovery rate for the second bond is the largest. This is because the probability firm B defaults is much larger than the probability firm A defaults and yet the two spreads are relatively close to each other. So to rationalize the two spreads, we need a large recovery rate for the second bond.

What would happen to the two credit spreads, once we assume that the recovery rates are the same, and equal to 50%? This question sheds additional light to the previous findings. If the recovery rates are the same and both equal 50%,

$$\begin{aligned}
\text{Spread}_A(2 \text{ years}) &= -\frac{1}{2} \ln [0.50\mathbb{Q}_A(2) + (1 - \mathbb{Q}_A(2))] \\
\text{Spread}_B(2 \text{ years}) &= -\frac{1}{2} \ln [0.50\mathbb{Q}_B(2) + (1 - \mathbb{Q}_B(2))]
\end{aligned}$$

Then, using the previously computed transition probabilities for two years, we obtain:

$$\text{Spread}_A(2 \text{ years}) = 1.52\%, \quad \text{Spread}_B(2 \text{ years}) = 8.12\%.$$

When the recovery rates are the same, the spread on the second bond diverges substantially from that on the first bond.

B. Collateralized debt obligations

Let us keep on using the same framework as before, but use different figures, so as to figure out the implications for CDOs pricing. Consider the following one year transition matrix, under the risk-neutral probability:

		To		
		A	B	Def
From	A	0.7	0.3	0
	B	0.1	0.6	0.3
	Def	0	0	1

where “Def” denotes default. Consider (i) 1 one-year bond issued by a company rated A, and (ii) 3 one-year bonds issued by a company rated B. Both bonds have face value equal to 100.

We assume that the recovery values in case of default of all these bonds are the same, and equal to 50. Finally, we assume the safe interest rate is taken to be equal to zero.

Consider a collateralized debt obligation (CDO, in the sequel), which gathers the previous four bonds. Therefore, the CDO has nominal value of 400, and pays off in one year. The CDO has (i) a senior tranche, with nominal value equal to 150; (ii) a mezzanine tranche, with nominal value equal to N_1 ; and (iii) a junior tranche, with nominal value equal to N_2 . We assume that the structure is such that $N_1 > 100$.

First, we determine the price and yields on all the four bonds. Since the safe interest rate is zero, and the company rated A is safe, up to the next year, the price of the A bond is 100, and its yield is zero. As for the three bonds rated B, we have:

$$P_B = 50 * 0.3 + 100 * 0.7 = 85.0, \quad Yield_B = -\ln 0.85 = 16.25\%.$$

Second, we determine the yield on the junior tranche, and derive the yield on the mezzanine, as a function of its nominal value N_1 . To determine the yield on the tranches, we need to figure out the following table:

No_Def	Pr	Π	π_0	π_1	π_2
0	0.7	400	150	N_1	N_2
1	0	NA	NA	NA	NA
2	0	NA	NA	NA	NA
3	0.3	250	150	100	0
4	4	NA	NA	NA	NA

where No_Def denotes the number of defaults, Pr is the probability of No_Def, Π is the pool payoff, defined as,

$$\Pi = \text{No_Def} * 50 + (4 - \text{No_Def}) * 100,$$

and, finally: π_0 is the payoff to the senior tranche, π_1 is the payoff to the mezzanine tranche, and, π_2 is the payoff to the junior tranche. Therefore, we have:

$$\text{price_mezzanine} = 0.70 * N_1 + 0.30 * 100, \quad \text{price_junior} = 0.70 * N_2,$$

such that:

$$\begin{aligned} \text{Yield_mezzanine} &= -\ln\left(\frac{0.70 * N_1 + 0.30 * 100}{N_1}\right) = -\ln\left(0.70 + 0.30 * \frac{100}{N_1}\right) \\ \text{Yield_junior} &= -\ln\left(\frac{0.70 * N_2}{N_2}\right) = 35.67\%. \end{aligned}$$

Naturally, we need to have that $\text{Yield_mezzanine} < \text{Yield_junior}$. It is simple to show this relation: it suffices to note that,

$$\text{Yield_junior} = -\ln(0.70) > -\ln\left(0.70 + 0.30 * \frac{100}{N_1}\right) = \text{Yield_mezzanine}.$$

A reverse engineering question is, now, to determine which nominal value of the mezzanine tranche N_1 is needed, to ensure that the yield on the mezzanine tranche is equal to or greater than the yields on the bonds issued by the company with credit rating B? The answer is $N_1 = 200$, for in this case, the mezzanine tranche would have the same payoff structure as the bond rated B: it would deliver (i) the face value, in the event the company rated B does not default; and (ii) half of its nominal value, 100, in the event the company rated B does default.

Finally, we ask which nominal value of the mezzanine tranche N_1 is needed, to ensure that the yield on the mezzanine is equal to 18%? And what is the corresponding nominal value of the junior tranche, N_2 ? To address these issues, we first want that:

$$\text{Yield_mezzanine} = -\ln\left(\frac{0.70 * N_1 + 0.30 * 100}{N_1}\right) = 18\%.$$

Solving for N_1 yields, $N_1 = 221.78$. Therefore, $N_2 = 400 - \text{Nominal_value_senior} - N_1 = 400 - 150 - 221.78 = 28.22$.

13.5 Procyclicality, credit crunches and quantitative easing

How is it that a relatively small loss in the banking system triggered by credit derivatives, such as one trillion or so US dollars, can have the power to lead to a spectacular financial crisis such as that experienced over the years 2007-2008, and spillover effects to the real economy, with one of the deepest recessions after World War II? One explanation might relate to “procyclicality.” We define procyclicality as the situation where a trend in asset prices feeds an automatic mechanism, which, in turn, reinforces the asset price trend, thereby creating a feedback loop between the asset price trend and the mechanism. The equilibriums in these markets might substantially differ from those we would observe absent any automatic mechanisms and, possibly, we would observe no equilibriums at all in the first place. The mechanism is automatic in that, once implemented, is not under discretion of any decision maker. One example of procyclicality includes the developments leading to the Black Monday crash of October 19th, 1987, most likely linked to program trading, as discussed in Chapter 10. The flash crash of May 6th, 2010 is the “high-tech” counterpart to the 1987 crash, as also discussed in Chapter 10.

A further instance of procyclicality, which is more closely connected to the topics studied in this chapter, relates to the amplification of business cycles determined by capital market frictions: in bad times, agency problems might entail an increase in the cost of external funds, thereby leading to a decrease in the availability of these funds and, then, to an amplification of shocks occurring in the real sectors of the economy. A leading example of this amplification mechanism is the financial accelerator hypothesis. Due to asymmetric information, financial intermediaries agree on lending plans based on the collateral made available by borrowers. The financial accelerator hypothesis holds that in bad times, financial intermediaries reduce their funding activities as collateral values are also reduced in bad times. The ensuing lending shrinkage contributes to further depress real economic activity and, hence, assets and collateral values, feeding a vicious feedback loop. Bernanke, Gertler and Gilchrist (1999) present a unified view of how agency problems make funding opportunities depend on firms’ collateral.⁶

Fisher (1933) is one of the earliest proponent of these procyclicality issues, in his attempt to explain the origins of the Great Depression through a debt-deflation spiral. In an economy with highly levered firms, such as that of the US during the 1930s, a negative productivity shock leads to bankruptcy of a fraction of these firms, which generates less investments and, hence, depresses aggregate demand and creates deflation. In turn, deflation boosts the real value of debt borne by firms, increasing the firms burden and leading a higher fraction of these firms to default. Such a debt-deflation spiral results in a deterioration of the balance sheet of

⁶Borio, Furfine and Lowe (2001) explain that the agents’ misperception of risk might constitute an additional amplification mechanism. For example, the credit/GDP ratio might be procyclical because financial intermediaries under-estimate risk in good times, and over-estimate risk in bad, thereby lending too much in good times and too less in bad.

financial intermediaries—banks obviously bleed money as their borrowers default—and to a default contagion, from firms to financial intermediaries. As a result, financial intermediation shrinks and the vicious feedback loop might go through a consistently long period.⁷

[Explain the connections with the “credit view” and the previous footnote on Friedman and Schwartz (1963)]

This section provides a basic discussion of these procyclicality problems, arising through the balance sheets of financial intermediaries. Section 13.5.1 is an overview of the extant regulatory framework, which is useful whilst framing procyclicality issues dealt with later in this section. Section 13.5.2 reviews a few institutional facts surrounding the 2007 subprime turmoil. Section 13.5.3 develops a few models where the amplification of small shocks occurs because financial intermediaries have concerns over the structure of their books. Thus, following a negative shock affecting the assets in the balance sheet, banks need to restore their Tier 1 and Tier 2 capital (in short, their “top tier capital”) and leverage ratios. Since they cannot raise fresh capital in the short-run, they cash-in by selling some of their assets. These sales create a vicious feedback loop where banks sell assets, contributing to a further drop in the value of these assets, triggering further sales into a depressed market. We may have situations where this loop leads to a complete market dry-up, which is even more likely to occur in the presence of capital market frictions, where some initially moderately low liquidity frictions can turn into spots of liquidity black holes. Even absent such extreme situations, the equilibriums in these markets can be those where an initial small loss in the banking system is amplified, to an extent determining a quite substantial shrinkage in lending activity, a credit crunch. Section 13.5.4 discusses the policy that monetary authorities have implemented in their attempt to mitigate the credit crunch originating from the 2007 subprime crisis. The standard policy action against a recession is to target low interest rates in the interbank markets for mandatory reserves. However, the cost of capital that matters to a recovery in the economic activity is that faced by firms whilst demanding new funds to banks (through loans) and/or the market (through issuance of corporate bonds). This cost can be substantially higher than the interest rates targeted by the monetary authority, due to the credit crunch. “Quantitative easing” is an unconventional policy action, where the monetary authority engages into the purchase of some of the assets held by banks (including the most illiquid ones), so as to give banks incentives to start lending again.

13.5.1 Regulatory framework

Banks have to set up capital buffers to guarantee the debt they issue against their risky activities of lending and investing. The Basel Committee on Banking Supervision (BCBS)⁸ drafts accords aiming to create an international standard for the capital necessary to cope with these risks, together with rigorous tools for risk measurement and management. Quite simply, the greater

⁷This view of the Great Depression was challenged by Friedman and Schwartz (1963), who proposed a “monetary view” instead. According to this view, the causes of the prolonged recession and the banking crises over the 1930s need to be linked to a non-accommodating monetary policy. Friedman and Schwartz examine the US economy from Civil War through 1960, and find a statistical relation between monetary policy and developments in the real macroeconomic aggregates: an expansionary monetary policy is associated with an expansion of the real economy. Friedman and Schwartz find that this linkage is particularly strong over the 1930s, and go further on, suggesting a causality from monetary policy to developments in the real economy. According to them, the only role banks might have played over the crisis was their contribution to the shrinkage in money supply through a lower money multiplier, defined in Section 13.6.3.

⁸The BCBS is a committee of banking supervisory authorities established by the central bank governors of the Group of Ten countries in 1975. It consists of senior representatives of bank supervisory authorities and central banks from Belgium, Canada, France, Germany, Italy, Japan, Luxembourg, the Netherlands, Spain, Sweden, Switzerland, the United Kingdom, and the United States. It usually meets at its permanent Secretariat, located at the Bank for International Settlements in Basel.

risk to which the bank is exposed, the greater the amount of capital the bank needs to hold to safeguard its solvency, in the interest of overall economic stability. The main issue, then, is to correctly measure this risk.

The first accord of 1988, known as Basel I, focussed on minimal capital requirements to cope with *credit risk*, and was enforced by law by the Group of Ten in 1992 and then by more than 100 countries. It relied on the so-called Cooke ratio (after Peter Cooke of the Bank of England), a minimum capital adequacy standard of 8% of the total risk-weighted assets. The accord was quite coarse, in that it considered five broad classes of credit risk with which to weigh the assets, which did not discriminate about the credit quality across classes. For example, corporate loans had 100% weightings and loans to OECD countries had zero weightings, independently of the ratings of the borrowing entities.

The first amendment to Basel I occurred in 1996, and aimed to include tools to cope with *market risk*. In 1999, a first consultative paper was drafted about a new accord, known as Basel II. One of the main issues under reform was the “one-size-fits-all” approach of Basel I—the fact that default risk could be substantially lower for some of the assets within the same class of risk in the banks accounts. For example, banks could have securitized the loans with default risk lower than that implied by the flat rate within the same class, and hold those loans with higher default risk. This might have led to an increase in the overall riskiness of financial institutions. In 1998, the Federal Reserve Chairman Alan Greenspan pointed to the possibility of regulatory arbitrage:

“Banks arbitrage away inappropriately high capital requirements on their safest assets by removing these assets from the balance sheet via securitization. The issue is not solely whether capital requirements on the bank’s residual risk in the securitized assets are appropriate. We should also be concerned with the sufficiency of regulatory capital requirements on the assets remaining on the book. In the extreme, such ‘cherry picking’ would leave on the balance sheet only those assets for which economic capital allocations are greater than the 8 percent regulatory standard.” [Greenspan, 1998 p. 166]

There is a consensus that Basel II eliminated these issues, by paying more attention to risk-sensitivity by means of a more precise set of indications about classes of risk and, also, distinguishing among credit risk, market risk and even operational risks.⁹ Moreover, the Basel II accords aimed to a flexible supervisory system whereby banks can adopt a “standardized” rating approach, or an “internal” rating approach, to measure and manage risks. The standardized approach is based on the ratings supplied by dedicated rating agencies approved by national supervisors. Capital incentives were also provided to move towards more advanced, and internal approaches, where banks could calculate weightings through their own models. However, moving out from mechanical rules to proprietary models certainly requires a particular care as to the quality of the models, which have to be approved by national supervisors.¹⁰ Therefore, Basel II encourages banks to improve their risk measurement and management systems.

Basel II relies on three “three pillars,” which impose increasingly stringent rules for capital adequacy:

⁹Operational risk is defined as the risk of losses resulting from inadequate or failed internal processes, people and systems, or external events. Examples of operational risk include two famous cases of “rogue trading”: Nick Leeson, who in 1995 led Baring Bank to bankruptcy, through a loss of £1.3bn, and Jérôme Kerviel, who in 2008 led Société Générale to a loss of €5bn.

¹⁰Note that within the internal rating approach, banks are not allowed to use internal models of credit risk. Banks that have received supervisory approval to use the internal approach may rely on their own internal estimates of risk components in determining the capital requirement for a given exposure. The risk components include: (i) measures of the probability of default, (ii) loss given default, (iii) exposure at default, (iv) effective maturity.

Pillar I: Minimal capital requirements. It defines a set of basic rules for measuring credit risk, market risk and operational risks (both “standardized” and “internal”).

