IMPERIAL

Non-parametric modelling of graph structure of milti variate time series

Gregg Laska,
Supervisor: Dr Ed Cohen[1]

We consider a p-vector-valued stationary time series $\{\mathbf{X_t}\}$ where $\mathbf{X_t} = [X_{1,t}, \cdots, X_{p,t}]^T$, $t \in Z$, and ^{T} denotes transposition. Without loss of generality we assume $\{\mathbf{X_t}\}$ has mean zero.

Spectral Matrix

Each element $s_{X_j X_k,\tau}$ of the cross-covariance matrix $\Gamma_{\mathbf{X}}(\tau)$ represents the covariance between the j-th component $X_{j,t}$ of $\mathbf{X}_t$ and the k-th component $X_{k,t+\tau}$ of $\mathbf{X}_{t+\tau}$:

$$s_{X_j X_k,\tau} = \text{cov}\left[X_{j,t}, X_{k,t+\tau}\right] = E\left[X_{j,t}X_{k,t+\tau}\right]$$

Cross-spectrum analysis is a method used to study the frequency domain relationships between time series. The cross spectra is given by:

$$S_{X_j X_k}(f) = \sum_{\tau=-\infty}^{\infty} s_{X_j X_k,\tau} e^{-i2\pi f\tau},$$

where $|f| \leq 1/2$. Spectral matrix is thus given by $S(f)$ where $(S(f))_{j,k} = S_{X_j X_k}(f)$. It plays crucial role in following analysis.

Periodogram

The well known estimator of the spectral matrix is periodogram. However, it tends to suffer from high variance thus we introduce frequency averaged periodogram defined as:

$$\hat{\boldsymbol{S}}\left(f_j\right) = \sum_{k=-M}^{M} w_k \hat{\boldsymbol{S}}^{(P)}\left(f_{j-k}\right),$$

where $f_j = j/N$ is the j th Fourier frequency and $\{w_k\}$ is symmetric positive weight sequence for $k = -M, \ldots, M$, with $\sum w_k = 1$.

Correct Graph

Let $(\boldsymbol{S}^{-1}(f))_{j,k} = S^{jk}(f)$. If (V, E) is true graphical model for $\{\mathbf{X_t}\}$, then (V, E') is correct for (V, E), if $S^{jk}(f) = 0, |f| \leq 1/2$, Note that complete graph (having all edges) is correct for any graphical model.

Test for Missing Edges

Given $\{\mathbf{X_t}\}$ with graph (V, E), and spectral matrix $\boldsymbol{S}(f)$ we consider graph (V, E') and corresponding matrix $\boldsymbol{T}(f)$ satisfying:

$$\begin{aligned} T_{jk} = S_{jk}(f) \quad &\text{if} \quad \{(j,k) \in E'\} \\ T^{jk} = 0 \quad &\text{if} \quad \{(j,k) \notin E'\} \end{aligned}$$

Matsuda(2006) proves that if $\boldsymbol{T}(f)$ satisfies above constrain then (V, E') is correct for (V, E) if and only if $\boldsymbol{T}(f) = \boldsymbol{S}(f)$.

We use determine whether graph (V, E_2) is correct, given (V, E_1) is correct, where $E_2 \subseteq E_1$, Assuming that $\boldsymbol{T_1}(f) = \boldsymbol{S}(f)$, we calculate the estimators $\hat{\boldsymbol{T}}_1(f), \hat{\boldsymbol{T}}_2(f)$ using observed data and then intuitively if there is a big difference between $\hat{\boldsymbol{T}}_2(f) \neq \hat{\boldsymbol{T}}_1(f) \approx \boldsymbol{S}(f)$, it would imply (V, E_2) is incorrect.

Test statistic

Estimators $\hat{\boldsymbol{T}}_1(f)$ and $\hat{\boldsymbol{T}}_2(f)$ can be found using Wermuth-Scheidt equations defined in Matsuda[1], which iteratively transforms the initial estimate to archive $T_i^{jk} = 0$ if $\{(j,k) \notin E_i\}$.

To measure the difference between $\hat{\boldsymbol{T}}_1(f)$ and $\hat{\boldsymbol{T}}_2(f)$ Matsuda [21] used the estimated Kullback-Leibler divergence, $\widehat{KL}\left(\boldsymbol{T}_1, \boldsymbol{T}_2\right)$. Under some assumptions, Matsuda [21] defined the test statistic as:

$$Z\left(\boldsymbol{T}_1, \boldsymbol{T}_2\right) \stackrel{\text{def}}{=} \left[\frac{2MN}{D\left(m_2 - m_1\right)}\right]^{1/2} \times \left[\widehat{KL}\left(\boldsymbol{T}_1, \boldsymbol{T}_2\right) - \frac{C\left(m_2 - m_1\right)}{2M}\right]$$

where $m_i = \#\{(j,k) : (j,k) \notin E_i, j < k\}$, and C, D are constants which depend on the weights. Under H_0 ((V, E_2) is correct):

$$Z\left(\boldsymbol{T}_1, \boldsymbol{T}_2\right) \to \mathcal{N}(0,1) \text{ as } N \to \infty$$

Matsuda's Algorithm

The algorithm identifies the best graphical model for $\{\mathbf{X_t}\}$. We start with a complete graph (V, E_0). Then:

- **Step 1**: Generate candidate models $(V, E^1_{k+1}), (V, E^2_{k+1}), \ldots, (V, E^{L_k}_{k+1})$ by removing one edge from (V, E_k).
- **Step 2**: Calculate test statistics $Z_i = Z(\boldsymbol{T}_k, \boldsymbol{T}^i_{k+1})$ for each candidate model.
- **Step 3**: Determine the critical value:
$$C_k(\alpha) = \Phi^{-1}\left((1-\alpha)^{1/L_k}\right),$$
where L_k is the number of candidate models.
- **Step 4**: Compare all test statistics Z_i to $C_k(\alpha)$:
  - If $Z_i > C_k(\alpha)$ for all i, stop and select the current graph (V, E_k).
  - Otherwise, update the graph by removing the edge corresponding to the smallest test statistic and repeat the process.

Further Work

While above approach shows existence of relationships between time series, it does not infer directional influence among vertices. To address this, I will work with the Hawkes processes, which allow for inferring the direction of influence, identifying both influencers and those influenced within the network.