Graphical models of time series: parameter estimation and topology selection

by

Jitkomut Songsiri

PROSPECTUS SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR PH.D. CANDIDACY IN ELECTRICAL ENGINEERING

Summer 2008

Contents

List of Figures

List of Tables

List of Symbols

Set notations

$\mathbf{R}$ ($\mathbf{C}$)	set of real (complex) numbers.
$\mathbb{Z}$	set of positive integers.
$\mathbf{R}_+$	set of nonnegative real numbers.
$\mathbf{R}^n$ ($\mathbf{C}^n$)	set of real (complex) n-tuple vectors.
$\mathbf{R}^{m\times n}$ ($\mathbf{C}^{m\times n}$)	set of real (complex) $m \times n$ matrices.
$\mathbf{S}^n$ ($\mathbf{H}^n$)	set of symmetric (hermitian) matrices of size n.
$\mathbf{S}^n_+$ ($\mathbf{H}^n_+$)	set of positive semidefinite matrices of size n.
$\mathbf{S}^n_{++}$ ($\mathbf{H}^n_{++}$)	set of positive definite matrices of size n.
$A \setminus B$	set difference; $A \setminus B = \{x \mid x \in A \text{ and } x \notin B\}$.

Numbers, vectors, matrices, operators

$A \succeq B$	$A - B$ is positive semidefinite.
$A \succ B$	$A - B$ is positive definite.
A^T	transpose of matrix A.
A^H	complex-conjugate (Hermitian) transpose of matrix A.
$\langle \cdot, \cdot \rangle$	an inner product of an inner product space.
$\| \cdot \|_2$	a Euclidean norm for a vector, or a spectral norm of a matrix.
$\mathbf{tr}(X)$	a trace of a square matrix X.

Chapter 1

Introduction

1.1 Overview

Let U, V, W be random variables with a joint density function f. U and V are said to be *conditionally independent* given W if

$$f_{UV|W}(u, v|w) = f_{U|W}(u|w) f_{V|W}(v|w).$$

In a graph representation of a multivariate random variable X, the nodes represent the components X_i and two nodes are connected by an undirected edge if the corresponding variables are conditionally independent given the other variables. There is a nice characterization of conditional independence for Gaussian random variables. Let X be an n-dimensional Gaussian random variable with covariance matrix Σ. We say X_i and X_j are conditionally independent given all other components if and only if $(\Sigma^{-1})_{ij} = 0$. The associated graph is called a Gaussian graphical model of the random variable. Graphical models are attractive for many reasons. They can visually represent the structure of the relationships among the variables. By exploiting the graph structure, they can also facilitate computations in large-scale problems of inferencing and estimation.

The notion of conditional independence can be extended to time series. Let $\{x(t), t \in \mathbb{Z}\}$ be a multivariate stationary Gaussian process with spectral density matrix $S(\omega)$. The components x_i and x_j are conditionally independent given the remaining variables if and only if $(S(\omega)^{-1})_{ij} = 0$ for all ω (see details in Chapter 2). This condition allows us to consider system identification problems with conditional

independence constraints by placing restrictions on the inverse of spectral density matrix. This is the main topic of the prospectus. We consider maximum-likelihood estimation of autoregressive models with conditional independence constraints. As we shall see, this can be formulated as a convex optimization problem and readily solved by efficient algorithms. With our method, one can learn the dependence structure of a time series by choosing the topology that best characterizes the data by minimizing some model selection criterion. The graphical models of time series that we describe in this prospectus have several applications. We present three examples of real data sets from various scientific fields to illustrate our method.

1.2 Related work and contributions

The concept of conditional independence between time series was first discussed in [Bri75], where the so-called *partial coherence function* was defined as a measure of linear dependence between two components in time series after removing linear effects from other variables. A graphical representation of dependencies in a stationary time series was investigated in [Bri96], [Dah00]. The latter showed that the conditional independence graphs can be represented by zeros in the inverse of spectral density matrix and derived a statistic for hypothesis tests examining whether an edge is present in the graph. The method was illustrated by the air pollution data to study interactions among polluted particles. The same approach was also applied to identification of functional neural connectivity in [DES97] and [EDS03]. This nonparametric approach based on a test in frequency domain has become a useful tool for many applications later on. For example, [TLH$^+$00] investigated the connection between the cortical activity and tremor in patients suffering from Parkinson's disease. [SSSB05] explored the correlated activities in human brain networks based on functional magnetic resonance imaging (fMRI) data. [GIF02] applied the technique to the haemodynamic system consisting of vital signs such as heart rate, or blood pressure, etc., which are crucial for detection of critical situation of patients in an intensive care unit. It can be also applied to the analysis of factors in therapy process from psychosomatic studies [FMM$^+$05].

An advantage of using the nonparametric approach in the above examples, is the ease of implementation. One can compute an empirical estimate of the inverse

spectrum and apply a criterion to detect its zeros elements. Alternatively, a parametric approach can be used to learn graphical models where the model parameters are constrained to each of possible graph structures. Therefore, the identification of conditional independence structures has become a model selection problem in which the best model minimizes a model selection criterion.

A natural parametric model is an autoregressive model. The most relevant study of this type is the work of [Eic06]. He considered an approximate maximum-likelihood estimation of autoregressive models with conditional independence constraints and used an iterative algorithm to solve the problem. The covariance functions equipped with conditional independence properties were estimated. Consequently, the model parameters were obtained from the Yule-Walker equations. Another related work was [VSSBLC+05]. Since the zero constraints on the inverse of spectral matrix can be translated to the restrictions on AR parameters, they considered a sparse AR model estimation by applying the regularization technique to recover the sparsity automatically. However, their problem formulation appears to correspond to Granger-causality graphs (see Chapter 5) rather than conditional independence graphs. Moreover, a first-order model was assumed, whereas the difficulties always arise from the models of higher order. It was also mentioned in [DE03] that numerical solutions to this problem have been under exploration.

Our main motivation is to propose a convex framework for maximum-likelihood estimation of autoregressive models with conditional independence constraints. More precisely, the zeros in the inverse of spectral matrix are equivalent to quadratic equality constraints on the AR parameters, which are generally nonconvex. We prove that, under some assumptions, a convex relaxation provides *exact* solutions for this problem, yielding polynomial-time algorithms. The results of this work can serve two purposes. Given a conditional independence graph of a time series, one can estimate the spectrum according to the graph structure. Furthermore, if the conditional independence is not specified a priori, the structure can be identified from the model selection problem.

1.3 Outline of the prospectus

The prospectus is organized as follows. Chapter 2 describes the overview of graphical models and the concept of conditional independence. The approximate maximum-likelihood estimation of a multivariate Gaussian autoregressive model is discussed in chapter 3. It explains the difficulty arising from considering the conditional independence constraints in the estimation. Section 3.2 in chapter 3 contains the main contribution of this work; it proposes a convex formulation which is a relaxation method to cope with the nonconvex constraints. An important result presented in this chapter is to prove that the relaxation provides exact optimal solutions. Chapter 4 illustrates the proposed method by demonstrating the examples of air pollution data, stock index returns, and fMRI data. We also present supplementary results on maximum-likelihood estimation of autoregressive models with Granger causality constraints in Chapter 5. The last chapter concludes the prospectus and discusses topics for further studies.

Chapter 2

Graphical models of time series

In recent years, graphical models have become a useful tool for many statistical applications. A probabilistic graphical model consists of a collection of probability distribution which can be factorized according to the graph structure. The graph $G = (V, E)$ contains a set of vertices, V identifying random variables and a collection of edges, E. Two nodes are connected by an edge if the corresponding variables are conditionally independent. This model combines probabilistic concept with a graph theory by representing dependencies among multivariate random variables in the graph. By exploiting the graph representation, basic statistical quantities such as marginal or conditional probabilities, or the likelihoods, can be computed with less complexity. This computational advantage can be found in numerous applications such as combinatoric optimization, bioinformatics, speech processing, or image processing. For an introduction to graphical models, one may refer to [Edw00, Whi90]. An extensive treatment can be found in [Lau96].

As mentioned above, graphical models have been initially developed for static multivariate random variables. A review of basic ideas and algorithms for probabilistic inference was discussed in [Jor04]. A remarkable result is the work of [Bri96] and [Dah00] who extended the concept of conditional independence to time series. This provides a method for identifying associations between entire time series. [BJ04] proposed an algorithm to learn graphical structures from the spectral representation of time series. [FD03] extended the work from [Dah00] to study the dependence structure of subprocesses when some of components in time series are not available. [DE03] discussed several ideas of causality and graphical models and defined the

global Markov properties. One type of such ideas is the integration between Granger causality and graphical models which was first introduced in Chapter 14 of [SWT06] and [Eic07]. The concept of Granger causality, which has been used extensively in econometrics, is based on the idea that one time series is said to be Granger causal for another series, if the prediction of the latter series can be improved by using information from the former series. In addition to applications in economics, this concept has become a promising technique for several applications in neuroscience (see Chapter 14,17,18 in [SWT06], [Eic05], and [FSGM$^+$07]). We refer to Chapter 5 for more details.

In this chapter, we mainly focus on the concept of conditional independence graphs for multivariate time series. We first discuss the definition of conditional independence for random variables and the associated graph. Then the idea will be generalized to the time series case.

2.1 Conditional independence of random variables

Definition 2.1. *Let X, Y, Z be random variables with a joint density function. We say that X and Y are conditionally independent given Z and write $X \perp\!\!\!\perp Y|Z$ if and only if*

$$f_{XY|Z}(x, y|z) = f_{X|Z}(x|z) f_{Y|Z}(y|z). \tag{2.1}$$

The condition (2.1) must hold for all z such that $f_Z(z) > 0$.

For Gaussian random variables, there is a simple characterization of this property.

Multivariate normal distribution

Let X be an n-dimension Gaussian random variable with mean μ and covariance Σ. The density function is given by

$$f(x) = \frac{1}{(2\pi)^{n/2} \det^{1/2}(\Sigma)} \exp\left\{ -\frac{1}{2} \left\langle \Sigma^{-1}(x-\mu), (x-\mu) \right\rangle \right\}.$$

Assume that the random vector X is partitioned into component Y and Z with the corresponding mean and variance;

$$X = \begin{bmatrix} Y \\ Z \end{bmatrix}, \quad \mu = \begin{bmatrix} \mu_y \\ \mu_z \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \Sigma_{yy} & \Sigma_{yz} \\ \Sigma_{zy} & \Sigma_{zz} \end{bmatrix}.$$

Proposition 2.2. *The conditional distribution of Y given $Z = z$ is also normal with mean*

$$\mu_{y|z} = \mu_y - \Sigma_{yz}\Sigma_{zz}^{-1}(z - \mu_z), \tag{2.2}$$

and covariance

$$\Sigma_{y|z} = \Sigma_{yy} - \Sigma_{yz}\Sigma_{zz}^{-1}\Sigma_{zy}. \tag{2.3}$$

The conditional mean is the linear least squares estimate of Y given $Z = z$ and is simply a linear transformation of z. More interestingly, the conditional covariance is constant for all values of z. Next, we will present an important result: the conditional independence between two variables produces zeros in the inverse of covariance matrix.

Corollary 2.3. *Let $X \sim \mathcal{N}(0, \Sigma)$ be an n-dimensional random variable and $V = \{1, 2, \ldots, n\}$. X_i and X_j are conditionally independent given the other variables, $X_{V\setminus\{i,j\}}$, if and only if*

$$X_i \perp\!\!\!\perp X_j | X_{V\setminus\{i,j\}} \iff (\Sigma^{-1})_{ij} = 0. \tag{2.4}$$

Proof. Without loss of generality, assume $i = 1$ and $j = 2$. From (2.3), the conditional covariance of (X_i, X_j) given the rest, is

$$\Sigma_{i,j|V\setminus\{i,j\}} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}^H \quad \triangleq \begin{bmatrix} \sigma_{ii} & \sigma_{ij} \\ \sigma_{ij} & \sigma_{jj} \end{bmatrix}_{2\times 2}.$$

Apply the Schur complement of Σ to obtain the concentration matrix

$$\Sigma^{-1} = \begin{bmatrix} (\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}^H)^{-1} & * \\ * & * \end{bmatrix},$$

where we neglect the $*$ terms as they are not relevant to the calculation. This shows that the $(1, 1)$ block of Σ^{-1} has size 2×2 and is infact the inverse of conditional covariance matrix. Thus, $(\Sigma^{-1})_{12} = 0 \Rightarrow \sigma_{ij} = 0$ and it implies that X_i and X_j are conditionally independent given the remaining variables. □.

Conditional independence graphs

A conditional independence graph is simply an undirected graph where the presence of an edge is encoded by the conditional independent constraint.

The conditional independence graph associated with a multivariate random variable X is the undirected graph $G(V, E)$ consisting of a set of vertices $V = \{1, 2, \ldots, n\}$ and a set of edges E such that

$$A—B \notin E \quad \Longleftrightarrow \quad X_A \perp\!\!\!\perp X_B | X_C,$$

for all $A \neq B$ and $C = V \setminus \{A, B\}$.

Corollary 2.3 shows that for Gaussian random variables, missing links in the graph can be read from zeros in the inverse of the covariance matrix. A problem of computing the estimates of the mean μ and covariance Σ of a multivariate Gaussian variable subject to conditional independence constraints is known as *covariance selection problems* [Dem72]. Numerous proposed algorithms for solving this problem can be found in [DVR08, BEG08, FHT07, YL07].

In the next section, the idea of conditional independence graph is generalized to a time series case.

2.2 Conditional independence of time series

Let $x(t) = \begin{bmatrix} x_1(t) & x_2(t) & \ldots & x_n(t) \end{bmatrix}^T, t \in \mathbb{Z}$, be a multivariate stationary Gaussian time series and (A, B, C) be the component indices of $x(t)$ where $A \neq B$ and $C = \{1, 2, \ldots, n\} \setminus \{A, B\}$. It is known that for a Gaussian process, zero correlation is equivalent to independence. Therefore, the two components x_A and x_B are *conditional independent* given x_C, denoted by $x_A \perp\!\!\!\perp x_B \mid x_C$, if and only if the partial correlation is zero for all time lags;

$$x_A \perp\!\!\!\perp x_B \mid x_C \quad \Longleftrightarrow \quad \mathbf{cov}\{\varepsilon_{A|C}(t), \varepsilon_{B|C}(t+k)\} = 0 \ , \forall k \in \mathbb{Z}, \tag{2.5}$$

where

$$\varepsilon_{A|C}(t) = x_A(t) - \mathbf{E}\left(x_A(t) \mid x_C(s) \ , \ s \in \mathbb{Z}\right),$$
$$\varepsilon_{B|C}(t) = x_B(t) - \mathbf{E}\left(x_A(t) \mid x_C(s) \ , \ s \in \mathbb{Z}\right),$$

are residual processes of x_A and x_B, respectively after removing the linear effects of x_C. This is the analogy of the static case. The conditional mean in (2.2), which is a linear function of the given information, must be removed from the two interested variables.

