August 19-21 2011
Trinity College Dublin
Ireland

YOUNG STATISTICIANS MEETING

Satellite Event of the 58th ISI World Congress

http://www.scss.tcd.ie/conferences/YSI2011

Invited Speakers:

Sir David Cox (Oxford)
Peter Guttorp (Washington)
Valerie Isham (UCL)
Rajan Patel (Google)
Adrian Raftery (Washington)
Jonathan Taylor (Stanford)
Martin Wainwright (Berkeley)

Scientific Programme Committee Chair
Victor Panaretos (EPFL)

Local Organizing Committee Chair
Claire Gormley (UCD)

Sponsors

Irish Aid
Department of Foreign Affairs
An Roinn Gnóthai Eachtracha

Google

Accenture
High performance. Delivered.
Contents

Contents 3
Céad míle fáilte 5
Programme 6
Local Information 8
Invited Speakers: abstracts. 10
Contributed posters: abstracts 18
ISI Jan Tinbergen Award: abstracts 113
List of delegates 117
Notes 122
Céad míle fáilte

Welcome to YSI 2011 and welcome to Dublin, Ireland!

Many thanks to our invited speakers who have kindly given their time to spend with us this weekend. Many thanks also to all of you for attending what we hope will be an interactive, enjoyable and learning experience for all of us.

Sincere thanks must go to our sponsors – without their vision and financial support this event would not have taken place. Many thanks to the Irish Statistical Association, Google, Accenture, Science Foundation Ireland, Irish Aid, the American Statistical Association, the Royal Statistical Society, the Bernoulli Society and the International Environmetrics Society.

Alison Murphy in MCI has been patient, helpful and professional and to her we owe many thanks. Thanks must also go to Trinity College Dublin and their conference organising team for all their assistance. Finally, to all the postgraduate students and colleagues in University College Dublin and Trinity College Dublin for all the (stressful!) behind the scenes help – it did not go unnoticed and we are very grateful for all your help.

The YSI 2011 Organising Committee
August 2011
Programme

Day 1  Friday August 19th 2011

13:00-15:00 Registration and Welcome Coffee.
15:00-15.30 Opening Address by Victor Panaretos and Claire Gormley.
15:30-16:15 Invited Lecture - Adrian Raftery.
16:15-16:30 Discussion.
16:30-17:15 Invited Lecture - Martin Wainwright.
17:15-17.30 Discussion.
17:30-18:30 Poster Session I and Refreshments.
18:30-19:00 Welcome Address by Jef Teugels (ISI President).
19:00 – Informal dinner arrangements.

Day 2  Saturday August 20th 2011

09.30-10.30 Poster Session II and Morning Coffee and Snacks.
11:30-11:45 Discussion
11:45-12:30 Invited Lecture - Rajan Patel
12:30-12:45 Discussion
12:45-15:00 Lunch (free for participants) and Poster Session III.
15:00-15:45 Invited Lecture - Peter Guttorp.
15:45-16:00 Discussion
16:00-16:45 Invited Lecture - Valerie Isham.
16:45-17:00 Discussion
17.00-18.00 Poster Session IV and Coffee Break
18:30– Dinner (for registered participants) in The Dining Hall, Trinity College Dublin.
Day 3  Sunday August 21st 2011

09:15-10:45  Lectures by the winners of the ISI Jan Tinbergen Award.
09:15-09.45 Kodzovi Senu Abalo
09.45-10.15 Kavitha Mehendale
10.15-10.45 Mahdi Roozbeh

10:45-11.15  Coffee break.
11:15-12:00  Invited Lecture - Sir David Cox.
12:00-12:15  Discussion.
12.15-12.30  Announcement of SFI poster winner(s) and close.
12.30-15.30  (Optional) Visit the Guinness Storehouse for a tour, lunch or a photograph with the William Sealy Gosset plaque. (A voucher for a 10% reduction on the admission price to the Guinness Storehouse is included in your conference pack.)
16:00  Grand opening ceremony of ISI 2011 in the Convention Centre Dublin.

Poster Session Details

Delegates will present their poster in one poster session only – posters are allocated to a specific poster session. If the first author of the poster has a surname beginning with the letter:

- A – E inclusive, they will display their poster during Poster Session I.
- F – K inclusive, they will display their poster during Poster Session II.
- L – R inclusive, they will display their poster during Poster Session III.
- S – Z inclusive, they will display their poster during Poster Session IV.

Please remove your poster from the poster board at the end of your poster session to facilitate delegates presenting during the next poster session.

There is a €1000 prize for the best poster(s) courtesy of Science Foundation Ireland. The winner(s) will be announced during the YSI 2011 closing ceremony on Sunday August 21st. The prize will be officially presented at the awards ceremony during the main ISI congress on Thursday August 25th from 15:00 in the Auditorium of the Convention Centre Dublin.
Local Information

All invited talks will be held in the JM Synge theatre in the Arts Building on the Nassau Street side of Trinity College Dublin.

The poster sessions will be held on the concourse outside the JM Synge theatre, as will all coffee breaks and lunch on Saturday.

On Saturday evening, the conference dinner will be held in the Dining Hall in Trinity College Dublin, on the College Street side of the campus.

Trinity College Dublin has the enviable location of Dublin city centre – many cafes, restaurants, bars and shops are located around Grafton Street which is easily accessible through the front gate, at the College Green end of campus.
ISI 2011

Following YSI 2011, the 58th World Statistics Congress of the International Statistical Institute (ISI) will be held in the Convention Centre Dublin from the 21st to the 26th August 2011.

The map below gives walking directions to the Convention Centre Dublin from Trinity College Dublin.
Invited speakers: abstracts
Probabilistic weather forecasting consists of finding a joint probability distribution for future weather quantities or events. Information about the uncertainty of weather forecasts can be important for decision-makers as well as the public, but currently is routinely provided only for the probability of precipitation, and not for other weather quantities such as temperature, wind or amount of precipitation. It is typically done using a numerical weather prediction model, perturbing the inputs to the model (initial conditions, physics parameters) in various ways, and running the model for each perturbed set of inputs. The result is viewed as an ensemble of forecasts, taken to be a sample from the joint probability distribution of the future weather quantities of interest. This is often uncalibrated, however.

We review a principled statistical method for postprocessing ensembles based on Bayesian Model Averaging (BMA), that models the predictive distribution conditionally on the ensemble by a finite mixture model. We describe applications to precipitation, wind speeds, wind directions, visibility and winter road maintenance, a multivariate decision problem. For probabilistic forecasting of an entire weather field, we describe a spatial extension of the BMA method that perturbs the outputs from the numerical weather prediction model rather than the inputs. Forecasts are available in real time at www.probcast.washington.edu, and the R packages ensembleBMA and ProbForecastGOP are available to implement the methods.

This is joint work with Tilmann Gneiting, Veronica Berrocal, McLean Sloughter, Le Bao, Chris Fraley, William Kleiber and Richard Chmielecki.
High-dimensional data: Some challenges and recent progress

Professor Martin Wainwright
University of California, Berkeley
USA.

High-dimensional data sets appear in many branches of science and engineering, and present interesting challenges (and opportunities) for statisticians. Many classical methods are known to break down for high-dimensional data, which motivates the development of new methods and theory. In this talk, we overview some areas in which high-dimensional statistical problems arise, including social networks, image analysis, recommender systems, and internet security, and discuss some recent progress.
The generalized LASSO and structured regression

Professor Jonathan Taylor
Stanford University
USA.

The LASSO and other L1 related methods have been extremely popular and useful in problems where the desired property of an estimator is sparsity. In this talk, I will introduce the generalized LASSO which focuses on possibly different properties of estimators such as piecewise linearity in a one-dimensional problem; piecewise constant in an imaging problem, etc. The generalized LASSO shares many similarities with the LASSO though it is a harder convex problem to solve. I will discuss an analogue of the LARS algorithm for the generalized LASSO as well as scalable first order descent methods that permit solving these problems on very large scales. This is joint work with Ryan Tibshirani.
Experimentation at Google

Dr. Rajan Patel
Google Research

Google runs thousands of experiments each year to help better understand whether changes to its search algorithm will help provide more useful and high quality results for our users. It’s likely that you’ve contributed to many of these experiments without consciously knowing it. In this talk I’ll describe some of the fundamental issues we’ve faced when designing live traffic experiments and human rater surveys so that we can more effectively measure the quality of thousands of experiments each year and allow engineers at Google to innovate more quickly than ever.
Climate science abounds with good statistical problems. Climate can be thought of as the distribution of weather (although climate models deal with much more than just weather). Since climate is changing, one needs ways to deal with estimates of marginal distribution in nonstationary models. From a theoretical point of view this entails studying empirical processes of non-identically distributed data, while from a practical point of view there is need for multivariate spatial tools generalizing qq-plots. Some of the problems are illustrated using Scandinavian weather data.
An introduction to the world of stochastic models: with examples from point processes, rainfall and epidemics

Professor Valerie Isham
University College London

In applications, stochastic models seek to represent an underlying physical phenomenon of interest, albeit often in a highly idealised way, and have parameters that are physically interpretable. Such models are sometimes termed mechanistic. In contrast, statistical models are descriptive, and represent the statistical properties of data and their dependence on covariates, without aiming to encapsulate the physical mechanisms involved. In this talk, I will use a range of examples to show how stochastic models can be developed, and to illustrate some of their properties, focussing particularly on applications in hydrology (rainfall, soil moisture) and the transmission dynamics of infection.
The interplay between theory and application in statistics

Sir David Cox
Nuffield College
University of Oxford

After some brief historical remarks about the development of statistical theory, a number of examples will be given illustrating the ways in which statistical ideas enter the discussion of scientific issues. The applications range from botany to particle physics.
Contributed posters: abstracts
Limiting genotype frequencies for a Y-linked two-sex branching process with blind choice

Alsmeyer, G.\textsuperscript{1}, Gutiérrez, C.\textsuperscript{2} and Martínez, R.\textsuperscript{2}

\textsuperscript{1} Institute of Mathematical Statistics, University of Münster.
\textsuperscript{2} Department of Mathematics, University of Extremadura.

In the paper [1], a two-dimensional bisexual branching model has been studied to analyze the evolution of the number of carriers of Y-linked genes from generation to generation in two-sex populations. In the model, genes which present two allelic forms are considered. Actually, one of these forms could mean the absence of the other. Moreover, monogamous mating with blind choice is assumed, which means that females choose their mate without caring about his genotype. The latter condition may be justified by the fact that Y-linked genes are typically not expressed in males, or, if they are, do not have any preferential impact on the mating process. Using this model, we first study extinction conditions for each particular allele when the probability of being female is equal to 0.5. In a second place, our goal is to describe the growth behavior of each genotype in the sets of coexistence and survival of one genotype.

It is deduced for this model that the average numbers of female and male offspring per mating unit of a genotype determine its growth rate, which does not depend on the behavior of the other genotype. Hence, the dominant genotype is found (these results have been published in [2]). Finally, we deal with the classical problem in population genetics of determining the limiting genotype frequencies and the limiting sex ratio. The main results are illustrated by means of examples.

Acknowledgement: The research was supported by the Ministerio de Ciencia e Innovación, the Junta de Extremadura and the FEDER, grants MTM2009-13248 and GR10118.

References


Modelling the regulation of systemic iron homeostasis at the molecular level

Anashua Banerji
University of Heidelberg

The importance of statistical analysis and modelling lies in its capacity to reveal the basis of interactions and mechanisms across nature. One of the most important applications is in the human body. Modelling the regulation of the iron master-regulatory hormone hepcidin is prime to our understanding of systemic iron homeostasis, especially because hepcidin expression is deregulated in diseases such as hemochromatosis which is the most rampant genetic disorder in the Western world with no successful medical treatments yet. At the molecular level, the hepcidin promoter, and therefore hepcidin transcription, is co-ordinately controlled by IL6/STAT and BMP/Smad signalling pathways, but the dynamics of interactions remain poorly understood. We will present methods of arriving at an optimal mathematical model, starting with statistical analysis of experimental data, which will reveal the underlying mechanisms of hepcidin regulation initiated by STAT and Smad signalling pathways, either independently or in combination. In particular, we apply statistical methods to identify the patterns in the data and aim to arrive at a robust model that can account for these patterns and provide a biological explanation.
Graphical Models for the analysis of a cohort study

Lorna Barclay
Department of Statistics
University of Warwick

Within a Bayesian framework we look at graphical models as an alternative approach for analysing parts of a cohort study. Bayesian Networks give a good visual representation of a complex problem, summarising the structure of the joint probability distribution in a way that provides a convenient framework for learning parameters and incorporating new information.

Furthermore, two Bayesian Networks are Markov equivalent when they encode the same conditional independence statements. However, when it is appropriate to assume a Causal Bayesian Network based on Pearl’s intervention formula [1], equivalent Bayesian Networks, but different directed graphs, give different causal interpretations. We illustrate these different classes of equivalence on a cohort study and contrast their distinct formally stated causal hypotheses with the often implicit causal informal interpretations made within a typical regression analysis of this genre of application.

Keywords: Bayesian Networks, Causal Bayesian Networks

References


Weighted Lasso for Data Integration

Linn Cecilie Bergersen
Department of Mathematics, University of Oslo

Ingrid K. Glad
Department of Mathematics, University of Oslo

Heidi Lyng
Department of Radiation Biology, Norwegian Radium Hospital

We propose a new method for data integration in high-dimensional prediction and variable selection based on the weighted lasso. The lasso is one of the most commonly used methods for high-dimensional regression, but can be unstable and lacks satisfactory asymptotic properties for variable selection. Our method integrates relevant external information on the covariates to guide the selection towards more stable results. Weighting the penalties with external information gives each regression coefficient a covariate specific amount of penalization and can improve upon standard methods, that do not use such information, by borrowing knowledge from the external material. The method is applied to two cancer data sets, with gene expressions as covariates. We find interesting gene signatures, which we are able to validate. We discuss various ideas on how the weights should be defined and illustrate how different types of investigations can utilize our method exploiting different sources of external data. By simulations we show that our method outperforms the lasso and the adaptive lasso when the external information is from relevant to partly relevant, in terms of both variable selection and prediction performance.
Test for Jumps in Semimartingale Models Using Clustering Criteria

Karthik Bharath
Department of Statistics, University of Connecticut

We propose a simple and intuitive test to determine the presence of jumps in discretely observed processes on a fixed time interval using methods from clustering techniques, as the sampling interval tends to 0. The test is based on the clustering criterion used in *k-means* clustering and provides a simple way to determine if jumps are present in high frequency financial time series modeled using a general class of Ito semimartingale processes. We show consistency of the parameter estimates obtained and examine their asymptotic distributions.
Modelling bovine tuberculosis and udder health in Irish dairy herds

Fiona Boland
School of Mathematical Sciences
University College Dublin

The association between bovine tuberculosis (TB) infection status and milk production has been previously described in dairy cows in TB-infected herds in Ireland. Study results showed that significantly less milk was produced by TB reactors in comparison to non-reactors during both the lactation of disclosure and all preceding lactations [1]. The biological basis is uncertain and further investigation is required to help understand the mechanisms underpinning this association. Hence, in this study, hierarchical generalised linear models (HGLMs) are used to examine the association between somatic cell counts (as a measure of udder health) and TB status. There is a defined hierarchical structure in the data, lactations are nested within cows and cows within herds, and HGLMs provide a useful class of models to incorporate these effects. In broad terms HGLMs extend generalised linear mixed models (GLMMs) by allowing more flexibility in the choice of the distribution for the random effects and also modelling of the dispersion of the random effects and error term. This practical example illustrates the application of HGLMs to model data with several random effects which occurs frequently in veterinary epidemiology.

Keywords: Bovine tuberculosis (TB), Somatic cell counts, Random effects, h-likelihood, Ireland.

References

An econometric Analysis of Capital Markets and Economic Growth the Venezuela Case.