Pillar II: Supervisory review of capital adequacy. It increases the role of banking supervisors, by setting rules of conduct national supervisors should maintain to ensure that banks have capital adequacy over and above the minimal capital requirements of Pillar 1.

Pillar III: Market disclosure. It increases the role of market discipline and disclosure, by relying on the publication of information by banks, relating to issues such as risk measurements, risk-rating processes, or risk-management systems.

Pillar I defines a clear separation of three types of risk: (a) credit risk; (b) operational risk and (c) market risk. The calculation rules rely on the concept of total risk-weighted assets (Total RWA), defined as

$$\text{Total RWA} = \text{Cr} + \text{MnO},$$

where

$$\begin{aligned} \text{Cr} &= \text{Risk-weighted assets for } \textit{credit} \text{ risk} \\ \text{MnO} &= \text{Assets weighted for } \textit{market} \text{ and } \textit{operational} \text{ risk} \\ &= (\text{Capital requirements for } \textit{market} \text{ and } \textit{operational} \text{ risk}) * \underbrace{12.5}_{=1/0.08} \end{aligned}$$

For example, the capital requirements for market risk can be determined through dedicated VaR models, such as (and, possibly, more sophisticated versions of) those surveyed in Section 13.7. The (total) minimum capital requirements are taken to be 8% of the total risk-weighted assets,

$$\frac{\text{Regulatory capital}}{\text{Total RWA}} \geq 8\%.$$

One immediate issue arising with Basel II is its heavy reliance on credit rating agencies for what it pertains the standardized approach to credit risk, which might be misguided due to conflicts of interest. At the time of writing, credit rating agencies are mostly unregulated, and the quality of their estimates is observable only with lags and, sometimes, only when mispricing issues are already in place. A second issue is “procyclicality”: in bad times, banks have to reduce lending, which exacerbates the current economic developments, which makes banks reduce lending even further, etc. Procyclicality is the theme dealt with in Sections 13.5.3 and 13.5.4.

After a number of additional consultative papers, national regulators indicated to the Financial Stability Institute (FSI)¹¹ that they would implement Basel II by 2015. During the process when the European Union was implementing Basel II through its EU Capital Adequacy Directive (CAD III), the global financial crisis following the 2007 subprime events determined a re-thinking of Basel II, leading to a new set of rules, known as Basel III. The main innovations of Basel III are summarized by the following points: (i) new capital requirements, such as those summarized in the table below, as well as a new mandatory capital conservation buffer of 2.5% of the Total RWA, to face economic stress; (ii) new rules that allow national supervisors to

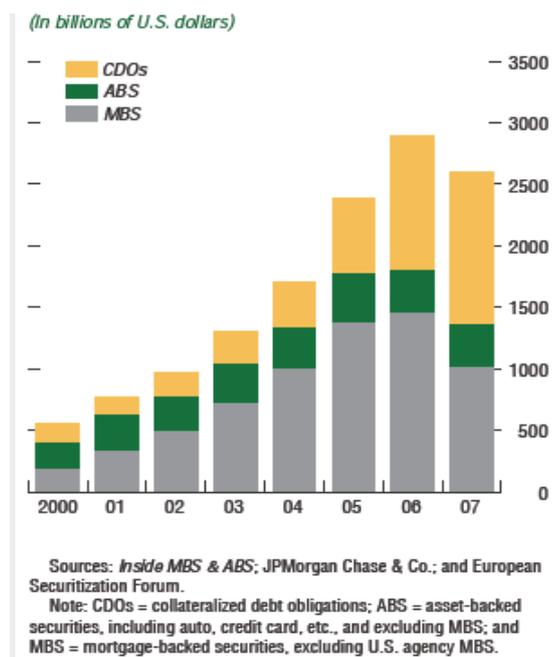
¹¹The FSI, headquartered by the Bank for International Settlements in Basel, was established in 1999 in response to the Asian crisis of 1997.

require banks to set up capital up to 2.5% of the Total RWA in times of high credit growth—countercyclical capital buffers; and (iii) a target for the leverage ratio, defined as the ratio of Tier 1 capital (i.e. equity plus reserves minus intangible assets) over total assets net of intangible assets, to be at least 3% ($\frac{\text{Tier 1}}{\text{Assets}} > 3\% \Leftrightarrow \text{Lev} \equiv \frac{\text{Non Tier 1}}{\text{Tier 1}} < \frac{1}{0.03} - 1$), as well as additional liquidity ratios. The following table summarizes the main differences in capital requirements that Basel III introduces against Basel II.

Capital requirements as a % of the Total RWA		
	Basel II	Basel III
Common equity	2%	4.5%
Tier 1	4%	6%
Total capital	8%	8%
Common equity (conservation)	–	2.5%

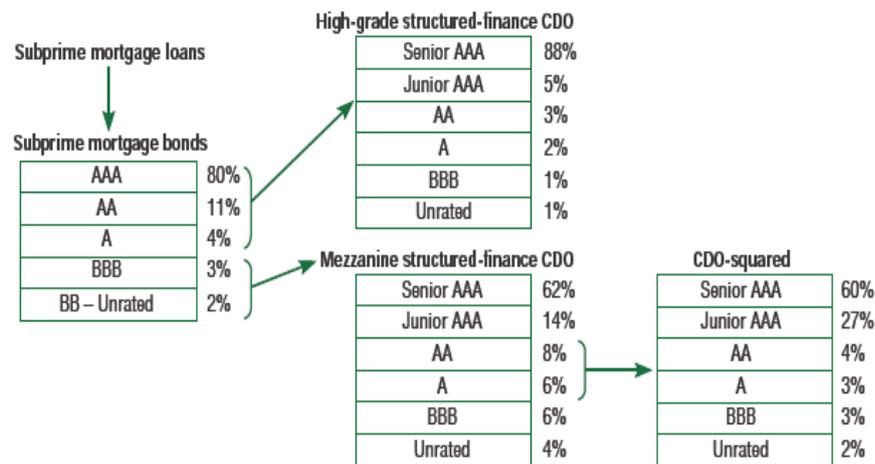
13.5.2 The 2007 subprime crisis

13.5.2.1 Issuance data



European and U.S. Structured Credit Insurance. Source: IMF, Global Financial Stability Report, April 2008.

Matryoshka — Russian Doll: Multi-Layered Structured Credit Products



Source: IMF staff estimates.
Note: CDO = collateralized debt obligation.

Outstanding U.S. Subprime issuance. Source: IMF, Global Financial Stability Report, April 2008.

13.5.2.2 Off-balance-sheet entities: “SIV,” “conduits,” and “SIV-lites”

[b. circa 1985]

On the funding side, a typical SIV (Structured Investment Vehicle) issues long-maturity notes. On the asset side, a SIV typically relies on assets that are more complex than those conduits rely on. SIVs tended to be more leveraged than conduits. Please remember: SPV = Special Purpose Vehicle, i.e. a vehicle that organizes securitization of assets; SIV = Structured Investment Vehicle, i.e. a *fund* that manages asset backed securities. In a sense, SIV were virtual banks, as they used to borrow through low-interest securities and invest the money in longer term securities yielding large rewards (and risk), as we discuss below. SIVs and conduits typically had an open-ended lifespan.

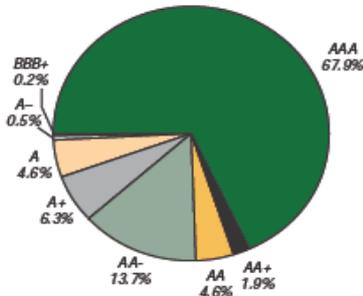
SIV-lites are less conservatively managed, and structured with greater leverage. Their portfolios are not much diversified, and are much smaller in size than SIVs. SIV-lites had a finite lifespan, with a one-off issuance vehicle. They were greatly exposed to the U.S. subprime market, more so than SIVs.

Off-balance-sheet entities borrow in the shorter term, typically through commercial paper or auction rate securities with average maturity of 90 days, as well as medium term notes with average maturity of a year. They purchase long-maturity debt, such as financial corporate bonds or asset-backed securities, which is high-yielding. Naturally, the profits made by these entities are paid to the capital note holders, and the investment managers. The capital note holders are, of course, the first-loss investors.

The obvious risk incurred by these entities is *solvency*, a risk that materializes when the value of long-term assets falls below the value of short-term liabilities. This risk has great chances to materialize when the pricing of the assets is “informal,” as argued below. A second risk is *funding liquidity*, the risk related to duration mismatch: refinancing occurs on a short-term frequency, but if short-term market conditions are bad, the entities need to sell the assets into a

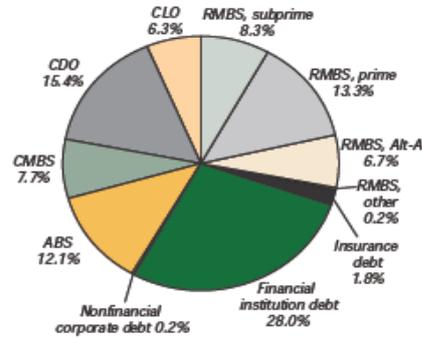
depressed market. To cope with this risk, the sponsoring banks would grant credit lines. Typical sponsors were: Citibank (\$100bn), JP Morgan Chase (\$77bn), Bank of America (\$60bn). In the European Union: HBOS (\$42bn), ABN Amro (\$40bn), HSBC (\$32bn).

Structured Investment Vehicle Portfolio by Ratings, October 2007



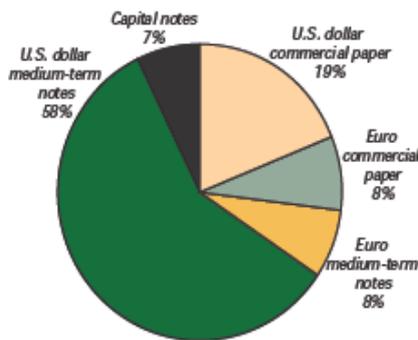
Source: Standard & Poor's.

Structured Investment Vehicles by Type of Assets, October 2007



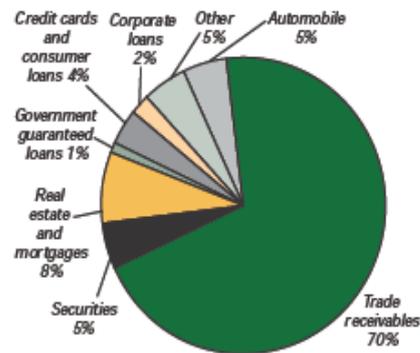
Source: Standard & Poor's.
 Note: ABS = asset-backed security; CDO = collateralized debt obligation; CLO = collateralized loan obligation; CMBS = commercial mortgage-backed security; RMBS = residential mortgage-backed security.

Funding Profile of Structured Investment Vehicles Held by Banks and Nonbanks, October 2007
 (Total liabilities: \$367.8 billion)



Source: Fitch Ratings, November 2007.

Asset-Backed Commercial Paper Conduits by Traditional Assets, May 2007
 (In percent of total)



Source: Moody's.

Source: IMF, Global Financial Stability Report, April 2008.

13.5.2.3 The 2007 meltdown

The first obvious issues to think about relate to pricing and the role played by credit ratings. Being illiquid, the pricing of structured credit products used to rely on that of similarly rated comparable products for which quotations were available. For example, the price of AAA ABX subindices would be used to estimate the values of AAA-rated tranches of MBS. Or, the price of BBB subindices would be used to value BBB-rated MBS tranches. This is the “mapping role” credit ratings played for the pricing of customized or illiquid structured credit products. However, it is well-known that the risk profile of structured products differs from that of corporate bonds. Even if a tranche has the same expected loss as an otherwise similar corporate bond, *unexpected loss or tail risk can* be much larger than that for corporate bonds. Therefore, it is

misleading to extrapolate structured products ratings from corporate bonds ratings. Typically, ratings used to capture only the first moments of the distribution. Moreover, credit rating inertia for bonds does not necessarily work for structured products, as illustrated in the picture below.

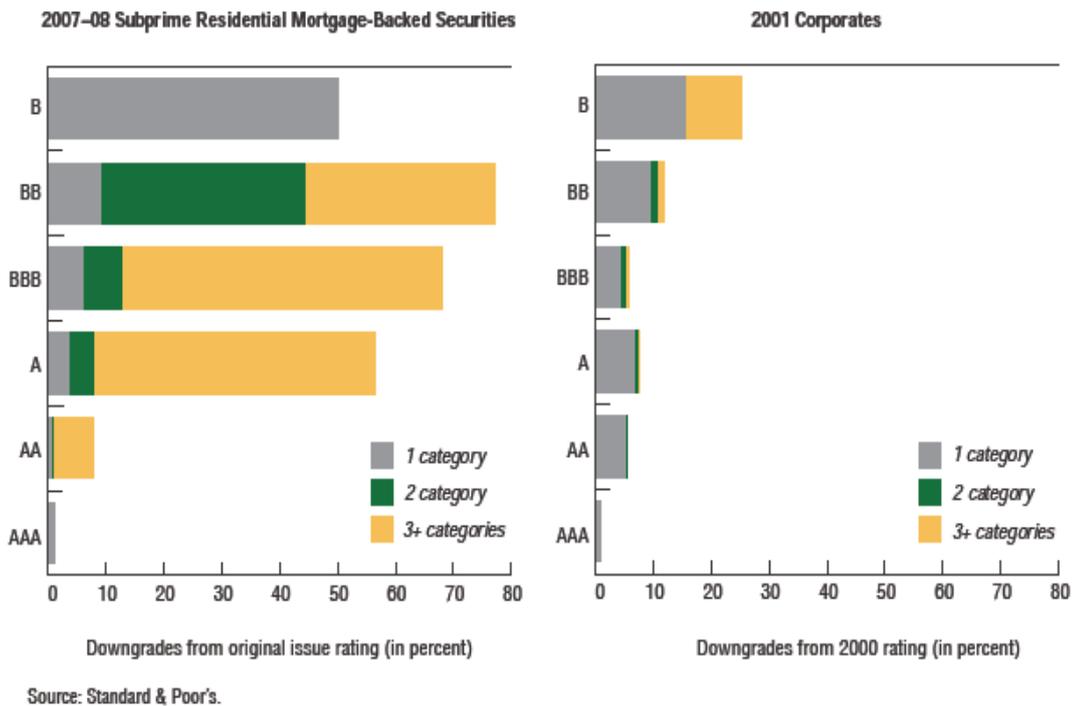
Two additional fundamental aspects contributing to the meltdown. First, there was an erosion in lending standards: statistical models were based on historically low mortgage default and delinquency rates that arose in a credit environment with tight credit standards. Second, there were correlation issues: past data suggested a quite weak correlation between regional mortgages, which made investors perceive a sense of “diversification.” However, the housing downturn turned up to be a nation-wide phenomenon.

