1 Graphical models of autoregressive processes

Jitkomut Songsiri, Joachim Dahl, and Lieven Vandenberghe

Jitkomut Songsiri is with the University of California, Los Angeles, USA.
Joachim Dahl is with Anybody Technology A/S, Denmark.
Lieven Vandenberghe is with the University of California, Los Angeles, USA.

We consider the problem of fitting a Gaussian autoregressive model to a time series, subject to conditional independence constraints. This is an extension of the classical covariance selection problem to times series. The conditional independence constraints impose a sparsity pattern on the inverse of the spectral density matrix, and result in nonconvex quadratic equality constraints in the maximum likelihood formulation of the model estimation problem. We present a semidefinite relaxation, and prove that the relaxation is exact when the sample covariance matrix is block-Toeplitz. We also give experimental results suggesting that the relaxation is often exact when the sample covariance matrix is not block-Toeplitz. In combination with model selection criteria the estimation method can be used for topology selection. Experiments with randomly generated and several real data sets are also included.

1.1 Introduction

Graphical models give a graph representation of relations between random variables. The simplest example is a *Gaussian graphical model*, in which an undirected graph with n nodes is used to describe conditional independence relations between the components of an n-dimensional random variable $x \sim N(0, \Sigma)$. The absence of an edge between two nodes of the graph indicates that the corresponding components of x are independent, conditional on the other components. Other common examples of graphical models include *contingency tables*, which describe conditional independence relations in multinomial distributions, and *Bayesian networks*, which use directed acyclic graphs to represent causal or temporal relations. Graphical models find applications in bioinformatics, speech and image processing, combinatorial optimization, coding theory, and many other fields. Graphical representations of probability distributions not only offer insight in the structure of the distribution, they can also be exploited to improve the efficiency of statistical calculations, such as the computation of conditional or marginal probabilities. For further background we refer the reader to several books and survey papers on the subject [1, 2, 3, 4, 5, 6, 7].

Estimation problems in graphical modeling can be divided in two classes, depending on whether the topology of the graph is given or not. In a Gaussian graphical model of $x \sim N(0, \Sigma)$, for example, the conditional independence relations between components of x correspond to zero entries in the inverse covariance matrix [8]. This follows from the fact that the conditional distribution of two variables x_i, x_j, given the remaining variables, is Gaussian, with covariance matrix

$$\begin{bmatrix} (\Sigma^{-1})_{ii} & (\Sigma^{-1})_{ij} \\ (\Sigma^{-1})_{ji} & (\Sigma^{-1})_{jj} \end{bmatrix}^{-1}.$$

Hence x_i and x_j are conditionally independent if and only if

$$(\Sigma^{-1})_{ij} = 0.$$

Specifying the graph topology of a Gaussian graphical model is therefore equivalent to specifying the sparsity pattern of the inverse covariance matrix. This property allows us to formulate the maximum likelihood (ML) estimation problem of a Gaussian graphical model, for a given graph topology, as

$$\begin{array}{ll} \text{maximize} & -\log\det\Sigma - \mathbf{tr}(C\Sigma^{-1}) \\ \text{subject to} & (\Sigma^{-1})_{ij} = 0, \quad (i,j) \in \mathcal{V}, \end{array} \tag{1.1}$$

where C is the sample covariance matrix, and $\mathcal{V}$ are the pairs of nodes (i, j) that are not connected by an edge, *i.e.*, for which x_i and x_j are conditionally independent. (Throughout the chapter we take as the domain of the function $\log\det X$ the set of positive definite matrices.) A change of variables $X = \Sigma^{-1}$ results in a convex problem

$$\begin{array}{ll} \text{maximize} & \log\det X - \mathbf{tr}(CX) \\ \text{subject to} & X_{ij} = 0, \quad (i,j) \in \mathcal{V}. \end{array} \tag{1.2}$$

This is known as the *covariance selection* problem [8], [2, Section 5.2]. The corresponding dual problem is

$$\begin{array}{ll} \text{minimize} & \log\det Z^{-1} \\ \text{subject to} & Z_{ij} = C_{ij}, \quad (i,j) \notin \mathcal{V}, \end{array} \tag{1.3}$$

with variable $Z \in \mathbf{S}^n$ (the set of symmetric matrices of order n). It can be shown that $Z = X^{-1} = \Sigma$ at the optimum of (1.1), (1.2), and (1.3). The ML estimate of the covariance matrix in a Gaussian graphical model is the maximum determinant (or maximum entropy) completion of the sample covariance matrix [9, 10].