To characterize an explicit condition in (2.5), two optimal filters $\{\mathbf{d}_A(k), k \in \mathbb{Z}\}, \{\mathbf{d}_B(k), k \in \mathbb{Z}\}$ and two optimal means μ_A, μ_B which minimize

$$\mathbf{E}\left[\left(x_A(t) - \mu_A - \sum_{k=-\infty}^{\infty} \mathbf{d}_A(t-k)x_C(k)\right)\left(x_A(t) - \mu_A - \sum_{k=-\infty}^{\infty} \mathbf{d}_A(t-k)x_C(k)\right)^H\right]$$
$$\mathbf{E}\left[\left(x_B(t) - \mu_B - \sum_{k=-\infty}^{\infty} \mathbf{d}_B(t-k)x_C(k)\right)\left(x_B(t) - \mu_B - \sum_{k=-\infty}^{\infty} \mathbf{d}_B(t-k)x_C(k)\right)^H\right]$$

must be determined and they are obtained by the following results from [Bri75].

Consider a multivariate stationary time series partitioned as

$$x(t) = \begin{bmatrix} x_1(t) \\ x_2(t) \\ \vdots \\ x_n(t) \end{bmatrix} \triangleq \begin{bmatrix} y(t) \\ z(t) \end{bmatrix} \tag{2.6}$$

with mean $\mathbf{E}[x(t)] = \mu_x$. Denote the covariance function

$$C(k) = \mathbf{E}[(x(t+k) - \mu_x)(x(t) - \mu_x)^H] = \begin{bmatrix} C_{YY}(k) & C_{YZ}(k) \\ C_{ZY}(k) & C_{ZZ}(k) \end{bmatrix}, \quad k = 0, \pm 1, \ldots,$$

and the spectral density

$$S(\omega) = \begin{bmatrix} S_{YY} & S_{YZ} \\ S_{ZY} & S_{ZZ} \end{bmatrix} = \sum_{k=-\infty}^{\infty} \begin{bmatrix} C_{YY} & C_{YZ} \\ C_{ZY} & C_{ZZ} \end{bmatrix} e^{i\omega k}. \tag{2.7}$$

Theorem 2.4. *Consider a multivariate time series defined above. Assume that $C_{YY}(k), C_{YZ}(k), C_{ZZ}(k)$ are absolutely summable and $S_{ZZ}(\omega)$ is nonsingular for all ω. Then the optimal μ and $\{\mathbf{a}(k), k \in \mathbb{Z}\}$ minimizing*

$$\mathbf{E}\left[\left(y(t) - \mu - \sum_{k=-\infty}^{\infty} \mathbf{a}(t-k)z(k)\right)\left(y(t) - \mu - \sum_{k=-\infty}^{\infty} \mathbf{a}(t-k)z(k)\right)^H\right]$$

are given by

$$\mu = \mu_y - \left(\sum_{k=-\infty}^{\infty} \mathbf{a}(k)\right)\mu_z = \mu_y - \mathbf{A}(0)\mu_z, \tag{2.8}$$

and $\mathbf{a}(k) = \frac{1}{2\pi}\int_0^{2\pi} \mathbf{A}(\omega)e^{ik\omega}d\omega$ where,

$$\mathbf{A}(\omega) = S_{YZ}(\omega)S_{ZZ}^{-1}(\omega). \tag{2.9}$$

Moreover, the cross spectrum of the residual error

$$\varepsilon(t) = y(t) - \mu - \sum_{k=-\infty}^{\infty} \mathbf{a}(t-k)z(k),$$

is given by

$$S_{\varepsilon\varepsilon}(\omega) = S_{YY}(\omega) - S_{YZ}(\omega)S_{ZZ}^{-1}(\omega)S_{ZY}(\omega). \tag{2.10}$$

Proof. The optimal mean and optimal filter in (2.8), (2.9) were proved in [Bri75]. We shall show (2.10) only. Since the residual error has zero mean, the covariance function is

$$\begin{aligned}
C_{\varepsilon\varepsilon}(\tau) &= \mathbf{E}[\varepsilon(t+\tau)\varepsilon^H(t)] \\
&= \mathbf{E}\left[(y(t+\tau)-\mu_y)(y(t)-\mu_y)^H\right] + \mathbf{E}\left[(y(t+\tau)-\mu_y)\left(\sum_{k=-\infty}^{\infty}\mathbf{a}(k)(\mu_z - z(t-k))\right)^H\right] \\
&+ \mathbf{E}\left[\left(\sum_{k=-\infty}^{\infty}\mathbf{a}(k)(\mu_z - z(t+\tau-k))\right)(y(t)-\mu_y)^H\right] \\
&+ \mathbf{E}\left[\left(\sum_{k=-\infty}^{\infty}\mathbf{a}(k)(\mu_z - z(t+\tau-k))\right)\left(\sum_{j=-\infty}^{\infty}\mathbf{a}(k)(\mu_x - \mathbf{z}(t-j))\right)^H\right].
\end{aligned}$$

Interchange the role of expectation and summation.

$$\begin{aligned}
C_{\varepsilon\varepsilon}(\tau) = C_{YY}(\tau) &- \sum_{k=-\infty}^{\infty} C_{YZ}(\tau-k)\mathbf{a}^H(k) - \sum_{k=-\infty}^{\infty}\mathbf{a}(k)C_{ZY}(\tau-k) \\
&+ \sum_{j=-\infty}^{\infty}\sum_{k=-\infty}^{\infty}\mathbf{a}(k)C_{ZZ}(\tau+j-k)\mathbf{a}^H(j).
\end{aligned} \tag{2.11}$$

By applying the discrete Fourier transform and substitute (2.9) to (2.11), we will obtain (2.10). □

Now we are ready to apply the above results and discuss about conditional independence. Note that for Gaussian time series, the residual error $\varepsilon_{A|C}, \varepsilon_{B|C}$ in (2.5) are then Gaussian. Therefore, the zero partial correlation is equivalent to conditional independence. The conditional independence property can be also easily characterized in the frequency domain. From (2.6), let $y(t) = (x_A(t), x_B(t))$ and $z(t) = x_C(t)$.

By using the result from (2.10), the *partial cross spectrum* of x_A and x_B given x_C, which is the cross spectrum between $\varepsilon_{A|C}(t)$ and $\varepsilon_{A|C}(t)$ is given by

$$S_{AB|C}(\omega) = S_{AB}(\omega) - S_{AC}(\omega)S_{CC}^{-1}(\omega)S_{CB}(\omega), \tag{2.12}$$

where S_{AB}, S_{CB}, S_{CC} are the cross spectra between the corresponding variables defined as in (2.7). The cross spectrum (2.10) is simply a Schur complement of $S(\omega)$ which is analogous to the conditional covariance matrix (2.3) in the static case. Moreover, the normalized version of $S_{AB|C}(\omega)$ called the *partial spectral coherence* of x_A and x_B given x_C is provided by

$$R_{AB|C}(\omega) = \frac{S_{AB|C}(\omega)}{\sqrt{S_{AA|C}(\omega)S_{BB|C}(\omega)}}. \tag{2.13}$$

These give us the orthogonality conditions of x_A and x_B which can be summarized as follows.

$$\boxed{\begin{aligned} x_A \perp\!\!\!\perp x_B|x_C \iff & \mathbf{cov}\{\varepsilon_{A|C}(t), \varepsilon_{B|C}(t+k)\} = 0 \ , \forall k \in \mathbb{Z} \\ \iff & S_{AB|C}(\omega) = 0 \ , \ \forall \omega \in [-\pi, \pi] \end{aligned}} \tag{2.14}$$

Next we will present an important result which is again an analogue of Corollary 2.3 for time series case.

Corollary 2.5. *Let $x(t)$ be a multivariate Gaussian time series with spectral density $S(\omega)$. Suppose $S(\omega)$ is nonsingular for all ω. Then the components x_A and x_B are conditionally independent given the other variables, denoted by x_C if and only if*

$$x_A \perp\!\!\!\perp x_B|x_C \iff (S^{-1}(\omega))_{AB} = 0 \ , \ \forall \omega \in [-\pi, \pi]. \tag{2.15}$$

The proof follows similarly to corollary 2.3 which is almost a direct consequence from (2.12). Let $x(t)$ be partitioned as in (2.6) and without loss of generality, we define, $y(t) = (x_A(t), x_B(t))$ and $z(t) = x_C(t)$. The spectral density matrix of $x(t)$ is

$$S(\omega) = \begin{bmatrix} S_{YY} & S_{YZ} \\ S_{ZY} & S_{ZZ} \end{bmatrix}.$$

The $(1,1)$ block of $S(\omega)^{-1}$ is obtained by taking the Schur complement on the $(1,1)$ block.

$$S(\omega)^{-1} = \begin{bmatrix} (S_{YY} - S_{YZ}S_{ZZ}^{-1}S_{ZY})^{-1} & * \\ * & * \end{bmatrix}.$$

The other blocks of S^{-1} can be neglected since we only need to show that $(S^{-1}(\omega))_{12} = 0$. This is obtained by the fact that the $(1,1)$ block of $S(\omega)^{-1}$ is diagonal if and only if the cross spectrum in (2.10) is diagonal, or equivalently, the partial cross spectrum of x_A and x_B given x_C is identically zero for all ω. □

In conclusion, for a graphical model of Gaussian random variables with covariance matrix Σ, x_i and x_j are conditionally independent if and only if $(\Sigma^{-1})_{ij} = 0$. Corollary 2.5 gives us an extended result to the time series case by replacing the covariance matrix with the spectral density matrix.

An interesting connection is proved by [Dah00]. If $S(\omega)^{-1}$ exists for all $\omega \in [-\pi, \pi]$ and let $G(\omega) = S(\omega)^{-1}$, then

$$R_{AB|C}(\omega) = -\frac{G_{AB}(\omega)}{\sqrt{G_{AA}(\omega)G_{BB}(\omega)}}. \tag{2.16}$$

This result can be verified easily from

$$\begin{aligned}\frac{G_{AB}(\omega)}{\sqrt{G_{AA}(\omega)G_{BB}(\omega)}} &= -\frac{S_{AB} - S_{AC}S_{CC}^{-1}S_{CB}}{\sqrt{S_{BB} - S_{BC}S_{CC}^{-1}S_{CB}}\sqrt{S_{AA} - S_{AC}S_{CC}^{-1}S_{CA}}} \\ &= -\frac{S_{AB|C}}{\sqrt{S_{BB|C} \cdot S_{AA|C}}} \quad = -R_{AB|C}(\omega).\end{aligned}$$

Conditional Independence Graph

We have described the definition of conditional independence and some specific results for Gaussian time series. This leads to the definition of a conditional independence graph as follows.

The conditional independence graph associated with a multivariate stationary process $\{x(t), t \in \mathbb{Z}\}$ is a graph $G = (V, E)$ which consists of a vertex set V and edge set E such that

$$A—B \notin E \quad \Longleftrightarrow \quad x_A \perp\!\!\!\perp x_B | x_C$$

for all $A \neq B$ and $C = V \setminus \{A, B\}$.

For Gaussian time series, the existence of an edge is determined by zeros in the inverse of spectral matrix. In order to learn the graph structure, one must estimate the spectrum subject to each of all possible conditional independence constraints and select the best topology by applying some model selection criterion. Next chapter will present a parametric estimation problem which combines this constraint with the formulation.

Chapter 3

Model estimation

This chapter considers conditional maximum-likelihood estimation of autoregressive models with conditional independence constraints. We show that the constraints can be characterized by quadratic equalities on the model parameters which results in a nonconvex problem. The main result is to propose a convex formulation with a change of variable and prove that our method provides optimal solutions to the original problem.

3.1 Maximum likelihood estimation

Consider a multivariate autoregressive model of order p

$$y_k = -A_1 y_{k-1} - A_2 y_{k-2} - \cdots - A_p y_{k-p} + w_k, \tag{3.1}$$

where w_k is a Gaussian white noise with covariance matrix Σ, $y_k \in \mathbf{R}^n$ and $A_k \in \mathbf{R}^{n\times n}$. Premultiplying $\Sigma^{-1/2}$ on both sides of (3.1) so as to normalize the covariance matrix of the input noise gives

$$B_0 y_k = -B_1 y_{k-1} - B_2 y_{k-2} - \cdots - B_p y_{k-p} + v_k, \tag{3.2}$$

where $v_k \sim \mathcal{N}(0, I)$ and $B_0 = \Sigma^{-1/2}$. A_k and B_k are related by $B_k = \Sigma^{-1/2} A_k$.

We are interested in maximum-likelihood estimation based on $N+p$ observations, $y_1, \ldots, y_{N+p}$. The exact likelihood function is highly nonlinear in A_k. Therefore, a simple approach is to condition on the first p observations, $y_1, \ldots, y_p$ and estimate based on the last N observations. To this end, we will write (3.1) in a vector form

and define $A = \begin{bmatrix} A_1 & A_2 & \dots & A_p \end{bmatrix}$ and the observation matrix

$$\begin{aligned} Y_{N+p} &\triangleq \left[\begin{array}{cccc|cccc} y_1 & y_2 & \dots & y_p & y_{p+1} & y_{p+2} & \dots & y_{N+p} \end{array}\right] \\ &= \left[\begin{array}{c|c} H_0 & H_1 \end{array}\right]. \end{aligned}$$

From (3.1),

$$\begin{aligned} H_1 &= -A \begin{bmatrix} y_p & y_{p+1} & \dots & y_{N+p-1} \\ y_{p-1} & y_p & \dots & y_{N+p-2} \\ \vdots & \vdots & \dots & \vdots \\ y_1 & y_2 & \dots & y_N \end{bmatrix} + \begin{bmatrix} w_{p+1} & w_{p+2} & \dots & w_{N+p} \end{bmatrix} \\ &= -AH_2 + W. \end{aligned}$$

The conditional density function of the last N observations given the first p initial states, is

$$f(H_1|H_0) = \frac{1}{(2\pi)^{nN/2}\det(\Sigma)^{N/2}} \exp\left\{-\frac{1}{2}\,\mathbf{tr}\left(\Sigma^{-1}(H_1 + AH_2)(H_1 + AH_2)^T\right)\right\}.$$

The log-likelihood function corresponding to the conditional density function is, up to a constant,

$$\log L(A, \Sigma) = -\frac{N}{2}\log\det\Sigma - \frac{1}{2}\,\mathbf{tr}\left[\Sigma^{-1}(H_1 + AH_2)(H_1 + AH_2)^T\right]. \tag{3.3}$$

By making change of variables, we can also derive the log-likelihood function in terms of B_k. Let $B = \begin{bmatrix} B_0 & B_1 & \dots & B_p \end{bmatrix}$ and note that $B_k = \Sigma^{-1/2}A_k$. It can be verified that (3.3) is equivalent to

$$\log L(B) = N\log\det B_0 - \frac{1}{2}\,\mathbf{tr}\left(BHH^TB^T\right), \tag{3.4}$$

where

$$H = \begin{bmatrix} H_1 \\ H_2 \end{bmatrix} = \begin{bmatrix} y_{p+1} & y_{p+2} & \dots & y_N \\ y_p & y_{p+1} & \dots & y_{N-1} \\ \vdots & \vdots & & \vdots \\ y_1 & y_2 & \dots & y_{N-p} \end{bmatrix}.$$

If we define

$$R = \frac{HH^T}{N}, \tag{3.5}$$

then the log-likelihood function in (3.4) becomes

$$\log L(B) = N \log\det B_0 - \frac{N}{2}\,\mathbf{tr}\left(RB^TB\right). \tag{3.6}$$

Without conditional independence constraints, we can solve the unconstrained problem either from (3.3) or (3.6). It is known that the optimal solution to the unconstrained ML problem yields the least-square solution.