The mechanics of the crisis started with fears of contagion from the rising level of defaults in subprime underlying instruments, many of which were incorporated in complex products. Fears of contagion concerned safer tranches as well. They came from the investors’ understanding the pricing models were misspecified, and their lack of trust vis-à-vis the rating agencies. Banks, on the other side, were affected for a number of reasons: (i) they had invested in subprime securities directly; (ii) they had provided credit lines to SIV (indebted through commercial paper) and conduits that held these securities, thereby creating a shadow banking system, which escaped accounting and supervision rules; and (iii) this very same shadow system generated banks’ loss of confidence in the ability of their counterparties to meet their contractual obligations. So the Asset Backed Commercial Paper market dried up, triggering credit lines. The result was a sell-off of anything related to structured finance, from junk to AAA, which led to a complete “liquidity black hole,” and a severe reappraisal of structured finance.

In turn, the reappraisal of structured finance determined severe writedowns, arising in part through the “liquidity black hole,” i.e. by the market participants expectations. Repricing was difficult indeed. In the absence of a liquid market, writedowns have to rely on marking to model. But investors did not trust the models and the rating process leading to them! Meanwhile, credit agencies proceeded to severe downgrades, confirming the investors’ beliefs that ratings were not entirely appropriate, a quite self-reinforcing mechanism. These events escalated to a complete dry up in September-October 2008, partly restored by painful bank bail-outs and

recapitalizations.

One-Year Cohort Rating Downgrades



Source: IMF, Global Financial Stability Report, April 2008.

13.5.3 Top tier capital ratio targets and endogenous volatility

Typically, we take volatility or credit events to be exogenous when in fact, they are not. Taking risk as an exogenous process might be an excellent approximation whilst living in good times. The quality of this approximation deteriorates in times of crisis. The implicit assumption made in many instances of this and previous chapters, is that one's own actions, based on a volatility forecast, do not affect future volatility, just like forecasting weather does not influence future weather. Arguably, the actions of many heterogeneous market participants should tend to cancel with each other, during periods of "calm." However, market participants tend to cluster their decision rules in periods of crisis. The literature on this "endogenous risk" is quite fascinating, as are the surveys in Shin (2010), or the recent modeling framework put forward by Danielsson, Shin and Zigrand (2011).

This section develops a simple model of endogenous risk, where markets can be destabilized by one instance of procyclicality, arising because financial institutions need to comply with a given "top tier capital" (i.e. the capital comprising Tier 1 and Tier 2, using the terminology of Section 13.5.1) ratio. After a negative shock in the value of the assets on the balance sheet, a financial institution needs to restore its top tier capital ratio. In the short-run, the institution can only restore this ratio through asset sales. Because every institution is doing the same, these asset sales have a market impact, collectively, determining a further fall in the value of the assets, and so on. The final outcome is an increased volatility of the risky asset price, as well as a disproportionate assets sell-off, if compared to the initial shocks triggering it. The model is useful to think about the subprime events in 2007, as well as the ensuing credit crunch and

the new solutions that monetary authorities have experimented to help mitigate these adverse developments, as we shall discuss.

13.5.3.1 Model

We consider a model with many identical financial institutions complying with regulation or, more generally, concerned with a pre-specified target of top tier capital ratio against risky assets. Each institution has the following balance sheet.

Balance sheet, Time 1	
A	E
C_b	D

The notation is a bit more elaborated than that used in Section 13.3: whilst we still define E as equity (including past retained earnings) and D as debt, we now define A as the value of risky assets, no matter how liquid these can be. Moreover, we define C_b to equal cash and reserves. We suppose the Cooke ratio is in place, or $E \geq 8\% * A$ and to simplify let $E = 8\% * A$. Note that a top tier capital rule does not determine a leverage rule: there are, obviously, many leverage ratios, $\frac{D}{E}$, consistent with a given top tier capital ratio $\frac{E}{A}$. In fact, the new Basel III explicitly considers leverage ratios, thereby innovating upon Basel II, as discussed in Section 13.5.1. We shall deal with the procyclicality induced by this new rule later in this section.

Next, assume that some exogenous shock takes place, which makes the value of the risky assets decrease by some amount, ΔA , after which each institution would have the following balance sheet.

Balance sheet, Time 2	
$A - \Delta A$	$E - \Delta A$
C_b	D

But, each institution has to comply with its top tier ratio target. Therefore, at time 2, the new top tier capital, $E - \Delta A$, must be at least 8% of the risky assets, $A - \Delta A$, or,

$$E - \Delta A = 8\% * (A - \Delta A). \quad (13.58)$$

At time 1, the financial institution had set $E = 8\% * A$. Therefore, Eq. (13.58) cannot hold, as a simple computation reveals. The intuition behind this impossibility is simple. As the value of the risky assets falls, the value of equity falls by a larger percentage than that of the risky assets, due to leverage—the value of risky assets falls by $-\frac{\Delta A}{A}$, whereas the value of equity falls by a larger percentage, $-\frac{\Delta A}{E}$, due to $E < A$.

Two solutions are available to the financial institution: (i) to inject fresh capital; (ii) to sell some of the risky assets. The first solution is not quite viable in the short-run. Let us analyze the second solution. We are looking for some quantity of the risky asset to sell, such that the reduction in value of the assets, say X^s , is able to meet the top tier capital ratio target. In terms of the balance sheet, we have the following situation.

Balance sheet, Time 3	
$A - \Delta A - X^s$	$E - \Delta A$
$C_b + X^s$	D

How much of the risky assets value should the financial institution precisely get rid off? To maintain the new top tier ratio, X^s must satisfy:

$$E - \Delta A = 8\% * (A - \Delta A - X^s).$$

Using $E = 8\% * A$, and solving for X^s yields,

$$X^s = \frac{1 - 8\%}{8\%} * \Delta A = 11.5 * \Delta A.$$

That is, roughly, the number of risky assets to sell is proportional to the percentage loss in their value. In general,

$$\frac{X^s}{A} = \frac{1 - \kappa}{\kappa} \frac{\Delta A}{A}, \quad \kappa \equiv \frac{E}{A}, \quad (13.59)$$

where κ denotes the top tier capital ratio against risky assets. This result is intuitive: following a negative shock affecting the value of the risky assets, the amount of asset sales is decreasing in the pre-specified top tier capital ratio, κ , because the closer κ is to 100%, the easier the adjustment is to maintain the same κ . Eq. (13.59) would end the description of this market, if the financial institution had no price impact.

We assume that financial institutions have a market impact, collectively: while the behavior of one single institution cannot affect the price of the risky asset, many institutions doing the same thing at the same time more likely could, thereby creating price pressures, with the price of assets falling and triggering new sales into a depressed market, over a vicious feedback loop. Is there an equilibrium for this loop? The answer relies on the way we think of selling pressure. We model selling pressure by assuming that there is a continuum of financial institutions with zero measure, and that the assets value changes according to:

$$\frac{\Delta A}{A} = \tilde{\epsilon} + \lambda \frac{X^s}{A}, \quad (13.60)$$

where λ is the “price impact,” as we say in market microstructure, surveyed in Chapter 9. In words, the percentage change in the assets value is the product of two components: (i) the initial shock, $\tilde{\epsilon}$, and (ii) a selling pressure component, which will pump up endogenous volatility, as we shall show. Naturally, if the market was perfectly liquid, $\lambda = 0$. Moreover, in a world where financial institutions had no concerns about top tier capital rules, we would have $X^s = 0$ in the first place, and so:

$$\frac{\Delta A}{A} = \tilde{\epsilon}. \quad (13.61)$$

Note that this solution is also that arising when the market is perfectly liquid, $\lambda = 0$. However, we assume the existence of price pressure, as formalized by Eq. (13.60), determined by the concern financial institutions have to comply with a top tier capital rule, leading them to a sale $\frac{X^s}{A}$ satisfying Eq. (13.59). The loop we have created is, then, the following. After an initial shock affecting the risky assets value, $\tilde{\epsilon}$, financial institutions sell risky assets, to an extent proportional to the percentage change in the assets value, according to Eq. (13.59). In turn, the sell-off entails a further percentage drop in the assets value, as determined by Eq. (13.60). An equilibrium, provided it exists, is a fixed point to this feedback-loop, i.e. a situation where the loop stops because the drops in the risky assets value do not happen anymore, and the assets sell-offs are interrupted as a result, thereby providing no reasons for the assets value to fall any further.

To find this fixed point, we replace Eq. (13.59) into Eq. (13.60) and solve for $\frac{\Delta A}{A}$, leaving:

$$\frac{\Delta A}{A} = c_A \cdot \tilde{\epsilon}, \quad c_A \equiv \frac{\kappa}{\kappa - \lambda(1 - \kappa)}. \quad (13.62)$$

Note that in order for this equilibrium to exist, we need that λ , the slope of the line $\frac{X^s}{A} \mapsto \frac{\Delta A}{A}$ in Eq. (13.60), be less than $\frac{\kappa}{1-\kappa}$, the slope of the line $\frac{X^s}{A} \mapsto \frac{\Delta A}{A}$ in Eq. (13.59), or that $\lambda(1-\kappa) < \kappa$. Intuitively, if the price impact λ was too large, the feedback from asset sales to the assets value drops would create a perverse spiral such that the market would collapse.

Eq. (13.62) shows, crucially, that the shock multiplier, $c_A > 1$. When financial institutions have a concern over top tier capital ratios, the ultimate change in the assets value resulting from an initial shock, is larger than that we would have observed otherwise, say in Eq. (13.61). For example, assuming a price impact $\lambda = 0.05$ implies a multiplier $c_A = 2.4$. Naturally, these effects become less important as the market becomes more liquid, and do not matter anymore in the limit case where the market is perfectly liquid, $\lambda = 0$.

What is the amount of asset sales resulting from this loop? Replacing Eq. (13.62) into Eq. (13.59) yields:

$$\frac{X^s}{A} = c_X \cdot \tilde{\epsilon}, \quad c_X \equiv \frac{1-\kappa}{\kappa - \lambda(1-\kappa)}. \quad (13.63)$$

Note that the feedback loop might exert quite substantial effects into the amount of asset sales. Assuming $\kappa = 0.08$, the multiplier c_X would equal 11.5 in the absence of feedback, $\lambda = 0$, as previously noted. The same multiplier more than doubles in the presence of feedbacks and a price impact of just $\lambda = 0.05$, attaining a value $c_X = 27.1$.

This model thus formalizes the idea that even a small shock affecting the risky assets held by financial institutions might lead to large sale adjustments and price corrections, similarly as for the developments inherent the subprime events described in the previous section. In the model, the concerns financial institutions have about top tier capital ratios leads them to substantial asset sales in response to a shock, which are even more amplified in the presence of feedback effects induced by liquidity frictions.

13.5.3.2 Multiple equilibria and market break-ups

In the model we analyze, an equilibrium exists under parameter restrictions that are independent of the realization of the initial shock, $\tilde{\epsilon}$. As noted, we simply need that the denominators of the multipliers c_A and c_X in Eqs. (13.62) and (13.63) be strictly positive. We now present a variant of the model, where an equilibrium fails to exist when the initial shock $\tilde{\epsilon}$ is sufficiently large. We simply assume that the price pressure is nonlinear, differently from the linearity assumption underlying Eq. (13.60). It is:

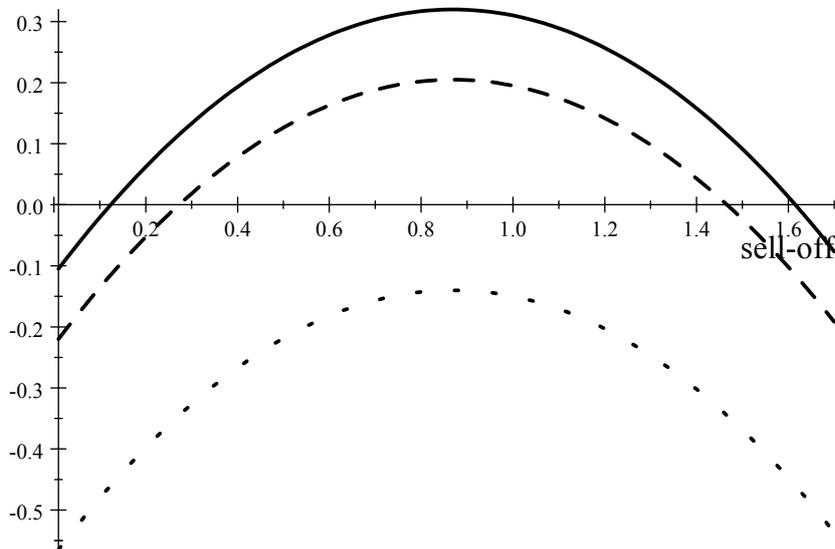
$$\frac{\Delta A}{A} = \tilde{\epsilon} + \lambda \left(\frac{X^s}{A} \right)^2. \quad (13.64)$$

The quadratic term in Eq. (13.64) formalizes the idea that the price impact of asset sales does not matter too much when the asset sales are limited, but becomes disproportionately high when the asset sales are at a large scale. This convexity translates into a non-linearity of the resulting feedback loop. Replacing Eq. (13.64) into Eq. (13.59), we find that in any equilibrium, the amount of asset sales is solution to the following quadratic equation

$$0 = \varphi \left(\frac{X^s}{A} \right) \equiv \frac{X^s}{A} - \frac{1-\kappa}{\kappa} \lambda \left(\frac{X^s}{A} \right)^2 - \frac{1-\kappa}{\kappa} \tilde{\epsilon}. \quad (13.65)$$

We hypothesize the market is hit by a series of shocks. Initially, we assume that $\tilde{\epsilon} = 0$, such that two equilibria are possible: one, where the sell-off is just zero, $\frac{X^s}{A} = 0$; and another,

where the sell-off is $\frac{\kappa}{\lambda(1-\kappa)}$. We assume the sell-off is zero. Then, we assume that a first positive shock hits the market, with $\tilde{\epsilon} = 1\%$. The solid line in the next picture is the graph of $\varphi\left(\frac{X^S}{A}\right)$ in this case. There are two equilibriums, arising for $\frac{X^S}{A} : \varphi\left(\frac{X^S}{A}\right) = 0$, but we assume that the market coordinates towards the leftmost one.



The three curves depict the graph of $\varphi\left(\frac{X^S}{A}\right)$ in Eq. (13.65), when the initial shock is $\tilde{\epsilon} = 1\%$ (solid line), $\tilde{\epsilon} = 2\%$ (dashed line), and $\tilde{\epsilon} = 5\%$ (dotted line), obtained assuming a top tier capital ratio against risky assets $\kappa = 8\%$, and a price impact $\lambda = 0.05$. The equilibrium sales are those where the curves intersect the horizontal axis, if any.