Maylix Brianto Contreras
Escuela de Estadística y Cómputo Científico
Universidad Simón Bolívar

This paper analyses the existence of the relationship between capital markets and economic growth through the design of a multiple linear regression model. The variables included in the model are the most representative of the Caracas Stock Exchange (BVC): Market Capitalization Index Caracas (IBC), market capitalization (CB) and total amount traded in fixed income (RFT) on non-oil GDP in real terms for the period 1989 - 2010, quarterly. Once tested the existence of the relationship between capital markets and growth, an error correction mechanism is applied to find the speed of convergence of this relationship in the long-term equilibrium. After, a Granger causality test is used to statistically detect the direction of the relationship and examine whether lagged values of stock variables help predict the LPIBNP. Finally, an exercise in the multivariate co-integration context of vector autoregressive models (VAR) is designed to estimate the long-term association of the variables that constitute the model.

Keywords: Capital market, economic growth, stock market, multiple linear regression model, error correction mechanism, causality, model vector autoregressive.

References


We consider models of the form $dY(t) = \beta(A - Y(t))dt + \sigma dW(t), \ Y(t_0) = y_0,$ to describe individual animal growth in random environments (see [1]), where $Y(t) = g(X(t))$ is a transformed size, $g$ is a strictly increasing $C^1$ function of the actual animal size $X(t)$ at time $t$, $\sigma$ measures the effect of random environmental fluctuations on growth, $W(t)$ is the standard Wiener process, $y_0$ is the transformed initial size, $A = g(S)$, where $S$ is the asymptotic size, and $\beta$ describes how fast the approach to $S$ is. For these stochastic differential equation (SDE) models, we are interested in the first passage time through a high threshold value $Q^* > y_0$ for animal size $X(t)$. This can be a desirable market sale size in livestock production, for instance. This is equivalent to determine the first passage time of $Y(t) = g(X(t))$ by $Q = g(Q^*)$. We denote this time by $T_Q = \inf\{t > 0 : Y(t) = Q\}$. For these models, explicit expressions for the mean and variance of $T_Q$ can be found in [1].

We will present the gSDE software, which includes maximum likelihood parameter estimation, data fitting and prediction, and other items still under development. In particular we will present a new module for the computation of the mean and variance of $T_Q$, solving simultaneously some problems of instability in the numerical quadrature methods traditionally used to compute the integrals appearing in the corresponding expressions.

References

Bayesian model selection for exponential random graph models

Alberto Caimo & Nial Friel
School of Mathematical Sciences
University College Dublin

Exponential random graph models are a class of widely used exponential family models for social networks. The topological structure of an observed network is modeled by the relative prevalence of a set of sub-graph configurations called network statistics which are regarded as random variables.

One of the key tasks in the application of these models is which network statistics to include in the model. This can be thought of as statistical model selection problem. This is a very challenging problem—the posterior distribution is often termed “doubly intractable” since computation of the likelihood is rarely available, but also, the evidence or marginal likelihood of the posterior is, as usual, also intractable.

We present a fully Bayesian model selection method based on a Markov chain Monte Carlo algorithm of [1] which returns a posterior distribution for each competing model as well as some possible approaches for computing the evidence.

References

Are There Clusters of Undergraduate Experiences?

Michael Carroll\textsuperscript{1}, Kathleen OSullivan\textsuperscript{1}, John OMullane\textsuperscript{2}, Jamie Madden\textsuperscript{1},
Sarah Murphy\textsuperscript{1}, Laura OSullivan\textsuperscript{1}
\textsuperscript{1} Department of Statistics, \textsuperscript{2} School of Computer Science & Information Technology
University College Cork, Ireland

The goal of this paper is to explore if distinct groups of undergraduate students can be identified based on: satisfaction with life in university; perceptions of good teaching practices, appropriate workload, and motivation; and confidence in their ability to study, interact, and attend formal teaching activities. We applied a hierarchical agglomerative cluster analysis using Wards method and the squared Euclidean distance to our data to determine the appropriate number of clusters and to obtain initial cluster centres. Subsequently we conducted a K-means cluster analysis using the cluster centres generated from the hierarchical analysis as the initial seed points to identify cluster membership. We confirmed the reliability of the cluster solution. We compared the derived clusters on the clustering variables and on additional variables. We visualised the data to illustrate cluster membership. Three student clusters emerged. Students in cluster 1 had mixed perceptions of their university experience, students in cluster 2 were content with their university experience and students in cluster 3 had a somewhat negative perception of their university experience. Cluster analysis identified that undergraduate students are heterogeneous with respect to satisfaction with life in university, teaching and learning, and academic confidence.
Dipping a toe into variance estimation for large data streams

Nicholas Chamandy
Google

Due to privacy and computational constraints, it is often not possible to hold an entire data set in memory when performing estimation at Google. Rather, raw data are usually ‘sharded’ across a number of machines. Typically, therefore, we are interested in estimators which can be computed in one pass over the data. Variance estimation can be problematic in this setting.

Consider for example a simplistic model for click rate:

\[ E[Y_i|X_i] = \beta X_i, \quad \text{Var}(Y_i|X_i) = \sigma^2 X_i, \quad X_i \sim (\mu, \tau^2), \]

where \( Y_i \) denotes the number of ad clicks and \( X_i \) the number of queries for the \( i \)th Google user over the course of a week. It is trivial to compute the unbiased GLS estimator of \( \beta \), namely \( \hat{\beta} = \sum_i Y_i / \sum_i X_i \), in one pass over the data. However, collecting second-order statistics would require joining a single user’s data across multiple days, a violation of Google’s privacy policy.

Two obvious single-pass strategies for obtaining the variance are (a) forming \( b \) independent, heterogeneous subgroups and computing the estimator over each subgroup (aggregation), and (b) computing the estimator in each of \( b \) streaming bootstrap resamples of the data (replication). We investigate the properties (relative efficiency, coverage probability and tail symmetry) of these two approaches and compare them to a gold standard estimator for a common class of statistics.
Depth functions in regional flood frequency analysis

F. Chebana* and Taha B.M.J. Ouarda
Canada Research Chair on the Estimation of Hydrometeorological Variables
Hydro-Quebec/NSERC Chair in Statistical Hydrology
INRS-ETE, University of Quebec
Canada

The usual practice in regional frequency analysis (RFA) consists in two main steps: the delineation of hydrological homogeneous regions and regional estimation. Methods adopting this two-step approach have some drawbacks. The restriction of the delineation step to a particular region excludes a number of sites which corresponds to a loss of a part of the available information. In addition, by defining a region, a border effect problem can be generated. To overcome these drawbacks, we propose a method incorporated in the regression model framework where hydrological and physio-meteorological variables are related. The border effect problem is treated using a weight function. A depth function is employed to evaluate the similarity between sites. An iterative procedure is also introduced to improve the estimation performance. It is shown that the depth-based approach estimation results are superior to those obtained by the classical CCA-based approach. Even though the method is applied for flood RFA, it is general and can be useful in other disciplines where the regression model is applicable.
Nash equilibriums in two-person red-and-black games

May-Ru Chen
Department of Applied Mathematics
National Sun Yat-sen University

A two-person red-and-black game is inspired by the famous red-and-black gambling problem presented by Dubins and Savage. There are two people gambling in stage with positive integer initial fortune and the total amount of fortune is fixed. At each stage, they simultaneously bid an integer portion of their current fortune, and each players win probability is a function of the ratio of his bet to the sum of the two bets. For each player, the gambling problem is to maximize the probability of winning the entire fortune of his opponent by bitting repeatedly with suitably chosen stakes. For this model, we find simple conditions on the win probability function such that the profiles (bold strategy, timid strategy), (bold strategy, bold strategy) and (timid strategy, timid strategy) are Nash equilibriums, respectively.

Key words: red-and-black games; bold strategy; timid strategy; win probability function; Nash equilibrium
Nonparametric Monitoring Procedures

Ondřej Chochola
Department of Statistics
Charles University in Prague
Czech Republic

The poster will concern sequential procedures for detection of a change in distribution when a training sample with no change is available. The procedure is based on empirical distribution functions. Both dependent and independent observations are considered.

The properties under both null and alternative hypothesis are studied and possible modifications are discussed. Theoretical results are accompanied by a simulation study.

References

The goal of this paper is to determine the underlying dimensions of a Teaching and Learning Instrument (19 items) distributed as part of an undergraduate student experience survey in 2011. A principal component factor analysis using a varimax rotation was conducted. A preliminary estimation of factors was determined by factor eigenvalues greater than 1.0 in conjunction with changes in the slope within the scree plot. Criteria for factor loadings included item values greater than 0.40 on the primary factor and values less than 0.40 on all other factors. Cronbach’s alpha was used to examine the scales internal consistency. Data visualisation techniques are used to illustrate the partitions amongst the items. Two items were excluded due to not loading or cross-loading. The results of the factor analysis applied to the remaining 17 items revealed the presence of three factors. Factor 1 is labelled Good Teaching (6 items, Cronbach’s alpha=0.85), factor 2 is labelled Motivation (5 items, Cronbach’s alpha=0.87) and factor 3 is labelled Appropriate Workload (6 items, Cronbach’s alpha=0.83). Factor analysis identified the presence of three distinct dimensions which represent students perceptions of good teaching practices, motivation within their programme of study, and manageable workloads.
Bayesian modeling of social networks: novel methodology and applications to longitudinal studies

Lilia Costa
Fabio Rigat
Statistics Department
The University of Warwick

Stochastic networks are often modeled by specifying sets of interactions among many partially exchangeable random variables [1]. From a subjective Bayesian perspective, this is the only viable approach when expert opinions about the marginal distributions of the network nodes are not available, but reliable assessments about their relations can be effectively elicited. In this study we consider methodological issues encountered when constructing dynamic network models [2] by specifying their transition probabilities only. We are mainly interested in likelihood approximations which allow the implementation of computationally efficient numerical Bayesian inference via Markov chain Monte Carlo and particle filtering. By simulating data using a computationally impractical network likelihood, we derive empirical coverage probabilities under a proposed approximated likelihood for various study population sample sizes and missing data patterns. We show that the likelihood approximation we propose provides a reliable computational workhorse for Multinomial classification networks.

Keywords: Dynamic Bayesian Networks, classification models, time series models, likelihood approximations, Bayesian inference, MCMC

References


Background: When planning clinical trials, biostatisticians ideally select designs based on sound statistical considerations, optimizing efficiency and interpretability. However, feasibility may affect design choice.

Scenario: Breast cancer patients adjuvant treatment decision is primarily based on certain prognostic factors. A clinical trial was planned to investigate whether results from a new test kit, estimating recurrence risk after hormonal treatment, would affect physicians treatment recommendations for their patients.

Discussion: Randomized controlled and cluster randomized designs were originally suggested. To reduce patient bias, stratified randomization or minimization method were considered, thus accounting for potential confounding factors. The latter method includes more factors, useful in smaller trials where several factors can affect outcome. In the cluster randomized design, physicians were the clusters, reducing physician bias by considering patients intercorrelated outcome. Both designs were unfeasible due to difficulties in randomizing physicians into clusters and the carry-over/learning effect. Thus, the primary endpoint was based on multidisciplinary tumour boards treatment recommendations before and after test kit results were available, reducing patient and physician bias.

Conclusion: Although, trial design choice is often hampered by feasibility, biostatisticians must meet this challenge.
A Short Term Wind Speed Forecast: an application in the Brazilian Northeast

Tiago Mendes Dantas
Reinaldo Castro Souza
Department of Electrical Engineering
Pontifical Catholic University of Rio de Janeiro (PUC-RIO)

The amount of energy generated by wind sources is increasing all over the world. Brazil has a huge potential due to its geographic localization and the Brazilian Government has given a clear signal that is prone to invest in this kind of energy. Accurate wind speed forecasts are essential in the operation planning for the electrical wind power system. This work aims to make more accurate forecasts in the northeast of Brazil. To do so, we use a model that takes into account the daily characteristics and the long memory behavior. The model applied in this particular region proved to be more accurate than the persistence model and other models (eg. Hybrid neuro-fuzzy model)

Keywords: Wind energy, Wind speed, forecasting, long memory

References


Nonparametric Estimation for Multivariate Extremes

Miguel de Carvalho & Anthony Davison
Ecole Polytechnique Fédérale de Lausanne

The RamosLedford spectral model has received considerable attention as an elegant tool for modelling multivariate extremes [2]. A major advantage over existing models is its ability to unify in a single framework the cases of asymptotic dependence and asymptotic independence. Despite its attractive features, for purposes of estimation only a parametric asymmetric logistic model is available, thus restricting the range of its applications. In this paper we propose nonparametric estimation and inference procedures for the RamosLedford spectral model by empirical likelihood techniques [1].

References


Fitting and Forecasting Binary Time Series with Long Memory Effects

Giovanni De Luca & Alfonso Carfora
Department of Statistics and Mathematics for Economic Research
University of Naples

We propose a novel model for binary data: the Binomial Heterogenous Autoregressive (BHAR) model. With respect to the models recently introduced for the analysis of binary time series, BHAR model allows to detect the phenomenon for a longer time, capturing long memory effects. After applying the model to the binary time series of U.S. recessions we have made in-sample and out-of-sample forecasts. The forecasts have been successfully compared to those obtained by the most popular models to predict U.S. recessions.

The results of descriptive statistics, in fact, are very positive in terms of MAE and RMSE. However, since simple descriptive statistics do not indicate whether one model is statistically better than another, we apply the Diebold-Mariano and Harvey-Leybourne-Newbold (1997) tests to evaluate the predictive accuracy. The results of the tests display a significant better forecasting performance of the BHAR model.

Keywords: Binary variables, long memory, forecasting

References

A measure of stationarity in locally stationary processes

Holger Dette, Philip Preuß, Mathias Vetter
Department of mathematical statistics
Ruhr-University Bochum, Germany

This paper is about testing the assumption of stationarity in locally stationary processes. Locally stationary processes possess a representation of the form

\[ X_{t,T} = \sum_{l=-\infty}^{\infty} \psi_{l,t,T} Z_{t-l}, \]  

(1)

with time-dependent coefficients \( \psi_{l,t,T} \). These processes can [in contrast to stationary processes, where \( \psi_{l,t,T} = \psi_l \)] change their covariance structure over time and therefore it is possible to model many processes more accurately than before.

Testing the assumption of stationarity in locally stationary processes is a very important problem and roughly speaking the testing problem is described by

\[ H_0 : X_{t,T} \text{ is stationary} \quad \text{vs.} \quad H_1 : X_{t,T} \text{ is locally stationary, but not stationary.} \]  

(2)

Mathematically we have to estimate terms of the form

\[ \int_0^1 \int_{-\pi}^{\pi} f(u, \lambda)^2 d\lambda du \]  

(3)

[where \( f \) is the so called time-varying spectral density] to develop a test for the stated hypothesis. In the last ten years several methods were derived, but they all have the disadvantage of choosing at least two smoothing-parameters, which can influence the result extremely. This paper presents a test for stationarity, which avoids the choice of a smoothing-bandwidth and which is therefore the first applicable result for the above testing problem.

Keywords: non stationary processes, goodness-of-fit tests, integrated periodogram, locally stationary processes

References

Survival and Movement of Cape Gannets

Greg Distiller
Department of Statistical Sciences
University of Cape Town

Res Altwegg
South African National Biodiversity Institute (SANBI)

Rob Crawford
Marine and Coastal Management (MCM)

Monitoring the population dynamics of seabirds can provide important information about the underlying oceanic ecosystem. This is particularly interesting for the Benguela upwelling region around the southern tip of Africa where the Agulhas and Benguela currents meet. The Benguela upwelling region is one of the world’s most productive marine ecosystems and appears to be subject to long term climatic changes.

Cape gannets (Morus capensis) breed at only six locations off the coast of South Africa and Namibia. A substantial change over time in the distribution of gannets to the south east has been observed but there is uncertainty about the extent to which differences in local reproduction and survival rates, as well as bird movements, have influenced this shift.