We characterize the conditional independence constraints in terms of AR parameters. Define polynomial matrix functions

$$\begin{aligned}\mathbf{A}(z) &= I + z^{-1}A_1 + \cdots + z^{-p}A_p,\\ \mathbf{B}(z) &= B_0 + z^{-1}B_1 + \cdots + z^{-p}B_p.\end{aligned}$$

The inverse z-spectrum of the output in (3.1) is

$$\begin{aligned}S(z)^{-1} &= \mathbf{A}(1/\bar{z})^H\Sigma^{-1}\mathbf{A}(z)\\ &= \mathbf{B}(1/\bar{z})^H\mathbf{B}(z)\\ &= Y_0 + \sum_{k=1}^{p}\left(z^{-k}Y_k + z^kY_k^T\right),\end{aligned} \tag{3.7}$$

where,

$$\begin{aligned}Y_k &= \sum_{i=0}^{p-k} A_i^T\Sigma^{-1}A_{i+k}\\ &= \sum_{i=0}^{p-k} B_i^TB_{i+k},\end{aligned} \tag{3.8}$$

for $k = 0, \ldots, p$.

The conditional independence constraints in (2.15) can be expressed as

$$(S(\omega)^{-1})_{ij} = 0, \forall\omega \in [-\pi,\pi] \quad\Longleftrightarrow\quad [Y_k]_{ij} = [Y_k]_{ji} = 0,\ \forall k = 0,\ldots,p, \tag{3.9}$$

for all $(i,j) \in \mathcal{V}$ where $\mathcal{V}$ is the index set of the sparsity pattern.

The expression of log-likelihood function in (3.6) will be chosen since it is clearly concave in B, whereas (3.3) is not obviously so, but it can be shown that it is concave jointly in (Σ, A) in a certain region (see Chapter 5). Furthermore, as we shall see in the next section, the constraints in (3.9) with the choice of Y_k in (3.8) can be cast as convex constraints.

The conditional ML estimation problem with the conditional independence constraints can therefore be expressed as

$$\boxed{\begin{array}{ll} \text{minimize} & -\log\det B_0 + \frac{1}{2}\,\mathbf{tr}(RB^TB) \\ \text{subject to} & Y_k = \displaystyle\sum_{i=0}^{p-k} B_i^T B_{i+k}\ ,\ k = 0,1,\ldots,p \\ & [Y_k]_{ij} = [Y_k]_{ji} = 0,\ \forall k = 0,\ldots,p\ \ \forall (i,j) \in \mathcal{V}. \end{array}} \tag{3.10}$$

The variables are $B_0 \in \mathbf{S}^n_{++}$ and $B_k \in \mathbf{R}^{n\times n}$, $k = 1,\ldots,p$. This problem is nonconvex due to the quadratic equalities from the sparsity constraints.

3.2 Convex formulation

This section presents the main contribution of the prospectus. The goal of this work is to provide a technique for solving (3.10) efficiently. We propose a convex relaxation problem where a change of variable is introduced. The main result is to prove that the optimal solution in the relaxed problem has low rank property, i.e., admits no gap with the true optimal value. Our method thus returns the optimal solution to (3.10). To this end, we will introduce some notations as follows.

The sparsity pattern of a sparse matrix $X \in \mathbf{S}^n$ will be characterized by specifying the set of indices $\mathcal{V} \subseteq \{1,\ldots,n\} \times \{1,\ldots,n\}$ of its zero entries. We assume $\mathcal{V}$ is symmetric, *i.e.*, if $(i,j) \in \mathcal{V}$ then $(j,i) \in \mathcal{V}$, and that it does not contain any diagonal entries, *i.e.*, $(i,i) \notin \mathcal{V}$ for $i = 1,\ldots,n$.

$P_{\mathcal{V}}(X)$ denotes the projection of a square symmetric or non-symmetric matrix X on $\mathcal{V}$:

$$P_{\mathcal{V}}(X)_{ij} = \begin{cases} X_{ij} & (i,j) \in \mathcal{V} \\ 0 & \text{otherwise.} \end{cases} \tag{3.11}$$

For ease of notation, we will drop the subscript $\mathcal{V}$ from $P_{\mathcal{V}}(X)$ and use only $P(X)$ throughout the text. We use the same notation for P as a mapping from $\mathbf{R}^{n\times n} \to \mathbf{R}^{n\times n}$ and as a mapping from $\mathbf{S}^n \to \mathbf{S}^n$. In both cases, P is self-adjoint. If X is a $p \times p$ block-matrix with i,j block X_{ij}, then we define $P(X)$ as $p \times p$ block matrix with i,j block $P(X)_{ij} = P(X_{ij})$.

With the above notations, the problem (3.10) is equivalent to

$$\boxed{\begin{array}{ll}\text{minimize} & -\log\det B_0 + \frac{1}{2}\,\mathbf{tr}(RB^TB) \\ \text{subject to} & P(\sum_{i=0}^{p-k} B_i^T B_{i+k}) = 0\ ,\ k = 0, 1, \ldots p\end{array}} \tag{3.12}$$

with variable $B = \begin{bmatrix} B_0 & B_1 & \ldots & B_p \end{bmatrix} \in \mathbf{S}^n_{++} \oplus \mathbf{R}^{n\times np}$.

The quadratic terms of B_k suggest a change of variable $X = B^TB$. We therefore propose a convex relaxed problem

$$\boxed{\begin{array}{ll}\text{minimize} & -\log\det X_{00} + \mathbf{tr}(RX) \\ \text{subject to} & P\left(\sum_{i=0}^{p-k} X_{i,i+k}\right) = 0, \quad k = 0, \ldots, p \\ & X \succeq 0\end{array}} \tag{3.13}$$

with variable

$$X = \begin{bmatrix} X_{00} & X_{01} & \cdots & X_{0p} \\ X_{01}^T & X_{11} & \cdots & X_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{0p}^T & X_{1p}^T & \cdots & X_{pp} \end{bmatrix} \in \mathbf{S}^{n(p+1)}.$$

A first observation from problem (3.13) is that its optimal value is less than or equal to the optimal value of (3.12), since we have dropped the rank constraint and minimize the same objective function over a larger set. We therefore say it is a relaxation. Second, we can conclude that if X has rank n at the optimum, then by factorizing $X = B^TB$, B must be optimal in (3.12). We will prove this result from the dual problem.

A special case of (3.13) is when $p = 0$ (no dynamic in (3.1) or the static case). As expected, it is the covariance selection problem

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

In problem (3.13), the matrix R is distinguished from (3.5). It is assumed to be block-Toeplitz and positive definite. We partition R as

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

with $R_0 \in \mathbf{S}^n$, $R_1, \ldots, R_p \in \mathbf{R}^{n\times n}$. As we shall see, this property is sufficient to conclude that X has rank n at the optimum.

Dual problem We introduce a Lagrange multiplier $Z_0 \in \mathbf{S}^n$ for the first equality constraints ($k = 0$), multipliers $2Z_k \in \mathbf{R}^{n\times n}$ for equality constraints $k = 1$ through p, and a multiplier

$$U = \begin{bmatrix} U_{00} & U_{01} & \cdots & U_{0p} \\ U_{01}^T & U_{11} & \cdots & U_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ U_{0p}^T & U_{1p}^T & \cdots & U_{pp} \end{bmatrix} \in \mathbf{S}^{n(p+1)}$$

for the inequality constraint. The multipliers Z_k will be interpreted as blocks of a Toeplitz matrix

$$Z = \begin{bmatrix} Z_0 & Z_1 & \cdots & Z_p \\ Z_1^T & Z_0 & \cdots & Z_{p-1} \\ \vdots & \vdots & \ddots & \vdots \\ Z_p^T & Z_{p-1}^T & \cdots & Z_0 \end{bmatrix}.$$

The Lagrangian is then

$$\begin{aligned} & L(X, Z, U) \\ &= -\log\det X_{00} + \mathbf{tr}(RX) + \mathbf{tr}\left(Z_0 P(\sum_{i=0}^{p} X_{ii})\right) + 2\,\mathbf{tr}\sum_{k=1}^{p}\left(Z_k^T P(\sum_{i=0}^{p-k} X_{i,i+k})\right) - \mathbf{tr}(UX) \\ &= -\log\det X_{00} + \mathbf{tr}(RX) + \sum_{i=0}^{p}\mathbf{tr}\left(P(Z_0)X_{ii}\right) + 2\sum_{k=1}^{p}\sum_{i=0}^{p-k}\mathbf{tr}\left(P(Z_k)^T X_{i,i+k}\right) - \mathbf{tr}(UX) \\ &= -\log\det X_{00} + \mathbf{tr}((R + P(Z) - U)X). \end{aligned}$$

The dual function is the infimum of L over all X with $X_{00} \succ 0$. L is bounded below if and only if the following conditions hold:

$$R_0 + P(Z_0) - U_{00} \succ 0, \qquad R_k + P(Z_k) - U_{i,i+k} = 0, \quad k = 1, \ldots, p, \quad i = 0, \ldots, p-k.$$

In other words, $R + P(Z) - U$ must be zero, except for the $(0, 0)$ block, which must be positive definite. If L is bounded below, it is minimized by

$$X = \begin{bmatrix} (R_0 + P(Z_0) - U_{00})^{-1} & 0 \\ 0 & 0 \end{bmatrix}.$$

Hence the dual function is

$$g(Z_0, \ldots, Z_p, U) = \log\det(R_0 + P(Z_0) - U_{00}) + n.$$

We have derived the dual problem

$$\begin{array}{ll} \text{maximize} & \log\det(R_0 + P(Z_0) - U_{00}) + n \\ \text{subject to} & R_k + P(Z_k) - U_{i,i+k} = 0, \quad k = 1, \ldots, p, \quad i = 0, \ldots, p-k \\ & U \succeq 0. \end{array}$$

If we define $W = R_0 + P(Z_0) - U_{00}$ and eliminate the slack variable U, we can write this more simply as

$$\boxed{\begin{array}{ll} \text{maximize} & \log\det W + n \\ \text{subject to} & \begin{bmatrix} W & 0 \\ 0 & 0 \end{bmatrix} \preceq R + P(Z). \end{array}} \tag{3.14}$$

Note that in the static case ($p = 0$) this reduces to the maximum determinant completion problem

$$\text{maximize} \quad \log\det(R + P(Z)) + n.$$

Strong duality and optimality conditions We note the following properties of the primal and dual problem.

- The primal problem is strictly feasible ($X = I$ is strictly feasible), so Slater's condition holds. This implies strong duality, and also that the dual optimum is attained if the optimal value is finite.

- We have assumed that $R \succ 0$, and this implies that the primal objective function is bounded below, and that the primal optimum is attained. This also follows from the fact that the dual is strictly feasible ($Z = 0$ is strictly feasible if we take W small enough), so Slater's condition holds for the dual.

Therefore, if $R \succ 0$, we have strong duality and the primal and dual optimal values are attained. The KKT conditions are therefore necessary and sufficient for optimality of X, Z, W. The KKT conditions are:

1. *Primal feasibility.*

$$X \succeq 0, \qquad X_{00} \succ 0, \qquad P(\sum_{i=0}^{p-k} X_{i,i+k}) = 0, \quad k = 0, \ldots, p. \tag{3.15}$$

2. *Dual feasibility.*

$$W \succ 0, \qquad R + P(Z) \succeq \begin{bmatrix} W & 0 \\ 0 & 0 \end{bmatrix}. \tag{3.16}$$

3. *Zero duality gap.*

$$X_{00}^{-1} = W, \qquad \mathbf{tr}\left(X\left(R + P(Z) - \begin{bmatrix} W & 0 \\ 0 & 0 \end{bmatrix}\right)\right) = 0 \tag{3.17}$$

   Note that $\mathbf{tr}(XP(Z)) = \mathbf{tr}(P(X)Z) = 0$ if X satisfies the primal feasibility constraints, so the inner product in (3.17) reduces to $\mathbf{tr}(RX) - \mathbf{tr}(X_{00}W) = \mathbf{tr}(RX) - n$.

The complementary slackness condition can also be written as

$$X\left(R + P(Z) - \begin{bmatrix} W & 0 \\ 0 & 0 \end{bmatrix}\right) = 0 \tag{3.18}$$

(If A, B are positive semidefinite matrices then $\mathbf{tr}(AB) = 0$ if and only if $AB = 0$.)

Low-rank property of X Assume X^*, W^*, Z^* are optimal. We will show that X^* has rank n at the optimum.

Proposition 3.1. *Let $R(p)$ be a symmetric block-Toeplitz matrix defined as*

$$R(p) = \begin{bmatrix} R_0 & R_1 & \cdots & R_p \\ R_1^T & R_0 & \cdots & R_{p-1} \\ \vdots & \vdots & \ddots & \vdots \\ R_p^T & R_{p-1}^T & \cdots & R_0 \end{bmatrix}, \quad R_k \in \mathbf{R}^{n\times n}, k = 0, 1, \ldots, p.$$

If $R(p)$ satisfies

$$R(p) \succeq \begin{bmatrix} I_n & 0 \\ 0 & 0 \end{bmatrix}, \tag{3.19}$$

then $R(p) \succ 0$.