As the risky assets value is hit by one additional shock, say with $\tilde{\epsilon} = 2\%$, the graph of $\varphi\left(\frac{X^S}{A}\right)$ shifts to South-East, to the dashed line, and the asset sales increase as a result, still being the leftmost zero of $\varphi\left(\frac{X^S}{A}\right)$. The market collapses when the shock is $\tilde{\epsilon} = 5\%$, leading the graph of $\varphi\left(\frac{X^S}{A}\right)$ to a further shift to South-East, where no equilibriums are left at all.

13.5.3.3 Deleveraging

The assumption made so far is that following a shock affecting the assets in the balance sheet, financial institutions sell additional assets to the extent their top tier capital ratios are restored. This section investigates additional adjustments, aiming to preserve leverage ratio targets. Denote the leverage ratio as $L \equiv \frac{D}{E}$. The timing of the shock and banks reaction is as in Section 13.5.3.1, with the exception that we now have one additional constraint: at time 3, financial institutions also wish to call portions of their debt, X^1 say, so as to comply with leverage ratio targets, and achieve the following balance sheet.

Balance sheet, Time 3
under a deleveraging scenario with deep liquidity buffers

$A - \Delta A - X^s$	$E - \Delta A$
$C_b + X^s - X^1$	$D - X^1$

The term X^s is the usual sell-off needed to maintain top tier capital targets. The financial institutions also target an amount of deleveraging $X^1 : \frac{D-X^1}{E-\Delta A} = L$, or:

$$\frac{X^1}{A} = L \cdot \frac{\Delta A}{A}, \quad (13.66)$$

and we initially assume that while doing so, they do not exhaust their liquidity buffers,

$$C_b + X^s - X^1 \geq 0. \quad (13.67)$$

Note that under the condition in (13.67), deleveraging does not have a price impact because it would imply banks are simply using cash to repay portions of their debt. In this case, X^s is the same as that in Eq. (13.63). If, instead, $C_b + X^s$ is not large enough, financial institutions would have to sell additional assets, ΔX^s say, so as to have sufficient cash with which to meet leverage ratio targets. Precisely, if the inequality in (13.67) does not hold, we need to have that, at least,

$$\Delta X^s : \Delta X^s = X^1 - (X^s + C_b), \quad (13.68)$$

such that the balance sheet faced by the financial institutions at time 3 would be as in the following alternative scenario:

$$\begin{array}{c} \text{Balance sheet, Time 3} \\ \text{under a deleveraging scenario leading to exhausting liquidity buffers} \\ \hline \begin{array}{l} A - \Delta A - X^s - \Delta X^s \\ C_b + X^s - X^1 + \Delta X^s \end{array} \quad \left| \quad \begin{array}{l} E - \Delta A \\ D - X^1 \end{array} \right. \end{array}$$

To determine the feedback effects of the asset sales, X^s and ΔX^s , replace the top tier capital ratio condition, Eq. (13.59), and the leverage condition, Eq. (13.66), into Eq. (13.68), leaving,

$$\frac{\Delta X^s}{A} = \frac{(1+L)\kappa - 1}{\kappa} \frac{\Delta A}{A} - \frac{C_b}{A}. \quad (13.69)$$

Note that this expression is positive by assumption—we are assuming that over the deleveraging process, banks would exhaust their liquidity buffers to the extent the condition in (13.67) does not hold. Such a situation arises precisely due to a high leverage, L . For example, assuming $\kappa = 0.08$, the loading for $\frac{\Delta A}{A}$ in Eq. (13.69) is positive only when $L > 11.5$. We would need to observe values of L larger than 11.5 (and state-dependent, i.e. depending on the realization of $\frac{\Delta A}{A}$), in order for the condition in (13.67) to break down.

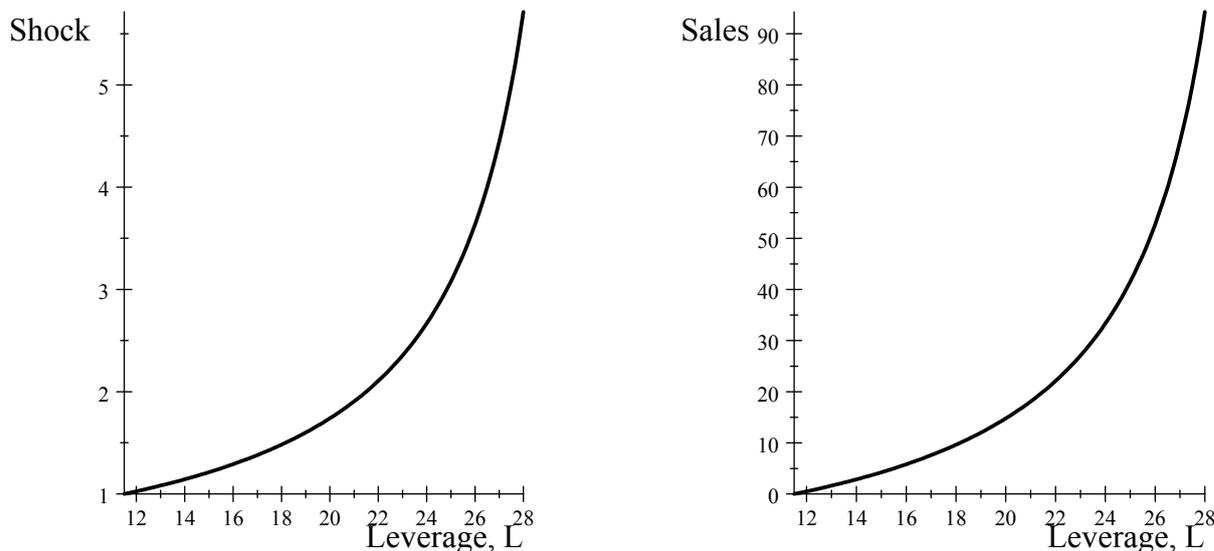
We determine an equilibrium for the feedback loop in this market. Replace Eq. (13.69) into Eq. (13.60), evaluated when the asset sales amount to ΔX^s , i.e. $\frac{\Delta A}{A} = \tilde{\epsilon} + \lambda \frac{\Delta X^s}{A}$, and solve for both the assets value drop and sales,

$$\frac{\Delta A}{A} = c_{L,a} \cdot \tilde{\epsilon} - \lambda c_{L,a} \cdot \frac{C_b}{A}, \quad \frac{\Delta X^s}{A} = c_{L,b} \cdot \tilde{\epsilon} - \frac{((1+L)\kappa - 1)\lambda c_{L,a} + \kappa}{\kappa} \cdot \frac{C_b}{A}, \quad (13.70)$$

and the constants $c_{L,a} \equiv \frac{\kappa}{\kappa - \lambda((1+L)\kappa - 1)}$ and $c_{L,b} \equiv \frac{(1+L)\kappa - 1}{\kappa - \lambda((1+L)\kappa - 1)}$. To ensure an equilibrium exists, we need that λ , the slope of the line $\frac{X^s}{A} \mapsto \frac{\Delta A}{A}$ in Eq. (13.60), be less than $\frac{\kappa}{(1+L)\kappa - 1}$, the slope of the line $\frac{X^s}{A} \mapsto \frac{\Delta A}{A}$ in Eq. (13.69), or that the denominators of $c_{L,a}$ and $c_{L,b}$ be strictly positive.

Finally, note that there is a third possibility available to banks: to sell additional assets even when the condition in (13.67) holds. This possibility is relevant when the financial institutions

also have concerns over maintaining a certain level of liquidity buffers, some of them possibly being mandatory. For example, if the liquidity target is at least C_b , we replace the inequality in (13.67) with the stricter inequality, $X^s - X^l \geq 0$. The solutions for the assets value drop and sales are the same as those in Eqs. (13.70), but with the terms involving C_b being dropped, $\frac{\Delta A}{A} = c_{L,a} \cdot \tilde{\epsilon}$, and $\frac{\Delta X^s}{A} = c_{L,b} \cdot \tilde{\epsilon}$. The next picture depicts the two multipliers of $\tilde{\epsilon}$ for $\frac{\Delta A}{A}$ and $\frac{\Delta X^s}{A}$ arising in this case, assuming the top tier capital ratio target is $\kappa = 0.08$, and the price impact of asset sales is $\lambda = 0.05$.



This picture depicts the graph of the two multipliers, $c_{L,a}$ and $c_{L,b}$ in Eqs. (13.70), as a function of banks leverage, L , when the top tier capital ratio is $\kappa = 0.08$, and the price impact of banks sell-offs is $\lambda = 0.05$. When banks have concerns over maintaining cash and reserves to their level before the shock, C_b , the overall assets value drop and sales equal $\frac{\Delta A}{A} = c_{L,a} \cdot \tilde{\epsilon}$ and $\frac{\Delta X^s}{A} = c_{L,b} \cdot \tilde{\epsilon}$. The left panel depicts the graph of $c_{L,a}$ and the right panel depicts that of $c_{L,b}$.

First, note that given the assumption that $\kappa = 0.08$, the relevant range of variation for leverage is that for $L \geq 11.5$, as we are studying situations where deleveraging leads to partial exhaustion of liquidity buffers. The effects can be quite substantial.

13.5.4 Credit crunches and quantitative easing

[Show pictures of (i) the balance sheet of FED, ECB and BoE over the last years. (ii) the FED fund rate, and EONIA, (iii) the corporate spreads]

13.5.4.1 The money multiplier

Loans make deposits! The well-known mechanics underlying the creation of money relies on the standard money multiplier, whereby new deposits made available to the banking system are partially used to extend new loans, which generate further deposits, and so on. Mathematically, the supply of money, say M1 aggregates, includes cash held by the public plus deposits, $M = C_p + D$, with straight forward notation. Instead, the monetary base, or “high potential” money, is made of cash held by the public plus banks cash and reserves, $H = C_p + C_b$, where C_b

denotes cash and reserves held by banks, as in the previous sections. Note the leakage banks create over the circuit of money creation: because banks face liquidity needs possibly arising in the short-term (vis-à-vis their clients and/or other banks), and possibly need to maintain mandatory reserves with the central banks of the countries where they operate, they hoard C_b , which escapes the loans-make-deposits loop.

We assume that the ratios $\frac{C_p}{D}$ and $\frac{C_b}{D}$ are constant and equal to c_p and θ , respectively, such that money supply equals a money multiplier times monetary base, viz

$$M = \frac{1 + c_p}{c_p + \theta} \cdot H. \quad (13.71)$$

The value of θ depends on a variety of factors, such as: (i) the discount rate at which a central bank lends money to banks; (ii) the interbank rates such as the LIBOR; (iii) the level of mandatory reserves banks have to keep with their central banks; (iv) the interbank rate for resources to allocate to mandatory reserves—the Federal Funds rate in the US; (v) the risk/return tradeoff prevailing in other markets. Clearly, the value of θ increases with the values of the items from (i) to (iv), and decreases when the tradeoff in (v) compares more favourably to banks.

13.5.4.2 Policy actions

Conventional policy strategies consist of actions aiming to affect the value of θ and H . For example, central banks can expand H through open market operations, by purchasing short-term Government bonds. Note that this action likely affects θ as well, as the opportunity costs of holding excess reserves increase as markets are flooded with more and more liquidity. There are, obviously, limits to this action, arising when short-term interest rates get close to zero. Consider the 2007 subprime events. We know that following a shock affecting the banks' books, the overall adjustments can be quite substantial. Consider, for example, the model in the previous section, where banks have concerns over the top tier capital ratio. After the shock takes place, banks aim to a shrinkage in the assets value equal to X^s . Moreover, the shrinkage can be even more substantial, $X^s + \Delta X^s$, in markets where banks are high levered, and concerned about not increasing their leverage even more as a result of the shock. The model is silent about which particular assets (liquid or not) or loans are involved into the shrinkage plans, X^s or $X^s + \Delta X^s$. We define a “credit crunch” as the situation where banks decide to cut on corporate loans and bonds—they hold more reserves, instead of lending money to the real sector.

A quite mechanical response to a credit crunch is to increase H through, say, open market operations. Put simply, a credit crunch entails a higher value of θ , among other things. Monetary aggregates, M , are destroyed as a result, but can be restored through an injection of high potential money. Precisely, by Eq. (13.71), the expansion of monetary base needed to maintain the same supply of money, M , is $\Delta H = \frac{M}{1+c_p} \Delta\theta$, where $\Delta\theta$ is the increase in θ determined by the credit crunch. This policy action is quite fundamental as it helps keep interest rates low, yet it may not be enough when the credit crunch is so particularly severe to lead to very substantial shrinkages in the economic activity, as we now explain.

The effects of a credit crunch on the real sector of the economy are quite obvious, with an increase in the cost of capital and a subsequent shrinkage in the economic activity. For example, the recession following the subprime events was spectacular, with industrial production falling by approximately 13% on a yearly basis in March 2009, the highest drop since World War II. The policy action was equally impressive: in less than two years after the subprime events, the FED was capable of pushing short-term rates close to zero although during those periods, it

was already clear that this policy would not be likely to prevent an even deeper recession. All in all, even if the Federal Funds rate and short-term rates on safe assets were close to zero, the cost of capital firms had to bear were quite substantial as a result of the credit crunch.¹² Note that at that time, the credit crunch was also exasperated by a freeze in the interbank lending market, arising from concerns financial institutions had about counterparty risk.¹³

The events following the 2007 turmoil can be described as those of a “liquidity trap,” where banks hold abundant liquidity and short-term rates are close to zero. Note that the nature of this liquidity trap is different from the standard Keynesian liquidity trap, as formalized in the Appendix of Chapter 1: the Keynesian trap arises when money demand is flat as a result of the expectations investors have that future interest rates can only increase. In this case, agents simply absorb any liquidity injections made by the monetary authority, and interest rates remain “trapped” at some minimum rate, coinciding with the lowest, shadow interest rate beyond which no investor is ready to bet against a decrease. The liquidity trap we are analyzing is different, and stems from the mere and mechanical circumstance that money supply is so abundant to have made short-term rates close to zero in the first place. However, in both cases, the economy is trapped in that a further increase in the monetary base would have no effects on short-term interest rates.

What was the initial policy reaction to this liquidity trap? Note that a further issue arising within the case we analyze in this section, is that the liquidity trap is accompanied by a surge in corporate spreads, as a result of the credit crunch. “Quantitative easing” is a policy action that aims to restore the shrinkage in credit supply, and possibly reduce these corporate spreads, so as to mitigate the adverse effects of the credit crunch on to the real economic activity. Consider the following balance sheets. The balance sheet on the left-hand side is that arising after financial institutions have reacted to a shock in their assets value, as formalized in previous sections. The portion of X^s that includes corporate loans and bonds is what we are terming credit crunch.