We combine multi-state capture-mark-recapture data with dead recoveries to estimate survival and movement between gannet colonies from 1989 to 2009 using a random effects model estimated with MCMC. The effects of environmental and fishing covariates on the parameters of interest are also explored.
The Normal Inverse Gaussian (NIG) and its Application in Finance and Palaeoclimate Reconstruction

Thinh Doan¹, Andrew Parnell² and John Haslett¹
¹ Trinity College Dublin
² University College Dublin

The NIG family is a variance-mean mixture of the Normal with the Inverse Gaussian. It is characterized by the first four moments (mean, variance, skewness and kurtosis). Its tails are heavier than those of the Normal. The remarkable properties of infinite divisibility and stability under convolution makes it an obvious candidate for modeling financial data, where the Normal distribution has been showed to underestimate the underlying random variation. From empirical illustration (ISEQ Index) we clarify whether the NIG performs better than the Normal.

In this work, we further propose the NIG as an application in palaeoclimate reconstruction. We use the NIG as a long-tailed random walk prior to filter out the noise in climate signal. We also use its bridging properties to create climate histories for the last 14000 years.
Variable Selection for Supervised Classification of High Dimensional Data

Katarina Domijan
Department of Mathematics and Statistics
NUI Maynooth

We propose a methodology for the general problem of variable selection in supervised classification of data where the number of available predictors exceeds ten to twenty-fold the number of samples. The approach is set in a Bayesian decision theoretic framework, where the decision space is the set of all possible subsets of predictors and the optimal decision maximizes the expected utility function. The utility function is chosen so that it measures the predictive ability of a sub-model and penalizes the inclusion of variables. Stochastic search algorithms are employed to search through the decision space. The task of optimization in discrete decision spaces of very high dimension is difficult, and it is unlikely that global optima can be identified, thus, the aim of the proposed method is to identify sparse sub-models that yield good classification rates. We apply our method to data sets from a variety of applications, as well as simulated data.
Development of an adjusted scan statistic for the efficient detection of road collision ‘hot-spots’ along the road network in Ireland

Erica Donnelly-Swift
Dept. Public Health & Primary Care
Trinity College Dublin

Road traffic collisions pose a major economic and personal burden on society. The purpose of this study was to develop and deploy an efficient statistical methodology for identifying road traffic collision ‘hot-spots’ in the Republic of Ireland. The statistical methodology is based on a set of cluster-detection techniques known as ‘Scan Statistics’ [1]. Findings suggest there are a diverse number of human, environmental and road characteristics associated road collision ‘hot-spots. For example, 6 ‘hot-spots were identified on the N2 route, which runs from Dublin to the Northern Ireland Border at Moy Bridge, Co. Monaghan, to connect Dublin with Derry via the A5. The main collision types identified at these ‘hot-spots were pedestrian, head-on, angle, rear-end and single-vehicle collisions. Fatal/serious injuries were sustained from collisions occurring at these ‘hot-spots. Findings from this analysis give meaningful insights into collision ‘hot-spots thus providing a greater understanding and ability to target safety measures/policies to specific areas.

References

Competing Risks in Survival Analysis

Cara Dooley
School of Maths, Stats and App Maths.
National University of Ireland, Galway.

Often in Survival Analysis, an event can occur which prevents the event of interest occurring, for example, we may be interested in death caused by cardiac issues but the patient may die from other causes meaning that the event of interest is never seen. This often means the data for the patient who had died from the other causes is treated as censored but as these events are rarely independent this introduces bias into the analysis. The two (or more) events can be treated as competing risks. By incorporating the competing risks into the analysis, the bias can be removed. The ideas of competing risks will be discussed and illustrated with an example.

References


A new class of test statistics for detecting Disease Associated Rare variants

John Ferguson
Yale Center for Statistical Genomics and Proteomics
Yale University

Many genetic polymorphisms, both rare and common, contribute to the heritability of complex genetic disease. For many such diseases, the common disease associated genetic variation has largely been identified via large scale Genome Wide Association Studies (GWAS). Unfortunately, rare variants that are disease associated are more difficult to find since the number of individuals carrying the rare allele, which is proportional to the effective sample size in a case control study, will be relatively small when the allele is rare. Recently, many techniques that collectively analyze multiple rare variants within a genomic region have been proposed in an effort to increase statistical power. Here we present a new class of test statistics designed for rare variant association analysis that unifies many of these existing approaches under a common framework. In particular, this class of test statistic encompasses and generalizes a large variety of both univariate and multivariate Score tests. Data-adaptive techniques are suggested for choosing a particular member of this class in real data situations. This is joint work with Professor Hongyu Zhao and Professor Judy Cho (from Yale) and Dr. Joshua Sampson (from NIH).
Analysis of the Brazilian health by DEA, data development analysis

Mauricio Vidotti Fernandes, Scarlett Sophia Rocha, Dr. José Angelo Belloni
Department Statistic, ICE, UnB

This work has the scope to study the performance of the health sector in Brazil, by proposing performance indicators using the indicators given by the Basic Data and Indicators (IDB2009), and observing its relevance for the subject studied in this paper. Also, aims to determine which actions can be taken to improve the health system in Goias. To achieve this, it was used the data envelopment analysis (DEA), which has the capacity to create efficiency scores based on input and outputs. To do so, three indicators were created: efficiency of the health sector (cover indicators versus indicators of appeal), the efficiency of prevention (morbidity indicators versus socioeconomic indicators) and, least, the quality provided by the health system (mortality indicators). The variables, which are present in the Database of the Unified Health System in Brazil (DATASUS). The main problem is that both the prevention and quality are far beyond the acceptable in Goias the quality 82% efficiency and prevention in 69.29%, even thought Brazilian hospitals operate near saturation. This is due to the great demand and dire conditions of quality and prevention amongst all hospitals in the Brazilian States evidenced by high capacity indicators.

References

A hybrid framework for traffic simulation in road networks

A. Fitzgerald, S. Moutari and A. H. Marshall
Centre for Statistical Science & Operational Research
Queen’s University Belfast

Due to the rapid growth of traffic density, the absolute necessity to increase the operational efficiency and capabilities of intelligent transportation systems (ITS) has led to the development of various traffic modelling theories. The major operational issue in developing traffic management models lies in the selection of a suitable methodological approach whilst making considerations, such as selection of the type of input and output data as well as the qualitative representation and the computational power of the model. Despite the considerable efforts in the area, there is still no approach which effectively models the various dynamically evolving features of traffic in road networks. On the one hand, deterministic continuum models based upon fluid dynamic equations are effective at modelling traffic dynamics along road sections but they fail to capture traffic patterns at junctions. On the other hand stochastic models based on historical observations successfully describe traffic patterns at junctions but they are inappropriate to describe traffic dynamics along road sections. This paper aims to introduce a new hybrid framework which combines the complementary features of continuum models e.g. [1], [2] with stochastic models e.g. [3] in order to effectively simulate traffic flow in road networks.

References


Tail dependence for two skew $t$ distributions

Thomas Fung
Department of Statistics
Macquarie University

We examine the tail dependence for two bivariate skew $t$ distributions, both of which generalise the bivariate symmetric $t$ distribution. The second, based on the bivariate skew-normal distribution, always displays positive asymptotic lower tail dependence, which suggests that this is the more appropriate skew extension of the symmetric bivariate $t$ in regard to tail dependence. The difference is associated with contrasting tail behaviour of the marginals. We will also present a new result regarding the rate of decay of the tail dependence coefficient for one of the skew $t$ distribution.

Keywords: Multivariate skew distribution, Skew normal distribution, Skew $t$ distribution, Tail dependence

References


Non-intrusive inference for stochastic simulators

Vincent Garreta, John Haslett
Department of Statistics
Trinity College Dublin

We review three methods of inference which do not require the computation of the likelihood but only the simulation of new data from the model; we describe these methods as non-intrusive. This is useful when the statistical model is available in the form of a computer model, seen as a black box "stochastic simulator". There is therefore no direct access to a "likelihood". Typically such simulators involve the implicit integration over very many variables that are internal to the simulator and are thus latent. The methods are: (i) Indirect inference, (ii) Approximate Bayesian computation and (iii) Emulation.

These methods have emerged in different fields of science and at different times, but they are very close in practice. In contrast to standard ‘likelihood-based inference’, classical or Bayesian, such methods require estimation of the likelihood from a necessarily finite number of simulations. This "extra-likelihood" estimation means that these methods are only an approximation to likelihood-based methods, were the likelihood to be available. Adopting a common formalism, we show similarities of practice and differences of interpretation underpinning these methods. We discuss potential improvements and apply the three methods on an example in climatology.

Keywords: Inference, Stochastic simulator, Approximate Bayesian computation, Indirect inference, Emulation of stochastic simulators
Improving Statistical Literacy through Graduate Programs in Ethiopia: the Case of North-South-South Collaborative Project in Jimma University

Yehenew Getachew¹, Kora Tushune¹, Luc Duchateau²
¹College of Agriculture & Veterinary Medicine, Jimma University, Ethiopia
²Faculty of Veterinary Medicine, Ghent University, Belgium

Introduction: Jimma University (JU) is a public university mandated to run graduate and undergraduate programs. Recently, JU has ranked first among all public and private universities in Ethiopia.

Challenges: In addressing the mandates, JU has faced chronic problems in the areas of Statistics, because of insufficient professional statisticians.

Opportunities: As a source of professionals, JU critically searched for statisticians working in statistics offices, research centers and other universities. The existence of an Inter-University collaborative program with Belgian universities was the second opportunity. With aim of pulling together all these opportunities, JU proposed a NSS project in Statistics, with South partners, Jimma, Hawassa, AddisAbaba, Gonder and Mekele Universities from Ethiopia and Eduardo Mondlane University, from Mozambique, and North partners, Ghent, Hasselt, Leuven Universities from Belgium.

A promising start: this collaborative initiative has resulted in running a joint graduate program, in various fields of applied statistics in JU and other sister universities in Ethiopia; and also enabled us to offer various refresher statistical trainings for the needy community in Ethiopia and east Africa in general (www.NSSbistat.ugent.be).

Keywords: North-South-South, NSSbiostat, Statistical Literacy, Jimma University, Inter-University Program
Deblurring signals with wavelets and lasso

Caroline Giacobino (joint work with Sylvain Sardy)
Section de Mathématiques
Université de Genève

Deblurring is an important class of inverse problems. The standard model is

\[ Y_n = Kf(t_n) + \epsilon_n, \quad n = 1, \ldots, N \]

where \( K \) is the blurring operator, \( f \) is the underlying function, \( \epsilon_n \) is the noise, and \( Y_n \) are the data. To detect sharp features such as narrow peaks in the underlying function \( f \), we derive a nonlinear estimator based on wavelets and the lasso [1]. The use of the lasso ensures the sparse wavelet representation of the estimated function. The estimator is controlled by regularization parameters selected to minimize the Stein unbiased risk estimate [2]. We obtain remarkable results in comparison with existing estimators in terms of mean squared errors, especially with \( \Sigma \)-rescaling of the predictors [3] and cycle spinning [4].

References


Modelling an Orthopaedic Integrated Clinical Assessment and Treatment Service (ICATS) using Analytic and Simulation Approaches

Jennifer Gillespie¹, Sally McClean¹, ², Francis FitzGibbon², Bryan Scotney¹, Frank Dobbs³ and Brian J. Meenan²

¹School of Computing and Information Engineering, University of Ulster
²Multidisciplinary Assessment of Technology Centre for Healthcare Programme, Nanotechnology and Integrated Bioengineering, University of Ulster
³Faculty of Life and Health Science, University of Ulster

Trauma and Orthopaedic departments were identified as having excessively long queues for diagnosis and treatment. This was created by a bottleneck when patients waited for an orthopaedic surgeon appointment before being placed on a treatment pathway. In 2005 the Department of Health introduced ICATS to England, Scotland and Wales, followed by Northern Ireland (NI) in 2007, to better manage orthopaedic referrals. The model initially implemented for ICATS implied that the process was deterministic. To analyse whether this was a realistic representation of the processes a stochastic model was created in Simul8 and compared with the original approach. This model was applied to data from the Southern Health and Social Care Trust (SHSCT) in NI, to assess the impact of variation on the real system. It was found that when variation was included, the system could not accommodate the number of referrals the department anticipated each year. Therefore, variation needs to be considered within orthopaedic processes otherwise, queues build up. We also modelled recent data from orthopaedic ICATS using a queueing network which allowed the problematic areas of the process to be found. Resources where reallocated within the department to ensure the system can cope with the anticipated number of referrals.
Rate Statistics and Simpson’s Paradox

Ryan Giordano, Google

Experimentally induced changes in aggregate rate statistics of interest to Google can be caused by changes in distribution among or changes within heterogeneous units. For example, a change in aggregate clicks per ad impression (click-through rate) can be caused both by changes in the distribution of impressions among heterogeneous users with different inherent click through rates or by individual users changing their propensity to click on ads. Changes in distribution can cause a phenomenon sometimes known as “Simspon’s Paradox”.

Similarly, some statistics, such as revenue per ad impression, can be decomposed into the changes in the products of independently meaningful quantities, such as in the identity

\[
\frac{\text{revenue}}{\text{ad impressions}} = \frac{\text{revenue}}{\text{clicks}} \times \frac{\text{clicks}}{\text{ad impressions}}
\]

These effects have qualitatively different implications for Google, and it is desirable to distinguish between them.

It is often the case in practice that the units of interest can be very numerous and may be missing data in one of the experimental arms. Furthermore, due to the large data set size, it is desirable to have online methods for measuring variance, since all the data may not fit simultaneously in memory on a single machine. Some practical techniques for dealing with these problems will be discussed.
Joint Modelling of Multiple Network Views

Isabella Gollini
School of Mathematical Sciences
University College Dublin

Recently there has been a growing interest in the modelling of network data. In many cases, different network views on the same set of nodes are available. It can therefore be useful to build a model that jointly summarises the information given by all the network views.

Latent space models (LSM) for network data were introduced by Hoff et al. (2002) under the basic assumption that each node has an unknown position in a \( D \)-dimensional Euclidean latent space: generally the smaller is the distance between two nodes in the latent space, the greater the probability of them being connected. In LSM the posterior distribution can not be calculated analytically, so we use a variational Bayes approach to estimate the model.

We propose a latent variable model that merges the information given by all the networks assuming that the probability of a node being connected with other nodes in each network are explained by the same latent variable.

To estimate this model we propose an EM algorithm: we use the parameters obtained from fitting a LSM for each network independently to find the joint posterior distribution. Then these results are used to update the estimate of the parameter in each LSM.

This model is demonstrated on an excerpt of 50 girls from ‘Teenage Friends and Lifestyle Study’ data.
Estimation of Parameters in DNA Mixture Analysis

Therese Graversen and Steffen Lauritzen
Department of Statistics
University of Oxford

In [1] a Bayesian Network for analysis of mixed traces of DNA was presented using Gamma-distributions for modelling peak sizes. It was suggested that a parameter in the variance of a peak size should be adapted to each new trace analysed. We discuss how the variance parameter can be estimated by maximum likelihood. The unknown proportions of DNA from each of the contributors can be estimated by maximum likelihood jointly with the variance-parameter. Furthermore we discuss how to incorporate prior beliefs about the parameters via Markov Chain Monte Carlo methods. The proposed estimation methods are illustrated through a few examples of applications of the model relevant for use in crime cases.