Proof. We will prove by induction. First of all, it is obvious that $R(0) = R_0 \succeq I \succ 0$. Next, we assume that $R(p) \succ 0$ and consider

$$R(p+1) = \begin{bmatrix} R_0 & \bar{R} \\ \bar{R}^T & R(p) \end{bmatrix}, \quad \bar{R} = \begin{bmatrix} R_1 & R_2 & \dots & R_{p+1} \end{bmatrix}.$$

The Schur complement of R_0 in $R(p+1)$ is

$$R_0 - \bar{R}R^{-1}(p)\bar{R}^T \succeq I \succ 0.$$

Therefore, with the assumption $R(p) \succ 0$, we can conclude that $R(p+1) \succ 0$. □

Proposition 3.2. *Consider the relaxation problem (3.13). Suppose R is block-Toeplitz and positive definite. Then there exists a solution X that has rank n at the optimum.*

proof. From the fact that $R + P(Z^*)$ having size $n(p+1) \times n(p+1)$ in (3.16) is block-Toeplitz, Proposition 3.1 implies that $R + P(Z^*) \succ 0$. Therefore the rank of the matrix

$$R + P(Z^*) - \begin{bmatrix} W^* & \mathbf{0}_{n\times np} \\ \mathbf{0}_{np\times n} & \mathbf{0}_{np\times np} \end{bmatrix}$$

is at least np, so its nullspace has dimension at most n. From (3.18) we see that $\mathbf{rank}(X^*) \leq n$ and the constraint $X^*_{00} \succ 0$ implies $\mathbf{rank}(X^*) = n$. □

This important result shows that under the assumptions that $R \succ 0$ and is block-Toeplitz, the optimal solution X can be factorized as $X = B^T B$ and B_0 is chosen such that $B_0 = X_{00}^{1/2}$. In other words, instead of solving the nonconvex problem (3.12), the convex problem (3.13) which can be solved efficiently, also provides an exact solution that achieves the true optimal value.

However, the matrix R in (3.5) used in the ML problem, is close to a block-Toeplitz matrix (in the norm sense), when N is relatively large compared to p. We conjecture that Proposition 3.2 is also true when R is almost-Toeplitz. This topic will be further studied in future work.

Chapter 4

Examples

In this chapter, we present three examples of real data sets from various fields to demonstrate how the proposed method can facilitate studies of interrelationships in multivariate time series. We discuss air pollution data, an example from chemistry, stock return data from economics, and fMRI data from neuroscience.

In order to learn a conditional independence graph, we estimate AR models of orders $p = 1$ to $p_{\max}$ subject to all possible sparsity constraints. Let n be the dimension of a time series. Therefore, the number of edges in the graph is $n(n-1)/2$ and the total number of all sparsity patterns is

$$\sum_{k=0}^{n(n-1)/2} \binom{n(n-1)/2}{k} = 2^{n(n-1)/2}. \tag{4.1}$$

Suppose, for a fixed p, we construct the matrix R from (3.5) based on the measurements from the process. For each sparsity constraint and each p, we solved (3.13) by using `CVX` [GB08a, GB08b] and decompose the optimal rank-n X to obtain AR parameters A_k. For each fitted model, we compute AIC (Akaike information criterion), second-order variant of AIC (AIC_c) or BIC (Bayesian information criterion) scores [BA02].

$$\text{AIC} = 2k - 2L, \tag{4.2}$$

$$\text{AIC}_c = 2k\left(\frac{N}{N-k-1}\right), \tag{4.3}$$

$$\text{BIC} = k\log N - 2L, \tag{4.4}$$

where L is the maximized log-likelihood, N is the sample size, and k is the effective

number of parameters. These scores are well-known criteria which are applicable in problems that the fitting is achieved by maximization of a log-likelihood. AIC may perform poorly if there are too many parameters compared to the size of sample. AIC_c will be recommended to use in this case. When N is large ($> e^2 \approx 7.4$), BIC tends to yield a simpler model as can be seen by the $\log N$ term which penalizes more heavily in complex models. We assume that the sample size, N in each data set may be large enough so that the approximate log-likelihood in (3.6) is close to the true value used in (4.2)-(4.4).

An autoregressive model of order p has $p+1$ parameters, $B_0 \in \mathbf{S}_+^n, B_1, \ldots B_p \in \mathbf{R}^{n\times n}$. Therefore, the number of independent parameters used in the experiment is

$$k = \frac{n(n+1)}{2} - |\mathcal{V}| + p(n^2 - 2|\mathcal{V}|). \tag{4.5}$$

Let p_{val} be the optimal value of (3.13) solved for each fitted model. From (3.6), (3.10), and (3.13), the maximized log-likelihood is

$$L = \frac{N p_{\mathrm{val}}}{2}. \tag{4.6}$$

By comparing these scores from all possible topologies, we can choose the best model that describes the data set. The spectrum and its inverse from the chosen model will be compared with the empirical estimate to illustrate the performance of our method. For this purpose, we will show the plots of normalized spectral density matrix *(coherence spectrum)* and normalized inverse of spectral density matrix *(partial coherence spectrum)*.

4.1 Air pollution data

In this section we will illustrate the proposed convex relaxation problem by application to a multivariate time series of air pollution data. The 5-dimensional time series consists of the concentration of four pollutants, CO, NO, NO_2, O_3 and the solar radiation intensity observed from Jan 1, 2006 to Dec 31, 2006 in Azusa, a city in Los Angeles county, California, USA.

This application has been studied previously in [Dah00] and [Eic06]. The first paper analyzed the data by a nonparametric frequency domain approach, i.e., determining the missing edge of the graph from the partial spectral coherence, while the latter learned the conditional independence graph by fitting the data to AR models whose parameters were solved by an iterative estimation. In this experiment, we will estimate AR models constrained by all combinations of sparsity patterns. The best model will be selected by applying BIC scores to compare those estimated models.

The data were collected in hourly basis from 12AM - 11PM except at 4AM (23 records per day). The original data with sample size $N = 8370$ contains missing values which is about 0.26% of the total values. We filled in the data by using linear interpolation method. The time series of the daily average over one year can be shown in Fig. 4.1 It agrees with the result from [Dah00] that CO and NO increase in the morning during rush hours. NO_2 consequently increases and O_3 also increases later due to the increase of the solar radiation and NO_2.

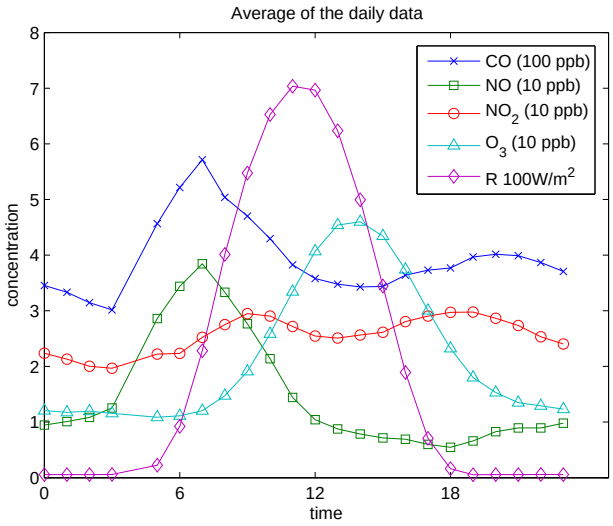


Figure 4.1: Average of daily data: CO, NO, NO_2, O_3, and the solar radiation (R).

We estimated the models of orders ranging from $p = 1$ to $p = 8$. In this example, $N = 8370$ is so large that L dominates the penalty term. Therefore, we opt to use BIC criteria and it was minimized when $p = 4$ as can be seen in Fig. 4.2 The BIC scores decrease dramatically from $p = 1$ to $p = 2$. The conditional independence graph that best characterizes the air pollution data is shown in Fig. 4.3

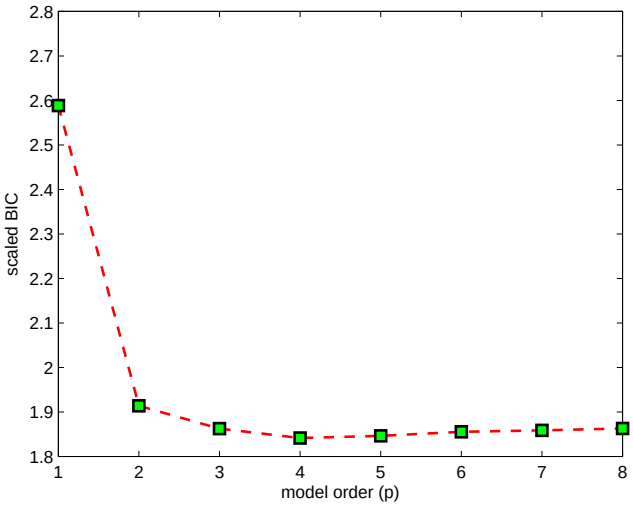


Figure 4.2: Minimized BIC scores (scaled by $1/N$) of p-order models of air pollution data.

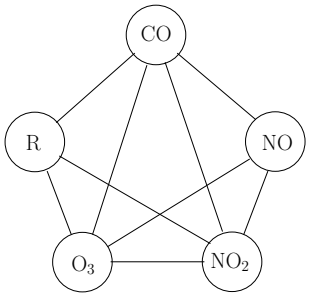


Figure 4.3: The conditional independence graph corresponding to the lowest BIC score for the air pollution data.

Fig. 4.4 shows the estimates of partial coherence and coherence spectrum obtained from the nonparametric approach and the ML estimates. The shape of spectra from both methods are fitted reasonably well.

In addition to the best topology, Table 4.1 also consider another best nine models since the first and the tenth BIC scores are different by only 0.84%. The notation

Table 4.1: Model selection results: index sets $\mathcal{V}$ of the 10-lowest BIC scores of estimated AR models for air pollution data.

rank	p	BIC scores	$\mathcal{V}$	description
1	4	15414	$\begin{pmatrix} 2 & 5 \end{pmatrix}$	NO-R
2	5	15455	$\begin{pmatrix} 2 & 5 \end{pmatrix}$	NO-R
3	4	15461	$\emptyset$	–
4	4	15494	$\begin{pmatrix} 1 & 4 \\ 1 & 5 \end{pmatrix}$	CO-O_3 CO-R
5	4	15502	$\begin{pmatrix} 1 & 5 \end{pmatrix}$	CO-R
6	5	15509	$\begin{pmatrix} 1 & 4 \\ 1 & 5 \end{pmatrix}$	CO-O_3 CO-R
7	5	15512	$\emptyset$	–
8	4	15527	$\begin{pmatrix} 1 & 4 \end{pmatrix}$	CO-O_3
9	6	15532	$\begin{pmatrix} 2 & 5 \end{pmatrix}$	NO-R
10	5	15544	$\begin{pmatrix} 1 & 5 \end{pmatrix}$	CO-R

$(i \;\; j)$ means that $(i, j) \in \mathcal{V}$, or x_i and x_j are conditionally independent.

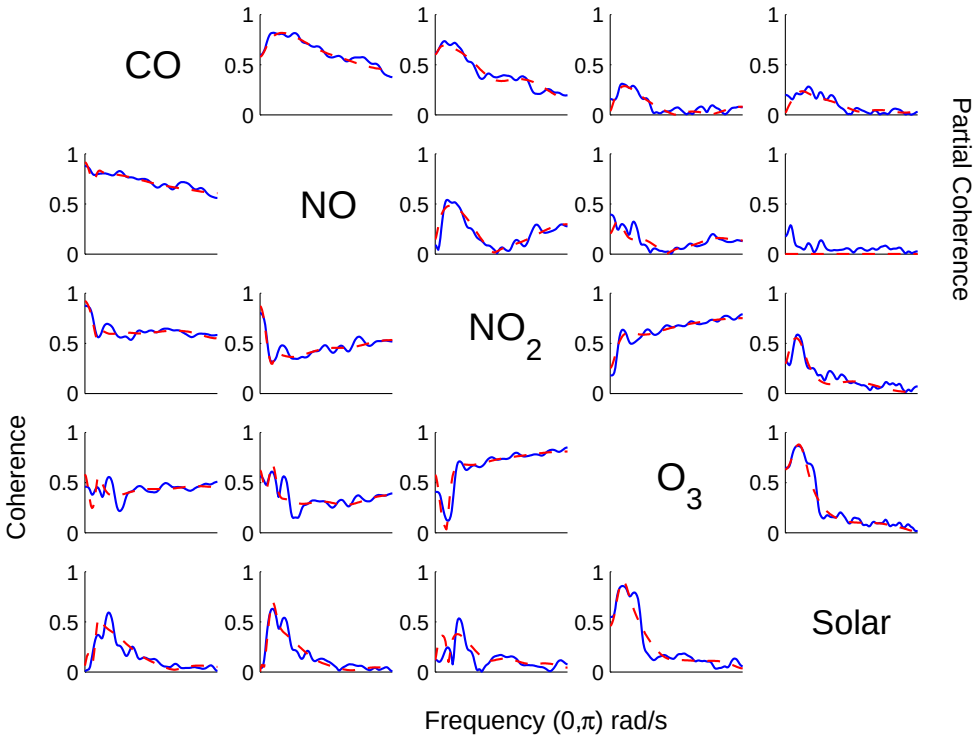


Figure 4.4: Partial coherence and coherence spectra of air pollution data: nonparametric estimates (solid blue lines) and ML estimates (dashed red lines).

From Table 4.1, the lowest BIC scores of each model of order $p = 4, 5, 6$ correspond to the missing edge between NO and the solar radiation. This agrees with the empirical partial coherence in Fig. 4.4 where the pair NO-R has visually the least magnitude. Table 4.1 also suggests that the most likely missing links are the combinations of $(1, 4), (1, 5)$ or CO-O_3, CO-R, respectively. In spite of the fact that the partial coherence spectra of these pairs are not identically zero, they tends to have small magnitudes compared to the other pairs.

The strong connections can be explained from [Dah00]. For example, The solar radiation plays a role in the photolysis of NO_2 and the generation of O_3. CO and NO are highly correlated because both of them are generated from cars. The edge between CO and NO_2 shows that the generation of NO_2 is mainly from the concentration of CO. The increase of O_3 is from the higher level of NO_2 and the radiation intensity.

4.2 Stock return data

4.2.1 International stock markets

In this experiment, we are interested in a multivariate time series of five major stock markets in the world. It consists of stock index closing prices of the markets in U.S., Japan, Hong Kong, United Kingdom and Germany. The study of dynamic interaction between the international stock markets has been of interest in economic literatures [ES89], [KP99], [BY03]. The goal of these works is to study how one stock market reacts to a change in other markets and how rapidly the movement in one market transmits to the others. The methodologies used in these works are based on firstly, forecast error covariance decomposition. It is an expression of mean-squared error of the prediction as a linear combination of variances in orthogonalized innovations. We can calculate the portion of the total variance of y_i due to the variance of the j^{th} shock (w_j). Secondly, an impulse response of AR process with orthogonalization of the noise can represent the reaction of the i^{th} variable due to a unit impulse of the j^{th} shock. With this method, they can explain how fast the movement in one market will affect the others. Last, the residuals or innovations represent the information that cannot be taken into account on the basis of all past data. Thus, the contemporaneous correlation matrix of the residual errors indicates the degree to which new information in one market is shared by the others.

