Inference on Multiple Volatile Time-series Given Irregular Observations

Thinh Doan, Dr. Andrew Parnell & Prof. John Haslett

TCD Statistics Seminar Series 2012/2013

11th Feb 2013
MULTIPLE IRREGULARLY OBSERVED TIME SERIES

observations
MULTIPLE IRREGULARLY OBSERVED TIME SERIES

observations

inference target
**BACKGROUND**

- Many **real world** datasets which are irregularly observed: weather / climate, astronomy, finance, clinical trial, etc.

- **Aim:** stochastic interpolation at arbitrary time scale and obtain summarized statistics of interest

- When data are irregular, using standard approach for regularly observed time series (e.g. **ARIMA**) might not be appropriate
Univariate Irregularly Observed Time Series

Reconstruction: to draw sample paths and compute summarized statistics of the true underlying process
**Simple Kriging**

- Unobserved continuous process: \( \mathbf{x} \)
- Irregularly observed data: \( \mathbf{y} \) at times \( \{t_1 = 2, t_2 = 10, t_3 = 14\} \)
- Inference target: \( \mathbf{x} \) at times \( \{0, 1, 2, ..., 13, 14\} \)

- This is known as **Kriging** where the objective is to estimate the predictor \( \hat{x} \) by minimizing the variance of the prediction error:

\[
Var[(\mathbf{x} - \hat{\mathbf{x}})]
\]

- This is related to the **BLUP** theory for multivariate Normal:

\[
\begin{align*}
(\mathbf{y} | \mathbf{x}) &\sim \mathcal{N} \left[ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \mathbf{C}_y & \mathbf{C}_{x,y}^T \\ \mathbf{C}_{x,y} & \mathbf{C}_x \end{pmatrix} \right] \\
(\mathbf{x} | \mathbf{y}) &\sim \mathcal{N}(\mathbf{C}_{x,y} \mathbf{C}_y^{-1} \mathbf{y}, \mathbf{C}_x - \mathbf{C}_{x,y} \mathbf{C}_y^{-1} \mathbf{C}_{x,y}^T)
\end{align*}
\]
POINT-WISE SUMMARY

- Kriging is a second-order technique which is to be used in the context of the classical assumption of Gaussianity.
- **Kriging might not be appropriate** if we are interested in sample paths and rate of change of the underlying process (more on this later).
INDEPENDENT INCREMENTS

Unobserved continuous process: $x$
Irregularly observed data: $y$ at times $\{t_1 = 2, t_2 = 10, t_3 = 14\}$
Inference target: $x$ at times $\{0, 1, 2, ..., 13, 14\}$

1. **Deterministically** varying variances:

   $y(t_j) \sim \mathcal{N}[x(t_j), \sigma^2]$

   $[x(t_{j+1}) - x(t_j)] \sim \mathcal{N}[0, (t_{j+1} - t_j) \times \sigma^2]$

2. **Stochasticity** varying variances:

   $y(t_j) \sim \mathcal{N}[x(t_j), \sigma^2]$

   $[x(t_{j+1}) - x(t_j)] \sim \mathcal{N}[0, \sigma^2(t_{j+1} - t_j)]$

$v$ defines **volatility** (rate of change of process $x$)
INTERPOLATION / SCALE OF INTERPOLATION

![Graph showing interpolation data]

**Graph Description:***
- The top graph displays the interpolated values of a variable (x) over time.
- The bottom graph shows the volatility (Vol) of the same variable over time, with reconstructed values indicated.

**Legend:**
- **Blue line** represents the reconstructed x values.
- **Green line** indicates the reconstructed volatility.