The problem of estimating the topology in a Gaussian graphical model is more involved. One approach is to formulate hypothesis testing problems to decide about the presence or absence of edges between two nodes [2, §5.3.3]. Another possibility is to enumerate different topologies, and use information-theoretic criteria (such as the Akaike or Bayes information criteria) to rank the models.

A more recent development is the use of convex methods based on ℓ_1-norm regularization to estimate sparse inverse covariance matrices; see [11, 12, 13].

In this chapter we address the extension of estimation methods for Gaussian graphical models to autoregressive (AR) Gaussian processes

$$x(t) = -\sum_{k=1}^{p} A_k x(t-k) + w(t), \tag{1.4}$$

where $x(t) \in \mathbf{R}^n$ and $w(t) \sim N(0, \Sigma)$ is Gaussian white noise. It is known that conditional independence between components of a multivariate stationary Gaussian process can be characterized in terms of the inverse of the spectral density matrix $S(\omega)$: Two components $x_i(t)$ and $x_j(t)$ are independent, conditional on the other components of $x(t)$, if and only if

$$(S(\omega)^{-1})_{ij} = 0$$

for all ω [14, 15]. This connection allows us to include conditional independence constraints in AR estimation methods by placing restrictions on the sparsity pattern of the inverse spectral density matrix. As we will see in section 1.3.1, the conditional independence constraints impose quadratic equality constraints on the AR parameters. The main contribution of the chapter is to show that under certain conditions the constrained estimation problem can be solved efficiently via a convex (semidefinite programming) relaxation. This convex formulations can be used to estimate graphical models where the AR parameters are constrained with respect to a given graph structure. In combination with model selection criteria they can also be used to identify the conditional independence structure of an AR process. In section 1.4 we present experimental results using randomly generated and real data sets.

Graphical models of AR processes have several applications; see [16, 17, 18, 19, 20, 21, 22]. Most previous work on this subject is concerned with statistical tests for topology selection. Dahlhaus [15] derives a statistical test for the existence of an edge in the graph, based on the maximum of a nonparametric estimate of the normalized inverse spectrum $S(\omega)^{-1}$; see [16, 17, 18, 19, 20, 21, 22] for applications of this approach. Eichler [23] presents a more general approach by introducing a hypothesis test based on the norm of some suitable function of the spectral density matrix. Related problems have also been studied in [24, 25]. Bach and Jordan [24] consider the problem of learning the structure of the graphical model of a time series from sample estimates of the joint spectral density matrix. Eichler [25] uses Whittle's approximation of the exact likelihood function, and imposes sparsity constraints on the inverse covariance functions via algorithms extended from covariance selection. Numerical algorithms for the estimation of graphical AR models have been explored in [22, 25, 26]. The convex framework proposed in this chapter provides an alternative and more direct approach and readily leads to efficient estimation algorithms.

Notation

$\mathbf{R}^{m\times n}$ denotes the set of real matrices of size $m \times n$, $\mathbf{S}^n$ is the set of real symmetric matrices of order n, and $\mathbf{M}^{n,p}$ is the set of matrices

$$X = \begin{bmatrix} X_0 & X_1 & \cdots & X_p \end{bmatrix}$$

with $X_0 \in \mathbf{S}^n$ and $X_1, \ldots, X_p \in \mathbf{R}^{n\times n}$. The standard trace inner product $\mathbf{tr}(X^T Y)$ is used on each of these three vector spaces. $\mathbf{S}^n_+$ ($\mathbf{S}^n_{++}$) is the set of symmetric positive semidefinite (positive definite) matrices of order n. X^H denotes the complex conjugate transpose of X.

The linear mapping $\mathrm{T} : \mathbf{M}^{n,p} \to \mathbf{S}^{n(p+1)}$ constructs a symmetric block Toeplitz matrix from its first block row: if $X \in \mathbf{M}^{n,p}$, then

$$\mathrm{T}(X) = \begin{bmatrix} X_0 & X_1 & \cdots & X_p \\ X_1^T & X_0 & \cdots & X_{p-1} \\ \vdots & \vdots & \ddots & \vdots \\ X_p^T & X_{p-1}^T & \cdots & X_0 \end{bmatrix}. \tag{1.5}$$

