Graphical models of time series: parameter estimation and topology selection

Jitkomut Songsiri

Electrical Engineering Department
University of California, Los Angeles

Sep 3, 2008

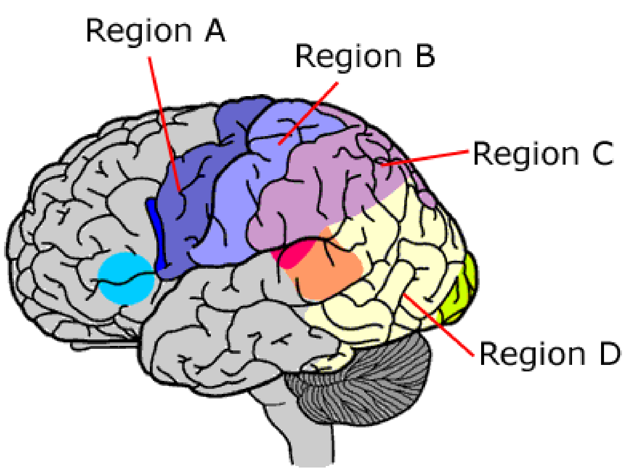
Outline

- Introduction
- Graphical Models and Conditional Independence
- Convex Formulation of ML Estimation of Autoregressive Models
- Examples
- Conclusions and Future Plans

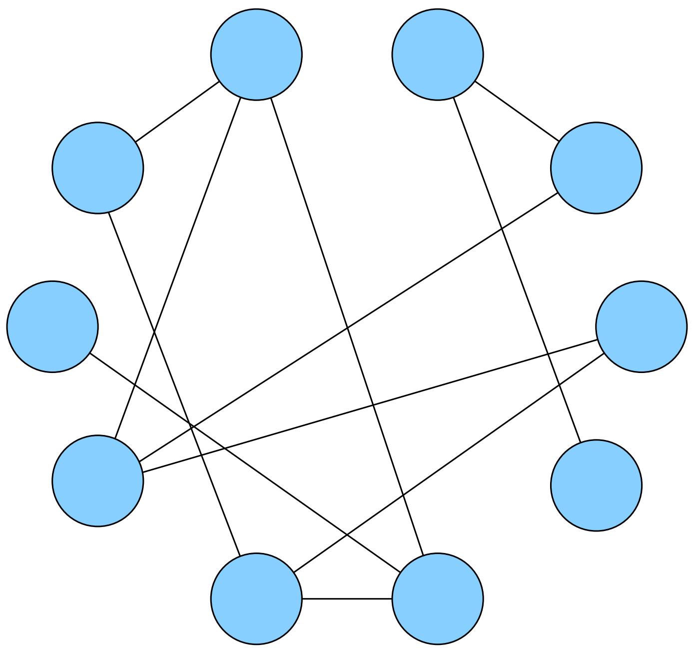
Graphical Models

represent dependency or causality structure between random variables

- economics (interexchange rates, stock prices, etc.)
- brain networks (functional connectivity between brain regions)
- haemodynamic systems (heart rate, blood pressure, etc.)
- ...



Conditional Independence Graph



- Nodes correspond to random variables X_i
- Link (i, j) is absent if X_i and X_j are **conditionally independent**

- visual representation of the relation between many variables
- by exploiting the graph structure, many large-scale problems can be solved with less complexity

Conditional Independence for Gaussian Time series

Gaussian random variable $X \sim \mathcal{N}(0, \Sigma)$

X_i and X_j are **conditionally independent** if $(\Sigma^{-1})_{ij} = 0$

(Demster (1972))

Gaussian time series $X(t) = (X_1(t), X_2(t), \ldots, X_n(t))\ ,\ t \in \mathbb{Z}$

X_i and X_j are **conditionally independent** if $(S(\omega)^{-1})_{ij} = 0\ ,\ \forall \omega$

$S(\omega)$ is the spectral density matrix of $x(t)$

(Brillinger (1996))

ML Estimation of Autoregressive Models
with
Conditional Independence Constraints

Multivariate Autoregressive Model

$$y(t) = -A_1 y(t-1) - A_2 y(t-2) - \cdots - A_p y(t-p) + w(t)$$

- $w(t) \sim \mathcal{N}(0, \Sigma)$
- $A_k \in \mathbf{R}^{n \times n}$
- Equivalent form :

$$B_0 y(t) = -B_1 y(t-1) - B_2 y(t-2) - \cdots - B_p y(t-p) + v(t)$$

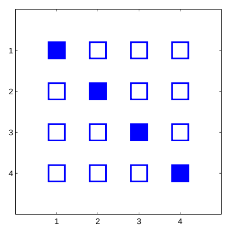
- $v(t) \sim \mathcal{N}(0, I)$
- $B_0 = \Sigma^{-1/2}$
- $B_k = \Sigma^{-1/2} A_k \ , \ k = 1, \dots, p$

Characterization of Conditional Independence

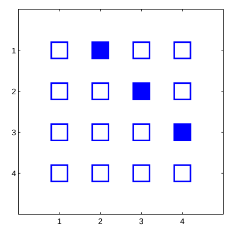
$$S(z)^{-1} = Y_0 + \sum_{k=1}^{p} \left(z^{-k} Y_k + z^k Y_k^T \right)$$

$$Y_k = \sum_{i=0}^{p-k} B_i^T B_{i+k} \, , \; k = 0, 1, \ldots, p$$

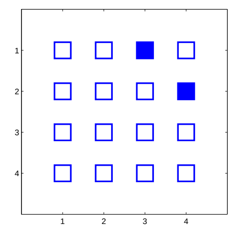
Y_k is the sum of k^{th}-off-diagonal blocks of $B^T B$, $B = \begin{bmatrix} B_0 & B_1 & \ldots & B_p \end{bmatrix}$