References

The project “combined firm data for Germany”  
Access to combined business micro data

Christopher Gürke  
Research Data Centre  
Federal Statistical Office Germany, Wiesbaden

In many European countries efforts are being made by official statistics to integrate business data from different sources. Sometimes such efforts aim at a combination of administrative data and survey data. In other cases the objective might be a linkage of survey data and process generated data or an integration of different registers. Two major reasons can be given for such activities: First of all, the matching of existing data sets occasionally constitutes a faster and cheaper way to gather certain information than the collection by means of a survey and it facilitates a reduction of the response burden of the reporting duties of enterprises and their establishments. Second, the integration of data from different sources often enables novel and more comprehensive economic analyses. There are usually some methodological problems to solve though: unique identifiers – for instance may not be available or relevant units may not be included in all data sources. A German project directly located in this context is the project “Combined Firm Data for Germany” (KombiFiD). It is carried out by the Federal Statistical Office, the Institute for Employment Research of the Federal Employment Agency, the German Federal Bank and the Leuphana University of Luebeck. The project is funded by the Federal Ministry of Education and Research (BMBF). One of the features of KombiFiD is the fact that official business data is collected independently by several institutions (Statistical Offices, Federal Employment Agency, German Federal Bank). The cross-institutional combined micro data, worked out in the project, offers new steps in accessing and analyzing business data not only in an economic sense also with respect to inspections of register versus survey data.

Keywords: record linkage, business micro data, official statistics

References


Learning in Bayesian hierarchical models for longitudinal and survival data

Laura Hatfield
Division of Biostatistics
University of Minnesota School of Public Health

In studying the evolution of disease and treatment effects on it, investigators may consider repeated measures of disease severity and time to occurrence of a clinical event. The development of joint models for such longitudinal and survival data has benefited from a latent variable framework. This approach posits an individual-level latent process evolving in time and contributing to the two outcome types. It flexibly incorporates association among outcomes, including across repeated measurements and between longitudinal and survival outcomes.

The literature has extended joint models to accommodate many features of real-life data. However, less attention has been paid to the properties of these models. In particular, we are interested in the contributions of each data type (longitudinal and survival) to Bayesian learning about individual- and population-level parameters.

As a motivating application, we consider a clinical trial for treatment of mesothelioma, a rapidly fatal lung disease. The trial protocol included longitudinal patient-reported outcome collection and followed patients until progression or death to determine progression-free survival times. We develop a joint model and quantify the contributions of the data sources to inference on parameters of scientific interest, comparing the real data results to a few small simulation studies.
A method of covariate adjustment in randomized clinical trials

Masayuki Henmi
Department of Mathematical Analysis and Statistical Inference
The Institute of Statistical Mathematics

In randomized clinical trials for comparison of two treatments, auxiliary baseline covariates are often measured for each patient, as well as response variables describing results of the treatments. Although a standard statistical analysis for the treatment effect uses only the response variables (and information of treatment assignment), it is well-recognized that the baseline covariates can be used to improve the efficiency of inferences. This paper proposes a method for this purpose, which is motivated by the remarkable phenomenon of the propensity scores in observational studies [1]. That is, it can be shown that estimating the assignment probability with the auxiliary covariates improves the efficiency of the estimator for the treatment effect. The performance of this method is investigated through the theory and simulation studies, and compared to some other methods including the classical ANCOVA and more recent ones such as Zhang et al [2].

Key Words: Covariate adjustment; Inverse probability weighting; Propensity scores; Randomized clinical trials.

References


The model web is a concept based on the publishing, discovery and consumption of models over the web, creating exciting opportunities for model reuse and sharing. The FP7 UncertWeb project (http://www.uncertweb.org/) extends the model web concept by accounting for uncertainty in model web inputs and outputs. We will consider some of the statistical challenges involved in addressing a complex problem by bringing together different component models to form a model chain that is both uncertainty-enabled and web-enabled.

To explore these statistical issues (including uncertainty quantification and propagation, computational efficiency, and spatial rescaling), we will concentrate on an example model chain for predicting land-use in arable farming and crop yield response to climatic and economic change. This model chain is composed of a set of models that consider land classification, the simulation of field usage and yield prediction in order to produce a regional estimate of future crop production. Considering the uncertainty throughout this model chain is important, because along with predictions about possible future scenarios for crop production, we want to have an appreciation about how likely the scenarios are to occur.

Keywords: Model chaining, Uncertainty propagation, Emulation, Land-use modelling.
Predicting poverty headcount using light household survey data, the case of WMS in Malawi

Willie Kachaka,
National Statistical Office,
Malawi

An individual is considered poor if his or her consumption or income falls below a certain threshold. This threshold defines the poverty line. We wish to predict the headcount ratio, i.e., the proportion of individuals with consumption below a given poverty line.

Consumption typically consist of the aggregated consumption of food items, non-food items, consumer durables and housing. Because the unit in the survey is the household, one needs to adjust total household consumption for the number of members in each household. The simplest solution is to adjust for the number of individuals living in the household.

Fortunately, Malawi, as probably the one of the countries in Sub-Sahara Africa, that has had a recent large scale household survey program. It began in 1998 and covers eight households expenditure surveys up to the most recent survey in 2010.

To test the models predictive ability it is critical that the consumption aggregates are comparable between the surveys and that there are sufficiently identical indicators (explanatory variables). As the Malawi household surveys rely on similar sampling procedures and questionnaires, and substantial work has been done to ensure comparability (see Appleton et al., 2001), they are suitable for our testing. The 1998 IHS1 survey, however, differs too much with respect to core indicators, and it is therefore not used in the analyses.
The comparison of population curves deserves much attention, including densities, regression curves, survival functions and some other characteristic functions of the variables of interest. This comparison can be done either in a parametric or a nonparametric way. In this article, we are interested in performing the comparison of several regression curves in a nonparametric context. We present a graphical device, called SiZer (SIgnificant ZERO crossing of the differences), for the test of the equality of several regression curves. Extensive literature exists on SiZer for applying in various fields. For the comparison of curves, these include references [1] and [2]. Our method is based on a scale-space approach and a ANOVA type statistics. It gives insightful information about the differences of the curves by combining statistical inference with visualization. A broad numerical study is conducted to demonstrate the sample performance of the proposed tool.

**References**


Wavelets offer relatively simply methods for the smoothing of signals, which adapt well and automatically to the form of the signal which is being smoothed. In recent years, wavelet based denoising methods have proven popular due to their highly adaptive nature and there has been a great deal of research interest concerning both wavelet denoising methodology and thresholds. However, much of this research has been upon the families of real-valued, compactly supported Daubechies wavelets. Conversely, there has been comparatively little work on denosing methods using complex-valued wavelets. Barber and Nason [1] showed complex wavelets to be an effective tool when denoising real-valued data. Moreover, most of the methods concerning the denoising of complex-valued data separate the signals into disconnected real and imaginary channels, electing to deal with them individually. We extend complex-valued wavelet methods for denoising of real-valued data to native denoising of complexvalued data, whereby we deal with real and imaginary parts together.

References

Bayesian Learning of Neural Networks using Genetic Algorithms

Ozan Kocadagli
Department of Decision Sciences
The George Washington University

The aim of this study is to estimate parameters and hyperparameters in the Bayesian Neural Networks (BBN) by Genetic Algorithms (GA) for regression problems. Therefore, the learning process is reduced to unconstrained optimization problem, and then regularization hyperparameters as well as an output noise hyperparameter are determined in order to minimize the objective function corresponding to maximizing posterior distribution of network parameters. In application part, example problems that confirm superiority of BBN with GA to traditional networks are given.

Keywords: Bayesian Neural Networks, Bayesian Learning, Genetic Algorithms, Stochastic Optimization.

References


Maximum t-statistics for combining multiple covariates

Osamu Komori
The Institute of Statistical Mathematics, Tokyo, Japan
Shinto Eguchi
The Institute of Statistical Mathematics, Tokyo, Japan

The Lasso type and the boosting methods have been widely paid attention and been shown to be useful in various scientific fields. The Lasso type method shrinks the number of covariates by imposing $L_1$ penalty; while the boosting is the forward stagewise procedure to combine multiple covariates in the iterative manner. This paper discusses t-statistic for a score function which linearly combines covariates, or functions of covariates. We propose the maximization method of square t-statistics for the score function in the linear form along with the forward and backward covariate selections, which we call t-Boost and t-Lasso, respectively. We show that the score function derived from the maximization of the square t-statistics is equivalent to the Bayes rule in a theoretical setting, which confirms one of the optimality for classification accuracy. We also investigate the performance of the proposed methods in simulation studies as well as real data sets in order to illustrate its utility.

Keywords: t-statistics; Boosting; Lasso.
Determining the Spreading Potential of a Disease from Incompletely Observed Counts of New Cases

Andrea Kvitkovičová and Victor M. Panaretos
Institute of Mathematics
Ecole Polytechnique Fédérale de Lausanne

We study the problem of estimating the spreading potential of a disease during the initial stages of an epidemic. We base the estimation on daily counts of new cases, taking into account that only a part of all cases is reported and that the reported individuals may be prevented from (or limited in) spreading the disease further.

We model the daily counts of new cases by an incompletely observed discrete-time Galton–Watson branching process. Limiting reported individuals in spreading the disease is translated into allowing the observed process to affect the unobserved one. We are interested in estimators for the first two moments of the offspring distribution, as the offspring mean is related to the spreading potential of the disease, and the second moment is connected with the variability of the mean estimators.

We propose non-parametric estimators analogous to those that can be used when all cases are reported, but adjusted to the different dependence structure in the observed process. The estimators are consistent and asymptotically normal conditionally on the explosion set, and serve as a basis for asymptotic confidence intervals.
Detection of critical points for the reliability of the physical tasks

Liubov Markovich
Moscow Institute of Physics and Technology State University

In many physical problems it is important to detect a critical point that is affecting some physical characteristics, e.g., electrons charge and temperature, Debye length and Curie temperature. For example, dependence of anodic current against the magnetic induction on the solenoid, theoretically looks like a step-wise function dropping to zero at some point that determines electrons charge. In practice this drop is continuous, so location of this point is indefinite.

We propose a method of its detection. Suppose we have a nonlinear curve pieces of underlying plots. The task is to estimate the change points in their convexity. We select points randomly from the $\epsilon$- neighborhood of a middle point between a minimal and maximal values of each of the plots. Sets of these points have to be approximated by linear models which can be averaged. A slope of the line obtained is estimated by the least squared method. This slope determines required characteristics. The linearity of the curve is verified by $\chi^2$ test. The accuracy of this line determines the validity of the points selection.

References


Using Principal Component Analysis and Cluster Analysis for the Reliability Problems in Reversed Items

Nur Fatihah Mat Yusoff
School of Mathematics, Statistics and Applied Mathematics
National University of Galway, Ireland

This study relates to an investigation of differential response to reversed items in multi-dimensional instruments. Data was collected at the University Malaysia Sarawak, Malaysia, where 899 students from the second year and above were randomly selected as respondents. A questionnaire was administered that included aspects of personality, motivation, learning strategies, academic achievement and background. We expected that all the statements which measure the same dimension would project in the same direction in the biplot, once the score of negative statements was reversed to agree with positive statements. While this expectation was almost true for the variables with no negative statements, for the variables with positive and negative statements, the biplot result from PCA showed contradictory patterns. These findings provide an insight into the existence of an inconsistency of responses, which leads to reliability and validity issues. In order to study this problem, further analysis was performed by using cluster analysis. Our aim from these analyses was to identify the group of respondents who gave inconsistent responses that caused reliability and validity problems.
A Comparison of Marginal and Multilevel models:
Modelling haemoglobin variability for haemodialysis patients

McCrink, Lisa[1]; Marshall, Dr Adele H.[2]; Cairns, Dr Karen J.[3]
Centre for Statistical Science and Operational Research (CenSSOR)
Queen’s University Belfast, United Kingdom

Previous research has found that variability of haemoglobin levels for individuals undergoing renal replacement therapy is related to a patient’s reduced survival [1]. This research focuses on modelling each patient’s haemoglobin level through time, using longitudinal data analysis techniques, with the aim to uncover factors that may influence haemoglobin fluctuation. Two longitudinal techniques, marginal models [2] and multilevel mixed effects models [3], are built using the statistical computer programme SAS and compared in terms of their predictive capabilities and ease of use. This research involves a complete case analysis.

The sample of data used for this analysis contains the repeated measurements of 5860 haemodialysis patients, with approximately 60,000 observations in total collected by the UK Renal Registry over a three year period.

**Keywords:** Longitudinal data; Marginal models; Multilevel models.

**References**


Clustering in networks with the collapsed stochastic blockmodel

Aaron McDaid, Brendan Murphy, Nial Friel and Neil Hurley
Clique Research Cluster
University College Dublin

An MCMC algorithm is created for the Stochastic Block Model [1]. Given a network, this model allows us to cluster the nodes of the network such that nodes with similar role in the network are clustered together.

Our algorithm is based on the allocation sampler [2] as we show that it is possible to integrate out many of the parameters in the SBM. This allows us to create a more efficient algorithm.

The algorithm and model directly provide us with an estimate of the number of clusters, as the model can naturally support a state space of varying dimension. We experiment with artificial and real data.

References


Ordinal data arise in many contexts and item response modelling is a long established method for analysing this type of data.

The ordinal response for individual $i$ on item $j$ is denoted $Y_{ij}$, where $i = 1, \ldots, N$ and $j = 1, \ldots, J$. Corresponding to each ordinal data point $Y_{ij}$ is a latent Gaussian variable $Z_{ij}$. The value of $Y_{ij}$ is observed to be level $k$ if the latent Gaussian variable $Z_{ij}$ lies within a specified interval. In addition, another latent Gaussian variable $\theta_i$, often called a latent trait, is used to model the underlying attributes of individual $i$. The mean of $Z_{ij}$ depends on $\theta_i$ which need not be uni-dimensional.

The extension to a mixture of item response models is explored here, these models provide clustering capabilities in the context of ordinal data. In this context the mean of $Z_{ij}$ depends on the group to which individual $i$ belongs through group specific parameters.

Many naturally arising data sets contain both continuous and categorical variables. Information is often lost by either treating the categorical variables as continuous or discretising the continuous variables. Quinn(1999) [1] proposed a model for mixed data which is an amalgamation of factor analysis and item response models. Quinn’s approach takes into account the nature of each of the variables which leads to more accurate results. An extension of this model to a mixture model is considered here. This extension would allow clustering of mixed data into homogeneous groups. These models are estimated within the Bayesian paradigm using a Markov Chain Monte Carlo algorithm and are applied to illustrative data sets.

References

Using Conditional Phase-Type Distributions to Model Hospital Length of Stay of Geriatric Patients-Identifying Patient Characteristics Influencing Discharge Destination.

Janette McQuillan, Adele H. Marshall and Karen J. Cairns
Centre for Statistical Science & Operational Research (CenSSOR)
Queen’s University Belfast

The increase in elderly patients in our population is putting a strain on already limited resources available for geriatric patient care. Coxian phase-type distributions are a special type of Markov model that describes duration until an event occurs in terms of a process consisting of a sequence of latent phases. Previous research has shown that Coxian phase-type distributions can suitably represent a patient’s length of stay in hospital [1]. The conditional phase-type (C-Ph) distribution extends the Coxian phase-type distribution by conditioning it on a Bayesian Network [2]. This paper models geriatric patient length of stay using a C-Ph distribution, based on a data set of 25,586 patients admitted to one of the six key acute hospitals in Northern Ireland. The C-Ph model identifies the key factors influencing a patient’s hospital discharge destination and models the length of stay corresponding to each discharge destination.

Keywords: elderly; Coxian/conditional phase-type distributions; Markov model; Bayesian network.

References


Ignoring overdispersion in hierarchical loglinear models: Possible problems and solutions

Elasma Milanzi, Ariel Alonso, Geert Molenberghs
Interuniversity Institute for Biostatistics and Statistical Bioinformatics
University of Hasselt

Poisson data frequently exhibit overdispersion and, for univariate models, many options exist to go around this problem. Nonetheless, in complex scenarios like, for example, in longitudinal studies, accounting for overdispersion is a more challenging task. Recently, [1] presented a model that accounts for overdispersion by combining two sets of random effects. However, introducing a new set of random effects implies additional distributional assumptions for intrinsically unobservable variables. Using the combined model as a framework, we explored the impact of ignoring overdispersion in complex longitudinal settings via simulations. Furthermore, we also evaluated the effect of misspecifying the random effects distribution on both the combined model and the classical Poisson hierarchical model. Our results indicate that even though inferences may be affected by ignored overdispersion, the combined model is a promising tool in this scenario.