In this prospectus, we would like to apply the conditional independence concept to study the interactions among international stock markets. This might not give full answers to the discussion in [ES89], [KP99], and [BY03], i.e., the role of leaders and followers in the markets cannot be identified. However, the interdependence between stock markets in two-way sense and the strength of these connections can be explained in the notion of conditional independence as well.

The data used in this experiment follows from [BY03], but the number of variables was reduced to five markets. They include

1. S & P 500 composite index (U.S.)
2. Nikkei 225 share index (Japan)
3. Hang Seng stock composite index (Hong Kong)

4. FTSE 100 share index (United Kingdom)

5. Frankfurt DAX 30 composite index (Germany)

The data were stock index closing prices recorded from June 4, 1997- June 15,1999 available from www.globalfinancial.com. We converted the data into US dollar to take the volatility of exchange rate into account. Due to nonsynchronous national holidays in each country, we complete the missing data by the most recent values. The return between trading day $k-1$ and k are defined as

$$r_k = 100(\log(p_k) - \log(p_{k-1})), \tag{4.7}$$

where p_k is the closing price on day k. This gives five-dimensional time series with sample size of 528 ($N = 528$) shown in Fig 4.5. It is known that the stock return data has high fluctuations due to the fact that it is generally sensitive to many factors such as news, political situation, or economic conditions.

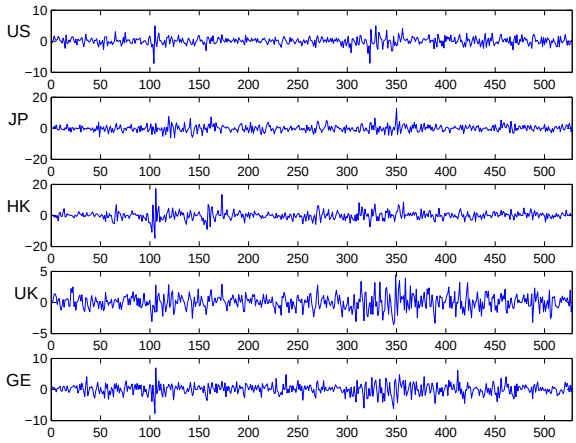


Figure 4.5: Daily data of international stock returns in US dollar from June 4, 1997 to June 15, 1999.

We estimate the models of orders ranging from $p = 1$ to $p = 9$. AIC_c score was used instead of AIC due to the small sample size. Fig. 4.6 shows that AIC_c and BIC criteria were minimized when the model order $p = 2$ and $p = 1$, respectively. The estimates of partial coherence and coherence spectra obtained from the nonparametric approach and the ML estimates are shown in Fig. 4.8. According to the AIC_c score, our method returns missing edges between US-JP and JP-GE while using BIC score

yields an additional edge between US-HK. The conditional independence graphs that best characterize the stock index return data are shown in Fig. 4.7. For AIC_c case, the shape of spectra appears to fitted in an acceptable level, but not in BIC case because the estimated model order (p) is too low to capture all characteristics in frequency domain. This is a nature of BIC as a parsimonious model selection method, which tends to choose a simple model.

For this application, it will give us more insight if the strength of connections are also considered. As shown in [Bri96], one can examine the strength of connection between two components by using the concept of *partial mutual information* between x_A and x_B given the remaining variables.

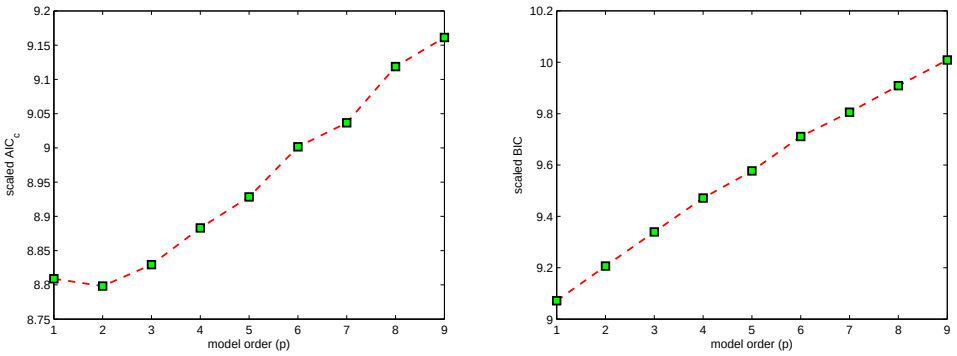
$$I_{AB|C} = -\frac{1}{2\pi}\int_0^{2\pi} \log\left(1 - |R_{AB|C}(\omega)|^2\right) d\omega. \tag{4.8}$$

We approximate the integral in (4.8) by the numerical Riemann sum as plotted in Fig. 4.9. It shows that UK and GE are highly correlated because there are much more transactions and business among the countries in Europe. US has the most interaction with GE and less with Asian countries.

From the methodologies in [BY03], more precisely, by using forcast error decompositions, they found that the German market has the most impact on the European markets. Especially, the UK market is substantially influenced by the German market. The Japanese market is highly exogenous, i.e., a change or information from other markets explains a price movement in Japan moderately. Also, a movement in Japan influences relatively little to other markets. These findings are consistent to a high $I_{AB|C}$ between GE-UK and a temperate $I_{AB|C}$ between JP and other markets in Fig. 4.9. The model selection results in Table 4.2 also support that the Japanese market tends to be isolated from the others, as the indices $(1,2),(2,5),(2,4)$ appear repeatedly in the best five models. Another result from [BY03] is that the contemporaneous structure of innovations showed that the volatility in the US market transmits to the world through Germany and Hong Kong. A similar result in our experiment can be seen from a relatively high level of partial mutual information between US-GE.

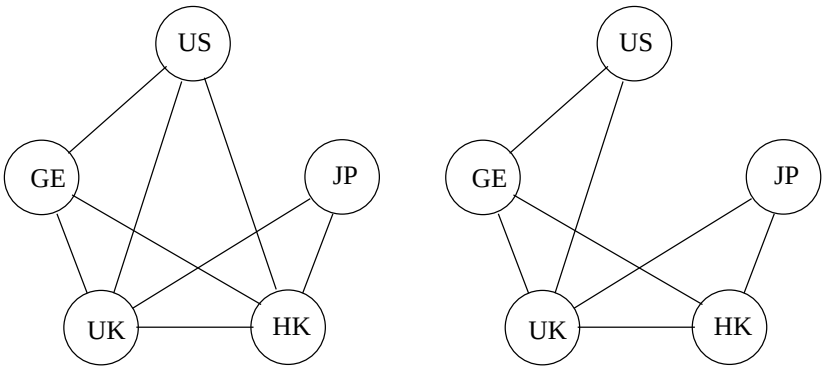
Our results may not fully support the previous works in all aspects since our methodology is different and the dimension of time series has been reduced. The dependence between two variables may still exist if they are correlated via a common

variable which has been removed. However, we select the major countries from Americas, Asia, and Europe and they should be able to represent the characteristic of each region in this particular period in a certain level.



(a) AIC$_c$ scores (b) BIC scores

Figure 4.6: Minimized AIC$_c$/BIC scores (scaled by $1/N$) of p-order models for international stock returns.

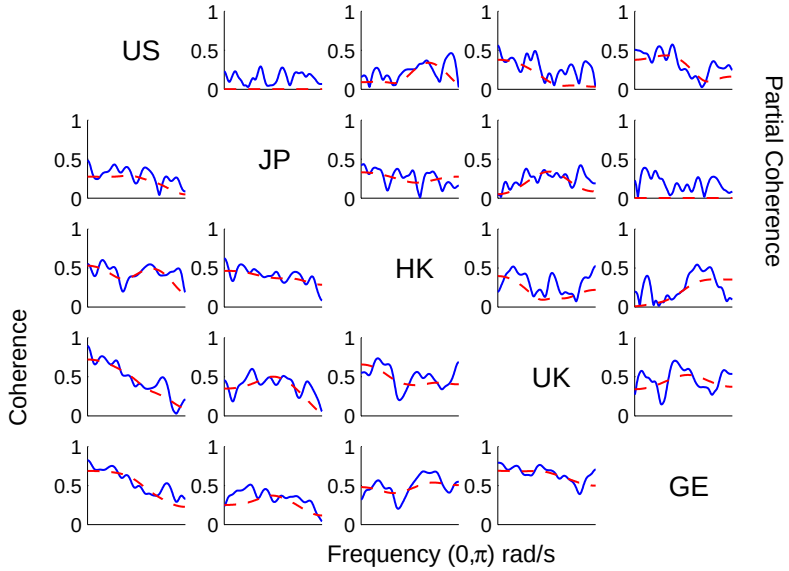


(a) AIC$_c$ (b) BIC

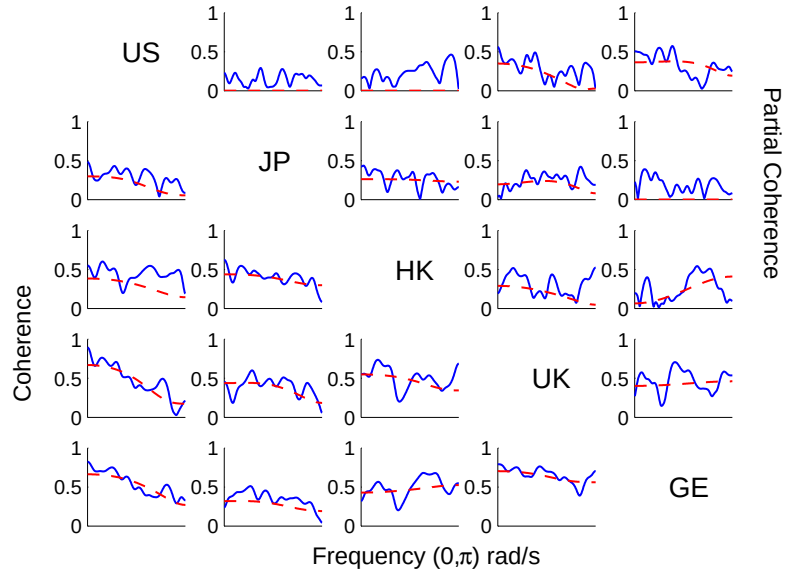
Figure 4.7: The conditional independence graph corresponding to the lowest AIC$_c$/BIC scores for international stock returns.

Table 4.2: Model selection results: index sets $\mathcal{V}$ of the 5-lowest AIC_c/BIC scores of estimated AR models for international stock returns.

rank	p	AIC_c scores	$\mathcal{V}$	countries	p	BIC scores	$\mathcal{V}$	countries
1	2	4645.5	$\begin{pmatrix} 1 & 2 \\ 2 & 5 \end{pmatrix}$	US-JP JP-GE	1	4789.65	$\begin{pmatrix} 1 & 2 \\ 1 & 3 \\ 2 & 5 \end{pmatrix}$	US-JP US-HK JP-GE
2	2	4648.0	$\begin{pmatrix} 1 & 2 \end{pmatrix}$	US-JP	1	4791.47	$\begin{pmatrix} 1 & 2 \\ 2 & 5 \end{pmatrix}$	US-JP JP-GE
3	1	4651.1	$\begin{pmatrix} 1 & 2 \\ 2 & 5 \end{pmatrix}$	US-JP JP-GE	1	4792.36	$\begin{pmatrix} 1 & 2 \\ 2 & 5 \\ 3 & 4 \end{pmatrix}$	US-JP JP-GE HK-UK
4	1	4651.6	$\begin{pmatrix} 1 & 2 \end{pmatrix}$	US-JP	1	4795.80	$\begin{pmatrix} 1 & 2 \\ 1 & 3 \\ 2 & 4 \end{pmatrix}$	US-JP US-HK JP-UK
5	2	4653.1	$\begin{pmatrix} 2 & 5 \end{pmatrix}$	JP-GE	1	4796.51	$\begin{pmatrix} 1 & 2 \\ 1 & 4 \\ 2 & 5 \end{pmatrix}$	US-JP US-UK JP-GE

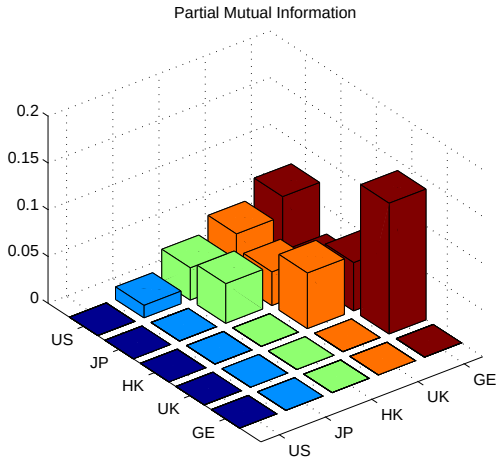


(a) AIC ($p = 2$)

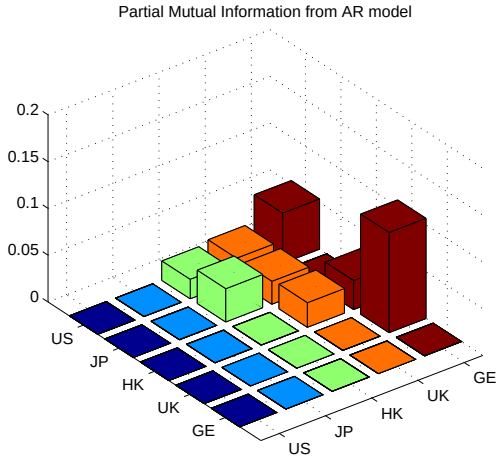


(b) BIC ($p = 1$)

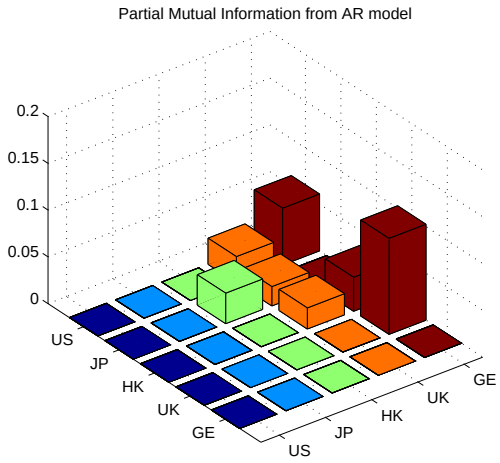
Figure 4.8: Partial coherence and coherence spectra of international stock returns: nonparametric estimates (solid blue lines) and ML estimates (dashed red lines) based on AIC_c and BIC criterions.



(a) Empirical estimate



(b) ML estimate based on AIC_c score



(c) ML estimate based on BIC score

Figure 4.9: Estimates of partial mutual information for international stock returns.