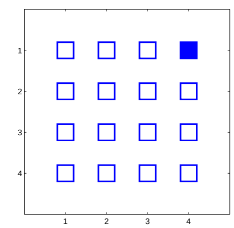
Y_0



Y_1



$\ldots$



Y_p

$$\boxed{[S(\omega)^{-1}]_{ij} = 0 \quad \Longleftrightarrow \quad [Y_k]_{ij} = [Y_k]_{ji} = 0 \, , \; k = 0, \ldots, p}$$

Conditional Maximum-likelihood Estimation

- $N + p$ measurements, $y_1, y_2, \ldots, y_{N+p}$
- Condition on the initial p states
- Log-likelihood function:

$$\log L(B) = N \log \det B_0 - \frac{N}{2} \mathbf{tr} \left(R B^T B \right),$$

where

$$R = \frac{H H^T}{N}, \qquad H = \begin{bmatrix} y_{p+1} & y_{p+2} & \cdots & y_{N+p} \\ y_p & y_{p+1} & \cdots & y_{N+p-1} \\ \vdots & \vdots & & \vdots \\ y_1 & y_2 & \cdots & y_N \end{bmatrix}$$

Summary

$$
\begin{array}{ll}
\text{minimize} & -\log\det B_0 + \frac{1}{2}\,\mathbf{tr}(RB^TB) \\
\text{subject to} & Y_k = \displaystyle\sum_{i=0}^{p-k} B_i^T B_{i+k}\ ,\ k = 0, 1, \ldots, p \\
& [Y_k]_{ij} = [Y_k]_{ji} = 0,\ k = 0, \ldots, p\ ,\ (i,j) \in \mathcal{V}.
\end{array}
$$

variables

- $B = (B_0, B_1, \ldots, B_p) \in \mathbf{S}^n \oplus \mathbf{R}^{n \times np}$
- $Y_0 \in \mathbf{S}^n\ ,\ Y_k \in \mathbf{R}^{n \times n}\ ,\ k = 1, \ldots, p$

Nonconvex because of quadratic equality constraints

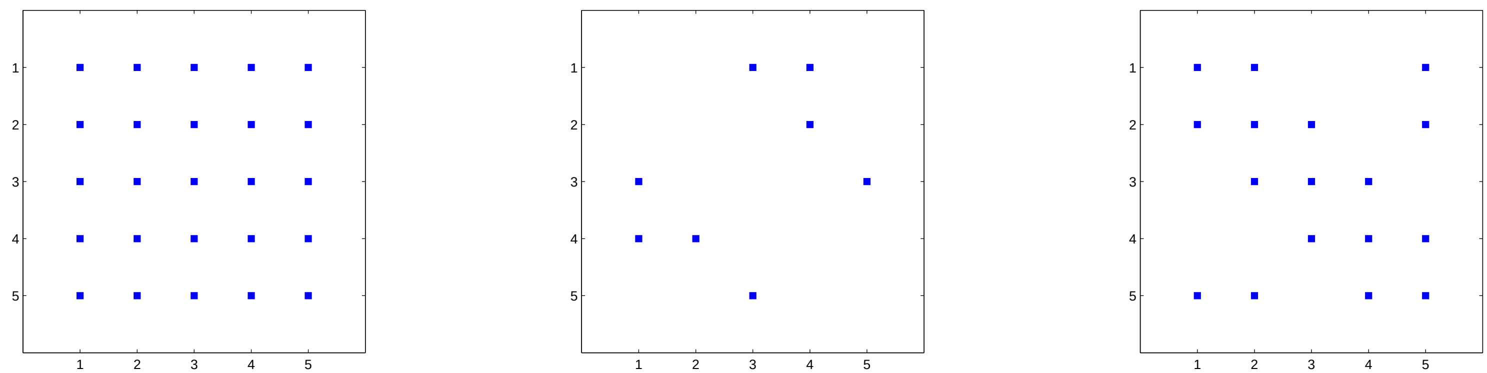
Convex Formulation

Notation

$P : \mathbf{S}^n \to \mathbf{S}^n_{\mathcal{V}}$ is a projection of X on $\mathcal{V}$

$$P(X)_{ij} = \begin{cases} X_{ij} & (i,j) \in \mathcal{V} \\ 0 & \text{otherwise.} \end{cases}$$

example $\mathcal{V} = \{(1,3),(1,4),(2,4),(3,5)\}$



(a) X (b) $P(X)$ (c) X s.t. $P(X) = 0$

Convex Formulation

Conditional ML Estimation

$$\begin{array}{ll} \text{minimize} & -\log\det B_0 + \frac{1}{2}\,\mathbf{tr}(RB^TB) \\ \text{subject to} & P(\displaystyle\sum_{i=0}^{p-k} B_i^T B_{i+k}) = 0\ ,\ \ k = 0,1,\ldots p \end{array} \qquad \text{(P1)}$$

variable $B = (B_0, B_1, \ldots, B_p) \in \mathbf{S}^n \oplus \mathbf{R}^{n\times np}$

Equivalent Form

$$\begin{array}{ll} \text{minimize} & -\log\det X_{00} + \mathbf{tr}(RX) \\ \text{subject to} & P\left(\displaystyle\sum_{i=0}^{p-k} X_{i,i+k}\right) = 0, \quad k = 0,1,\ldots,p \\ & X \succeq 0\ ,\ \mathbf{rank}(X) = n \end{array} \qquad \text{(P2)}$$

variable $X \in \mathbf{S}^{n(p+1)}$ with $X = B^TB$

Relaxation

$$\begin{array}{ll} \text{minimize} & -\log\det X_{00} + \mathbf{tr}(RX) \\ \text{subject to} & P\left(\sum_{i=0}^{p-k} X_{i,i+k}\right) = 0, \quad k = 0, 1, \ldots, p \\ & X \succeq 0 \end{array} \qquad \text{(P3)}$$

variable $X \in \mathbf{S}^{n(p+1)}$

- The optimal value of (P3) is less than or equal to the optimal value of (P2), since we minimize on a larger set
- If X^* has rank n, then by factorizing $X^* = B^T B$, B must be optimal in (P1)
- The relaxation is exact if X^* always has rank n