The adjoint of T is a mapping $\mathrm{D} : \mathbf{S}^{n(p+1)} \to \mathbf{M}^{n,p}$ defined as follows. If $S \in \mathbf{S}^{n(p+1)}$ is partitioned as

$$S = \begin{bmatrix} S_{00} & S_{01} & \cdots & S_{0p} \\ S_{10}^T & S_{11} & \cdots & S_{1p} \\ \vdots & \vdots & & \vdots \\ S_{p0}^T & S_{p1}^T & \cdots & S_{pp} \end{bmatrix},$$

then $\mathrm{D}(S) = \begin{bmatrix} \mathrm{D}_0(S) & \mathrm{D}_1(S) & \cdots & \mathrm{D}_p(S) \end{bmatrix}$ where

$$\mathrm{D}_0(S) = \sum_{i=0}^{p} S_{ii}, \qquad \mathrm{D}_k(S) = 2\sum_{i=0}^{p-k} S_{i,i+k}, \quad k = 1, \ldots, p. \tag{1.6}$$

A symmetric sparsity pattern of a sparse matrix X of order n will be defined by giving the set of indices $\mathcal{V} \subseteq \{1, \ldots, n\} \times \{1, \ldots, n\}$ of its zero entries. $\mathrm{P}_{\mathcal{V}}(X)$ denotes the projection of a matrix $X \in \mathbf{S}^n$ or $X \in \mathbf{R}^{n\times n}$ on the complement of the sparsity pattern $\mathcal{V}$:

$$\mathrm{P}_{\mathcal{V}}(X)_{ij} = \begin{cases} X_{ij} & (i,j) \in \mathcal{V} \\ 0 & \text{otherwise.} \end{cases} \tag{1.7}$$

The same notation will be used for $\mathrm{P}_{\mathcal{V}}$ 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, $\mathrm{P}_{\mathcal{V}}$ is self-adjoint. If X is a $p \times q$ block matrix with i, j block X_{ij}, and each block is square of order n, then $\mathrm{P}_{\mathcal{V}}(X)$ denotes the $p \times q$ block matrix with i, j block $\mathrm{P}_{\mathcal{V}}(X)_{ij} = \mathrm{P}_{\mathcal{V}}(X_{ij})$. The subscript of $\mathrm{P}_{\mathcal{V}}$ is omitted if the sparsity pattern $\mathcal{V}$ is clear from the context.

1.2 Autoregressive processes

This section provides some necessary background on AR processes and AR estimation methods. The material is standard and can be found in many textbooks [27, 28, 29, 30, 31].

We use the notation (1.4) for an AR model of order p. Occasionally the equivalent model

$$B_0 x(t) = -\sum_{k=1}^{p} B_k x(t-k) + v(t), \tag{1.8}$$

with $v(t) \sim N(0, I)$, will also be useful. The coefficients in the two models are related by $B_0 = \Sigma^{-1/2}$, $B_k = \Sigma^{-1/2} A_k$ for $k = 1, \ldots, p$.

The autocovariance sequence of the AR process is defined as

$$R_k = \mathbf{E}\, x(t+k)x(t)^T,$$

where $\mathbf{E}$ denotes the expected value. We have $R_{-k} = R_k^T$ since $x(t)$ is real. It is easily shown that the AR model parameters A_k, Σ, and the first $p+1$ covariance matrices R_k are related by the linear equations

$$\begin{bmatrix} R_0 & R_1 & \cdots & R_p \\ R_1^T & R_0 & \cdots & R_{p-1} \\ \vdots & \vdots & \ddots & \vdots \\ R_p^T & R_{p-1}^T & \cdots & R_0 \end{bmatrix} \begin{bmatrix} I \\ A_1^T \\ \vdots \\ A_p^T \end{bmatrix} = \begin{bmatrix} \Sigma \\ 0 \\ \vdots \\ 0 \end{bmatrix}. \tag{1.9}$$

These equations are called the *Yule-Walker equations* or *normal equations*.

The transfer function from w to x is $\mathbf{A}(z)^{-1}$ where

$$\mathbf{A}(z) = I + z^{-1}A_1 + \cdots + z^{-p}A_p.$$

The AR process is stationary if the poles of $\mathbf{A}$ are inside the unit circle. The spectral density matrix is defined as the Fourier transform of the autocovariance sequence,

$$S(\omega) = \sum_{k=-\infty}^{\infty} R_k e^{-\mathrm{j}k\omega}$$

(where $\mathrm{j} = \sqrt{-1}$), and can be expressed as $S(\omega) = \mathbf{A}(e^{\mathrm{j}\omega})^{-1}\Sigma\mathbf{A}(e^{\mathrm{j}\omega})^{-H}$. The inverse spectrum of an AR process is therefore a trigonometric matrix polynomial

$$S(\omega)^{-1} = \mathbf{A}(e^{\mathrm{j}\omega})^H \Sigma^{-1} \mathbf{A}(e^{\mathrm{j}\omega}) = Y_0 + \sum_{k=1}^{p} (e^{-\mathrm{j}k\omega} Y_k + e^{\mathrm{j}k\omega} Y_k^T) \tag{1.10}$$

where

$$Y_k = \sum_{i=0}^{p-k} A_i^T \Sigma^{-1} A_{i+k} = \sum_{i=0}^{p-k} B_i^T B_{i+k} \tag{1.11}$$

(with $A_0 = I$).