Balance sheets, Time 3 and beyond

$$\begin{array}{c|c} A - \Delta A - X^s & E - \Delta A \\ \hline C_b + X^s & D \end{array} \quad \begin{array}{c} \rightarrow \\ \leftarrow \end{array} \quad \begin{array}{c|c} A - \Delta A - X^s + Q^s & E - \Delta A \\ \hline C_b + X^s - Q^s & D \end{array}$$

The effects of quantitative easing can be seen through the balance sheet on the right-hand side. The ideal, spontaneous, resolution of a credit crunch is when financial institutions are willing to extend corporate loans or purchase corporate bonds, to restore their credit shrinkage, at least partially, by some amount Q^s . However, we know this resolution cannot occur until the value of the assets remains depressed. However, the central bank could step in to purchase the assets the banking system is disliking, to an extent equal to at least Q^s , so as to leave banks with the excess reserves they wish and comply with their top tier capital ratio targets. This action is the essence of quantitative easing, with assets typically involved being ABS and even long-term bonds. Its effects include both (i) an increase in H , equal to the extent of the

¹²A high cost of capital to firms might occur during a credit crunch, as a result of one additional effect: a credit crunch makes determines a contraction in aggregate demand, which makes defaults more likely.

¹³A shrinkage in the economy activity suggests a natural extension of the models in the previous section, with one additional element of procyclicality: as the real economy plummets as a result of the credit crunch, the value of the assets decreases even more, thereby deepening the credit crunch, over a vicious feedback loop. Note that this feedback mechanism would be, quite naturally, part of the financial accelerator hypothesis, although distinct from the mechanism mentioned at the beginning of this section, and surveyed by Bernanke, Gertler and Gilchrist (1999). In the version we suggest, the credit crunch is determined by concerns financial intermediaries have about their top tier capital ratios and leverage exposures, rather than agency problems occurring over their relationships with clients. We do not examine this additional source of procyclicality to keep the analysis as simple as we can.

liquidity injection, and (ii) higher incentives given to banks to refinance the real sector, due to the liquidity buffers supplied by the central bank.

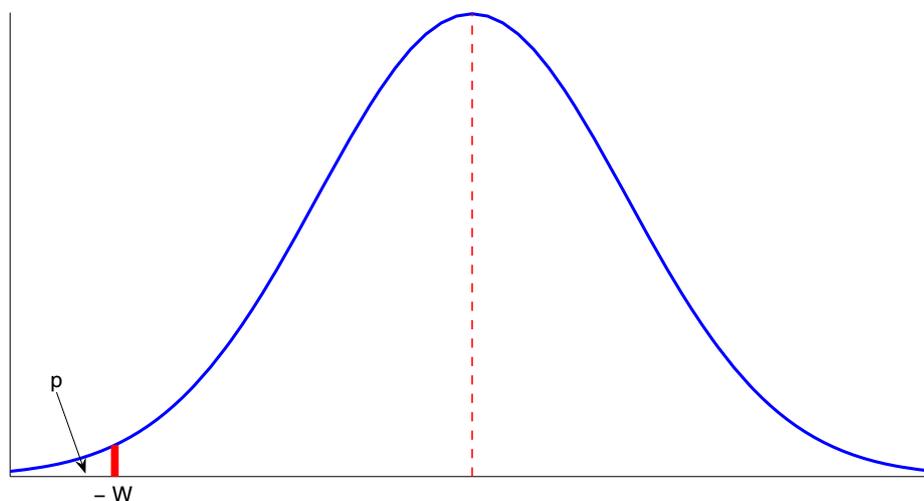
13.6 A few hints on the risk-management practice

13.6.1 Value at Risk (VaR)

We need to review Value at Risk (VaR), in general. VaR is a method of assessing risk that uses statistical techniques. Useful for supervision and management of financial risks. Origins: reaction to financial disasters in the early 1990s involving Orange County, Barings, Metallgesellschaft, Daiwa, etc.

DEFINITION I: *VaR measures the worst expected loss over a given horizon under normal market conditions at a given confidence level.*

DEFINITION II: *We are $(1 - p)\%$ certain that a given portfolio will not suffer of a loss larger than $\mathcal{L}W$ over the next N weeks, $\Pr(\text{Loss} < -W) = p$. That is, $\mathcal{L}\text{VaR}_p = \mathcal{L}W$.*



Equivalently, note that

$$\frac{\text{Loss}}{V_0} = \frac{\Delta V}{V_0} = \text{portfolio return}$$

where ΔV denotes the change in value of the portfolio over the next N days, and $\mathcal{L}V_0$ is the current value of the portfolio. Hence,

$$p = \Pr(\text{Loss} < -\text{VaR}_p) = \Pr\left(\frac{\Delta V}{V_0} < -\frac{\text{VaR}_p}{V_0}\right).$$

This formulation leads us to the following alternative definition:

DEFINITION III: We are $(1 - p)\%$ certain that a given portfolio will not experience a relative loss larger than $\frac{\text{VaR}_p}{V_0}$ over the next N weeks.

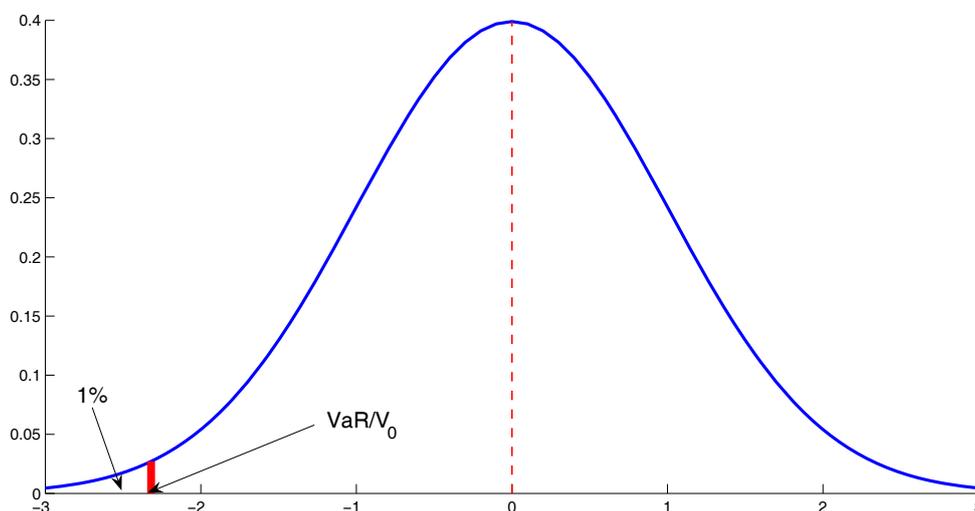
So in practice, we shall have to find the relative loss, ℓ_p , for a given confidence p , as follows:

$$p = \Pr\left(\frac{\Delta V}{V_0} < -\ell_p\right), \quad \text{where } \ell_p = \frac{\text{VaR}_p}{V_0}.$$

The corresponding VaR_p is just

$$\text{VaR}_p = \ell_p \cdot V_0$$

For example, suppose that the portfolio return over the next 2 weeks, $\frac{\Delta V}{V_0}$, is normally distributed with mean zero and unit variance. We know that $0.01 = \Pr(\frac{\Delta V}{V_0} < -2.32)$. Hence, $\text{VaR}_p = 2.32 \cdot V_0$.



We are 99% certain that our portfolio will not suffer of a *loss* larger than -2.32 times its current value over the next 2 weeks. We are 99% certain that our portfolio will not experience a *relative loss* larger than -2.32 over the next 2 weeks.

As a second example note that the previous assumption about the portfolio return was extreme. Assume, instead, the portfolio return over the next 2 weeks, $\frac{\Delta V}{V_0}$, is normally distributed with mean zero and variance $\sigma^2 = \frac{2}{52}\sigma_{\text{year}}^2$, where σ_{year}^2 is the annualized variance. We assume that $\sigma_{\text{year}}^2 = 0.15^2$. We have to re-scale the previous formulas, as follows. First, we introduce a variable $\tilde{\epsilon} \sim N(0, 1)$, i.e. $\tilde{\epsilon}$ is normally distributed with mean zero and variance = 1. So we can write,

$$\frac{\Delta V}{V_0} \stackrel{d}{=} \tilde{\epsilon} \cdot \sigma \sim N(0, \sigma^2),$$

and, hence,

$$0.01 = \Pr(\tilde{\epsilon} < -2.32) = \Pr(\Delta V < -2.32 \cdot V_0 \cdot \sigma),$$

whence, $\text{VaR}_p = 2.32 \cdot V_0 \cdot \sigma$. We know the annualized variance, $\sigma_{\text{year}}^2 = 0.15^2$, from which we can derive the two-week standard deviation, $\sigma^2 = \frac{2}{52} \sigma_{\text{year}}^2 \approx 0.03^2$, and, hence, $\frac{\text{VaR}_p}{V_0} = 2.32 \cdot \sigma = 2.32 \cdot 0.03 \approx 7\%$. That is, we are 99% certain that our portfolio will not suffer of a *loss* larger than 7% times its current value over the next 2 weeks. We are 99% certain that our portfolio will not experience a *relative loss* larger than 7% over the next 2 weeks.

More generally, we may assume the portfolio return over the next 2 weeks, $\frac{\Delta V}{V_0}$, is normally distributed with mean μ and variance σ^2 . In this case,

$$\frac{\Delta V}{V_0} \stackrel{d}{=} \mu + \tilde{\epsilon} \cdot \sigma \sim N(\mu, \sigma^2),$$

and, hence,

$$0.01 = \Pr(\tilde{\epsilon} < -2.32) = \Pr(\Delta V < -V_0 \cdot (2.32 \cdot \sigma - \mu))$$

whence, $\text{VaR}_p = V_0 \cdot (2.32 \cdot \sigma - \mu)$. In practice, μ is very small if the horizon is as short as two weeks.

13.6.1.1 Challenges to VaR

Challenges related to *distributional assumptions, nonlinearities, or conceptual difficulties*.

Distributional assumptions

The assumption that data are generated by a normal distribution does not describe asset returns well. Chapters 10 and 11 explain that we need ARCH effects, stochastic volatility and multifactor models. More generally, data can exhibit changes in regimes, nonlinearities and fat tails. Fat tails are particularly important to understand, since this is what we're interested in after all. More in general, it is quite challenging to understand what the data generating process is, especially in so far as we consider portfolios of assets. Asset returns and volatilities are typically correlated, with correlation rising in bad times—correlation is stochastic.

We may make distributional assumptions but then, these assumptions have to be carefully assessed through, for example, backtesting (to be explained below). We may proceed with nonparametric methods, and this is indeed a promising avenue, but with its caveats.

How do nonparametric methods work? These methods rely on an old and idea, which is to estimate the data distribution through histograms. These histograms can be readily used to compute VaR. This approach is nonparametric in nature, as it does not rely on any model. A more refined method replaces “rough” histograms with “smoothed” histograms, as follows. Suppose to have access to a time series of data x_n , which are drawn from a certain probability law, with density $f(x)$. We may define the following estimate of the density $f(x)$,

$$\hat{f}_N(x) = \frac{1}{N} \sum_{n=1}^N \frac{1}{\lambda} K\left(\frac{x - x_n}{\lambda}\right),$$

where N is the sample size, and K is some symmetric function integrating to one. We may think of $\hat{f}_N(x)$ as a smoothed histogram, with window bin equal to λ . It is possible to show that as N goes to infinity and λ goes to zero at a certain rate, $\hat{f}_N(x)$ converges “in probability” to $f(x)$, for all x . But we are not done, since there are not obvious rules to choose λ and K ? The choice of λ is notoriously difficult. Unfortunately, the “bias,” $\hat{f}_N(x) - f(x)$, tends to be large exactly on the tails of $f(x)$, which do represent the region we're interested in. In general, we can use Montecarlo simulations out of a smoothed density like this to compute VaR.

Nonlinearities

Finally, portfolios of assets can behave in a nonlinear fashion, especially when the portfolio contains derivatives. In general, the value of a portfolio including M assets is,

$$P = \sum_{j=1}^M \alpha_j S_j,$$

where α_i is the number of the i -th asset in the portfolio, and S_i is the price of the i -th asset in the portfolio. Holding α_i constant, the variation on the portfolio return is simply a weighed average of all the asset returns,

$$\Delta P_T \equiv P_T - P_t = \sum_{j=1}^M \alpha_j \Delta S_j \iff \frac{\Delta P}{P} = \sum_{j=1}^M \left(\frac{\alpha_j S_j}{P} \right) \frac{\Delta S_j}{S_j},$$

where the variations relate to any time interval. Often, the prices S_i are rational functions of the state variables, or are interlinked through arbitrage restrictions. Use factors to determine the risk associated with fixed income securities. When the horizon of the VaR is large, it is unlikely that α_i is constant. Typically, we shall need to go for numerical methods, based, for example, on Monte Carlo simulations. So all in all, we need to have a careful understanding of the derivatives in the book, and proceed with back testing and stress testing.

VaR as an appropriate measure of risk

There are technical difficulties with the very definition of VaR. VaR suffers from some statistic-theoretic foundation. VaR tells us that 1% of the time, losses will exceed the VaR figure, but it does not tell us the entity of the loss. So we need to compute the *expected shortfall*. Any risk measure should enjoy a number of sensible properties. Artzner et al. (1999) have noted a number of properties, and showed that VaR does not enjoy the so-called *subadditivity* property, according to which the sum of the risk measures for any two portfolios should be larger than the risk measure for the sum of the two portfolios. VaR doesn't satisfy the subadditivity property, but expected shortfall does satisfy the subadditivity property.

13.6.2 Backtesting

How well the VaR estimate would have performed in the past? How often the loss in a given sample exceeded the reference-period 99% VaR? If the exceptions occur more than 1% of the time, there is evidence that the models leading to VaR estimates are “misspecified”—a nice word for saying “bad” models.

The mechanics of backtesting is as follows. Suppose the models leading to the VaR are “good”. By construction, the probability the VaR number is exceeded in any reference period is p , where p is the coverage rate for the VaR. Next, we go to our sample, which we assume it comprises N days, and let M be the number of days the VaR is exceeded. We wish to test whether the number of exceptions we observe in the sample “conforms” to the expected number of exceptions based on the VaR. For example, it might be that the number of exceptions we have observed, M , is larger than the expected number of exceptions, $p \cdot N$. We want to make sure this circumstance arose due to sample variability, rather than model misspecification. A simple one-tail test is described below.