Exactness of Relaxation

- The low-rank property of X^* can be proved for block-Toeplitz and positive definite R

$$R = \begin{bmatrix} R_0 & R_1 & \cdots & R_p \\ R_1^T & R_0 & \cdots & R_{p-1} \\ \vdots & \vdots & \ddots & \vdots \\ R_p^T & R_{p-1}^T & \cdots & R_0 \end{bmatrix}$$

- For almost-Toeplitz

$$R = \frac{HH^T}{N},$$

X^* has low rank in the experiments. R is close to a block-Toeplitz matrix when $N \to \infty$

Dual Problem

$$\boxed{\begin{array}{ll} \text{maximize} & \log\det W + n \\ \text{subject to} & \begin{bmatrix} W & 0 \\ 0 & 0 \end{bmatrix} \preceq R + P(Z) \end{array}} \qquad \text{(D3)}$$

variables $W \in \mathbf{S}^n$ and $Z = \begin{bmatrix} Z_0 & Z_1 & \cdots & Z_p \\ Z_1^T & Z_0 & \cdots & Z_{p-1} \\ \vdots & \vdots & \ddots & \vdots \\ Z_p^T & Z_{p-1}^T & \cdots & Z_0 \end{bmatrix}$, $Z_k \in \mathbf{R}^{n\times n}$

$P(Z)$ is the blockwise projection of Z

- $X = I$ is strictly feasible $\Rightarrow$ Slater's condition holds $\Rightarrow$ Strong duality holds and the dual optimum is attained if the optimal value is finite
- $Z = 0$ is strictly feasible $\Rightarrow$ the primal optimum is attained

Karush-Kuhn-Tucker (KKT) Conditions

1. *Primal feasibility.*

$$X \succeq 0, \qquad X_{00} \succ 0, \qquad P(\sum_{i=0}^{p-k} X_{i,i+k}) = 0, \quad k = 0, \ldots, p.$$

2. *Dual feasibility.*

$$W \succ 0, \qquad R + P(Z) \succeq \begin{bmatrix} W & 0 \\ 0 & 0 \end{bmatrix}.$$

3. *Zero duality gap.*

$$X_{00}^{-1} = W, \qquad \mathbf{tr}\left(X\left(R + P(Z) - \begin{bmatrix} W & 0 \\ 0 & 0 \end{bmatrix}\right)\right) = 0.$$

Low-rank Property of X^*

Let R be a symmetric block-Toeplitz matrix.

$$R \succeq \begin{bmatrix} I_n & 0 \\ 0 & 0 \end{bmatrix} \quad \Longrightarrow \quad R \succ 0$$

The low-rank property of X^* follows from

$$R + P(Z^*) \succeq \begin{bmatrix} W^* & 0 \\ 0 & 0 \end{bmatrix} \quad \Longrightarrow \quad R + P(Z^*) \succ 0$$

and

$$X^* \left(R + P(Z^*) - \left[\begin{array}{cc} W^* & 0 \\ 0 & 0 \end{array} \right] \right) = 0$$

Primal and Dual Problems

Conditional ML estimation

$$\begin{array}{ll} \min & -\log\det B_0 + \frac{1}{2}\,\mathbf{tr}(RB^TB) \\ \text{s.t.} & P(\sum_{i=0}^{p-k} B_i^T B_{i+k}) = 0\ ,\ \forall k \end{array}$$

$X = B^TB$

Relaxation

$$\begin{array}{ll} \min & -\log\det X_{00} + \mathbf{tr}(RX) \\ \text{s.t.} & P(\sum_{i=0}^{p-k} X_{i,i+k}) = 0\ ,\ \forall k \\ & X \succeq 0 \end{array}$$

Dual

Dual Problem

$$\begin{array}{ll} \max & \log\det W + n \\ \text{s.t.} & \begin{bmatrix} W & 0 \\ 0 & 0 \end{bmatrix} \preceq R + P(Z) \end{array}$$

static case $(p = 0)$

static case $(p = 0)$

Covariance Selection Problem

$$\begin{array}{ll} \min & -\log\det X + \mathbf{tr}(RX) \\ \text{s.t.} & P(X) = 0 \end{array}$$

Dual

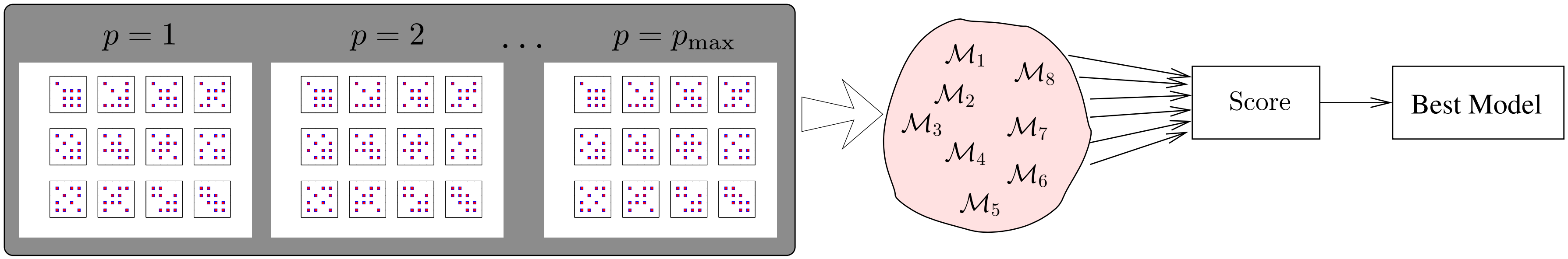
Maximum Determinant Completion

$$\max \quad \log\det(R + P(Z)) + n$$

Examples

- Air pollution data
- Stock return data
- fMRI data

Model Selection Problem



- L : maximized log-likelihood
- N : sample size
- k : number of effective parameters

$$\begin{aligned} \mathrm{AIC} &= 2k - 2L \\ \mathrm{AIC}_c &= 2k\left(\frac{N}{N-k-1}\right) - 2L \\ \mathrm{BIC} &= k \log N - 2L \end{aligned}$$