Keywords: Poisson-normal model; Overdispersion; Hierarchical; Combined model; Type I error.

References

The Effect Of Observations on the Bayes Factor of Bayesian Principal Component Analysis

Rofizah Mohammad  
Department of Mathematics  
University of Surrey, UK

One of the methods for Bayesian model comparison is via the pair wise Bayes factor [2]. The purpose of the study is to see the effect of observations on a Bayes factor of Bayesian principal component analysis. We use a fully Bayesian principal component analysis model proposed by Oh and Kim (2010) [3]. We use Gibbs sampling to estimate the parameters of the model. The marginal likelihood also estimated from the Gibbs output using the approach proposed by Chib (1995)[1]. The method was illustrated by using data generated from the multivariate normal distribution.

Keywords: Bayesian principal component analysis, Bayes factor, Gibbs sampling, influential observation, model choice

References


F tests with random sample size - Applications

Elsa Moreira and João T. Mexia
CMA - Research center in Mathematics and Applications
Faculdade de Ciências e Tecnologia
Universidade Nova de Lisboa

Christoph Minder
University of Bern

When we cannot consider as known the dimensions $n_1, \ldots, n_k$ of the samples for the $k$ levels of the one-way layout, it is more correct to consider these as realizations of poisson random variables with parameters $\lambda_1, \ldots, \lambda_k$. This situation arises mostly when we have a given time span for collecting the observations and this collecting is carried out separately for the different levels. A good example is the obtention of data during a time span for the comparison of pathologies. The number of patients that arrive to an hospital with different pathologies during a period, is not known in advance. The data is obtained from the patients with each pathology as soon as they present themselves. Once the data are in hand, an F test statistics with F distribution conditional to the number of observations can be obtained [1]. The approach can also be used in a phase of planning studied, before the data are in. Given a predefined test power $1 - \beta$ and a significance level $\alpha$, the F distribution assuming random sample sizes is used in order to obtain the minimal sample size and correspondent duration a study needs. The application of F tests with random sample sizes in this situations is further demonstrated through practical examples.

Keywords: ANOVA; F tests; F distribution; Random sample sizes; Power analysis

References

A Bayesian Analysis Of Partial Orders

Alexis Muir Watt
Department of Statistics
University of Oxford

A time series of rankings or “total orders” is analysed to infer the presence and timing of a potential change-point reflecting a sudden change in hierarchy between nodes. This is done by considering an underlying process of partial orders from which the total order observations are sampled. A prior model for partial orders is adapted from the k-dimensional random orders reviewed in [1] to include the time series aspect.

Allowing for uncertainty in the sequence ordering in time of the observations, a hybrid MCMC/particle filtering approach based on [2] simulates the posterior distribution on partial orders. Related work include the maximum likelihood partial orders for conjunctive Bayesian networks [3], but there is no well-developed Bayesian framework for partial orders.

This work is joint with Dr Geoff Nicholls.

References


Nonprofit organisations provide important public goods and noncash benefits to the community. The need to produce estimates of the size and growth of the sector in Hungary was created internally by transformation of the former socialist regime and externally by the desire to conform to the SNA that defines a sector of nonprofit organizations supplying services to households.

The HCSO has been carrying out annual surveys on the nonprofit sector since 1993. During this period a special method of data collection and processing has been developed. The presentation reflects the first steps, the problems of various definitions used in nonprofit sector, and the different registers providing fundamental information on nonprofit organisations. Hungarian experts worked out a special hot-deck imputation method, the so called donor-finder program to make the database comprehensive. The presentation points out not only the methodology, but its function in the practice describing the features of the Hungarian civil sphere. The data on the size, composition, finance and human resources of the sector can show a satisfying picture to foreign experts having interest in nonprofit research.

References

Distribution of the Largest Eigenvalue of Multivariate Beta Distribution and Gaussian Ensembles

Rajendran Narayanan
Department of Statistical Science
Cornell University

The largest eigenvalue of the multivariate beta distribution (also called the greatest root statistic) plays an important role in classical multivariate analysis. It describes the null distribution of the union intersection test for many classical problems such as canonical correlations, testing for equality of covariance matrices, multiple response multivariate regression and MANOVA among others. Historically, the evaluation of its exact distribution has required extensive tables or use of specialised software. More recently, Johnstone [1] proved that the asymptotic distribution of the logit transform of the greatest root statistic, under suitable centring and scaling, converges to the Tracy-Widom law. We derive the exact c.d.f of the largest eigenvalue of the multivariate beta distribution in terms of the elements of a matrix that have an explicit scalar analytical form and as such are computationally tractable. Using a similar approach, we also present the exact c.d.f of the largest eigenvalue of real and complex Gaussian ensembles that has been a long standing open problem in random matrix theory.

Keywords: Canonical correlations, extreme eigenvalue, multivariate beta distribution, greatest root statistic, MANOVA, random matrix theory, Gaussian ensemble, Jacobi Ensemble

References

A Flexible Joint-Modelling Framework for Longitudinal and Time-to-Event Data with Overdispersion

Edmund Njeru Njagi
Interuniversity Institute for Biostatistics and statistical Bioinformatics
University of Hasselt

It is common, in medical research, for outcomes of different types to be conjunctively collected from the study subjects. A longitudinal (continuous, binary or count) outcome may be recorded, alongside the time to a certain event. Moreover, the time-to-event outcome may or may not be repeatedly collected, with the former occurring when the event of interest is recurrent. In modelling repeated binary, count and time-to-event outcomes, [3] propose a flexible framework in which both conjugate and normal random effects are included, in order to simultaneously relax the mean-variance prescription and address data hierarchy. They present the framework in the context of a single repeated outcome. We hereby recast this versatile framework into the context of joint modelling. We formulate joint models for pairs of jointly collected outcomes, where, for each type of outcome, a model including both conjugate and (shared) normal random effects is considered. We derive the joint marginal distributions for all the settings, and, by using partial marginalization [2], estimate the models by the maximum likelihood method. We exemplify the application of these models using data collected from a study on liver cirrhosis.

Keywords: Weibull-Gamma-Normal model, Poisson-Gamma-Normal model, Probit-Beta-Normal model, Partial marginalization.

References


The Millennium Development Goal in Nigeria: How far have we gone?

Nwosu C.P., Babatunde O.O., Aliyu Saidu
Central Bank of Nigeria

The Millennium Development Goals (MDGs) embodies the key human development expressed first in a millennium declaration adopted by 189 countries in September 2000 and are set to be achieved by 2015. The paper examined the progress of Nigeria towards achieving the MDGs by 2015. It covers the eight goals relating to poverty reduction, education, gender equality, child and maternal health, combating HIV/AIDS and other diseases, environmental sustainability and global partnership for development. Recently implemented policies are accelerating the achievement of these Goals. These gains are based on sustained economic growth, improvements in planning and policy, and growing government investment in the social sector. Challenges still remain that many families still do not have access to safe water and sanitation; although children start primary school, many of them do not complete their primary education. Even though more mothers and children are surviving, avoidable deaths are still unacceptably common. None of the MDGs are certain to be achieved by 2015 and so more needs to be done on each.

Keywords: MDG, Goals, Achievements, Indicators, Way forward
Longitudinal Metabolomic Data Analysis

Gift Nyamundanda$^1$, Claire Gormley$^1$ and Lorraine Brennan$^2$

$^1$ School of Mathematical Sciences
$^2$ School of Agriculture, Food Science and Veterinary Medicine
University College Dublin

The term metabolomics is used to describe the study of small molecules present in biological samples. Data sets from metabolomic studies are typically high-dimensional and complex. In a longitudinal metabolomic study, multiple metabolites are measured from subjects at multiple time points.

Principal component analysis (PCA) is currently the most widely used statistical technique for analyzing metabolomic data. However, the application of PCA to longitudinal metabolomic studies is limited since it assumes independence of the repeated measurements and it is not based on a statistical model. Probabilistic principal component analysis (PPCA) [1, 2] addresses some of the limitations of PCA. Here, we propose an extension of PPCA called dynamic PPCA (DPPCA) which allows PPCA to model longitudinal metabolomic data by taking into account the correlation due to repeated measurements. This is achieved by assuming an autoregressive model on some of the parameters of the PPCA model.

The DPPCA model allows us to observe the change in position of subjects in the latent principal subspace and to identify the spectral regions responsible for the data structure at each time point.

References


Pre- and Post-Enrolment Factors Influencing The Academic Success Of Third Level Students

Leanne O’Callaghan¹, Kathleen O’Sullivan¹, Noel Woods², Dermot Keane¹, Gillian Kissane¹, Eoghan O’Brien¹, Peter O’Driscoll¹
¹Department of Statistics, ²Centre for Policy Studies, University College Cork, Ireland

This study examined if pre- and post-enrolment factors contribute to academic success, defined as whether or not a student achieved his/her desired grade. Participants were 157 second-year university students enrolled in a business economics module. Logistic regression analysis of pre-enrolment factors identified that two contributed significantly: Leaving Certificate (LC) points \((p = 0.001)\) and highest LC grade subject \((p = 0.02)\). Similar analysis on post-enrolment factors revealed that three were significant: lecture attendance \((p = 0.007)\), number of study hours \((p = 0.001)\) and living at home \((p = 0.02)\). When both groups of factors were considered together, logistic regression analysis identified the same pre- and post-enrolment factors. The results indicate that a student who is most likely to achieve his/her desired grade has some of the following characteristics: achieved more than 490 LC points, attained their highest grade in LC Mathematics, lives at home, attended more than 70% of their lectures, and studied for less than two hour per week. The influence of study hours is an anomaly, possibly due to under reporting and timing of the survey. Educators can use these findings to inform students about the best approach to maximise their chances of academic success.
Clustering approach to autonomous robot navigation

Lukas Pastorek, Tomas Vintr and Hana Rezankova
Department of Statistics and Probability, Faculty of informatics and statistics
University of Economics, Prague, Czech Republic

Clustering, as a powerful statistical tool for the multivariate data analysis, has been used as an untraditional approach for autonomous robot navigation issue. The Speeded-Up Robust Features (SURF) algorithm, which creates map of the surrounding environment, consisting of high-dimensional feature vectors (descriptors), enables to navigate robot in different environments. The aim of unsupervised classification of SURF descriptors, across all the recorded images, is to construct map of cluster prototypes. Computed prototypes are consequently used in the robot decision making. Clustering includes nature-inspired Self-Organizing Map (SOM), Neural Gas (NG) and fuzzy set theory-inspired Fuzzy C-Mean (FCM) algorithms.

Keywords: SURF, Autonomous robot navigation, Clustering, FCM, SOM, Neural Gas

Acknowledgement: This work is supported by project IGA VSE F4/5/2011.

References


On MCMC Methods for Conditioned Diffusion Processes and Indirectly Observed Diffusions

Erik Pazos
Department of Statistical Sciences
University College London

MCMC algorithms provide for a way to simulate diffusion processes that might not have an analytical solution. We examine HMC (see [1]) and use it to solve a class of problems. Our discussion starts at using HMC to simulate conditioned diffusion processes. In the past HMC has been modified to simulate conditioned processes in Hilbert Space [2], for diffusion processes with constant diffusion coefficient. We expand this by modifying HMC to work on diffusion processes with Non-Constant diffusion coefficient by using the Wilkinson-Golightly transformation (see [3]). We continue our discussion by looking at inferring parameters from indirectly observed diffusion processes. In this case we use as an example the case where real data is explained by a stochastic volatility model where we wish to infer on the volatility of the process.

References


Predicting Censored Semiconductor Lifetimes with Bayesian Regression Models Using Mixtures of Experts.

Kathrin Plankensteiner, Olivia Bluder
KAI Kompetenzzentrum Automobil- und Industrie-Elektronik
Alpen-Adria University of Klagenfurt

In this paper different Bayesian regression based Mixtures of Expert models (MoE) are proposed to predict censored semiconductor lifetime data obtained under different electrical and thermal conditions. For all models it is assumed that the distribution of the lifetime data is log-normal [1] and the mixture weights depend on the test settings. As a first approach independent normal and inverse gamma priors are applied. Due to censored observations these priors are not conjugate and MCMC simulations are necessary to determine the posterior density distribution. Therefore the statistical toolbox bayesf Version 2.0 [2] for MATLAB has been extended to censored data. For the model selection goodness of fit criteria such as Bayesian Information Criterion (BIC), marginal likelihoods, predictive density distributions and Probability Integral Transform (PIT) are used [3]. For the prediction of censored semiconductor lifetime, physical models (e.g. Coffin-Manson) and linear regression models based on test settings have been used. The results indicate that both models are valid, but the physical approach is more accurate.

References


The impact of water in economic growth and social welfare: case study of Lesotho

Katiso Ramalebo
Lesotho Bureau of Statistics

Water development is essential to growing economies as industries and agricultural projects need water for their day-to-day operations. Lack of safe and reliable access to clean water is a critical constraint to economic growth and poverty reduction. Unclean water has serious implications for the world’s poor in relation to improving basic health and productivity.

As Peter et al. (2002) advised, water is important to bolster the process of economic development essential for life and health.

Does water affect economic growth and social welfare?

The HBS 2002/03 shows that poverty was high among households headed by subsistence farmers. About 10.9% of males are subsistence farmers in Lesotho (CMS 2009/10). In Lesotho, irrigation is only responsible for 1% (0.6 Mm3) of water use each year (Aquastat Lesotho, 2009).

References

Social capital and human capital for measuring sustainable development

Emanuela Recchini
ISTAT Istituto Nazionale di Statistica
Italia

Sustainable development is one of the general objectives of the European Union for the improvement of life quality and the well-being of present and future generations. In this perspective, all the dimensions of progress—economic, environmental and social—are equally important. Unfortunately, the social dimension is perhaps the weakest pillar of sustainable development, due to contrasting theoretical and analytical bases and the insufficient understanding for interactions with economic and environmental dimensions.

Many studies concentrate on the theories, definitions and measurement of social and human capital, two major concepts that strive to explain the socio-economic development of a society. These, however, are intangible and multi-dimensional and cannot be measured directly.

In addition to presenting measurement attempts made by some national statistical institutions, this paper deals with the relationships between social and human capital and the activities carried out by the Italian National Institute of Statistics using the approach suggested by the OECD and the Stiglitz Commission, with the aim of supplying the country with a shared framework of the evolution of the main economic, social and environmental phenomena.

Keywords: Social capital, human capital, sustainable development, well-being indicators.

References

The effect of individual and household factors on income inequality in Italy, 1998–2008

Andrea Regoli, Rosalba Manna
Department of Statistics and Mathematics for Economic Research
University of Naples

The regression-based decomposition method combined with the Shapley value approach gives the opportunity of quantifying the contribution to the inequality of a set of factors, while taking the correlations among them into account. The aim of this paper is to measure the relative contributions of individual as well as household factors to the explanation of the inequality in individual disposable incomes. The factors are introduced as explanatory variables in an income generating model that is estimated through a time-invariant unobserved random effects model on panel data (Baltagi, 2002) from the Survey of Household Income and Wealth (SHIW, version 6.0) conducted every two years by the Bank of Italy. The factors that play a dominant role in explaining the observed inequality are educational level and gender.

Keywords: Inequality decomposition; regression-based methods; Shapley value; Gini index; panel data.