Let us compute the probability that in N days, the VaR is exceeded for M or more days. Assuming exceptions are binomially distributed, this probability is,

$$\Pi_p = \sum_{k=M}^N \frac{N!}{k!(N-k)!} p^k (1-p)^{N-k}.$$

Then, we can say the following. If $\Pi_p \leq 5\%$ (say), we reject the hypothesis that the probability of exceptions is p at the 5% level—the models we're using are misspecified. If $\Pi_p > 5\%$ (say), we cannot reject the hypothesis that the probability of exceptions is p at the 5% level—we can't say the models we're using are misspecified. This test is reviewed in more detail by Hull (2007, p. 208). Other tests are reviewed by Christoffersen (2003, p. 184).

13.6.3 Stress testing

Stress testing is a technique through which we generate artificial data from a range of possible *scenarios*. Stress scenarios help cover a range of factors that can create extraordinary losses or gains in trading portfolios, or make the control of risk in those portfolios very difficult. These factors include low-probability events in all major types of risks, including the various components of credit, market, and operational risks. Stress scenarios need to shed light on the impact of such events on positions that display both linear and nonlinear price characteristics (i.e. options and instruments that have options-like characteristics).

Possible scenarios include simulating (i) shocks that although rare or even absent from the historical database at hand, are likely to happen anyway; and (ii) shocks leading to structural breaks and/or smooth transition in the data generating mechanism. One possible example is to set the percentage changes in all market variables in the portfolio equal to the worst percentage changes having occurred in ten days in a row during the subprime crisis 2007-2008.

This example on the subprime crisis is related to the *historical simulation approach* to generate scenarios. This approach consists can be explained through a single formula. Let v_t the value of some market variable i in day t in our sample, where $t = 0, \dots, T$ (say). We can generate T scenarios for the next day, $T + 1$, as follows.

- (i) The first scenario is that in which each variable grows by the same amount it grew at time 1,

$$v_{T+1} = v_T \cdot \frac{v_1}{v_0}.$$

- (ii) The second scenario is that in which each variable grows by the same amount it grew at time 2,

$$v_{T+1} = v_T \cdot \frac{v_2}{v_1}.$$

- (iii) ...

- (iv) The T -th scenario is that in which each variable grows by the same amount it grew at time T ,

$$v_{T+1} = v_T \cdot \frac{v_T}{v_{T-1}}.$$

- (v) The T scenarios are generated for all the market variables, which would give us an artificial multivariate sample of T observations. We can use this sample for many things, including VaR.

13.6.4 Credit risk and VaR

We can use the tools in Section 13.2 to assess the likelihood of default for a given name. The important thing to do is to use the physical probability of default, not the risk neutral one. The risk neutral probability of default is likely to be larger than the physical one. Therefore, using the risk neutral probability leads to too conservative estimates.

VaR for credit risks pose delicate issues as well. The key issue is the presence of default correlation. In practice, defaults among names or loans are likely to be correlated, for many reasons. First, there might be direct relationships or, more generally, network effects, among names. Second, firms performance could be driven by common economic conditions, as in the one factor model which we now describe. This one factor model, developed by Vasicek (1987), is at the heart of Basel II. In the appendix, we provide additional technical details about how this model is related to a modeling tool known as copulae functions. We now proceed to develop this model in an intuitive manner. Let us define the following variable:

$$z_i = \sqrt{\rho}F + \sqrt{1 - \rho}\epsilon_i, \quad (13.72)$$

where F is a common factor among the names in the portfolio, ϵ_i is an idiosyncratic term, and $F \sim N(0, 1)$, $\epsilon_i \sim N(0, 1)$. As we explain in the Appendix, $\rho \geq 0$ is meant to capture the default correlation among the names.

Next, assume that the *physical* probability each firm defaults, by T , say $\mathbb{P}(T)$, is the same for each firm within the same class of risk, and given by,

$$\mathbb{P}(T) = \Phi(\zeta_{\text{PD}}) \equiv \text{PD},$$

where PD is the probability of default, and Φ is the cumulative distribution of a standard normal variable. That is, by time T , each firm defaults any time that,

$$z_i < \zeta_{\text{PD}} \equiv \Phi^{-1}(\text{PD}),$$

where Φ^{-1} denotes the inverse of Φ . One economic interpretation of Eq. (13.72) is that z_i is the value of a firm and that the firm defaults whenever this value hits some exogenously given barrier ζ_{PD} .

Conditionally upon the realization of the macroeconomic factor F , the probability of default for each firm is,

$$p(F) \equiv \Pr(\text{Default} | F) = \Phi\left(\frac{\Phi^{-1}(\text{PD}) - \sqrt{\rho}F}{\sqrt{1 - \rho}}\right). \quad (13.73)$$

By the law of large numbers, this is quite a good approximation to the default rate for a portfolio of a large number of assets falling within the same class of risk.

We see that this conditional probability is decreasing in F : the larger the level of the common macroeconomic factor, the smaller the probability each firm defaults. Hence, we can fix a value of F such that $\Pr(\text{Default} | F) = \text{Default rate}$ is what we want. Note, the probability F is larger than $-\Phi^{-1}(x)$ is just x ! Formally,

$$\Pr(F > -\Phi^{-1}(x)) = \Pr(-F < \Phi^{-1}(x)) = \Phi(\Phi^{-1}(x)) = x.$$

Then, with probability x , the default rate will not exceed

$$\text{VaR}_{\text{Credit Risk}}(x) = \Phi\left(\frac{\Phi^{-1}(\text{PD}) + \sqrt{\rho}\Phi^{-1}(x)}{\sqrt{1 - \rho}}\right).$$

It is easy to see that $\text{VaR}_{\text{Credit Risk}}(x)$ increases with ρ . Basel II sets $x = 0.999$ and, accordingly, it imposes a capital requirement equal to,

$$\text{Loss-given-default} * [\text{VaR}_{\text{Credit Risk}}(0.999) - \text{PD}] * \text{Maturity adjustment}.$$

The reason Basel II requires the term $\text{VaR}_{\text{Credit Risk}}(0.999) - \text{PD}$, rather than just $\text{VaR}_{\text{Credit Risk}}$, is that what is really needed here is the capital in excess of the 99.9% worst case loss over the expected idiosyncratic loss, PD. Well functioning capital markets should already discount the idiosyncratic losses.

Finally, Basel II requires banks to compute ρ through a formula in which ρ is inversely related to PD. The formula is based on empirical research (see Lopez, 2004): for a firm which becomes less creditworthy, the PD increases and its probability of default becomes less affected by market conditions. Basel II requires banks to compute a maturity adjustment factor that takes into account that the longer the maturity the more likely it is a given name might eventually migrate towards a more risky asset class.

The previous model can be further elaborated. We ask: (i) What is the unconditional probability of defaults, and (ii) what is the density function of the fraction of defaulting loans?

First, note that conditionally upon the realization of the macroeconomic factor F , defaults are obviously independent, being then driven by the idiosyncratic terms ϵ_i in Eq. (13.72). Given N loans, and the realization of the macroeconomic factor F , these defaults are binomially distributed as:

$$\Pr(\text{No of defaults} = n | F) = \binom{N}{n} p(F)^n (1 - p(F))^{N-n},$$

where $p(F)$ is as in Eq. (13.73). Therefore, the unconditional probability of n defaults is:

$$\Pr(\text{No of defaults} = n) = \int_{-\infty}^{\infty} \Pr(\text{No of defaults} = n | F) \phi(F) dF,$$

where ϕ denotes the standard normal density. This formula provides a valuable tool analysis in risk-management. It can be shown that VaR levels increase with the correlation ρ .

Next, let ω denote the fraction of defaulting loans. For a large portfolio of loans, $\omega = p(F)$, such that:

$$\Pr(\omega \leq x) = \int_{-\infty}^{\infty} \Pr(\omega \leq x | F) \phi(F) dF = \int_{-\infty}^{\infty} \mathbb{I}_{p(F) \leq x} \phi(F) dF = \Phi(F^*), \quad (13.74)$$

where \mathbb{I} denotes the indicator function, and F^* satisfies, by Eq. (13.73), $-F^* : x = p(-F^*) = \Phi\left(\frac{\Phi^{-1}(\text{PD}) + \sqrt{\rho}F^*}{\sqrt{1-\rho}}\right)$. Solving for F^* leaves:

$$F^* = \frac{\sqrt{1-\rho}\Phi^{-1}(x) - \Phi^{-1}(\text{PD})}{\sqrt{\rho}}.$$

It is the threshold value taken by the macroeconomic factor that guarantees a frequency of defaults ω less than x . Replacing F^* into Eq. (13.74) delivers the cumulative distribution function for ω . The density function $f(x)$ for the frequency of defaults is then:

$$f(x) = \sqrt{\frac{1-\rho}{\rho}} e^{\frac{1}{2}(\Phi^{-1}(x))^2 - \frac{1}{2\rho}(\sqrt{1-\rho}\Phi^{-1}(x) - \Phi^{-1}(\text{PD}))^2}.$$

13.7 Appendix 1: Present values contingent on future bankruptcy

The value of debt in Leland's (1994) model can be written as:

$$D(A) = \mathbb{E} \left(\int_0^{T_B} e^{-rs} C ds \right) + \mathbb{E} [e^{-rT_B} (1 - \alpha) A^B], \quad (13A.1)$$

where T_B is the time at which the firm is liquidated. Eq. (13A.1) simply says that the value of debt equals the expected coupon payments plus the expected liquidation value of the bond. We have:

$$\mathbb{E} (e^{-rT_B}) = \int_0^\infty e^{-rt} f(t; A, A^B) dt \equiv p_B(A), \quad (13A.2)$$

where $f(t; A, A^B)$ denotes the density of the first passage time from A to A^B . It can be shown that $p_B(A)$ is exactly as in Eq. (13.11) of the main text. Similarly,

$$\begin{aligned} \mathbb{E} \left(\int_0^{T_B} e^{-rs} C ds \right) &= C \cdot \mathbb{E} \left(\int_0^{T_B} e^{-rs} ds \right) \\ &= C \cdot \int_0^\infty \left(\int_0^t e^{-rs} ds \right) f(t; A, A^B) dt \\ &= C \cdot \int_0^\infty \frac{1 - e^{-rt}}{r} f(t; A, A^B) dt \\ &= \frac{C}{r} \cdot (1 - p_B(A)). \end{aligned} \quad (13A.3)$$

Replacing Eq. (13A.2)-(13A.3) into Eq. (13A.1) yields Eq. (13.10).

13.8 Appendix 2: Proof of selected results

ALTERNATIVE DERIVATION OF EQ. (13.21). Under the risk-neutral probability, the expected change of any bond price must equal zero when the safe short-term rate is zero,

$$\frac{\partial B(t)}{\partial t} + \lambda(\text{Rec} - B(t)) = rB = 0, \quad \text{with } B(T) = N,$$

where the first term, $\frac{\partial B(t)}{\partial t}$, reflects the change in the bond price arising from the mere passage of time, and $\lambda(\text{Rec} - B(t))$ is the expected change in the bond price, arising from the event of default, i.e. the probability of a sudden default arrival, λ , times the consequent jump in the bond price, $\text{Rec} - B(t)$.

The solution to the previous equation is,

$$B(0) = \int_0^T \text{Rec} \cdot \underbrace{\lambda e^{-\lambda t}}_{=\text{Pr}\{\text{Default at } t\}} dt + Ne^{-\lambda T},$$

which is Eq. (13.21).

PROOF OF EQ. (13.22). The spread is given by:

$$s(T) = -\frac{1}{T} \ln \left(\frac{\text{Rec}_T (1 - e^{-\lambda T}) + Ne^{-\lambda T}}{N} \right).$$

With $N = 1$, and $\text{Rec}_T = R \cdot e^{-\kappa T}$, we have,

$$s(T) = -\frac{1}{T} \ln \left(Re^{-\kappa T} (1 - e^{-\lambda T}) + e^{-\lambda T} \right) = \lambda - \frac{1}{T} \ln \left(Re^{-(\kappa-\lambda)T} (1 - e^{-\lambda T}) + 1 \right),$$

or equivalently,

$$s(T) = -\frac{1}{T} \ln \left(Re^{-\kappa T} (1 - e^{-\lambda T}) + e^{-\lambda T} \right) = \kappa - \frac{1}{T} \ln \left(R (1 - e^{-\lambda T}) + e^{-(\lambda-\kappa)T} \right),$$

Therefore, if $\kappa \geq \lambda$, then, $\lim_{T \rightarrow \infty} s(T) = \lambda$, and if $\kappa \leq \lambda$, $\lim_{T \rightarrow \infty} s(T) = \kappa$.

13.9 Appendix 3: Transition probability matrices and pricing

Consider the matrix $P(T - t)$ for $T - t \equiv \Delta t$, $P(\Delta t)$, and write,

$$P(\Delta t)_{ij} \equiv \begin{cases} 1 + \lambda_{ij}\Delta t, & i = j \\ \lambda_{ij}\Delta t, & i \neq j \end{cases} \quad (13A.4)$$

We are defining the constants λ_{ij} as they were the counterparts of the intensity of the Poisson process in Eq. (13.20). Accordingly, these constants are simply interpreted as the instantaneous probabilities of migration from rating i to rating j over the time interval Δt . Naturally, for each i , we have that $\sum_{j=1}^Z P(\Delta t)_{ij} = 1$, and using into Eq. (13A.4), we obtain,

$$\lambda_{ii} = - \sum_{j=1, j \neq i}^Z \lambda_{ij}. \quad (13A.5)$$

The matrix Λ containing the elements λ_{ij} defined in Eqs. (13A.4) and (13A.5) is called the *generating matrix*.

Next, let us rewrite Eq. (13A.4) in matrix form,

$$P(\Delta t) = I + \Lambda \Delta t.$$

Suppose we have a time interval $[0, T]$, which we chop into n pieces, so to have $\Delta t = \frac{T}{n}$. We have,

$$P(T) = P(\Delta t)^n = \left(I + \Lambda \frac{T}{n} \right)^n.$$

For large n ,

$$P(T) = \exp(\Lambda T), \quad (13A.6)$$

the matrix exponential, defined as, $\exp(\Lambda T) \equiv \sum_{n=0}^{\infty} \frac{(\Lambda T)^n}{n!}$.

To evaluate derivatives “written on states,” we proceed as follows. Suppose F_i is the price of derivative in state $i \in \{1, \dots, Z\}$. Suppose the Markov chain is the only source of uncertainty relevant for the evaluation of this derivative. Then,

$$dF_i = \frac{\partial F_i}{\partial t} dt + [F_{\tilde{R}} - F_i],$$

where $\tilde{R} \in \{1, \dots, Z\}$, with the usual conditional probabilities. In words, the instantaneous change in the derivative value, dF_i , is the sum of two components: one, $\frac{\partial F_i}{\partial t} dt$, related to the mere passage of time, and the other, $[F_{\tilde{R}} - F_i]$, related to the discrete change arising from a change in the rating.