An autoregressive model of order p has $p+1$ parameters, $B_0, \ldots, B_p$

$$k = \frac{n(n+1)}{2} - |\mathcal{V}| + p(n^2 - 2|\mathcal{V}|)$$

Terminology

Coherence spectrum

$$\bar{S}(\omega) = U(\omega) S(\omega) U^H(\omega) \quad , \quad U(\omega) = \begin{bmatrix} S_{11}^{-1/2}(\omega) & & \\ & \ddots & \\ & & S_{nn}^{-1/2}(\omega) \end{bmatrix}$$

i.e., normalized spectral density matrix

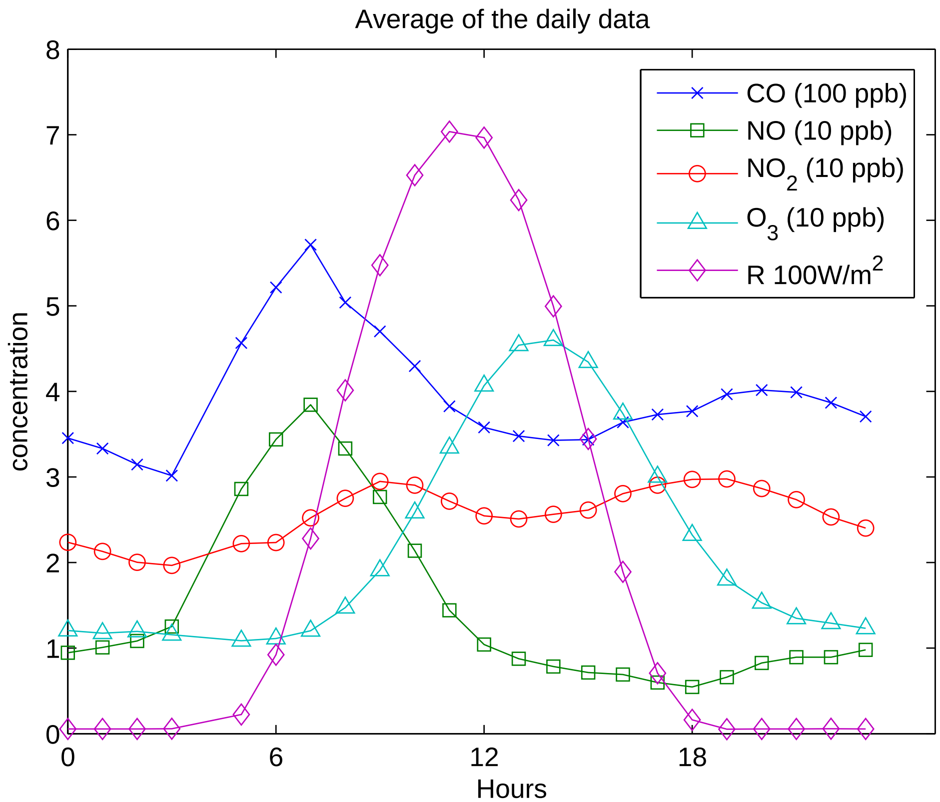
Partial coherence spectrum (with $G(\omega) = S(\omega)^{-1}$)

$$\bar{G}(\omega) = V(\omega) G(\omega) V^H(\omega) \quad , \quad V(\omega) = \begin{bmatrix} G_{11}^{-1/2}(\omega) & & \\ & \ddots & \\ & & G_{nn}^{-1/2}(\omega) \end{bmatrix}$$