References

Currently, 75% of the material used to create a crown is discarded. This study examined the viability of reusing heat-pressed glass-ceramic in the production of crowns without affecting the crowns physical properties. The most important property is biaxial flexural strength. An ingot was heatpressed to form a disc and this disc was detached from the button and sprue. The button and sprue were re-pressed to produce a second disc. This procedure was repeated twice more, to produce a total of four discs from the ingot. Ten such ingots were processed for a total of 40 discs. Biaxial flexural strength (MPa) was calculated for each disc. The data was analysed by fitting a general linear model. The effect of re-pressing was not statistically significant ($p = 0.18$). The pairwise differences were assessed for potential clinical significance; four comparisons indicated that their differences may be clinically significant. However, a power analysis identified that this study did not have adequate power to detect clinically meaningful differences. In addition, other material properties such as Vickers hardness and fracture toughness, which reflect a crown's durability and wear, should also be considered when evaluating the effects of repeated processing.
Simulation and Analysis of Genotype by Environment Interactions and QTL by Environment Interactions

Paulo Canas Rodrigues
Faculty of Sciences and Technology
Nova University of Lisbon

The study and understanding of the interactions between genotypes and environments (location x year combinations), and between QTL (quantitative trait loci) and environment, are major challenges in plant breeding and genetics. Within the last years simulation tools such as genotype to phenotype models (i.e. non-linear functions of purely genotypic components and environmental inputs which integrates over time) models have been used to better understand these interactions [1].

In this paper a physiological genotype to phenotype model with seven physiological parameters is presented. This simple model simulates two-way tables, with genotypes and environments, of a complex trait such as yield. Moreover, it was able to simulate crossover interactions. A sensitivity analysis based on factorial regression methods [2] is developed to investigate which interactions between environmental (e.g. radiation, temperature, CO$_2$) and genotypic (physiological model parameters, marker information) covariables are significantly important to explain the interaction.

References


Kisses are Better than Hugs: An Analysis of Sentiment Bearing Terms in Twitter

Michael Salter-Townshend
School of Mathematical Sciences
University College Dublin

We have a large volume of tweets from the micro-blogging site twitter.com upon which we perform sentiment analysis. We employ a set of commonly occurring words and emoticons to seed a semi-supervised algorithm. These words and ascii expressions such as :-) have been manually labelled as carrying positive or negative sentiment.

We present a simple model in which the number of positive words per tweet is Binomially distributed with the number of trials being the number of words in the tweet. Each tweet has a latent sentiment index such that the expected number of positive words in a tweet is this index times the number of words in the tweet. Our goal is then to obtain estimates for the unknown tweet sentiment indices and the unobserved labels for all words not appearing in the seed list.

We employ an Expectation Maximisation algorithm where the E step calculates the expected sentiment for the unlabelled words and the M step finds maximum-likelihood estimates for the tweet sentiments. We present results and discuss extensions including geographic heterogeneity of sentiment, multiple word meanings and fuzzy labelling of words.
Variable Screening and Parameter Estimation for High-Dimensional Generalized Linear Mixed Models Using $\ell_1$-Penalization

Jürg Schelldorfer and Peter Bühlmann
Seminar für Statistik
Departement of Mathematics
ETH Zurich

We propose a two-stage procedure for modeling high-dimensional generalized linear mixed models. Generalized linear mixed models [2, 3, 5] are an extension of generalized linear models for clustered observations. In the first step, we perform a Lasso-type [1] variable screening algorithm to reduce the number of active variables. In the second step, we perform ordinary maximum likelihood estimation including only the selected variables of the first step.

We present the methodology for fitting high-dimensional generalized linear mixed models and the implementation in the corresponding R package glmmlasso. Moreover, we demonstrate the performance of the suggested two-stage approach on simulated as well as on real data examples.

This work can be seen as an extension of [4], which illustrates the procedure for modeling high-dimensional Gaussian linear mixed models.

Key words: GLMM, Laplace approximation, Lasso, random-effects model, variable selection

References


Sliced Inverse Regression for Cox’s proportional hazard model

Maya Shevlyakova
Chair of Applied Statistics
Ecole Polytechnique Fédérale de Lausanne

Dimension reduction is one of the many popular techniques for handling multivariate data. We consider one of its versions, sliced inverse regression (SIR) and we study its performance on survival data. Although we perform a similar method as in ([1]), our approach is different. The first time we slice the survival times directly and we use the results to estimate the hazards. The censored observations are taken into account at this point, during slicing, as each of them is assigned (under the equal weights) to all of the remaining slices. The original version of the algorithm ([2]) allows us then to estimate the effective dimension-reduction directions, which we plug in the Cox’s model to get the hazards. The subsequent run of the algorithm performs slicing on the estimated hazards, which corresponds better to the theory of the SIR since the linear relationship in the Cox’s model is established via hazards. The results of the second run are final and allow to analyze the original data structure.

A number of simulations were performed to evaluate how well the method estimates the coefficients of the underlying Cox’s model, depending on the percentage of censoring and sample size. Although it never recovers their true values, the model’s structure can be deduced successfully. Larger components are easier to recover, while true zero components remain of the order $1/\sqrt{n}$. A large sample size can compensate for a severe case of censoring. Unlike popular regularization methods, the response does not necessarily have to be sparse, which allows a more general application of a method.

References


A comparative simulation study of heteroscedasticity consistent covariance matrix estimators in the linear regression model

Yegnanew Alem Shiferaw
Hawassa University
Ethiopia.

In the context of econometric methods of estimation OLS estimates derived under the assumption of homoscedasticity are not consistent when there is heteroscedasticity and their use can lead to incorrect inferences. Thus this paper sets out to examine the performance of several modified versions of heteroscedasticity consistent covariance matrix (hereafter HCCM) estimator of White (1980) and white and Mackinnon (1985) over a range of sample sizes. Most applications that use HCCM appear to rely on HCO, yet tests based on HCCM was found to be consistent even in the presence of heteroscedasticity of an unknown form. However, based on Monte Carlo experiments, which compare the performance of t statistic, HC2 and HC3 estimators precisely perform better in small samples than the rest. HC3 estimator for samples less than 100 was found to be better than the other HCCM estimators; when samples are 250 or larger, other versions of the HCCM can be used. Added to that, it was cost advantageous to employ HC3 instead of OLSCM even when there is little evidence of heteroscedasticity.

Key words: White estimator, Monte Carlo Simulation, Linear Regression, Heteroscedasticity
Self Assessment of Health in Rural Malawi Women: A Dynamic Probit Model

Jupiter Simbeye
School of Politics and International Relations
University College Dublin
and Mathematical Sciences Department
University of Malawi

This study investigates the role of household shocks and other important household factors on rural Malawi women's self-assessed health. We exploit a panel data consisting of two waves collected in 2007 and 2009 using a dynamic probit model. We also account for individual-specific random effects. The results show that women whose households had experienced serious illness/injury of a household member had an increased probability of about 20% into reporting poor health. In addition, respondents from households which suffered property theft had an increased probability of about 8% into reporting poor health. Poor health state dependence is also evident. The accumulation of wealth has the potential of reducing the probability of assessing oneself into poor health. This study demonstrates that interventions targeted at reducing the general morbidity of household members have promising returns in improving the health of women caregivers.

Keywords: caregiver, health, household shocks, self-assessment, women

References

HIV-1 Incidence among Patients attending the Sexually Transmitted Infections Clinic at Queen Elizabeth Central Hospital in Blantyre-Malawi

Isaac Singini, MSc1; Fatima Zulu, MSc1; Esther Gondwe, PhD1; Johnstone Kumwenda, FRCP2
1 Johns Hopkins Research Project.
2 The College of Medicine, University of Malawi

Sexually transmitted infections are well known risk factors for HIV acquisition. Data is lacking on determinants of drivers of HIV infections in populations with generalized HIV epidemic. We decided to determine the incidence of HIV-1 infection in patients attending the STI clinic in Blantyre Malawi and the cofactors for HIV infection among this population.

Patients 18 years and older attending the STI clinic were counseled for HIV testing. HIV antibody negative patients are recruited and followed up every 3 months. At each visit demographic, socio-economic, sexual behavior, clinical and laboratory data is collected. Data is analysed using Poisson regression modeling for each co-factor in a step-wise forward selection until an adequate model is fitted. Departure from Poisson assumptions will be assessed using formal statistical diagnostic tests (overdispersion).

Out of a total 400 participants recruited between February 2010 and April 2011 approximately 40% are male. The HIV-1 incidence rate for the recruited participants is 84 per 1000 person years. A measure of goodness of fit of the Poisson regression model was obtained using deviance statistic of a base-line model against a fuller model and it showed an adequate fit for the data captured. The rest of the Poisson assumption are validated by the study design e.g. independence.

The HIV incidence in this group of patients is more than twice than seen pregnant women in Blantyre. The Poisson regression model fits adequately the data that has been captured so far. It would, however, be interesting to investigate if the data would be better explained by other regression models like the Cox proportional hazards model after a full follow-up period.
Extreme Value Theory approach in modelling
Exchange Rates of V4 countries

Alena Tartálová
Faculty of Economics
Technical University of Košice
Slovakia

This paper proposed an extreme value approach to estimate the exchange rates volatility. We analyzed exchange rates of Visegrad countries CZK/EUR, PLN/EUR and HUF/EUR. We applied two models, one is the Block Maxima Method based on Fisher-Tippet theorem and the other is the Peaks over Threshold Model which models the observed values exceeding a large threshold. In the first method is important to choose the number of blocks which we made by using graphical methods. In the second method is important to find suitable threshold. Using both methods we calculated high quantiles of exchange rates and compared results.

Keywords: Extreme Value Theory, Method of block maxima, Peaks Over Threshold Method.

Acknowledgement: This work was supported by the Slovak Scientific Grant Agency as part of the research project VEGA 1/0127/11 Spatial Distribution of Poverty in the European Union.

References


We consider the problem of hypothesis testing when the data consist of random curves. Our study is motivated by the quest to understand the shape characteristics of short strands of DNA imaged at random configurations in solution. This is posed as an inference problem in an infinite dimensional Hilbert space, and as such, gives rise to special challenges involving ill-posed problems [1, 2]. Even the usually simple task of testing a simple hypothesis on the mean of a sample of iid observations involves an inverse problem in this setting, and methods incorporating appropriate regularization are necessary. We review procedures based on spectral truncation and Tikhonov regularization, and apply them to the analysis of DNA shape.

References


The relationship between scarcity of natural resources and their real prices

Roland Tóth
Hungarian Central Statistical Office
Economic Statistician
Budapest

There has been a long running concern about resource depletion. Some argue this concern is misplaced, while others consider it to be an urgent problem requiring immediate action. Economists suggest that long term prices, adjusted for inflation (real prices), provide a useful and effective indicator of resource scarcity. This study tests this hypothesis in consideration of the accepted theory that traditional price deflators, such as the US consumer price index, overestimate inflation-, and accordingly-, are likely to underestimate long term commodity prices.

To investigate the usefulness of real prices as an indicator of scarcity, a case study of two metals considered to be expensive (platinum and rhodium) and two considered to be relatively inexpensive (copper and lead) was used. Real long term price indices were constructed and econometric analysis used to determine the direction and significance of long-term price trends and whether real prices were correlated with other scarcity indicators such as the Reserves-to-production ratio.

The results show, when an appropriate adjustment is made to the deflator, long-run trends in real metal prices are all upward, and there is a significant relationship between the real prices and scarcity indicators, such as the reserves-to-production ratios, for platinum and rhodium, but not for copper and lead. These findings suggest that real prices of platinum and rhodium are more affected by their scarcity, while copper and lead prices are likely to be more dependent on other factors such as high substitutability with other virgin and recycled materials.

Key words: non-renewable resources; depletion; real price; inflation bias
Bayesian Inference with Shape Mixtures of Log-Normal Distributions

Catalina Vallejos
Department of Statistics
University of Warwick

A new family of distributions for positive random variables is being introduced. It can be interpreted either as the exponential of a random variable distributed as a scale mixture of normals or as an infinite shape mixture of log-normal distributions. This new class covers a wide range of shapes, in particular cases in which the lognormal cannot capture the tail behaviour of the data. It includes the already studied log-Laplace, log-logistic, log exponential power and log Student-t distributions among others. Some properties such as the existence of moments and the Fisher Information Matrix are studied, and Bayesian inference under the commonly used Jeffreys style prior distributions is examined.

Keywords: log-normal, log-Laplace, log-logistic, log exponential power, scale mixtures of normals, Bayesian inference, life distributions.

References


An application of Structural Equation Modeling to test, in persons with chronic disease, the effect mediator of Optimism between Personality and Quality of Life

Estela Vilhena, IPCA-Barcelos; ICBAS-University of Porto, Portugal
José Luís Pais-Ribeiro, FPCE-University of Porto; UIPES-Lisbon, Portugal
Isabel Silva, University of Fernando Pessoa, Portugal
Luísa Pedro, UIPES; ESTeSL IP, Lisbon, Portugal
Rute Meneses, University of Fernando Pessoa, Portugal
Helena Cardoso, ICBAS-University of Porto; Hospital Center of Porto, Portugal
Madalena Abreu, FPCE-University of Porto, Portugal
Vera Melo, FPCE-University of Porto, Portugal
Ana Silva, Hospital Center of Porto, Portugal
António Martins da Silva, ICBAS-University of Porto;
Hospital Center of Porto, Portugal
Denisa Mendonça, ICBAS, ISPUP-University of Porto, Portugal

Optimism is an important variable that has consistently been shown to affect adjustment to quality of life in chronic disease. The aim of present study was to clarify if dispositional optimism exerts a mediating influence on the personality traits (neuroticism and extraversion) - quality of life association (QoL - measured by the general well-being and mental health) in Portuguese patients with chronic diseases. Structural Equation Models (SEM) were used to test the quality of the hypothesized model, estimating the simultaneous effects of personality traits on QoL, and evaluated the mediator effect of dispositional optimism. A sample of 729 patients was recruited from central hospitals in various districts of Portugal. All completes self-reported questionnaires assessing personality, optimism and QoL. The hypothesized model fitted the data reasonably well. It was found that dispositional optimism exerts a simultaneous mediating effect between neuroticism - general well-being and neuroticism - mental health association and, between extraversion and general well-being. The results suggest ”the expectation that good things will happen” contributes to a better quality of life. Interventions in patients should be in order to achieve and maintain a level of optimism that helps to facilitate and improve the quality of life.

Keywords
Chronic disease, Mediator, Optimism, SEM
The Effect of Zinc Phosphate Cement Composition on Diametral Tensile Strength

Amanda Wall\textsuperscript{1}, Kathleen OSullivan\textsuperscript{1}, John OSullivan\textsuperscript{1}, Craig Keane\textsuperscript{1}, Philip O’Mahoney\textsuperscript{1}, Frank Burke\textsuperscript{2}, Noel Ray\textsuperscript{2}, Gerry McKenna\textsuperscript{2}, Jill McKenna\textsuperscript{2}

\textsuperscript{1}Department of Statistics, \textsuperscript{2}Cork University Dental School & Hospital
University College Cork, Ireland

This study investigated if variations in the proportioning of powder and liquid affect the diametral tensile strength of the cement used in crowning. Thirty eight students followed the manufacturers instructions to proportion the powder and liquid and recorded the amount of powder and liquid they dispensed. The maximum and minimum values for powder and liquid were identified. A single operator then created 34 specimens for each of three compositions: maximum powder, minimum liquid; minimum powder, maximum liquid; and the manufacturers composition. The diametral tensile strength of each specimen was measured and analysed using a general linear model. Analysis of the students recorded data for powder and liquid identified that the amounts of each dispensed varied widely from the manufacturers guidelines. Composition did have a significant effect on diametral tensile strength ($p < 0.001$). Although the difference between maximum powder minimum liquid and the manufacturers composition was statistically significantly, clinical significance was not achieved. Both had significantly higher diametral tensile strengths compared to minimum powder maximum liquid. The current packaging of powder and liquid together with the manufacturers instructions incurs variability in the quantities of each ingredient dispensed. The packaging should be revised to allow for more accurate dispensing.
Using Node resampling in Classification and Regression Trees as an alternative to Random Forest approaches

Deirdre Wall, Alberto Alvarez, John Hinde and John Newell
School of Mathematics, Statistics and Applied Mathematics, National University of Ireland, Galway

CART (Classification and Regression Trees) construct models which are obtained by recursively partitioning the data and fitting a simple prediction model to each partition [1]. Regression trees involve response variables that take continuous or ordered discrete values, with prediction error typically measured by the square difference between the observed and predicted values.