Suppose that $r = 0$. Then,

$$rF_i = 0 = \frac{E(dF_i)}{dt} = \frac{\partial F_i}{\partial t} + \sum_{j=1}^Z \lambda_{ij} [F_j - F_i] = \frac{\partial F_i}{\partial t} + \sum_{j \neq i} \lambda_{ij} [F_j - F_i],$$

with the appropriate boundary conditions.

As an example, consider defaultable bonds. In this case, we may be looking for pricing functions having the following form,

$$F_i(T - t) = xQ_i(T - t) + 1 - Q_i(T - t),$$

and then solve for $Q_i(T - t)$, for all $i \in \{1, \dots, Z\}$. Naturally, we have

$$\begin{aligned} 0 &= xQ'_i - Q'_i + \sum_{j \neq i} \lambda_{ij} [x(Q_j - Q_i) - (Q_j - Q_i)] \\ &= x \left[Q'_i + \sum_{j \neq i} \lambda_{ij} (Q_j - Q_i) \right] - \left[Q'_i + \sum_{j \neq i} \lambda_{ij} (Q_j - Q_i) \right], \end{aligned}$$

which holds if and only if,

$$Q'_i = - \sum_{j \neq i} \lambda_{ij} (Q_j - Q_i) = - \sum_{j \neq i} \lambda_{ij} Q_j + \sum_{j \neq i} \lambda_{ij} Q_i = - \left[\sum_{j \neq i} \lambda_{ij} Q_j + \lambda_{ii} Q_i \right].$$

That is, $Q' = -\Lambda Q$, which solved through the appropriate boundary conditions, yields precisely Eq. (13A.6).

13.10 Appendix 4: Bond spreads in markets with stochastic default intensity

We derive Eq. (13.34), by relying on the pricing formulae of Chapter 12. If the short-term is constant, the price of a defaultable bond derived in Section 13.4.7 of Chapter 12 can easily be extended to, with the notation of the present chapter,

$$P(y, N) = e^{-rN} \mathbb{E} \left[e^{-\int_0^N \lambda(t) dt} \right] + \int_0^N e^{-rt} \underbrace{\mathbb{E} \left[\lambda(t) e^{-\int_0^t \lambda(u) du} \right]}_{=\Pr\{\text{Default} \in (t, t+dt)\}} \text{Rec}(t) dt. \quad (13A.7)$$

The term indicated inside the integral of the second term, is indeed the density of default time at t , because,

$$P_{\text{default by time } t}(\lambda) = 1 - \mathbb{E} \left[e^{-\int_0^t \lambda(s) ds} \right],$$

such that by differentiating with respect to t , yields, under the appropriate regularity conditions, that $\Pr\{\text{Default} \in (t, t+dt)\}$ is just the term indicated in Eq. (13A.7). So Eq. (13.34) follows. Naturally,

$$\Pr\{\text{Default} \in (t, t+dt)\} = -\frac{\partial}{\partial t} P_{\text{surv}}(\lambda, t).$$

Replacing this into Eq. (13A.7),

$$\begin{aligned} P(y, N) &= e^{-rN} \mathbb{E} \left[e^{-\int_0^N \lambda(t) dt} \right] + \text{Rec} \int_0^N e^{-rt} \left[-\frac{\partial}{\partial t} P_{\text{surv}}(\lambda, t) \right] dt \\ &= 1 - \text{LGD} (1 - e^{-rN} P_{\text{surv}}(\lambda, N)) - (1 - \text{LGD}) \int_0^N r e^{-rt} P_{\text{surv}}(\lambda, t) dt, \end{aligned}$$

where the second equality follows by integration by parts and the assumption of constant recovery rates. Setting $r = 0$, produces Eq. (13.35).

13.11 Appendix 5: Conditional probabilities of survival

We prove Eqs. (13.37)-(13.39). First, for (t_{i-1}, t_i) small, the numerator in Eq. (13.36) can be replaced by

$$-\frac{\partial}{\partial t} P_{\text{surv}}(\lambda, t) \equiv \mathbb{E} \left[\lambda(t) e^{-\int_0^t \lambda(s) ds} \right],$$

and rescaled by dt . Regularity conditions under which we can perform this differentiation can be found in a related context developed in Mele (2003). Eqs. (13.37)-(13.38) follow.

As for Eq. (13.39), the proof follows the same lines of reasoning as that in Appendix 3 of Chapter 12. That is, we can define a density process,

$$\eta_T(\tau) = \frac{e^{-\int_0^\tau \lambda(s) ds} P_{\text{surv}}(\lambda(\tau), \tau, T)}{\mathbb{E} \left[e^{-\int_0^T \lambda(s) ds} \right]}, \quad P_{\text{surv}}(\lambda(\tau), \tau, T) \equiv \mathbb{E} \left[e^{-\int_\tau^T \lambda(s) ds} \middle| \mathcal{F}_\tau \right].$$

It is easy to show that the drift of P_{surv} is $\lambda(\tau) d\tau$, such that by Itô's lemma,

$$\frac{d\eta_T(\tau)}{\eta_T(\tau)} = -[-\text{Vol}(P_{\text{surv}}(\lambda(\tau), \tau, T))] dW(\tau),$$

where,

$$-\text{Vol}(P_{\text{surv}}(\lambda(\tau), \tau, T)) \equiv -\frac{\frac{\partial}{\partial \lambda} P_{\text{surv}}(\lambda(\tau), \tau, T)}{P_{\text{surv}}(\lambda(\tau), \tau, T)} \sigma \sqrt{\lambda(\tau)} = B(T - \tau) \sigma \sqrt{\lambda(\tau)},$$

where the second line follows by the closed-form expression of P_{surv} in Eq. (13.31). Therefore, $W_\lambda(\tau)$ is a Brownian motion under Q_λ , where

$$dW_\lambda(\tau) = dW(\tau) + B(T - \tau) \sigma \sqrt{\lambda(\tau)} d\tau,$$

and Eq. (13.39) follows.

13.12 Appendix 6: Details regarding CDS index swaps and swaptions

We prove Eq. (13.51). Note that the expectation of the first term in Eq. (13.48), conditional on the information set τ , for $\tau \leq t_0$ is, now, for $i = 1, \dots, 4M$,

$$\begin{aligned} & \text{LGD} \frac{1}{n} \sum_{j=1}^n \mathbb{E}_\tau \left[e^{-\int_\tau^{t_i} r(\tau) d\tau} \mathbb{I}_{\{\text{Surv}_j \text{ at } \tau\}} \mathbb{I}_{\{\text{Def}_j \in (t_{i-1}, t_i)\}} \right] \\ &= \text{LGD} \frac{1}{n} \sum_{j=1}^n \mathbb{I}_{\{\text{Surv}_j \text{ at } \tau\}} \mathbb{E}_\tau \left[e^{-\int_\tau^{t_i} r(\tau) d\tau} \mathbb{I}_{\{\text{Def}_j \in (t_{i-1}, t_i)\}} \right] = \text{LGD} \cdot N(\tau) \mathbb{E}_\tau \left[e^{-\int_\tau^{t_i} r(\tau) d\tau} \mathbb{I}_{\{\text{Def}_j \in (t_{i-1}, t_i)\}} \right], \end{aligned} \quad (13A.8)$$

where the last equality follows by the definition of the outstanding notional value in Eq. (13.52), and the fact that the expectation in the first equality is the same for each name j , due to the assumption that all names in the index have the same credit quality. Summing over the reset dates, $i = 1, \dots, 4M$, delivers the first term in Eq. (13.51). The second term in Eq. (13.51) follows by elaborating the time τ expectation of the second term in Eq. (13.48),

$$\begin{aligned} & \mathbb{E}_\tau \left[e^{-\int_\tau^{t_i} r(\tau) d\tau} \cdot \mathbb{I}_{\{\text{Surv}_j \text{ at } t_i\}} \right] \\ &= \mathbb{E}_\tau \left[e^{-\int_\tau^{t_i} r(\tau) d\tau} \cdot \mathbb{I}_{\{\text{Surv}_j \text{ at } \tau\}} \mathbb{I}_{\{\text{Surv}_j \text{ at } t_i | \text{Surv}_j \text{ at } \tau\}} \right] = \mathbb{I}_{\{\text{Surv}_j \text{ at } \tau\}} \mathbb{E}_\tau \left[e^{-\int_\tau^{t_i} r(\tau) d\tau} \mathbb{I}_{\{\text{Surv}_j \text{ at } t_i | \text{Surv}_j \text{ at } \tau\}} \right], \end{aligned}$$

and summing over the reset dates and all names j , using the definition of V_{1t}^x in Eq. (13.50), and noting, again, that the expectation in the second equality is the same for all the assets.

We now derive the value of the *front-end protection* in Eq. (13.53). Note that the derivation of Eq. (13A.8) relies on default events occurring after the swap origination, i.e. over the reset dates, after $T = t_0$. In evaluating the front-end protection, we need to price securities that pay off over defaults possibly occurring over the life of the swaption, i.e. before time $T = t_0$. We have,

$$\begin{aligned} V_\tau^F &= \mathbb{E}_\tau \left[e^{-\int_\tau^T r(u) du} F_T \right] \\ &= \text{LGD} \frac{1}{n} \mathbb{E}_\tau \left[e^{-\int_\tau^T r(u) du} \sum_{j=1}^n \left(\mathbb{I}_{\{\text{Def}_j \in (t, \tau)\}} + \mathbb{I}_{\{\text{Surv}_j \text{ at } \tau\}} \mathbb{I}_{\{\text{Def}_j \in (\tau, T)\}} \right) \right] \\ &= \text{LGD} \frac{1}{n} \mathcal{D}(t, \tau) P(\tau, T) + \text{LGD} \frac{1}{n} \sum_{j=1}^n \mathbb{I}_{\{\text{Surv}_j \text{ at } \tau\}} \mathbb{E}_\tau \left[e^{-\int_\tau^T r(u) du} \left(1 - \mathbb{I}_{\{\text{Surv}_j \text{ at } T | \text{Surv}_j \text{ at } \tau\}} \right) \right] \\ &= \text{LGD} \left(\frac{1}{n} \mathcal{D}(t, \tau) P(\tau, T) + N(\tau) (P(\tau, T) - P_{\text{def}}(\tau, T)) \right), \end{aligned}$$

where the third equality holds by the assumption that the names have the same credit quality, and $P(\tau, T) = \mathbb{E}_\tau[e^{-\int_\tau^T r(u) du}]$ and $P_{\text{def}}(\tau, T) = \mathbb{E}_\tau[e^{-\int_\tau^T (r(u) + \lambda(u)) du}]$. Note that the first term in the brackets of the second equality is, obviously, always zero, when the timing of possible defaults does not overlap with the evaluation horizon, as for Eq. (13A.8).

Next, we show that the survival contingent measure \hat{Q}_{sc} defined in Eq. (13.55) does integrate to one, and that $\text{CDX}_t(M) = \text{LGD} \frac{V_{0\tau}}{V_{1\tau}} + \frac{V_\tau^F}{N(\tau)V_{1\tau}}$ in Eq. (13.54), is a martingale under \hat{Q}_{sc} . We shall need the equality summarized by the following lemma:

LEMMA 13A.1. *The following equality holds true: $\mathbb{E}_\tau [N(T) | \mathbb{F}^r(T)] = N(\tau) e^{-\int_\tau^T \lambda(u) du}$.*

PROOF. We have:

$$\begin{aligned}\mathbb{E}_\tau [N(T) | \mathbb{F}^r(T)] &= \mathbb{E}_\tau \left[\frac{1}{n} \sum_{j=1}^n \mathbb{I}_{\{\text{Surv}_j \text{ at } T\}} \middle| \mathbb{F}^r(T) \right] \\ &= \frac{1}{n} \sum_{j=1}^n \mathbb{I}_{\{\text{Surv}_j \text{ at } \tau\}} \mathbb{E}_\tau \left[\mathbb{I}_{\{\text{Surv}_j \text{ at } T | \text{Surv}_j \text{ at } \tau\}} \middle| \mathbb{F}^r(T) \right] \\ &= N(\tau) e^{-\int_\tau^T \lambda(u) du}. \quad \parallel\end{aligned}$$

As for the survival contingent measure \hat{Q}_{sc} , we have that under regularity conditions,

$$\begin{aligned}\mathbb{E}_\tau \left[e^{-\int_\tau^T r(u) du} N(T) V_{1T} \right] &= \mathbb{E}_\tau \left[\mathbb{E}_\tau \left(e^{-\int_\tau^T r(u) du} N(T) V_{1T} \middle| \mathbb{F}^r(T) \right) \right] \\ &= N(\tau) \mathbb{E}_\tau \left[e^{-\int_\tau^T (r(u) + \lambda(u)) du} V_{1T} \right] \\ &= N(\tau) V_{1\tau},\end{aligned}$$

where the last equality follow by the definition of Q_{sc} in Eq. (13.46). Therefore, $\int d\hat{Q}_{\text{sc}} = 1$.