i.e., normalized inverse of spectral density matrix

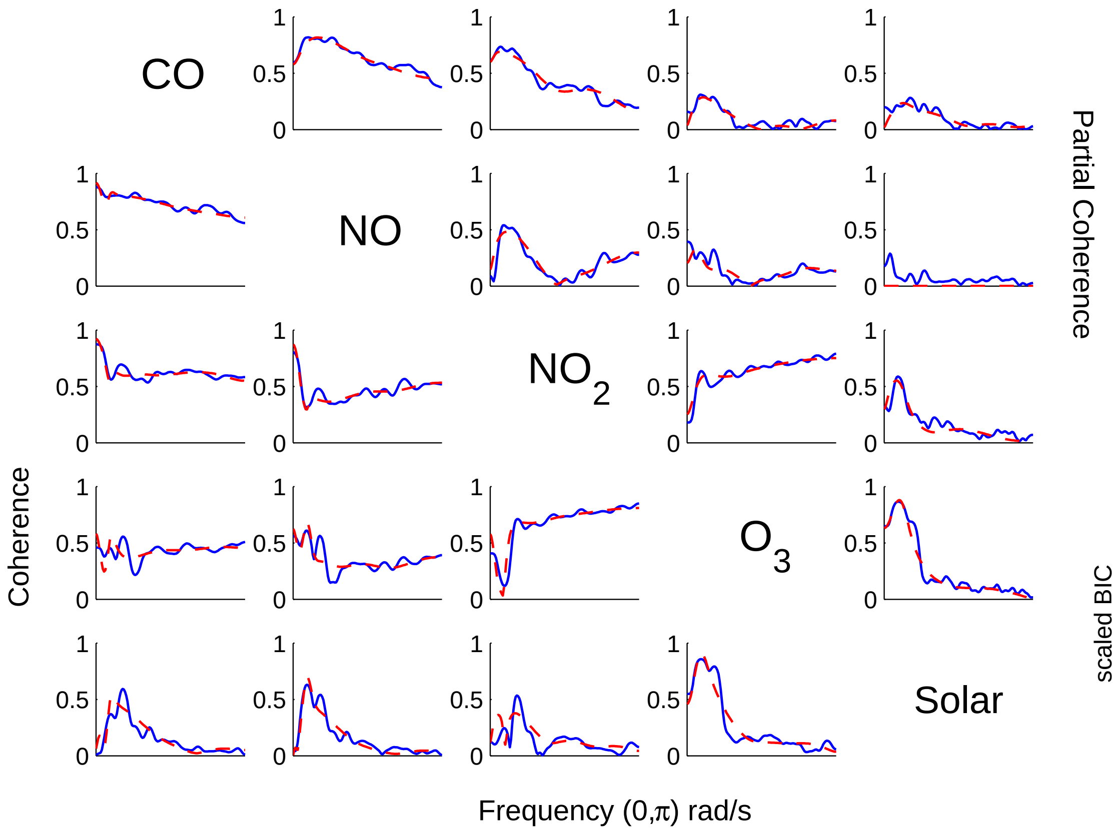
Example I : Air Pollution Data

- CO, NO, NO_2, O_3, and solar radiation intensity
- recorded from Jan 1 to Dec 31, 2006 from Azusa, Los Angeles

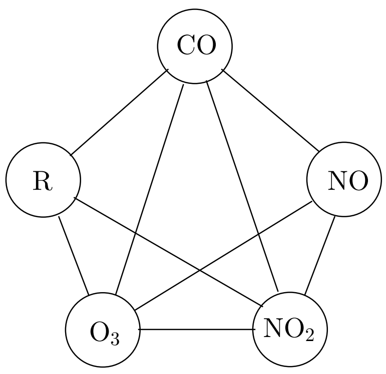


Average of daily data

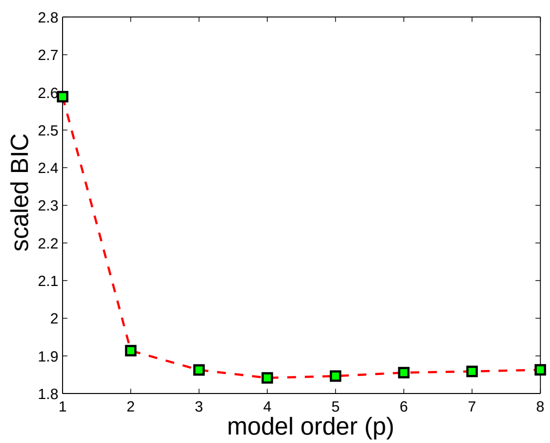
Example I : Air Pollution Data



spectrum estimates



the best topology



BIC scores $(p^* = 4)$

Example II : Stock Return Data

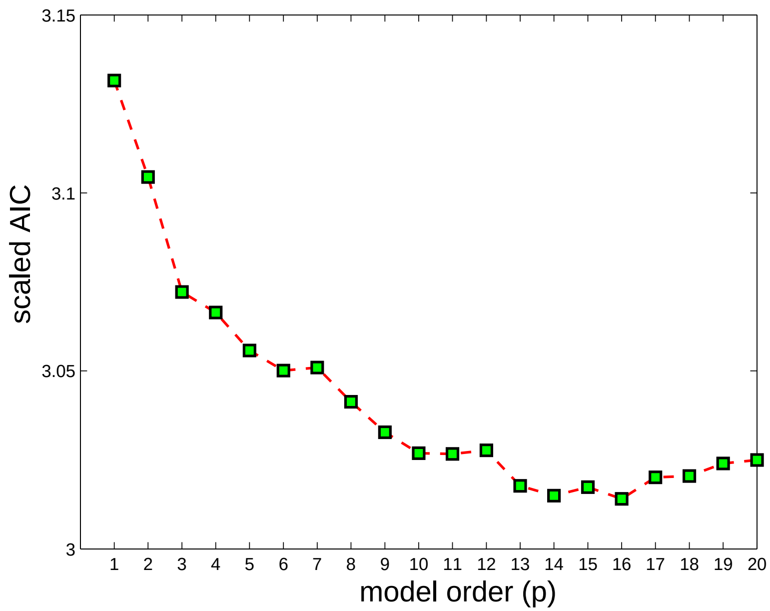
Stock closing prices of 5 markets in Europe:

- FTSE 100 share index (United Kingdom)
- CAC 40 (France)
- Frankfurt DAX 30 composite index (Germany)
- MIBTEL (Italy)
- Austrian Traded index ATX (Austria)