4.2.2 European Markets

In this example, we particularly focus on European stock markets. It consists of stock index closing prices of the markets in United Kingdom, France, Germany, Italy and Austria during Jan 1,1999 to Jul 31, 2008. All of these countries except UK have joined European Monetary Union (EMU) introduced since 1990. EMU is the agreement among the participating member states of the European Union to adopt a single hard currency and monetary system. On Jan 1,1999 the Euro currency became a legal currency. Its advantage is to eliminate the currency exchange fees from the cost of doing business between the European states. Therefore, by considering the data from the above period, we expect an integration of economics which results in highly dependencies among these countries. Despite of not being a member in EMU, UK was included in this model since it is known to be the leading market in Europe and expected to have strong linkages with EMU markets. The first four countries are considered large markets in terms of market capitalization. Austria, on the other hand, is classified as a small market and will be used to investigate how the large and small markets would affect to each other.

The relationships among European stock markets has been discussed in several papers. For example, [FS97] examined the data from 1988-1994 and found that the large markets (UK, France, Germany, and Netherlands) are highly correlated, but the smaller markets such as Belgium and Denmark are more independent. They showed that UK is the leading market which affects France, Germany, and Netherlands. The effect of EMU on the European market was studied in [YML03]. Their results indicate that the large EMU markets (Germany, France, Italy, Netherlands) became more correlated while the small markets (Austria, Belgium, and Ireland) became more independent from other EMU markets after the EMU has been introduced. They also found that the EMU markets appear to become less integrated with the non-member country (UK). The integration of stock markets within EMU members was also found in [KMW05]. They claimed that the overall comovement was a result from the introduction of the euro. Additionally, this effect is only significant for the small EMU members whose backgrounds in economic structure are different.

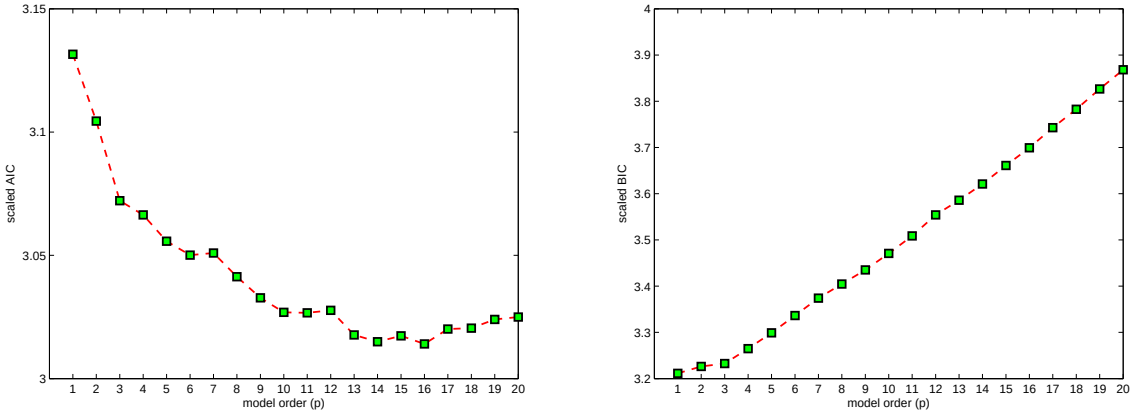
We analyzed the data in the same way as presented in section 4.2.1. The 5-dimensional time series includes the stock index prices from the following markets.

1. FTSE 100 share index (United Kingdom)
2. CAC 40 (France)
3. Frankfurt DAX 30 composite index (Germany)
4. MIBTEL (Italy)
5. Austrian Traded Index ATX (Austria)

The data were stock index closing prices recorded from Jan 1,1999-Jul 31, 2008 available from www.globalfinancial.com. The stock index price of UK was converted into Euro currency. The stock returns was computed from (4.7) and this yields a five-dimensional time series of European stock index returns with sample size of 2458.

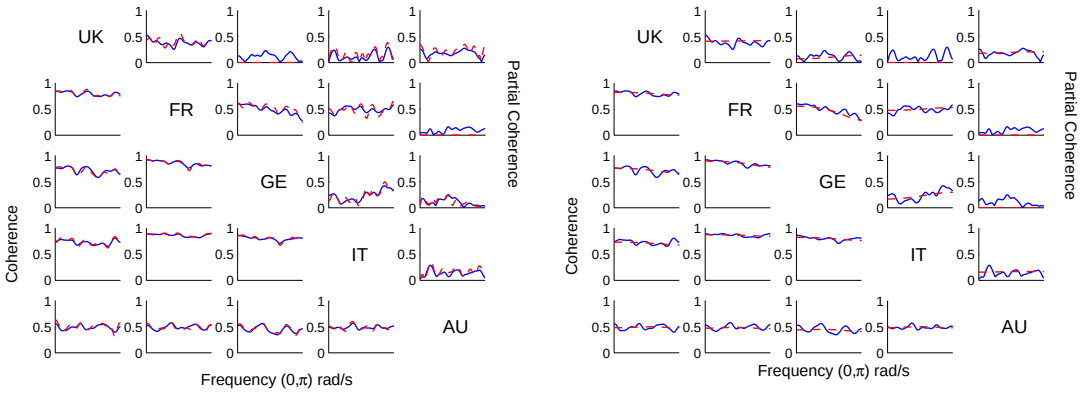
The model orders (p) range from 1 to 20 and we found that the optimal model order is 16 and 1, based on AIC and BIC scores respectively, shown in Fig 4.10. The best topology from AIC scores has missing edges between UK-GE and FR-AU. However, the model of order 16 seems to overfit the spectrum. We decided to select the best model of order $p = 14$, which has the same topology as $p = 16$, to represent the spectral estimates. The best model based on BIC scores explains that UK-IT, FR-AU, and GE-AU are conditionally independent. These topologies are graphically summarized in Fig. 4.12 and the corresponding spectral estimates are shown in Fig. 4.11. The topologies from both scores are slightly different and have only one common missing link between FR-AU. Nevertheless, both of them are understandable since their missing edges agree with small magnitudes of the empirical partial coherence estimates.

The plots of partial mutual information in Fig. 4.13 show that UK-FR, FR-GE, and FR-IT are highly dependent and IT-GE are correlated in a moderate level. This supports the hypothesis that the EMU members, especially the large markets are integrated significantly. The strong linkages among those countries may be partially explained from the impact of EMU, where the common currency and a single monetary policy will strengthen the market integration. It is interesting to see that UK has a very strong connection with EMU markets via France only. We can observe that the small market, Austria is likely to be isolated from other large markets. These results agree with [YML03] whose data were sampled after EMU launched as in our case and their results were mentioned previously.



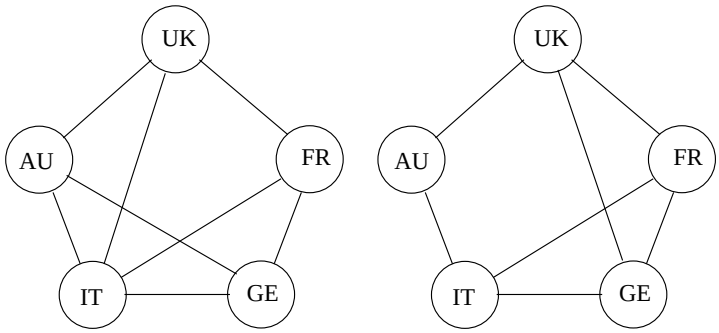
(a) AIC (b) BIC

Figure 4.10: Minimized AIC/BIC scores (scaled by $1/N$) of p-order models for European stock returns.



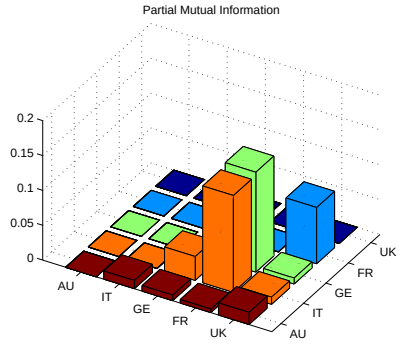
(a) AIC ($p = 9$) (b) BIC ($p = 1$)

Figure 4.11: Partial coherence and coherence spectra of European stock returns: nonparametric estimates (solid blue lines) and ML estimates (dashed red lines) based on AIC and BIC criterions.

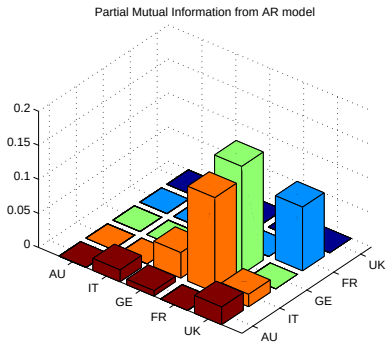


(a) AIC (b) BIC

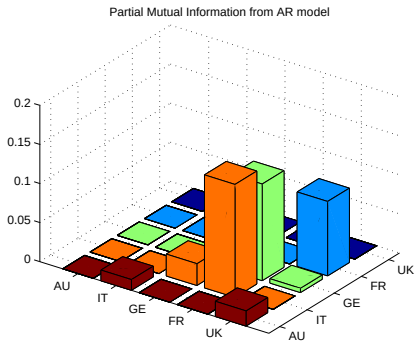
Figure 4.12: The conditional independence graphs corresponding to the lowest AIC/BIC scores for European stock returns.



(a) empirical estimate



(b) ML estimate based on AIC score



(c) ML estimate based on BIC score

Figure 4.13: Estimates of partial mutual information for European stock returns.

4.3 fMRI data

Recent researches in neuroscience have focused on investigating interactions between brain areas that are either stimulated by certain tasks or in resting states. Analyzing associations between interested regions could bring some insight understanding on the brain function to neuroscientists. It is widely accepted that the functional activity of each subregion can be demonstrated by human functional magnetic resonance imaging (fMRI) time series in which most cases, depicts blood oxygenation level. It is based on the assumption that the more activities the brain has, the higher level of oxygen will be used. Inference about the functional connectivity can be explained from the underlying dependence structure of the system. A graph-theoretical approach has been suggested to accommodate such analysis (see [SSSB05], [Eic05], Chapter 14 in [SWT06]). This section aims to illustrate how graphical models can be applied in analysis of fMRI data and demonstrate some preliminary results.

A brain is mainly divided into four anatomical regions, abbreviated by IFG, IFS, LOT, and STS. Four visual stimuli involved images of pictures and words changing randomly at a fixed rate. The volunteers were asked to response to these inputs while the brains were being scanned. Regional mean time series were estimated for each subject simply by averaging the data over all voxels in each region. This yields 4-dimensional time series associated with four condition codes of stimuli.

The fMRI data is clearly subject to the condition codes. Whereas a reasonable approach is to include the inputs in the modeling, such as using autoregressive model with exogenous input (ARX), at the initial state of this work, we simply ignore the input and keep using an autoregressive model (AR) to describe the dynamics of fMRI time series. While this model could be brain-unrealistic, it will serve our purpose of developing methods for further investigation. Examples of using AR models for fMRI data can be also found in [HPF03] and [VSSBLC$^+$05].

A time series of fMRI data measured from subject A is shown in Fig. 4.14. The empirical spectra appear to be different from subjects to subjects. We therefore selected some preliminary results of model selections from three volunteers and they are illustrated in Fig. 4.16. We fitted AR models of orders ranging from $p = 1$ to $p = 9$ for all 64 sparsity patterns. The best model were chosen from AIC_c criteria since the data have small sample sizes (N ranges from 540 to 648) in relation to the number

of estimate parameters. We present two similar results from the first two subjects showing that IFG-LOT and IFS-LOT are uncorrelated given the remaining variables. Moreover, they both have a strong dependency between IFG-IFS. In contrast, the plot of the last subject indicates no correlation between IFG and STS.

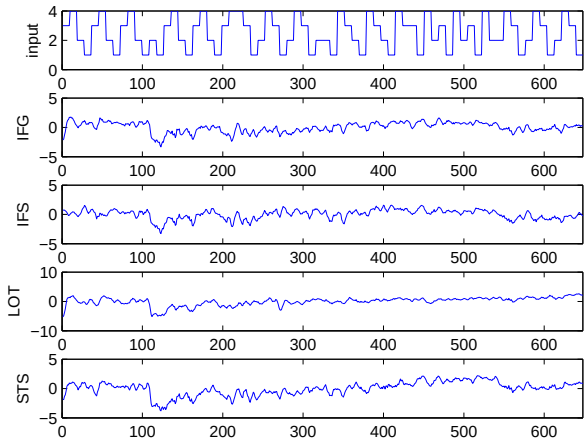
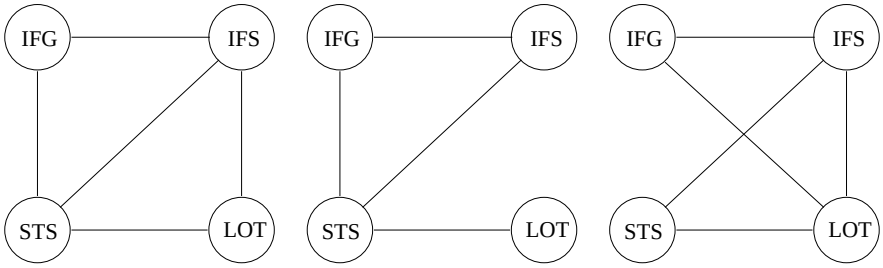


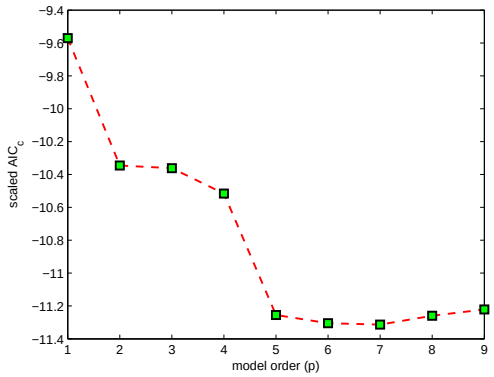
Figure 4.14: Detrended time series of average fMRI data over all voxels in each of four brain regions that are activated by four condition codes



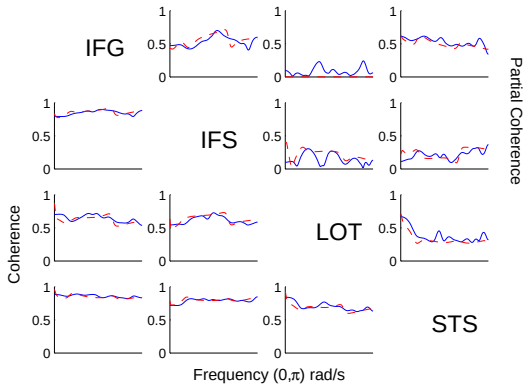
(a) subject A (b) subject B (c) subject C

Figure 4.15: The conditional independence graphs corresponding to the lowest AIC scores for fMRI data.