As for the martingale property of $\text{CDX}_t(M)$ under \hat{Q}_{sc} , let $\hat{\mathbb{E}}_\tau^{\text{sc}}[\cdot]$ the time τ conditional expectation operator under the the survival contingent measure \hat{Q}_{sc} . We have, using the definition of \hat{Q}_{sc} in Eq. (13.55),

$$\begin{aligned}\hat{\mathbb{E}}_\tau^{\text{sc}} [\text{CDX}_T(M)] &= \text{LGD} \cdot \hat{\mathbb{E}}_\tau^{\text{sc}} \left[\frac{V_{0T}}{V_{1T}} \right] + \hat{\mathbb{E}}_\tau^{\text{sc}} \left[\frac{V_T^{\text{F}}}{N(T) V_{1T}} \right] \\ &= \text{LGD} \cdot \mathbb{E}_\tau \left[e^{-\int_\tau^T r(u) du} \frac{N(T) V_{1T}}{N(\tau) V_{1\tau}} \frac{V_{0T}}{V_{1T}} \right] + \mathbb{E}_\tau^{\text{sc}} \left[e^{-\int_\tau^T r(u) du} \frac{N(T) V_{1T}}{N(\tau) V_{1\tau}} \frac{V_T^{\text{F}}}{N(T) V_{1T}} \right] \\ &= \text{LGD} \frac{1}{V_{1\tau}} \cdot \mathbb{E}_\tau \left[e^{-\int_\tau^T r(u) du} \frac{N(T)}{N(\tau)} V_{0T} \right] + \frac{1}{N(\tau) V_{1\tau}} \mathbb{E}_\tau \left[e^{-\int_\tau^T r(u) du} V_T^{\text{F}} \right] \\ &= \text{LGD} \frac{V_{0\tau}}{V_{1\tau}} + \frac{V_\tau^{\text{F}}}{N(\tau) V_{1\tau}},\end{aligned}$$

where the third equality follows by the Law of Iterated Expectations, Lemma 13A.1, and Eq. (13.47),

$$\begin{aligned}\mathbb{E}_\tau \left[e^{-\int_\tau^T r(u) du} \frac{N(T)}{N(\tau)} V_{0T} \right] &= \mathbb{E}_\tau \left[\mathbb{E}_\tau \left(e^{-\int_\tau^T r(u) du} \frac{N(T)}{N(\tau)} V_{0T} \middle| \mathbb{F}^r(T) \right) \right] \\ &= \mathbb{E}_\tau \left[e^{-\int_\tau^T (r(u) + \lambda(u)) du} V_{0T} \right] \\ &= V_{0\tau}.\end{aligned}$$

13.13 Appendix 7: Modeling correlation with copulae functions

A. Statistical independence and correlation

Two random variables are always uncorrelated, provided they are independently distributed. Yet there might be situations where two random variables are *not* correlated and still exhibit statistical dependence. As an example, suppose a random variable y relates to another, x , through $y = kx^3$, for some constant k , and x can take on $2N + 1$ values, $x \in \{-x_N, x_{N-1}, \dots, -x_1, 0, x_1, \dots, x_{N-1}, x_N\}$, and $\Pr\{x_j\} = \frac{1}{2N+1}$. Then, we have that $\text{Cov}(x, y) \propto \sum_{j=1}^N (-x_j) x_j^3 + \sum_{j=1}^N (x_j) x_j^3 = 0$ and yet, y and x are obviously dependent. This example might be interpreted, economically, as one where y and x are two returns on two asset classes. These two returns are not correlated, overall. Yet they comove in the same direction in both very bad and in very good times. This appendix is a succinct introduction to copulae, which are an important tool to cope with these issues.

Consider two random variables Y_1 and Y_2 . We may relate Y_1 to another random variable Z_1 and we may relate Y_2 to a second random variable Z_2 , on a percentile-to-percentile basis, viz

$$F_i(y_i) = G_i(z_i), \quad i = 1, 2, \quad (13A.8)$$

where F_i are the cumulative marginal distributions of Y_i , and G_i are the cumulative marginal distributions of Z_i . That is, for each y_i , we look for the value of z_i such that the percentiles arising through the mapping in Eq. (13A.8) are the same. Then, we may assume that Z_1 and Z_2 have a joint distribution and model the correlation between Y_1 and Y_2 through the correlation between Z_1 and Z_2 . This indirect way to model the correlation between Y_1 and Y_2 is particularly helpful. It might be used to model the correlation of default times, as in the main text of this chapter. We now explain.

B. Copulae functions

We begin with the simple case of two random variables. This simple case shall be generalized to the multivariate one with a mere change in notation. Given two uniform random variables U_1 and U_2 , consider the function $C(u_1, u_2) = \Pr(U_1 \leq u_1, U_2 \leq u_2)$, which is the joint cumulative distribution of the two uniforms. A copula function is any such function C , with the property of being capable to aggregate the marginals F_i into a summary of them, in the following natural way:

$$C(F_1(y_1), F_2(y_2)) = F(y_1, y_2), \quad (13A.9)$$

where $F(y_1, y_2)$ is the joint distribution of (y_1, y_2) . Thus, a copula function is simply a cumulative bivariate distribution function, as $F(Y_1)$ and $F(Y_2)$ are obviously uniformly distributed. To prove Eq. (13A.9), note that

$$\begin{aligned} C(F_1(y_1), F_2(y_2)) &= \Pr(U_1 \leq F_1(y_1), U_2 \leq F_2(y_2)) \\ &= \Pr(F_1^{-1}(U_1) \leq y_1, F_2^{-1}(U_2) \leq y_2) \\ &= \Pr(Y_1 \leq y_1, Y_2 \leq y_2) \\ &= F(y_1, y_2). \end{aligned} \quad (13A.10)$$

That is, a copula function evaluated at the marginals $F_1(y_1)$ and $F_2(y_2)$ returns the joint density $F(y_1, y_2)$. In fact, Sklar (1959) proves that, conversely, any multivariate distribution function F can be represented through some copula function.

The most known copula function is the Gaussian copula, which has the following form:

$$C(u_1, u_2) = \Phi(\Phi_1^{-1}(u_1), \Phi_2^{-1}(u_2)), \quad (13A.11)$$

where Φ denotes the joint cumulative Normal distribution, and Φ_i denotes marginal cumulative Normal distributions. So we have,

$$F(y_1, y_2) = C(F_1(y_1), F_2(y_2)) = \Phi(\Phi_1^{-1}(F_1(y_1)), \Phi_2^{-1}(F_2(y_2))), \quad (13A.12)$$

where the first equality follows by Eq. (13A.10) and the second equality follows by Eq. (13A.11).

As an example, we may interpret Y_1 and Y_2 as the times by which two names default. A simple assumption is to set:

$$F_i(y_i) = \Phi_i(z_i), \quad i = 1, 2, \quad (13A.13)$$

for two random variables Z_i that are “stretched” as explained in Part A of this appendix. By replacing Eq. (13A.13) into Eq. (13A.12),

$$F(y_1, y_2) = \Phi(z_1, z_2).$$

This reasoning can be easily generalized to the N -dimensional case, where:

$$F(y_1, \dots, y_N) = C(F_1(y_1), \dots, F_N(y_N)) = \Phi(z_1, \dots, z_N),$$

where

$$z_i : F_i(y_i) = \Phi_i(z_i).$$

We use this approach to model default correlation among names, as explained in the main text, and in the next appendix.

13.14 Appendix 8: Details on CDO pricing with imperfect correlation

We follow the copula approach to price the stylized CDOs in the main text of this chapter. For each name, create the following random variable,

$$z_i = \sqrt{\rho}F + \sqrt{1 - \rho}\epsilon_i, \quad i = 1, 2, 3, \quad (13A.14)$$

where F is a common factor among the three names, ϵ_i is an idiosyncratic term, and $F \sim N(0, 1)$, $\epsilon_i \sim N(0, 1)$. Finally, $\rho \geq 0$ is meant to capture the default correlation among the names, as follows. Assume that the *risk-neutral* probability each firm defaults, by T , is given by,

$$Q_i(T) = \Phi(\zeta_{0.10}) \equiv 10\%,$$

where Φ is the cumulative distribution of a standard normal variable. That is, by time T , each firm defaults any time that,

$$z_i < \zeta_{0.10} \equiv \Phi^{-1}(10\%).$$

Therefore, ρ is the default correlation among the assets in the CDO.

We can now simulate Eq. (13A.14), build up payoffs for each simulation, and price the tranches by just averaging over the simulations, as explained below. Naturally, the same simulation technique can be used to price tranches on CDOs with an arbitrary number of assets. Precisely, simulate Eq. (13A.14), and obtain values $\tilde{z}_{i,s}$, $s = 1, \dots, S$, where S is the number of simulations and $i = 1, 2, 3$. At simulation no s , we have

$$\tilde{z}_{1,s}, \tilde{z}_{2,s}, \tilde{z}_{3,s}, \quad s \in \{1, \dots, S\}.$$

We use the previously simulated values as follows:

- For each simulation s , count the number of defaults across the three names, defined as the number of times that $\tilde{z}_{i,s} < \zeta_{0.10}$, for $i = 1, 2, 3$. Denote the number of defaults as of simulation s with Def_s .
- For each simulation s , compute the total realized payoff of the asset pool, defined as,

$$\tilde{\pi}_s = \text{Def}_s \cdot 40 + (3 - \text{Def}_s) \cdot 100.$$

- For each simulation s , compute recursively the payoffs to each tranche, $\pi_{i,s}$,

$$\pi_{i,s} = \min \left\{ \max \left\{ \tilde{\pi}_s - \sum_{k=1}^{i-1} \pi_{k,s}, 0 \right\}, N_i \right\},$$

where N_i is the nominal value of each tranche ($N_1 = 140$, $N_2 = 90$, $N_3 = 70$).

- Estimate the price of each tranche by averaging across the simulations,

$$\text{Price Senior} = e^{-r} \frac{1}{S} \sum_{s=1}^S \pi_{1,s}, \quad \text{Price Mezzanine} = e^{-r} \frac{1}{S} \sum_{s=1}^S \pi_{2,s}, \quad \text{Price Junior} = e^{-r} \frac{1}{S} \sum_{s=1}^S \pi_{3,s}.$$

Note, the previous computations have to be performed under the risk-neutral probability Q . Using the probability P in the previous algorithm can only lead to something useful for risk-management and VaR calculations at best

Note, this model, can be generalized to a multifactor model where,

$$z_i = \sqrt{\rho_{i1}}F_1 + \dots + \sqrt{\rho_{id}}F_d + \sqrt{1 - \rho_{i1} - \dots - \rho_{id}}\epsilon_i,$$

with obvious notation.

References

- Amato, J. D. (2005): "Risk Aversion and Risk Premia in the CDS Market." *BIS Quarterly Review*, September, 55-68.
- Anderson, R. W. and S. Sundaresan (1996): "Design and Valuation of Debt Contracts." *Review of Financial Studies* 9, 37-68.
- Artzner, P., F. Delbaen, J.-M. Eber, and D. Heath (1999): "Coherent Measures of Risk." *Mathematical Finance* 9, 203-228.
- Bernanke, B. S., M. Gertler and S. Gilchrist (1999): "The Financial Accelerator in a Quantitative Business Cycle Framework." In J. B. Taylor and M. Woodford (Eds.): *Handbook of Macroeconomics*, Vol. 1C, Chapter 21, 1341-1393.
- Berndt, A., R. Douglas, D. Duffie, M. Ferguson and D. Schranz (2005): "Measuring Default Risk-Premia from Default Swap Rates and EDFs." *BIS Working Papers* no. 173.
- Black, F. (1976): "The Pricing of Commodity Contracts." *Journal of Financial Economics* 3, 167-179.
- Black, F. and J. Cox (1976): "Valuing Corporate Securities: Some Effect of Bond Indenture Provisions." *Journal of Finance* 31, 351-367.
- Black, F. and M. Scholes (1973): "The Pricing of Options and Corporate Liabilities." *Journal of Political Economy* 81, 637-659.
- Borio, C., C. Furfine, and P. Lowe (2001): "Procyclicality of the Financial System and Financial Stability: Issues and Policy Options." Bank for International Settlements working paper no. 1.
- Broadie, M., M. Chernov and S. Sundaresan (2007): "Optimal Debt and Equity Values in the Presence of Chapter 7 and Chapter 11." *Journal of Finance* 62, 1341-1377.
- Christoffersen, P. F. (2003): *Elements of Financial Risk Management*. Academic Press.
- Cox, J. C., J. E. Ingersoll and S. A. Ross (1985): "A Theory of the Term Structure of Interest Rates." *Econometrica* 53, 385-407.
- Danielsson, J., Shin, H. S. and J.-P. Zigrand (2011): "Balance Sheet Capacity and Endogenous Risk." Working paper London School of Economics and Princeton.
- Duffie, D. and D. Lando (2001): "Term Structure of Credit Spreads with Incomplete Accounting Information." *Econometrica* 69, 633-664.
- Fender, I. and P. Hördahl (2007): "Overview: Credit Retrenchement Triggers Liquidity Squeeze." *BIS Quarterly Review* (September), 1-16.
- Fisher, I. (1933): "The Debt-Deflation Theory of Great Depressions." *Econometrica* 1, 337-57.
- Friedman, M. and A. J. Schwartz (1963): *A Monetary History of the United States: 1867-1960*. Princeton, NJ: Princeton University Press.

- Greenspan, A. (1998): “The Role of Capital in Optimal Banking Supervision and Regulation.” FRBNY Economic Policy Review, October, 163-168.
- Hull, J. C. (2007): *Risk Management and Financial Institutions*. Pearson Education International.
- Ingersoll, J. E. (1977): “A Contingent-Claims Valuation of Convertible Securities.” *Journal of Financial Economics* 5, 289-321.
- International Monetary Fund, (2008): Global Financial Stability Report. April 2008.
- Jamshidian, F. (1989): “An Exact Bond Option Pricing Formula.” *Journal of Finance* 44, 205-209.
- Jarrow, R. A., D. Lando and S. M. Turnbull (1997): “A Markov Model for the Term-Structure of Credit Risk Spreads.” *Review of Financial Studies* 10, 481-523.
- Jorion, Ph. (2008): *Value at Risk*. New York: McGraw Hill.
- Lando, David, 2004. *Credit Risk Modeling—Theory and Applications*. Princeton: Princeton University Press.
- Leland, H. E. (1994): “Corporate Debt Value, Bond Covenants and Optimal Capital Structure.” *Journal of Finance* 49, 1213-1252.
- Leland, H. E. and K. B. Toft (1994): “Optimal Capital Structure, Endogenous Bankruptcy, and the Term Structure of Credit Spreads.” *Journal of Finance* 51, 987-1019.
- Lopez, J. (2004): “The Empirical Relationship Between Average Asset Correlation, Firm Probability of Default and Asset Size.” *Journal of Financial Intermediation* 13, 265-283.
- McDonald, R. L. (2006): *Derivatives Markets*, Boston: Pearson International Edition.
- Mele, A. (2003): “Fundamental Properties of Bond Prices in Models of the Short-Term Rate.” *Review of Financial Studies* 16, 679-716.
- Merton, R. C. (1974): “On the Pricing of Corporate Debt: The Risk-Structure of Interest Rates.” *Journal of Finance* 29, 449-470.
- Modigliani, F. and M. Miller (1958): “The Cost of Capital, Corporation Finance and the Theory of Investment.” *American Economic Review* 48, 261-297.
- Morini, M. and D. Brigo (2011): “No-Armageddon Arbitrage-Free Equivalent Measure for Index Options in a Credit Crisis.” *Mathematical Finance* 21, 573-593.
- Rutkowski, M. and A. Armstrong (2009): “Valuation of Credit Default Swaptions and Credit Default Index Swaptions.” *International Journal of Theoretical and Applied Finance* 12, 1027-1053.
- Schönbucher, Ph J., 2003. *Credit Risk Pricing Models—Models, Pricing and Implementation*. Chichester, UK: Wiley Finance.

- Shin, H. S. (2010): *Risk and Liquidity*. Clarendon Lectures in Finance, Oxford University Press.
- Sklar, A. (1959): “Fonction de Répartition à N dimensions et Leurs Marges.” *Publications de l’Institut Statistique de l’Université de Paris* 8, 229-231.
- Vasicek, O. (1987): “Probability of Loss on Loan Portfolio.” Working paper KMV, published in: *Risk* (December 2002) under the title “Loan Portfolio Value.”