The graphs illustrate how interpolation techniques are used to estimate values at intermediate points in time, providing a clearer understanding of the data's behavior over time.
Cumulated (integrated) variances between $t_j$ and $t_k$

1. Deterministically varying (cumulative) variance:
   \[(t_k - t_l)v^2 \text{ or } \int_{t_j}^{t_k} v^2 \delta u\]

2. Stochastically varying (cumulative) variance:
   \[\sum_{k,l} v^2(t) \text{ or } \int_{t_j}^{t_k} v^2(u)\delta u\]
MODEL CHOICE FOR THE STOCHASTIC VARIANCES

- **Independent increments**: variance of the sum is sum of the variances

- The **Inverse Gaussian Levy (IG)** process has a nice property: **infinitely divisible** (it could be represented as the sum of n iid random variables)

- Conditioned on the stochastic variances, \( x \) is Gaussian, i.e. \( x \) is marginally a **Normal Inverse Gaussian** Levy process which is also infinitely divisible and fits nicely into the framework of arbitrarily dense interpolation

- (G)ARCH & other SVM models: **not** infinitely divisible, mostly applicable to **regularly** observed data set
Stochastic Interpolation

The graph shows two sets of data points plotted over time. The top graph illustrates a trend in the variable $x(A)$, with error bars indicating variability. The bottom graph represents $Vol(A)$, also with error bars.
MULTIPLE VOLATILE TIME-SERIES GIVEN IRREGULAR OBS

![Chart 1: Time-series comparison](chart1.png)

![Chart 2: Detailed observation points](chart2.png)
MULTIVARIATE INDEPENDENT INCREMENT PROCESSES

\[
\begin{pmatrix}
y_A(t_j) \\
y_B(t_j)
\end{pmatrix}
\sim \mathcal{N} \left[ \begin{pmatrix} x_A(t_j) \\ x_B(t_j) \end{pmatrix}, \begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix} \right] \tag{1}
\]

\[
\begin{pmatrix}
x_A(t_{j+1}) - x_A(t_j) \\
x_B(t_{j+1}) - x_B(t_j)
\end{pmatrix}
\sim \mathcal{N} \left[ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \nu^2(t_{j+1} - t_j) \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right] \tag{2}
\]

\[
\nu^2(t_{j+1} - t_j) \sim \mathcal{IG}[\mu(t_{j+1} - t_j), \phi(t_{j+1} - t_j)] \tag{3}
\]

\[
(x_A(t_1) - x_B(t_1)) \sim \mathcal{N}(0, \tau^2) \tag{4}
\]

▶ \( \nu \) is the common latent volatility process

▶ \( \rho \) is the correlation co-efficient describing the relationship between the two time series A and B
JOINT RECONSTRUCTION
EVALUATION OF THE BIVARIATE MODEL

- Two time series: A (3 data points) and B (12 data points).

- To examine the influence of $\rho$:
  1. Perform independent inference of $x$ conditioned on the variance vector $v$ and other hyper-parameters
  2. Perform joint inference using varying values of $\rho$
  3. Compare the prediction uncertainty of 1 and 2
Evaluation of the Bivariate Model

Independent reconstruction

Joint reconstruction
Prediction uncertainty as a function of $\rho$

- $\rho = 0.01$
- $\rho = 0.6$
- $\rho = 0.85$ (the truth)
- $\rho = 0.99$
Ratio of prediction uncertainty as a function of $\rho$
Discussion

- Joint inference permits **crossed-borrowing of information** (correlation co-efficient, quantity and quality of observed data).

- Other parameters that might influence the strength of joint modelling:
  - **Quality** of data (measurement errors, locations of data points)
  - **Quantity** of data (number of data points)

- Other approaches/metrics for evaluation, frequentist hypothesis tests
Parameter Estimation

- Parameter space: \( x \) (e.g. latent climate process), \( v \) (latent volatility), and other hyper-parameters

- Off-line methods, e.g.
  - **MLE**: could be cumbersome, main reasons being the large parameter space
  - **INLA**: required Gaussian noise for the volatility components (or their transformation)

- Online update (e.g. **sequential MC**): difficult if (for many real world data set) the model is non-linear, non-Gaussian and has a large number of hyper-parameters

- Our current approach: using **Empirical Bayes** to estimate the hyper-parameter of the stochastic variances, and **MCMC** to infer the variance sample paths. **Further research: the EM algorithm?**
**HIGHER DIMENSIONAL CASE**

1. Parameter space increases with the number of time series, hence dimension of the matrices and the *big-n problem*. Further research: Finn Lindgren’s SPDE approach?

![Graph showing observations over time for multiple series A, B, C.](image)
ACKNOWLEDGEMENT: Dr. Andrew Parnell and Prof. John Haslett