Random Forests, an extension to CART, are an ensemble method involving resampling of the data [2]. In addition to constructing each tree using a different bootstrap sample of the data, each node is split using the best split from a small subset of variables randomly chosen at that node, instead of using the best split from all the variables. The advantage in Random Forests is a reduction in prediction error however no information relating to the underlying structure in the covariate space is available.

We will investigate the performance of resampling data at the node level, which results in a single tree, thereby retaining structure, while retaining the robustness of Random Forests.

A practical example involving biomedical data will be given in order compare the performance of the node resampling and forest approaches.

References

[1] Wei-Yin Loh: *Classification and Regression Trees*

Completion Time Dynamics of Doctoral Studies at Makerere University: A Hazard Model Evaluation

Robert Wamala
School of Statistics and Applied Economics
Makerere University

Issues related to completion time of doctoral studies are certainly an internationally challenging and important area of higher education literature. In this paper, completion time dynamics of doctoral studies at Makerere University were investigated based on data extracted for all 295 candidates in the commencement cohorts for 2000 to 2005. The total elapsed time, from first enrollment to submission of a final copy of a thesis, was adopted as a measure of completion time; event history (survival) analysis methodology was applied. Results reveal a median completion time of 5.0 years. Following a Cox model, in a range of candidate, candidature, discipline and institutional variables, the rate of completion was higher for candidates at the younger ages during commencement, international students, registered in science-related disciplines and those in the commencement cohorts for 2000 to 2002. The model correctly identified the order of completion times by about 72% of the time.

Key words: Time-to-completion, Higher Degree studies, Makerere University
Modeling maize price volatility in the East African Market

Jonesmus Wambua
International Livestock Research Institute
Nairobi

The global food system has become more susceptible to periods of extreme price volatility that impact on food security especially among the vulnerable groups. Highly volatile food prices subject market players to difficulties in planning ahead and adjusting to fluctuating market signals. This paper investigates maize prices volatility in two East African (EA) countries; Kenya and Tanzania. The exponential GARCH model was used on monthly maize price data for the period 2003 to 2010, obtained from national institutions of the countries. The findings show strong evidence that in both markets; price volatility is not prone to new market information (good or bad news) but sensitive to market events. Further, price volatility was found to be greater and more persistent in Kenya than in Tanzania throughout the period under investigation. Interestingly, price volatility in Kenya reduced significantly after 2005 when the country withdrew the import tariff on maize from other EAC member countries. The findings of this study show that prices of staples can be stabilized through increased cross border trade.

Keywords: Exponential GARCH model, volatility, maize price, Kenya and Tanzania.
Dirichlet Processes for Quantal Response Equilibrium Models - Examination of Pegged Exchange Rate Policies

Anton H. Westveld
Department of Mathematical Sciences
University of Nevada - Las Vegas

Building on work of McKelvey and Palfrey ([2]) on quantal response equilibrium models, and Newton, Czado, and Chappell ([3]) on semiparametric Bayesian methods, we develop a Bayesian inferential approach for a particular class of semiparametric strategic choice models. In particular, we consider modifications of the centrally standardized finite approximation to the Dirichlet process prior of Newton, Czado, and Chappell ([3]) in order to accommodate both dynamic (extensive form) and simultaneous (normal form) strategic choice models in the context of both observational and experimental data. We apply the methodology to examine strategic decision processes of governments with pegged exchange rates in relation to speculative currency attacks utilizing a data set and strategic game outlined by David Leblang ([1]).

References


Detection of User Roles in Message Board Forums
Using Variational Bayes Methods

Arthur White
School of Mathematical Sciences
University College Dublin

Forums are a form of social media which allows high volume user interaction in a complex manner on a variety of topics. In many cases, insight into a forum may be gained by examining user behaviour. However, due to the size and complexity of interaction, manually grouping users can be problematic. Here, we develop techniques to cluster user behaviour [1] to incorporate a fully probabilistic model.

User behaviour is categorised into several features, such as persistence, popularity and reciprocity, which are then summarized using descriptive statistics such as degree, response rate and post-length. We model the data using a mixed membership model which makes the assumption that user behaviour can be described by a fixed number of extreme profiles. This model extends previous mixed membership models for binary data [3], [2] and Gaussian data [4] to data types generated from a combination of members of an exponential family. The model is fitted using a variational EM algorithm.

Methods to identify the correct number of extreme profiles are also examined. Illustrations of this approach, with both simulated data and data from the boards.ie web forum are discussed.

References


Post-model-selection Estimators of Entropy
Functional
Małgorzata Wojtyś
and Jan Mielniczuk
Department of Mathematics and Information Sciences
Warsaw University of Technology, Poland

We study the properties of post-model-selection estimators of entropy $K(f) = \int f \log f$ based on such model selection rules as Bayesian Information Criterion, minimal p-value criterion and maximal p-value criterion in the case when considered list of models is of the form

$$\mathcal{M}_i = \left\{ f_{\theta}(x) = \exp\left( \sum_{j \in i} \theta_j b_j(x) - \psi_i(\theta) \right) I(x \in [0,1]), \ \theta_j \in \mathbb{R} \text{ for } j \in i \right\},$$

(1)

where $i \subset \{1, \ldots, m\}$ for some fixed $m \in \mathbb{N}$, $(b_j(x))_{j=1}^m$ is a vector of orthonormal functions and $\psi_i(\theta)$ is a normalising constant.

Let $\hat{i}$ be any model selection rule taking values in $2^{\{1, \ldots, m\}}$ and based on a random sample $X_1, \ldots, X_n \sim f$. Given $\hat{i}$, the post-model-selection estimator of entropy is defined as $\hat{K}(f) = \int_{\hat{i}} \hat{f}_i(s) \log \hat{f}_i(s) \, ds$, where $\hat{f}_i$ is maximum likelihood estimator of $f$ in family $\mathcal{M}_i$.

In the case of exponential families of densities of the form (1) $\hat{K}(f)$ simplifies to the form

$$\sum_{j \in \hat{i}} \hat{\theta}_j^{\text{ML}} b_j - \psi_i(\hat{\theta}_i^{\text{ML}}),$$

where $\hat{\theta}_j = \frac{1}{n} \sum_{i=1}^n b_j(X_i)$.

We prove that $\hat{K}(f)$ based on any conservative model selection rule is consistent and asymptotically normal under appropriate assumptions. The results are generalized to the case when the logarithm of $f$ has infinite expansion with respect to $(b_j(\cdot))_{j=1}^\infty$. We give the rate of convergence of $\hat{K}(f)$ to $K(f)$ or, in case when the true model is not on the list, to $K(f^*)$, where $f^*$ is an information projection of $f$ onto the model which is nearest to $f$ in the sense of Kullback-Leibler distance, among the considered models.

We also present results of simulation study showing the advantage in mean integrated squared error coming from incorporating a model selection rule in the estimation process.

References


Solving the Sum of Euclidean Distances Clustering Problem by the Hyperbolic Smoothing Method

Vinicius Layter Xavier
Systems Engineering and Computer Science
Department of the Federal University of Rio de Janeiro – UFRJ
Ilha do Fundão, Centro Tecnológico

The considered clustering problem corresponds to the minimization of the sum of the euclidean distances of observations to their cluster centroids. The mathematical modeling of this problem leads to min-sum-min formulations which in addition to its intrinsic bi-level nature, has the significant characteristic of being strongly non-differentiable and non-convex problem with a large number of local minima. The hyperbolic smoothing strategy solves a sequence of low dimension differentiable unconstrained optimization sub-problems, which gradually approaches the original problem. The reliability and efficiency of the method are illustrated via a set of computational experiments. It must be emphasized that the proposed methodology can be analogously applied to the solving of the Fermat-Weber location problem.

Keywords: Cluster analysis, Fermat-Weber location, Global Optimization, Min-Sum-Min Problem

References


Cost Ratios for Cost and ROC Curves

Hyun Sang Yoo, Chong Sun Hong
Department of Statistics
Sungkyunkwan University

For binary classification problems on mixture distribution, a threshold based on cost functions is an optimal from the viewpoint of minimum expected cost. Assuming that there is no cost information, we propose cost ratios in the expected cost corresponding to thresholds where the total accuracy and the true rate are maximized, and explain the relation of these cost ratios minimizing the expected cost. With comparing the normalized expected costs when classification accuracy is maximized, another cost ratios are also proposed. Values of these cost ratios are located between two cost ratios for the expected costs based on classification accuracies, and converge to that of the minimum expected cost. This work suggests two cost ratios: one is minimized both the expected cost and the normalized expected cost, and the other in the expected cost and the normalized expected cost functions which are maximized classification accuracies. We discuss their compatibility based on the relation of these cost ratios.

Keywords: Classification accuracy, Credit evaluation, Default, Expected cost, Discriminative power, Threshold.
Horvitz-Thompson-like estimators with random inclusion probabilities for abundance estimation

Yuan Yuan, David L. Borchers, Stephen T. Buckland, Joanne M. Potts, Len Thomas
The Centre for Research into Ecological and Environmental Modelling
School of Mathematics and Statistics
University of St Andrews
Elizabeth D. Clarke and Paul G. Fernandes
Marine Scotland - Science, Marine Laboratory

Abstract: Random-effects models for detection or capture probability are frequently used in the analysis of mark-recapture surveys and are sometimes used for other surveys, particularly in ecology. They are appropriate when capture/detection probability depends on unobserved variables, and when it depends on observed variables but inferences are to be made about a broader population from which the observed variable values are drawn. We develop Horvitz-Thompson-like abundance estimators in the context of wildlife surveys. The estimators are applicable to any survey in which estimated inclusion probabilities have random effects, including surveys of human populations. We apply them to a variety of wildlife survey datasets, using a variety of survey methods, and investigate their performance by simulation. We find that a generalization of the form of estimator typically used on line transect surveys performs best overall. It has low bias, and also the lowest bias and mean squared error among all the estimators we considered.

Keywords: random-effects, detection/capture probability, simulation study, mark-recapture, distance sampling, trapping point transect, trawl survey
Approaches to Subjective Poverty Lines Estimation

Tomáš Želinský
Faculty of Economics
Technical University of Košice
Slovakia

Subjective poverty lines are inherently subjective judgements people make about what is a socially accepted minimum standard of living in a particular society [1]. Estimates of subjective poverty lines are usually based on so called “discrete information” [2] or minimum income question. Assuming a binary variable (0: not poor; 1: poor) allows us to use (among other methods) logistic regression and sensitivity/specificity curves. Assuming a minimum income question linear regression approach can be used. Another possibility is to use price sensitivity meter [3] which is a technique for determining consumer price preferences primarily used in marketing.

Keywords: Subjective poverty, poverty lines, logistic regression, sensitivity/specificity curves, MIQ, PSM.

Acknowledgement: This work was supported by the Slovak Scientific Grant Agency as part of the research project VEGA 1/0127/11 Spatial Distribution of Poverty in the European Union.

References


High-dimensional covariance estimation based on Gaussian graphical models

Shuheng Zhou, Department of Statistics, University of Michigan
Philipp Rütimann, Seminar for Statistics, ETH Zürich
Min Xu, Machine Learning Department, Carnegie Mellon University
Peter Bühlmann, Seminar for Statistics, ETH Zürich

Undirected graphs are often used to describe high-dimensional distributions. Under sparsity conditions, such graphs can be estimated using $\ell_1$-penalization methods. We propose and study the following high dimensional covariance estimation method called Gelato (Graph estimation with Lasso and Thresholding). We combine a multiple regression approach with ideas of thresholding and refitting: first we infer a sparse undirected graphical model structure via thresholding of each among many $\ell_1$-norm penalized regression functions; with this graphical structure we then estimate the covariance matrix and its inverse using the maximum likelihood estimator. We show under suitable conditions that this approach yields consistent estimation in terms of graphical structure and fast convergence rates with respect to the Frobenius norm for the covariance matrix and its inverse. We compare our method Gelato to the GLasso [2] and Space [3] in simulations and real data examples.

Keywords: Graphical model selection, covariance estimation, Lasso, node-wise regression, thresholding

References


ISI Jan Tinbergen Award: abstracts
Comparative study of working conditions of children from farm households in Burkina Faso, Côte d’Ivoire and Mali

Kodzovi Senu Abalo
ENSEA, CERMMA
Abidjan, Côte D’Ivoire

Based on a series of surveys on the living and migration conditions of children, this study highlights the magnitude and characteristics of child labour in Burkina Faso, Côte D’Ivoire and Mali. Four major lessons were learnt: (I) the overwhelming majority of children working in violation of the international norms as issued by the International Labour Organisation (ILO); (II) the level of poverty influences child labour only in Côte D’Ivoire. The direction of this influence confirms the luxury axiom since poor households relatively send their children to work; (III) whatever the level of education of the head of the household, early entry of children into the labour market appears to remain a stable deal impacted by other factors. In addition, school does not seem to be a viable alternative discriminating child labour, as most working children are in school; (IV) finally, there is interdependence between the decisions to participate in the different activities studied.

Keywords : Child work ; child labor ; Harkin & Engel protocol ; logistic model ; multivariate probit regression model ; sequential model ; multidimensional poverty index ; luxury axiom.

Manjula Kalluraya
Department of Statistics, Mangalore University,
Mangalagangothri, Karnataka, India
and Kavitha, M.
Tata consultancy services, Bangalore, India

Students t-test is commonly used to test for equality of means of two independent normal populations. This test has an inherent assumption that although the variances of the two normal distributions are unknown, they are equal. The test is uniformly most powerful unbiased. When no knowledge is available regarding the variances of the normal distributions, Behren(1929) and Fisher(1935) provided an initial solution to the problem of testing for equality of means. Since then many researchers have worked on this problem. The commonly used procedures are due to Welch(1947) and Cochran(1964).

This paper compares these two procedures in terms of small sample coverage probability of the confidence interval for the difference in means and average length of the confidence interval. New tests are proposed for the Behren-Fisher problem which are distribution free; The tests are compared with the Welch(1947) and the Cochran(1964) tests in terms of coverage probability and the average length of the confidence interval. The proposed tests perform well compared to the Welch(1947) and Cochran(1964) tests.
Feasible Ridge Estimator in Partially Linear Models

M. Roozbeh
Department of Statistics, School of Mathematical Sciences
Ferdowsi University of Mashhad, Iran

In a partial linear model, some non-stochastic linear restrictions are imposed under a multicollinearity setting. Semi-parametric ridge that was proposed in the 1970’s to combat the multicollinearity and non-ridge type estimators, in a restricted manifold are defined. For practical aspects, it is assumed that the covariance matrix of the error term is unknown and thus feasible estimators are replaced and their asymptotic distributional properties are derived. Also, necessary and sufficient conditions for the superiority of the ridge type estimator over its counterpart for selecting the ridge parameter $k$ are derived. Lastly, a Monte Carlo simulation study is conducted to estimate the parametric and non-parametric parts. In this regard, kernel smoothing and cross validation methods for estimating the non-parametric function are used.