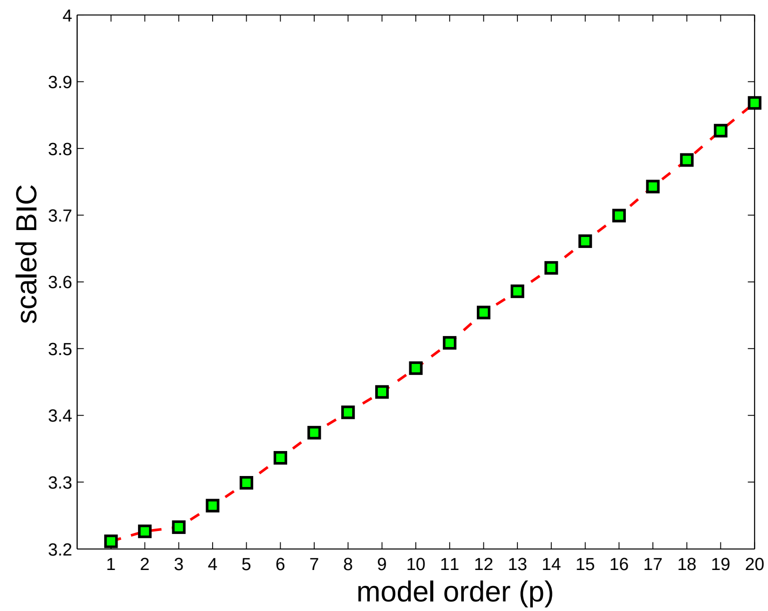
recorded from Jan 1, 1999- Jul 31, 2008

Markets	EMU	Non-EMU
LARGE	FR,GE,IT	UK
SMALL	AU	

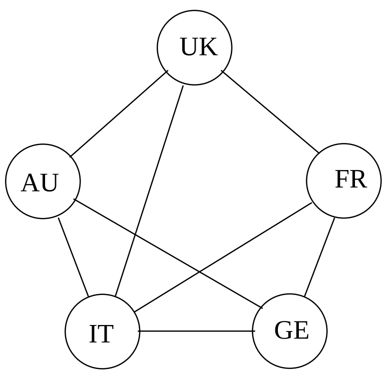
Example II : Stock Return Data



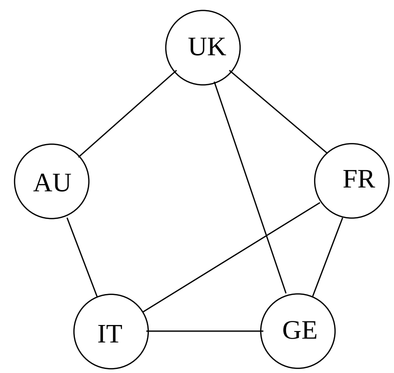
(a) AIC scores ($p^* = 16$)



(b) BIC scores ($p^* = 1$)

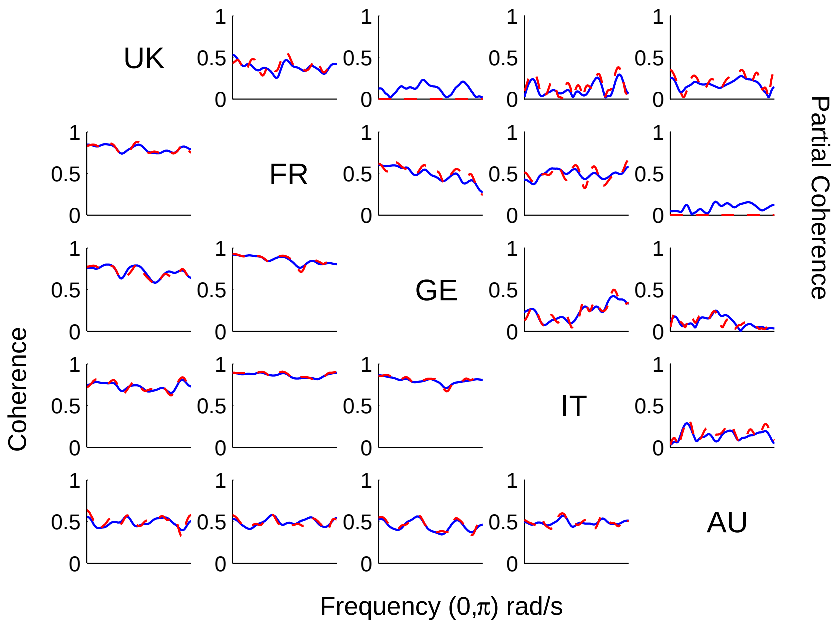


(c) AIC

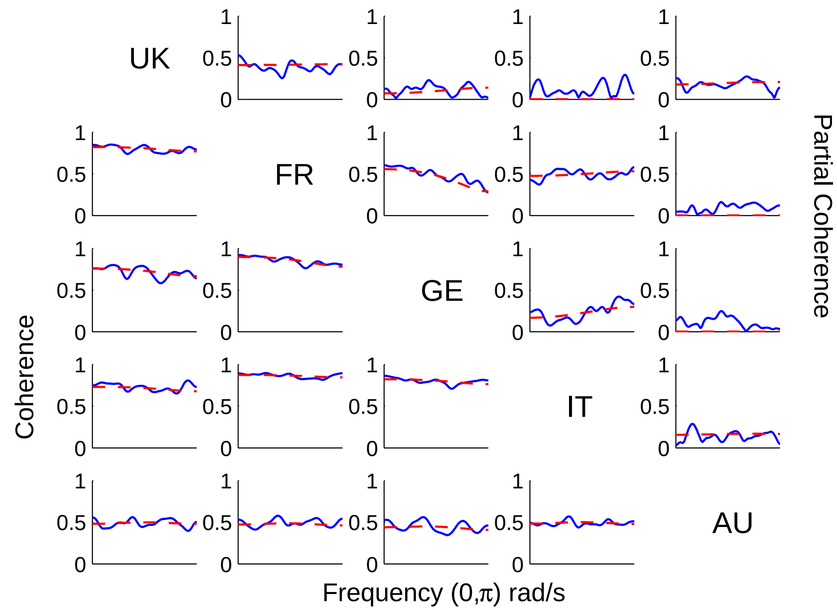


(d) BIC

Example II: Stock Return Data



(e) spectrum estimates (AIC, $p = 14$)



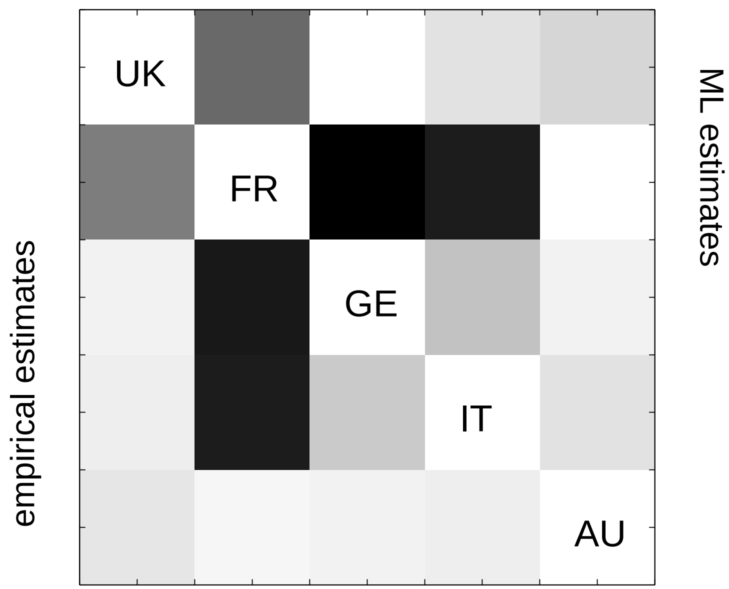
(f) spectrum estimates (BIC, $p = 1$)