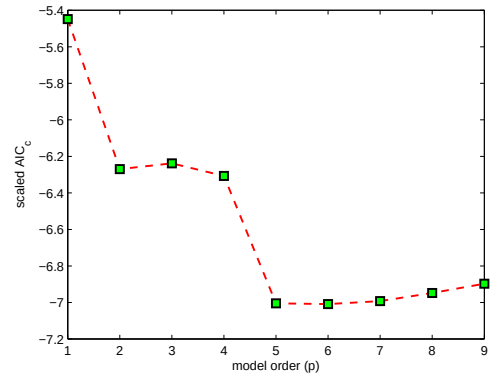
The preliminary results show that for each individual data, our method generally selects the best topology of which spectrum acceptably fits the empirical estimate. Our ultimate goal on this application is to have consistent results for most subjects and improve the modelling, in which the inputs must be included in the model.



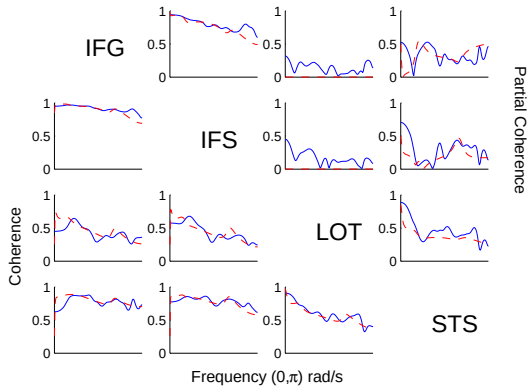
(a) AIC$_c$ scores of model order p



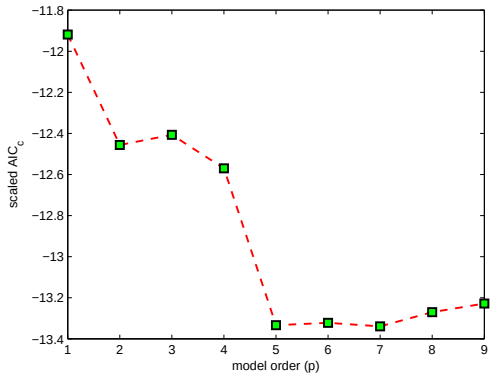
(b) spectrum estimates of the best model ($p = 7$)



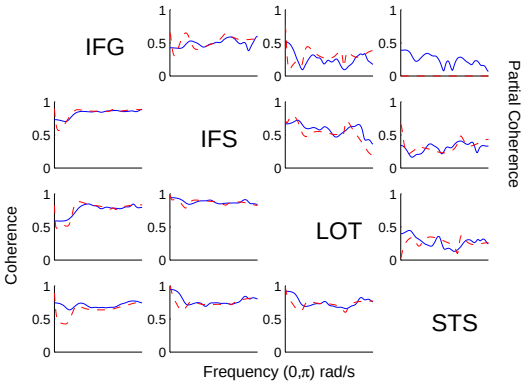
(c) AIC$_c$ scores of model order p



(d) spectrum estimates of the best model ($p = 6$)



(e) AIC$_c$ scores of model order p



(f) spectrum estimates of the best model ($p = 7$)

Figure 4.16: Spectral estimation results of fMRI data from subject A, B, and C shown in each row. The right column shows nonparametric estimates (solid blue lines) and ML estimates (dashed red lines) based on AIC$_c$.

Chapter 5

Granger causality

Chapter 2 has described the integration between a graph theory and the concept of conditional independence for identifying the interactions among variables. In this chapter we investigate another concept of such interaction and discuss graphical models in which the direction of connections is exploited for causal inference. We discuss the concept of Granger causality which has been extensively used for economic time series and follow a graphical framework from [Eic07] and Chapter 14 in [SWT06]. Our result is to provide a convex formulation for learning graphical models encoded by Granger causality constraints.

5.1 Definition

The idea of Granger causality is based on the concept that a cause cannot come after the effect. If a variable x affects a variable z, the former should help to improve the predictions of the latter variable [Lüt93]. To define this mathematically, suppose $\Omega(t)$ is the information set containing all available information up to the present time t. Let $z_t(h|\Omega(t))$ be the optimal h-step predictor of the process z_t at the origin t, based on the information in $\Omega(t)$. By optimal, we use mean-squared error (MSE) as a criterion. Let $\Sigma_z(h|\Omega(t))$ be the corresponding error covariance matrix. The process x_t is said *cause* z_t *in Granger's sense* if

$$\Sigma_z(h|\Omega(t)) < \Sigma_z(h|\Omega(t) \setminus \{x_s|s \leq t\}) \quad \text{for at least one } h = 1, 2, \ldots$$

In other words, x_t *Granger-causes* z_t if z_t can be predicted more efficiently when the information of x_t is taken into account in addition to all other information in the universe.

Furthermore, we say that there is *instantaneous causality* between z_t and x_t if

$$\Sigma_z(1|\Omega(t) \cup \{x_{t+1}\}) \neq \Sigma_z(1|\Omega(t)).$$

In another word, consider the period t, if we add x_{t+1} to the data set, it will make a difference in the prediction of z_{t+1}. It can be shown that this definition of causality is symmetric, i.e., if there is instantaneous causality between z_t and x_t, there is also instantaneous causality between x_t and z_t.

Consider a multivariate autoregressive model of order p.

$$y_k = -A_1 y_{k-1} - A_2 y_{k-2} - \cdots - A_p y_{k-p} + w_k, \tag{5.1}$$

where w_k is a Gaussian white noise with the covariance matrix Σ, $y_k \in \mathbf{R}^n$ and $A_k \in \mathbf{R}^{n\times n}$. The concept of Granger-causality can be characterized nicely as follows.

Proposition 5.1. *Let y_k be a multivariate AR process described in (5.1). Then we say y_j Granger-causes y_i if and only if*

$$[A_k]_{ij} = 0 \quad , \quad \forall k = 1, 2, \ldots p. \tag{5.2}$$

Moreover, there is no instantaneous causality between y_i and y_j if and only if

$$\Sigma_{ij} = 0. \tag{5.3}$$

Proof. See [Lüt93].

[Eic07] and [SWT06] (Chapter 14) have provided a graph-theoretical framework to represent causal inference under Granger-causality conditions. Path diagrams illustrating the dependence structure of a multivariate time series were introduced as follows.

Definition 5.2. *Let y be a multivariate time series with autoregressive description in (3.1). The path diagram associated with y in a graph $G(V, E)$ with a vertex set $V = \{1, 2, \ldots, n\}$ and an edge set E such that*

1. $j \to i \notin E \iff [A_k]_{ij} = 0$ *for all* $k = 1, \ldots, p$ *and*

2. $i \not\!\!- j \notin E \iff \Sigma_{ij} = 0.$

The path diagrams defined aboved contain two types of edges. The presence of directed edges $j \to i$ implies that y_j Granger-causes y_i. The graph contains an undirected edge $i \!\!- j$ if and only if y_i and y_j are instantaneously correlated.

Like conditional independence graphs, inference on causal structures in multivariate time series encoded by Granger causality can be determined by fitting graphical vector autoregressive models. If a graph $G(V, E)$ is given, we estimate an AR model subject to the constraints in Definition 5.2. Chapter 14 in [SWT06] suggested one iterative algorithm for fitting such model, but we will show in the next section that this can be done efficiently in a convex framework as well.

5.2 Maximum likelihood estimation

We are interested in a maximum-likelihood estimation based on $N + p$ observations, $y_1, \ldots, y_{N+p}$. From (3.3), the approximate log-likelihood function is, up to a constant,

$$\log L(A, \Sigma) = -\frac{N}{2} \log\det\Sigma - \frac{1}{2}\,\mathbf{tr}\left(\Sigma^{-1}(H_1 + AH_2)(H_1 + AH_2)^T\right). \tag{5.4}$$

Setting the gradient of (5.4) with respect to A and Σ gives

$$2H_2(H_1 + AH_2)^T\Sigma^{-1} = 0, \tag{5.5}$$

$$N\Sigma - (H_1 + AH_2)(H_1 + AH_2)^T = 0. \tag{5.6}$$

Assume that all rows of H_2 are linearly independent. The unconstrained ML estimates of A and Σ are

$$\boxed{\begin{array}{rcl} \bar{\Sigma} & = & \frac{1}{N}(H_1 + AH_2)(H_1 + AH_2)^T, \\ \bar{A} & = & -H_1H_2^T(H_2H_2^T)^{-1}, \end{array}} \tag{5.7}$$

which are just the estimates from the least-square problem. Now suppose the constraints on Granger causality (5.2), (5.3) are imposed. We will show that it can be formulated as a convex optimization problem. We first express the log-likelihood function in (5.4) as

$$\log L(A, \Sigma) = -\frac{N}{2}\left(\log\det\Sigma + \mathbf{tr}(\Sigma^{-1}\bar{\Sigma}) + \mathbf{tr}[\Sigma^{-1}(A - \bar{A})S(A - \bar{A})^T]\right). \tag{5.8}$$

By making use of (5.5),

$$\begin{aligned}(H_1 + AH_2)(H_1 + AH_2)^T \\ &= \left[(H_1 + \bar{A}H_2) + (A - \bar{A})H_2\right]\left[(H_1 + \bar{A}H_2) + (A - \bar{A})H_2\right]^T \\ &= (H_1 + \bar{A}H_2)(H_1 + \bar{A}H_2)^T + (A - \bar{A})H_2H_2^T(A - \bar{A})^T. \quad (5.9)\end{aligned}$$

Define

$$S = \frac{H_2H_2^T}{N}. \quad (5.10)$$

The last term in (5.4) can be expressed as

$$\mathbf{tr}\left[\Sigma^{-1}(H_1 + AH_2)(H_1 + AH_2)^T\right] = N\,\mathbf{tr}(\Sigma^{-1}\bar{\Sigma}) + N\,\mathbf{tr}[\Sigma^{-1}(A - \bar{A})S(A - \bar{A})^T], \quad (5.11)$$

and (5.8) will be readily obtained.

The log-likelihood function in (5.8) includes a convex term $(-\log\det\Sigma)$, so obviously it is not concave. However, we can show that $\log L$ is concave, jointly in Σ and A [BV04], in the region defined by

$$\Sigma \preceq 2\bar{\Sigma}.$$

This shows that we can use convex optimization to compute ML estimates of Σ and A, subject to convex constraints, as long as the constraints include $\Sigma \preceq 2\bar{\Sigma}$. We can make justification that the estimate Σ should not exceed twice the unconstrained ML estimates. In order to prove the convexity of the log-likelihood function, we need the following result.

Proposition 5.3. *The function $f : \mathbf{S}^n_{++} \to \mathbf{R}$ defined as*

$$f(X) = \log\det(X) + \mathbf{tr}(X^{-1}Y),$$

is convex on $\mathbf{dom}\, f = \mathbf{S}^n_{++}$ if $X \preceq 2Y$, where $Y \in \mathbf{S}^n_+$ is given.

proof. Define the first and second directional derivative of $f(X)$ in the direction V:

$$\begin{aligned}\left.\frac{d}{dt}f(X + tV)\right|_{t=0} &= \mathbf{tr}(\nabla f(X)^T V), \\ \left.\frac{d^2}{dt^2}f(X + tV)\right|_{t=0} &= \mathbf{tr}(V\nabla^2 f(X)[V]),\end{aligned}$$

where

$$\nabla^2 f(X)[V] = \frac{d}{dt}\nabla f(X+tV)\Big|_{t=0}.$$

If a function is twice differentiable, it is convex if and only if its Hessian is positive definite. The gradient and Hessian of f are given by

$$\begin{aligned}\nabla f(X) &= X^{-1} - X^{-1}YX^{-1},\\ \nabla^2 f(X)[V] &= -X^{-1}VX^{-1} + X^{-1}VX^{-1}YX^{-1} + X^{-1}YX^{-1}VX^{-1}.\end{aligned}$$

We will show that $\mathbf{tr}(V\nabla^2 f(X)[V]) \succeq 0$ for all V:

$$\begin{aligned}\mathbf{tr}(V\nabla^2 f(X)[V]) &= \mathbf{tr}\left[-VX^{-1}VX^{-1} + VX^{-1}VX^{-1}YX^{-1} + VX^{-1}YX^{-1}VX^{-1}\right]\\ &= \mathbf{tr}\left[-VX^{-1}VX^{-1} + 2VX^{-1}VX^{-1}YX^{-1}\right]\\ &= \mathbf{tr}\left[-X^{-1/2}VX^{-1/2}X^{-1/2}VX^{-1/2}\right]\\ &\quad + \mathbf{tr}\left[2X^{-1/2}VX^{-1/2}X^{-1/2}VX^{-1/2}X^{-1/2}YX^{-1/2}\right]\\ &= \mathbf{tr}\left[X^{-1/2}VX^{-1/2}(2X^{-1/2}YX^{-1/2} - I)X^{-1/2}VX^{-1/2}\right].\end{aligned}$$

If $X \preceq 2Y$, then $I \preceq 2X^{-1/2}YX^{-1/2}$. Hence, $\mathbf{tr}(V^T\nabla^2 f(X)V \succeq 0$ for all V, which implies that $f(X)$ is convex. □

As a result, the first two terms in (5.8) is convex if $\Sigma \preceq 2\bar{\Sigma}$. The last term is a matrix fractional function in Σ and A, which is convex.

Suppose the Granger-causality constraints are given. The sparsity patterns are characterized by two sets. Let $\mathcal{V}_1$ and $\mathcal{V}_2$ be the index sets of Granger-causality constraints (5.2) and instantaneous causality constraints (5.3), respectively. In this problem, $\mathcal{V}_1$ is not necessarily symmetric. The projection operators defined as in (3.11) on sets $\mathcal{V}_1$ and $\mathcal{V}_2$ are P_1 and P_2, respectively. We can formulate a convex optimization problem of maximum-likelihood estimations with Granger causality constraints as follows:

$$\boxed{\begin{array}{ll}\text{minimize} & \log\det\Sigma + \mathbf{tr}(\Sigma^{-1}\bar{\Sigma}) + \mathbf{tr}[\Sigma^{-1}(A-\bar{A})S(A-\bar{A})^T]\\ \text{subject to} & 0 \preceq \Sigma \preceq 2\bar{\Sigma}\\ & P_1(A_k) = 0\ ,\ k = 0, 1, \ldots p\\ & P_2(\Sigma) = 0.\end{array}} \tag{5.12}$$

The variables are $A_k \in \mathbf{R}^{n\times n}$ for $k = 1, \ldots, p$ and $\Sigma \in \mathbf{S}^n_{++}$.

Zero constraints on AR parameters which are one form of linear constriants were fully treated in [Ham94] and [Lüt93]. A least squares estimator and its extensions with various asymptotic properties were derived. Estimation under zero restrictions for Σ is often performed by decomposing $\Sigma = U\Lambda U^H$ and making change of variables in the recursive equation. Some patterns of zero constraints on Σ are then equivalent to zero constraints on the new variable, which can be imposed in the least squares problems.