Keywords and Phrases: Linear restrictions; Kernel smoothing; Multi-collinearity; Feasible ridge estimator; Partial linear model
List of delegates
<table>
<thead>
<tr>
<th>Last Name</th>
<th>First Name</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Abalo</td>
<td>Kodzovi Senu</td>
<td>REPUBLIC OF TOGO</td>
</tr>
<tr>
<td>2 Alvarez-Iglesias</td>
<td>Alberto</td>
<td>IRELAND</td>
</tr>
<tr>
<td>3 Amann</td>
<td>Markus</td>
<td>GERMANY</td>
</tr>
<tr>
<td>4 Aslett</td>
<td>Louis</td>
<td>IRELAND</td>
</tr>
<tr>
<td>5 Banerji</td>
<td>Anashua</td>
<td>GERMANY</td>
</tr>
<tr>
<td>6 Barclay</td>
<td>Lorna</td>
<td>U.K.</td>
</tr>
<tr>
<td>7 Barry</td>
<td>Paul</td>
<td>IRELAND</td>
</tr>
<tr>
<td>8 Bergersen</td>
<td>Linn Cecilie</td>
<td>NORWAY</td>
</tr>
<tr>
<td>9 Bharath</td>
<td>Karthik</td>
<td>U.S.A.</td>
</tr>
<tr>
<td>10 Bhattacharya</td>
<td>Arnab</td>
<td>IRELAND</td>
</tr>
<tr>
<td>11 Boland</td>
<td>Fiona</td>
<td>IRELAND</td>
</tr>
<tr>
<td>12 Bolt</td>
<td>Sarah</td>
<td>AUSTRALIA</td>
</tr>
<tr>
<td>13 Brianto Contreras</td>
<td>Maylix</td>
<td>VENEZUELA</td>
</tr>
<tr>
<td>14 Brites</td>
<td>Nuno</td>
<td>PORTUGAL</td>
</tr>
<tr>
<td>15 Caimo</td>
<td>Alberto</td>
<td>IRELAND</td>
</tr>
<tr>
<td>16 Carfora</td>
<td>Alfonso</td>
<td>ITALY</td>
</tr>
<tr>
<td>17 Caruana</td>
<td>Mark Anthony</td>
<td>MALTA</td>
</tr>
<tr>
<td>18 Chamandy</td>
<td>Nicholas</td>
<td>U.S.A.</td>
</tr>
<tr>
<td>19 Chan</td>
<td>Edith L S</td>
<td>HONG KONG</td>
</tr>
<tr>
<td>20 Chebana</td>
<td>Fateh</td>
<td>CANADA</td>
</tr>
<tr>
<td>21 Chen</td>
<td>May-Ru</td>
<td>TAIWAN, R.O.C.</td>
</tr>
<tr>
<td>22 Chochola</td>
<td>Ondrej</td>
<td>CZECH REPUBLIC</td>
</tr>
<tr>
<td>23 Corkery</td>
<td>Elaine</td>
<td>IRELAND</td>
</tr>
<tr>
<td>24 Costa</td>
<td>Lilia</td>
<td>ENGLAND</td>
</tr>
<tr>
<td>25 Cox</td>
<td>David</td>
<td>U.K.</td>
</tr>
<tr>
<td>26 Crowe</td>
<td>Susanne</td>
<td>SWITZERLAND</td>
</tr>
<tr>
<td>27 Dahyot</td>
<td>Rozenn</td>
<td>IRELAND</td>
</tr>
<tr>
<td>28 Dantas</td>
<td>Tiago</td>
<td>BRAZIL</td>
</tr>
<tr>
<td>29 De Carvalho</td>
<td>Miguel</td>
<td>PORTUGAL</td>
</tr>
<tr>
<td>30 Dean</td>
<td>Nema</td>
<td>U.K.</td>
</tr>
<tr>
<td>31 Distiller</td>
<td>Greg</td>
<td>SOUTH AFRICA</td>
</tr>
<tr>
<td>32 Doan</td>
<td>Thinh</td>
<td>IRELAND</td>
</tr>
<tr>
<td>33 Doherty</td>
<td>Cathal</td>
<td>IRELAND</td>
</tr>
<tr>
<td>34 Domijan</td>
<td>Katarina</td>
<td>IRELAND</td>
</tr>
<tr>
<td>35 Donnelly-Swift</td>
<td>Erica</td>
<td>IRELAND</td>
</tr>
<tr>
<td>36 Dooley</td>
<td>Áine</td>
<td>IRELAND</td>
</tr>
<tr>
<td>37 Dooley</td>
<td>Cara</td>
<td>IRELAND</td>
</tr>
<tr>
<td>38 Ferguson</td>
<td>John</td>
<td>U.S.A.</td>
</tr>
<tr>
<td>39 Fernandes</td>
<td>Mauricio Vidotti</td>
<td>BRAZIL</td>
</tr>
<tr>
<td>40 Fitzgerald</td>
<td>Aidan</td>
<td>U.K.</td>
</tr>
<tr>
<td>Last Name</td>
<td>First Name</td>
<td>Country</td>
</tr>
<tr>
<td>-----------</td>
<td>------------</td>
<td>-----------------</td>
</tr>
<tr>
<td>Fung</td>
<td>Thomas</td>
<td>AUSTRALIA</td>
</tr>
<tr>
<td>Garreta</td>
<td>Vincent</td>
<td>IRELAND</td>
</tr>
<tr>
<td>Getachew</td>
<td>Yehenew</td>
<td>ETHIOPIA</td>
</tr>
<tr>
<td>Giacobino</td>
<td>Caroline</td>
<td>SWITZERLAND</td>
</tr>
<tr>
<td>Gillespie</td>
<td>Jennifer</td>
<td>NORTHERN IRELAND</td>
</tr>
<tr>
<td>Giordano</td>
<td>Ryan</td>
<td>U.S.A.</td>
</tr>
<tr>
<td>Gollini</td>
<td>Isabella</td>
<td>IRELAND</td>
</tr>
<tr>
<td>Gormley</td>
<td>Claire</td>
<td>IRELAND</td>
</tr>
<tr>
<td>Graversen</td>
<td>Therese</td>
<td>U.K.</td>
</tr>
<tr>
<td>Gutiérrez</td>
<td>Cristina</td>
<td>SPAIN</td>
</tr>
<tr>
<td>Guttorp</td>
<td>Peter</td>
<td>U.S.A.</td>
</tr>
<tr>
<td>Hatfield</td>
<td>Laura</td>
<td>U.S.A.</td>
</tr>
<tr>
<td>Healy</td>
<td>Mark</td>
<td>IRELAND</td>
</tr>
<tr>
<td>Henmi</td>
<td>Masayuki</td>
<td>JAPAN</td>
</tr>
<tr>
<td>Hernández</td>
<td>Belinda</td>
<td>IRELAND</td>
</tr>
<tr>
<td>Honchar</td>
<td>Oksana</td>
<td>UKRAINE</td>
</tr>
<tr>
<td>Houlding</td>
<td>Brett</td>
<td>IRELAND</td>
</tr>
<tr>
<td>Hurley</td>
<td>Catherine</td>
<td>IRELAND</td>
</tr>
<tr>
<td>Isham</td>
<td>Valerie</td>
<td>U.K.</td>
</tr>
<tr>
<td>Jamela</td>
<td>Thubelihle</td>
<td>SOUTH AFRICA</td>
</tr>
<tr>
<td>Jin</td>
<td>Renhao</td>
<td>IRELAND</td>
</tr>
<tr>
<td>Johnson</td>
<td>Jill</td>
<td>U.K.</td>
</tr>
<tr>
<td>Kachaka</td>
<td>Willie</td>
<td>MALAWI</td>
</tr>
<tr>
<td>Kang</td>
<td>Keehoon</td>
<td>KOREA</td>
</tr>
<tr>
<td>Kartsonaki</td>
<td>Christiana</td>
<td>U.K.</td>
</tr>
<tr>
<td>Kirdruang</td>
<td>Phatta</td>
<td>U.S.A.</td>
</tr>
<tr>
<td>Klapper</td>
<td>Jennifer</td>
<td>U.K.</td>
</tr>
<tr>
<td>Kocadaghi</td>
<td>Ozan</td>
<td>TURKEY</td>
</tr>
<tr>
<td>Komori</td>
<td>Osamu</td>
<td>JAPAN</td>
</tr>
<tr>
<td>Korhonen</td>
<td>Minna</td>
<td>FINLAND</td>
</tr>
<tr>
<td>Kvitkovicova</td>
<td>Andrea</td>
<td>SWITZERLAND</td>
</tr>
<tr>
<td>Madden</td>
<td>Jamie</td>
<td>IRELAND</td>
</tr>
<tr>
<td>Mai</td>
<td>Tiep</td>
<td>IRELAND</td>
</tr>
<tr>
<td>Manna</td>
<td>Rosalba</td>
<td>ITALY</td>
</tr>
<tr>
<td>Markovich</td>
<td>Liubov</td>
<td>RUSSIA</td>
</tr>
<tr>
<td>Marshall</td>
<td>Adele</td>
<td>NORTHERN IRELAND</td>
</tr>
<tr>
<td>Mat Yusoff</td>
<td>Nur Fatihah</td>
<td>IRELAND</td>
</tr>
<tr>
<td>McCarthy</td>
<td>Shane</td>
<td>IRELAND</td>
</tr>
<tr>
<td>McCrink</td>
<td>Lisa</td>
<td>NORTHERN IRELAND</td>
</tr>
<tr>
<td>McDaid</td>
<td>Aaron</td>
<td>IRELAND</td>
</tr>
<tr>
<td></td>
<td>Last Name</td>
<td>First Name</td>
</tr>
<tr>
<td>---</td>
<td>--------------</td>
<td>-------------</td>
</tr>
<tr>
<td>81</td>
<td>McMahon</td>
<td>Donal</td>
</tr>
<tr>
<td>82</td>
<td>McParland</td>
<td>Damien</td>
</tr>
<tr>
<td>83</td>
<td>McQuillan</td>
<td>Janette</td>
</tr>
<tr>
<td>84</td>
<td>Mehendale</td>
<td>Kavitha</td>
</tr>
<tr>
<td>85</td>
<td>Miettinen</td>
<td>Arto</td>
</tr>
<tr>
<td>86</td>
<td>Milanzi</td>
<td>Elasma</td>
</tr>
<tr>
<td>87</td>
<td>Mohammad</td>
<td>Rofizah</td>
</tr>
<tr>
<td>88</td>
<td>Moreira</td>
<td>Elsa</td>
</tr>
<tr>
<td>89</td>
<td>Muir Watt</td>
<td>Alexis</td>
</tr>
<tr>
<td>90</td>
<td>Murphy</td>
<td>Brendan</td>
</tr>
<tr>
<td>91</td>
<td>Murphy</td>
<td>Sarah</td>
</tr>
<tr>
<td>92</td>
<td>Nagy</td>
<td>Renáta</td>
</tr>
<tr>
<td>93</td>
<td>Narayanan</td>
<td>Rajendran</td>
</tr>
<tr>
<td>94</td>
<td>Njagi</td>
<td>Edmund Njeru</td>
</tr>
<tr>
<td>95</td>
<td>Nugent</td>
<td>Rebecca</td>
</tr>
<tr>
<td>96</td>
<td>Nwosu</td>
<td>Chioma</td>
</tr>
<tr>
<td>97</td>
<td>Nyamundanda</td>
<td>Gift</td>
</tr>
<tr>
<td>98</td>
<td>O’Sullivan</td>
<td>Laura</td>
</tr>
<tr>
<td>99</td>
<td>O’Brien</td>
<td>Eoghan</td>
</tr>
<tr>
<td>100</td>
<td>O’Loughlin</td>
<td>Eamonn</td>
</tr>
<tr>
<td>101</td>
<td>O’Mahoney</td>
<td>Philip</td>
</tr>
<tr>
<td>102</td>
<td>O’Sullivan</td>
<td>Kathleen</td>
</tr>
<tr>
<td>103</td>
<td>Panaretos</td>
<td>Victor</td>
</tr>
<tr>
<td>104</td>
<td>Pastorek</td>
<td>Lukas</td>
</tr>
<tr>
<td>105</td>
<td>Patel</td>
<td>Rajan</td>
</tr>
<tr>
<td>106</td>
<td>Pazos</td>
<td>Erik</td>
</tr>
<tr>
<td>107</td>
<td>Piela</td>
<td>Pasi</td>
</tr>
<tr>
<td>108</td>
<td>Plankensteiner</td>
<td>Kathrin</td>
</tr>
<tr>
<td>109</td>
<td>Portier</td>
<td>Francois</td>
</tr>
<tr>
<td>110</td>
<td>Preuss</td>
<td>Philip</td>
</tr>
<tr>
<td>111</td>
<td>Raftery</td>
<td>Adrian</td>
</tr>
<tr>
<td>112</td>
<td>Ramalebo</td>
<td>Katiso</td>
</tr>
<tr>
<td>113</td>
<td>Razpotnik</td>
<td>Barica</td>
</tr>
<tr>
<td>114</td>
<td>Recchini</td>
<td>Emanuela</td>
</tr>
<tr>
<td>115</td>
<td>Rocha</td>
<td>Scarlett</td>
</tr>
<tr>
<td>116</td>
<td>Rodrigues</td>
<td>Paulo</td>
</tr>
<tr>
<td>117</td>
<td>Roozbeh</td>
<td>Mahdi</td>
</tr>
<tr>
<td>118</td>
<td>Rütimann</td>
<td>Philipp</td>
</tr>
<tr>
<td>119</td>
<td>Salloum</td>
<td>Wael</td>
</tr>
<tr>
<td>120</td>
<td>Salter-Townshend</td>
<td>Michael</td>
</tr>
<tr>
<td></td>
<td>Last Name</td>
<td>First Name</td>
</tr>
<tr>
<td>---</td>
<td>----------------</td>
<td>---------------</td>
</tr>
<tr>
<td>121</td>
<td>Schelldorfer</td>
<td>Jörg</td>
</tr>
<tr>
<td>122</td>
<td>Sheerin</td>
<td>Padraic</td>
</tr>
<tr>
<td>123</td>
<td>Shevlyakova</td>
<td>Maya</td>
</tr>
<tr>
<td>124</td>
<td>Shiferaw</td>
<td>Yegnanew Alem</td>
</tr>
<tr>
<td>125</td>
<td>Simbeye</td>
<td>Jupiter</td>
</tr>
<tr>
<td>126</td>
<td>Simpson</td>
<td>Daniel</td>
</tr>
<tr>
<td>127</td>
<td>Singini</td>
<td>Isaac</td>
</tr>
<tr>
<td>128</td>
<td>Southern</td>
<td>Richard</td>
</tr>
<tr>
<td>129</td>
<td>Sweeney</td>
<td>James</td>
</tr>
<tr>
<td>130</td>
<td>Tartalova</td>
<td>Alena</td>
</tr>
<tr>
<td>131</td>
<td>Tavakoli</td>
<td>Shahin</td>
</tr>
<tr>
<td>132</td>
<td>Taylor</td>
<td>Jonathan</td>
</tr>
<tr>
<td>133</td>
<td>Tóth</td>
<td>Roland</td>
</tr>
<tr>
<td>134</td>
<td>Towey</td>
<td>Kieran</td>
</tr>
<tr>
<td>135</td>
<td>Tripathi</td>
<td>Shailesh</td>
</tr>
<tr>
<td>136</td>
<td>Vallejos Meneses</td>
<td>Catalina</td>
</tr>
<tr>
<td>137</td>
<td>Vilhena</td>
<td>Estela</td>
</tr>
<tr>
<td>138</td>
<td>Vintr</td>
<td>Tomas</td>
</tr>
<tr>
<td>139</td>
<td>Wainwright</td>
<td>Martin</td>
</tr>
<tr>
<td>140</td>
<td>Wall</td>
<td>Deirdre</td>
</tr>
<tr>
<td>141</td>
<td>Wall</td>
<td>Amanda</td>
</tr>
<tr>
<td>142</td>
<td>Wamala</td>
<td>Robert</td>
</tr>
<tr>
<td>143</td>
<td>Wambua</td>
<td>Jonesmus</td>
</tr>
<tr>
<td>144</td>
<td>Weiss</td>
<td>Emanuel</td>
</tr>
<tr>
<td>145</td>
<td>Westveld</td>
<td>Anton</td>
</tr>
<tr>
<td>146</td>
<td>White</td>
<td>Arthur</td>
</tr>
<tr>
<td>147</td>
<td>Wojtys</td>
<td>Malgorzata</td>
</tr>
<tr>
<td>148</td>
<td>Xavier</td>
<td>Vinicius Layter</td>
</tr>
<tr>
<td>149</td>
<td>Yoo</td>
<td>Hyun Sang</td>
</tr>
<tr>
<td>150</td>
<td>Yuan</td>
<td>Yuan</td>
</tr>
<tr>
<td>151</td>
<td>Zelinsky</td>
<td>Tomas</td>
</tr>
</tbody>
</table>