- The large markets are highly correlated
- UK has a strong connection via France only
- The small market, AU is likely to be isolated from the others

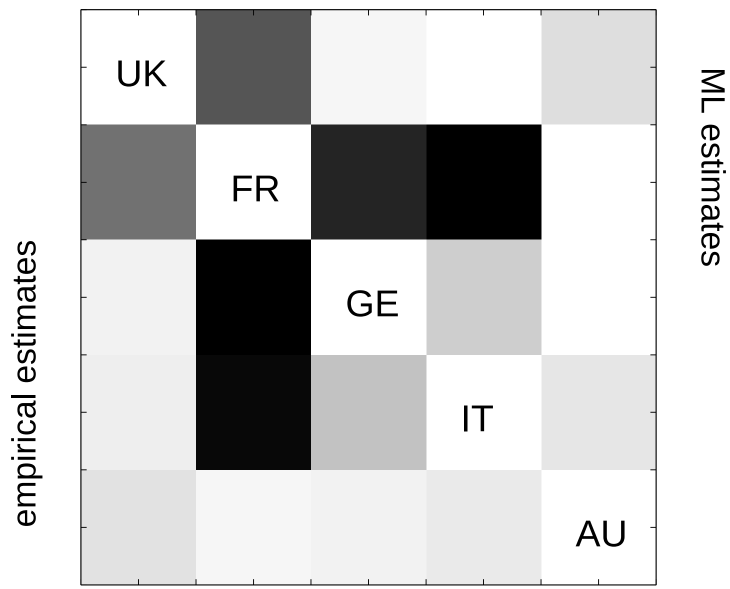
Example II : Stock Return Data

Partial mutual information

$$I = -\frac{1}{2\pi}\int_0^{2\pi} \log\left(1 - |\bar{G}(\omega)|^2\right) d\omega.$$