However, if both restrictions on AR parameters and covariance matrix Σ have to be imposed in the estimation simultaeneously, it is not evident in [Lüt93] how this can be carried out. Closest to this goal is the work of Eichler in [SWT06] where an iterative algorithm of the two steps has been used. Our formulation in (5.12) can be an alternative approach to serve this purpose.

If no a priori knowledge of sparsity constraints is available, one can compare all possible models and determine the best topology by minimizing a model selection criterion such as AIC or BIC as described in Chapter 4.

Chapter 6

Conclusions

6.1 Summary

We have considered a parametric approach for maximum-likelihood estimation of autoregressive models with conditional independence constraints. The zero constraints on the inverse of spectral density matrix result in nonconvex constraints, leading to a nonconvex problem which is generally difficult to guarantee a global solution. We made change of variables and proposed a convex framework which can be solved efficiently by interior-point algorithms. This allows us to solve a graphical inference problem by fitting autoregressive models according to all possible topologies and applying a model selection criterion to determine the best fitted graph. We have also applied our approach to three sets of real data and they illustrate that our method performs reasonably well.

6.2 Future plans

There are several interesting issues related to our work. The topics for further studies are described as follows.

Extension of the proof to non-Toeplitz R

The matrix R defined in (3.5) is in general, a non-block-Toeplitz matrix, while an assumption in our proof requires this property to conclude the low-rank property of

the solutions to the relax problem (3.13) . Denote R_T a sample covariance matrix which is of the form

$$R_T = \begin{bmatrix} R_0 & R_1 & \cdots & R_p \\ R_1^T & R_0 & \cdots & R_{p-1} \\ \vdots & \vdots & \ddots & \vdots \\ R_p^T & R_{p-1}^T & \cdots & R_0 \end{bmatrix} \quad , R_k = \sum_{j=1}^{N-k} y_{j+k} y_j^T$$

It is known that R_T has block-Toeplitz structure. From the expression of R in (3.5),

$$R_T = R + \Delta R.$$

We can claim that when the sample size (N) is relatively large compared to the model order (p), $\|\Delta R\|$ will be small and the result in Proposition 3.1 should still be true. This is confirmed by the experimental results that the numerical solutions to the convex problem (3.13) generally have low rank. We try to seek a formal proof to affirm this result.

l_1-Penalization

To detect the sparsity pattern of the inverse of spectral density matrix that best describes the data, the number of subset models that need to be estimated is

$$\sum_{k=0}^{n(n-1)/2} \binom{n(n-1)/2}{k} = 2^{n(n-1)/2}.$$

For example, if $n = 5$, we have 1024 candidate models of order p. In practice, the dimension of a process will often be greater than in this example. For instance, a brain network typically has a large number of nodes which in the order of hundreds. As a result, the practicability of this procedure is limited by its computational cost. We would like to investigate a method that is capable of dealing with this difficulty.

It is well-known that a regularization with an l_1-norm can be used as a heuristic method to find a sparse solution [BV04]. Consider a simple problem

$$\text{minimize} \quad \|Ax - b\|_2 + \gamma \|x\|_1,$$

where γ is the penalty parameter controlling a tradeoff between the residual error and the sparsity of x.

Consider the convex problem (3.13) in an equivalent form,

$$\begin{aligned} \text{minimize} \quad & -\log\det X_{00} + \mathbf{tr}(RX) \\ \text{subject to} \quad & Y_k = \sum_{i=0}^{p-k} X_{i,i+k}\ ,\ k = 0, 1, \ldots, p \\ & [Y_k]_{ij} = [Y_k]_{ji} = 0,\ \forall k = 0, \ldots, p\ \ \forall (i,j) \in \mathcal{V}. \end{aligned} \tag{6.1}$$

The goal is to recover the sparsity pattern in Y_k automatically. Moreover, the location of zeros in all matrices Y_k must be the same (refer to (3.9)). To accomplish this we propose the following problem:

$$\boxed{\begin{aligned} \text{maximize} \quad & \log\det X_{00} - \mathbf{tr}(RX) + \gamma\|W\|_1 \\ \text{subject to} \quad & Y_k = \sum_{i=0}^{p-k} X_{i,i+k}\ ,\ k = 0, 1, \ldots, p \\ & -W_{ij} \leq [Y_k]_{ij} \leq W_{ij}, \quad \forall i \neq j, k = 0, 1, \ldots p \\ & X \succeq 0, \quad W_{ij} \geq 0, \quad \forall i \neq j. \end{aligned}} \tag{6.2}$$

γ is the regularization parameter and W is introduced as a maximum modulus of Y_k. It is obvious that the sparsity of all matrices Y_k are the same as the sparsity of W.

By varying γ and solve (6.2), we can plot a trade-off curve between $\log\det X_{00} - \mathbf{tr}(RX)$ and $\|W\|_1$. We then find the smallest γ at which the log-likelihood term tends to increase slightly. The formulation and the choice of γ may need some improvements to obtain more satisfactory results.

Applications

Discovering and characterizing functional brain connectivity is an important and very active research topic in neuroscience. However, the data and methods used in the analysis vary enormously. Even defining brain connectivity has proven to be difficult [FFK05, Hor03]. One possible approach is to use the statistical definition of conditional independence of time series. However, several improvements are needed to apply our methodology to fMRI data. First, a realistic model must include the inputs used to stimulate the brain. Second, the inputs which are generally visual images are categorical inputs. Typically, a dynamical model requires all variables to have numeric values. It thus seems important to construct an encoding scheme

that maps these inputs to numeric fields. This can be intuitively modeled as a set of binary vectors or multiple value thresholds for discriminating one category from another. However, this approach is often dealt with a basic difficulty that there is no meaningful correspondence between the generated numbers and the catogorical inputs.

There is a vast literature on modeling of fMRI data that is relevant for further studies and is based on linear time-invariant models [FJT94, BEGH96, Coh97, JTF97, RKMVC98]. The haemodynamic response in fMRI (HRF) is modeled as a convolution of a stimulus function and a linear filter. These approaches impose the shape of the HRF by choosing several types of filter or modeling the stimulus as an impulse-like function. A recent survey paper discussed wide ranges of models used in imaging neuroscience [Fri05], starting from anatomical models to some statistical models.

We also plan to investigate some applications of graphical models with Granger causality constraints. There is an example of economic data in [Eic07] that we would like to replicate, since it represents important elements in macroeconomic systems, and it has been addressed in many studies. Applications on neural systems shown in [Eic05] [FSGM$^+$07], or [VSSBLC$^+$05] will also be of our interest.

Bibliography

[BA02] K.P. Burnham and D.R. Anderson. *Model Selection and Multimodel Inference: A Practical Information-theoretic Approach*. Springer, 2002.

[BEG08] O. Banerjee and L. El Ghaoui. Model Selection Through Sparse Maximum Likelihood Estimation for Multivariate Gaussian or Binary Data. *Journal of Machine Learning Research*, 9:485–516, 2008.

[BEGH96] G.M. Boynton, S.A. Engel, G.H. Glover, and D.J. Heeger. Linear Systems Analysis of Functional Magnetic Resonance Imaging in Human V1. *Journal of Neuroscience*, 16(13):4207–4221, 1996.

[BJ04] FR Bach and MI Jordan. Learning graphical models for stationary time series. *Signal Processing, IEEE Transactions on [see also Acoustics, Speech, and Signal Processing, IEEE Transactions on]*, 52(8):2189–2199, 2004.

[Bri75] D.R. Brillinger. *Time Series Analysis: Data Analysis and Theory*. Holt, Rinehart & Winston, Inc., 1975.

[Bri96] D.R. Brillinger. Remarks concerning graphical models for time series and point processes. *Revista de Econometria*, 16:1–23, 1996.

[BV04] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[BY03] D.A. Bessler and J. Yang. The structure of interdependence in international stock markets. *Journal of International Money and Finance*, 22(2):261–287, 2003.

[Coh97] M.S. Cohen. Parametric Analysis of fMRI Data Using Linear Systems Methods. *Neuroimage*, 6(2):93–103, 1997.

[Dah00] R. Dahlhaus. Graphical interaction models for multivariate time series. *Metrika*, 51(2):157–172, 2000.

[DE03] R. Dahlhaus and M. Eichler. Causality and graphical models in time series analysis. *Highly Structured Stochastic Systems*, 27:115–144, 2003.

[Dem72] A.P. Dempster. Covariance selection. *Biometrics*, 28(1):157–175, 1972.

[DES97] R. Dahlhaus, M. Eichler, and J. Sandkuhler. Identification of synaptic connections in neural ensembles by graphical models. *Journal of Neuroscience Methods*, 77(1):93–107, 1997.

[DVR08] J. Dahl, L. Vandenberghe, and V. Roychowdhury. Covariance selection for non-chordal graphs via chordal embedding. *Optimization methods and software*, 23(4):501–520, 2008.

[EDS03] M. Eichler, R. Dahlhaus, and J. Sandkühler. Partial correlation analysis for the identification of synaptic connections. *Biological Cybernetics*, 89(4):289–302, 2003.

[Edw00] D. Edwards. *Introduction to Graphical Modelling*. Springer, 2000.

[Eic05] M. Eichler. A graphical approach for evaluating effective connectivity in neural systems. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1457):953, 2005.

[Eic06] Michael Eichler. Fitting graphical interaction models to multivariate time serie. *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence*, 2006.

[Eic07] M. Eichler. Granger causality and path diagrams for multivariate time series. *Journal of Econometrics*, 137(2):334–353, 2007.

[ES89] C.S. Eun and S. Shim. International transmission of stock market movements. *Journal of Financial and Quantitative Analysis*, 24(2):241–256, 1989.

[FD03] R. Fried and V. Didelez. Decomposability and selection of graphical models for multivariate time series. *Biometrika*, 90(2):251, 2003.

[FFK05] A.A. Fingelkurts, A.A. Fingelkurts, and S. Kähkönen. Functional connectivity in the brainis it an elusive concept? *Neuroscience and Biobehavioral Reviews*, 28(8):827–836, 2005.

[FHT07] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 2007.

[FJT94] KJ Friston, P. Jezzard, and R. Turner. Analysis of functional MRI time-series. *Human Brain Mapping*, 1(2):153–171, 1994.

[FMM+05] S. Feiler, KG Muller, A. Muller, R. Dahlhaus, and W. Eich. Using Interaction Graphs for Analysing the Therapy Process. *Psychother Psychosom*, 74(2):93–99, 2005.

[Fri05] K.J. Friston. Models of Brain Function in Neuroimaging. *Annual Review of Psychology*, 56(1):57–87, 2005.

[FS97] J. Friedman and Y. Shachmurove. Co-movements of major European community stock markets: A vector autoregression analysis. *Global Finance Journal*, 8(2):257–277, 1997.

[FSGM+07] A. Fujita, J.R. Sato, H.M. Garay-Malpartida, R. Yamaguchi, S. Miyano, M.C. Sogayar, and C.E. Ferreira. Modeling gene expression regulatory networks with the sparse vector autoregressive model. *BMC Systems Biology*, 1(1):39, 2007.

[GB08a] M. Grant and S. Boyd. CVX: Matlab software for disciplined convex programming (web page and software). http://stanford.edu/ boyd/cvx, August 2008.

[GB08b] M. Grant and S. Boyd. Graph Implementations for Nonsmooth Convex Programs. In V. Blondel, S. Boyd, and H. Kimura, editors, *Recent Advances in Learning and Control (a tribute to M. Vidyasagar)*. Springer Verlag, 2008.

[GIF02] U. Gather, M. Imhoff, and R. Fried. Graphical models for multivariate time series from intensive care monitoring. *Statistics in Medicine*, 21(18):2685–2701, 2002.

[Ham94] J.D. Hamilton. *Time Series Analysis*. Princeton University Press, 1994.

[Hor03] B. Horwitz. The elusive concept of brain connectivity. *Neuroimage*, 19(2):466–470, 2003.

[HPF03] L. Harrison, WD Penny, and K. Friston. Multivariate autoregressive modeling of fMRI time series. *Neuroimage*, 19(4):1477–1491, 2003.

[Jor04] M.I. Jordan. Graphical models. *Statistical Science*, 19, 2004.

[JTF97] O. Josephs, R. Turner, and K. Friston. Event-Related fMRI. *Human Brain Mapping*, 5(4):243–248, 1997.

[KMW05] S.J. Kim, F. Moshirian, and E. Wu. Dynamic stock market integration driven by the European Monetary Union: An empirical analysis. *Journal of Banking and Finance*, 29(10):2475–2502, 2005.

[KP99] J. Knif and S. Pynnönen. Local and global price memory of international stock markets. *Journal of International Financial Markets, Institutions & Money*, 9(2):129–147, 1999.

[Lau96] S.L. Lauritzen. *Graphical Models*. Oxford University Press, USA, 1996.

[Lüt93] H. Lütkepohl. *Introduction to Multiple Time Series Analysis*. Springer, 1993.

[RKMVC98] J.C. Rajapakse, F. Kruggel, J.M. Maisog, and D.Y. Von Cramon. Modeling hemodynamic response for analysis of functional MRI time-series. *Human Brain Mapping*, 6(4):283–300, 1998.

[SSSB05] R. Salvador, J. Suckling, C. Schwarzbauer, and E. Bullmore. Undirected graphs of frequency-dependent functional connectivity in whole brain networks. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1457):937–946, 2005.

[SWT06] B. Schelter, M. Winterhalder, and J. Timmer. *Handbook of Time Series Analysis: Recent Theoretical Developments and Applications*. Wiley-VCH Verlag GmbH & Co. KGaA, 2006.

[TLH+00] J. Timmer, M. Lauk, S. Häußler, V. Radt, B. Köster, B. Hellwig, B. Guschlbauer, CH Lücking, M. Eichler, and G. Deuschl. Cross-spectral analysis of tremor time series. *Internation Journal of Bifurcation and Chaos*, 10:2595–2610, 2000.

[VSSBLC+05] P.A. Valdés-Sosa, J.M. Sánchez-Bornot, A. Lage-Castellanos, M. Vega-Hernández, J. Bosch-Bayard, L. Melie-García, and E. Canales-Rodríguez. Estimating brain functional connectivity with sparse multivariate autoregression. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1457):969–981, 2005.

[Whi90] J. Whittaker. *Graphical models in applied multivariate statistics*. Wiley New York, 1990.

[YL07] M. Yuan and Y. Lin. Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94(1):19, 2007.

[YML03] J. Yang, I. Min, and Q. Li. European Stock Market Integration: Does EMU Matter? *Journal of Business Finance & Accounting*, 30(9-10):1253–1276, 2003.