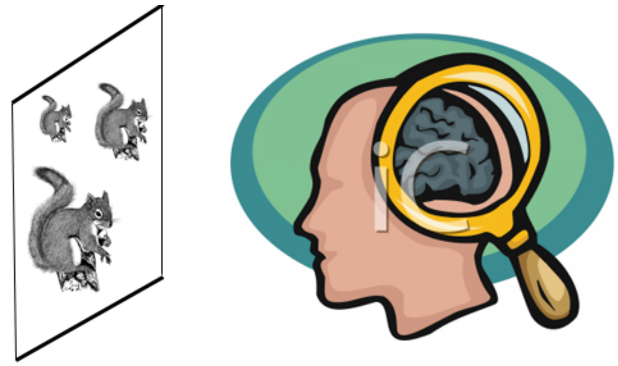


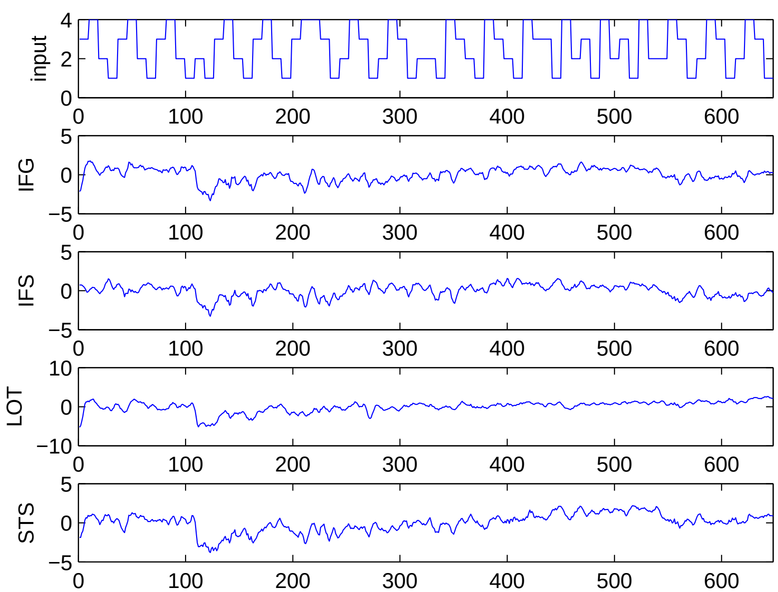
(g) AIC, $p = 14$



(h) BIC, $p = 1$

Example III: fMRI Data

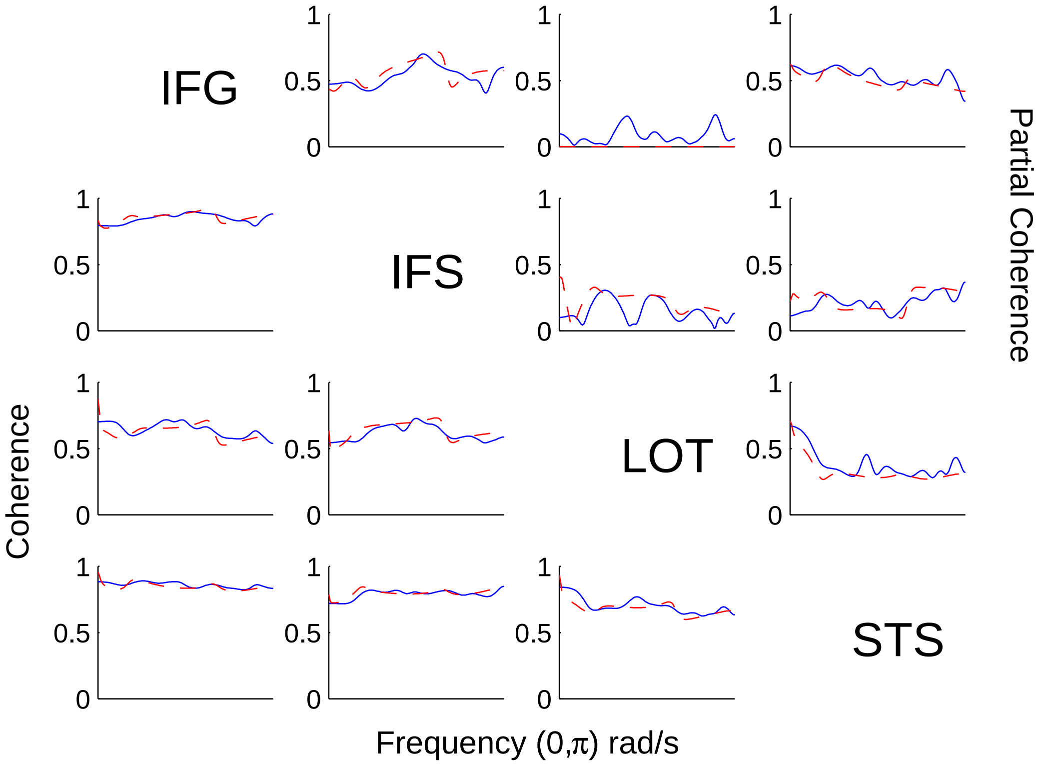




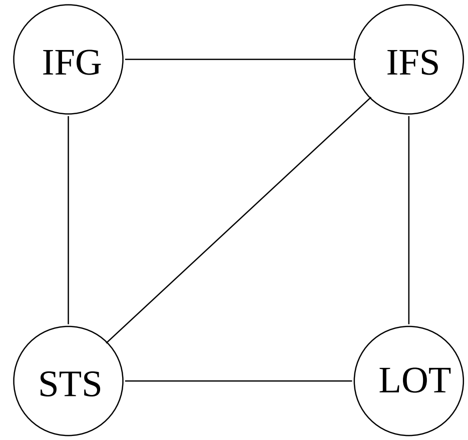
Average of fMRI time series over all voxels

- Four subregions (IFG, IFS, LOT, STS) are activated by 4 visual stimuli
- The stimuli involve images of pictures and words
- Average the data over all voxels in each region

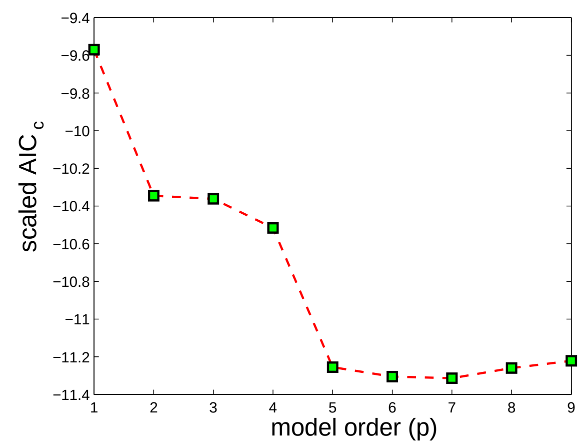
Example III: fMRI Data



spectrum estimates



the best topology



AIC$_c$ scores $(p^* = 7)$

Summaries and Future Plans

Summaries

- We consider conditional independence of multivariate Gaussian time series and its graphical representation

- Maximum-likelihood estimation of AR models with conditional independence constraints leads to a nonconvex problem

- A convex formulation provides exact solutions to ML problem by showing that the optimal solution has low rank

- Graphical inference problems can be solved by fitting AR models according to all possible sparsity constraints

- The best topogology is selected by applying some model selection criterion such as AIC, BIC

- The method is applied to air pollution data, stock index returns, and fMRI data

Future Plans

Model and topology selection

- The goal is to recover the sparsity pattern in Y_k automatically
- The location of zeros in all matrices Y_k must be the same

$$
\begin{array}{ll}
\text{maximize} & \log\det X_{00} - \mathbf{tr}(RX) + \gamma\|W\|_1 \\
\text{subject to} & \displaystyle Y_k = \sum_{i=0}^{p-k} X_{i,i+k}\ ,\ k = 0, 1, \ldots, p \\
& -W_{ij} \leq [Y_k]_{ij} \leq W_{ij}, \quad \forall i \neq j, k = 0, 1, \ldots p \\
& X \succeq 0, \quad W_{ij} \geq 0, \quad \forall i \neq j.
\end{array}
$$

- γ is the regularization parameter
- W is the maximum modulus of all matrices Y_k except diagonal elements

Future Plans

Extension of the proof to non-Toeplitz R

- The matrix R in the ML problem is close to a block-Toeplitz matrix if the sample size (N) is relatively large
- Relax the assumption in the proof to almost-Toeplitz R

Granger causality

- defined in terms of predictibility. The cause should improve the predictions of the effect
- correspond to sparse AR coefficients and sparse covariance matrix of the input noise
- has a convex formulation for solving maximum-likelihood estimation of AR models with Granger causality constraints
- has wide applications in economic time series and neural systems (Eicheler (2005), Valdes-Sosa et.al (2005), Fujita et.al (2007), etc.)

Future Plans

fMRI application

- requires refinements of AR model
  - categorical inputs
  - switching
  - dependence on subjects
- Vast literatures on functional connectivity

  (Friston (1994), Cohen (1997), Boynton (1996), Josephs (1997), Rajapakse (1998), Friston (2005))