1.2.1 Least squares linear prediction

Suppose $x(t)$ is a stationary process (not necessarily autoregressive). Consider the problem of finding an optimal linear prediction

$$\hat{x}(t) = -\sum_{k=1}^{p} A_k x(t-k),$$

of $x(t)$, based on past values $x(t-1)$, ..., $x(t-p)$. This problem can also be interpreted as approximating the process $x(t)$ by the AR model with coefficients A_k. The prediction error between $x(t)$ and $\hat{x}(t)$ is

$$e(t) = x(t) - \hat{x}(t) = x(t) + \sum_{k=1}^{p} A_k x(t-k).$$

To find the coefficients A_1, ..., A_p, we can minimize the mean squared prediction error $\mathbf{E}\,\|e(t)\|_2^2$. The mean squared error can be expressed in terms of the coefficients A_k and the covariance function of x as $\mathbf{E}\,\|e(t)\|_2^2 = \mathbf{tr}(A\,\mathrm{T}(R)A^T)$ where

$$A = \begin{bmatrix} I & A_1 & \cdots & A_p \end{bmatrix}, \qquad R = \begin{bmatrix} R_0 & R_1 & \cdots & R_p \end{bmatrix},$$

$R_k = \mathbf{E}\,x(t+k)x(t)^T$, and $\mathrm{T}(R)$ is the block-Toeplitz matrix with R as its first block row (see the Notation section at the end of section 1.1). Minimizing the prediction error is therefore equivalent to the quadratic optimization problem

$$\text{minimize } \mathbf{tr}(A\,\mathrm{T}(R)A^T) \tag{1.12}$$

with variables A_1, ..., A_p.

In practice, the covariance matrix $\mathrm{T}(R)$ in (1.12) is replaced by an estimate C computed from samples of $x(t)$. Two common choices are as follows. Suppose samples $x(1)$, $x(2)$, ..., $x(N)$ are available.

- The *autocorrelation method* uses the *windowed* estimate

$$C = \frac{1}{N} HH^T, \tag{1.13}$$

where

$$H = \begin{bmatrix} x(1) & x(2) & \cdots & x(p+1) & \cdots & x(N) & 0 & \cdots & 0 \\ 0 & x(1) & \cdots & x(p) & \cdots & x(N-1) & x(N) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & x(1) & \cdots & x(N-p) & x(N-p+1) & \cdots & x(N) \end{bmatrix}. \tag{1.14}$$

  Note that the matrix C is block-Toeplitz, and that it is positive definite (unless the sequence $x(1)$, ..., $x(N)$ is identically zero).
- The *covariance method* uses the *non-windowed* estimate

$$C = \frac{1}{N-p} HH^T, \tag{1.15}$$

where

$$H = \begin{bmatrix} x(p+1) & x(p+2) & \cdots & x(N) \\ x(p) & x(p+1) & \cdots & x(N-1) \\ \vdots & \vdots & & \vdots \\ x(1) & x(2) & \cdots & x(N-p) \end{bmatrix}. \tag{1.16}$$

In this case the matrix C is not block-Toeplitz.

To summarize, least-squares estimation of AR models reduces to an unconstrained quadratic optimization problem

$$\text{minimize } \mathbf{tr}(ACA^T). \tag{1.17}$$

Here, C is the exact covariance matrix, if available, or one of the two sample estimates (1.13) and (1.15). The first of these estimates is a block-Toeplitz matrix, while the second one is in general not block-Toeplitz. The covariance method is known to be slightly more accurate in practice if N is small [31, page 94]. The correlation method on the other hand has some important theoretical and practical properties, that are easily explained from the optimality conditions of (1.17). If we define $\hat{\Sigma} = ACA^T$ (*i.e.*, the estimate of the prediction error $\mathbf{E}\,\|e(t)\|_2^2$ obtained by substituting C for $\mathrm{T}(R)$), then the optimality conditions can be expressed as

$$\begin{bmatrix} C_{00} & C_{01} & \cdots & C_{pp} \\ C_{10} & C_{11} & \cdots & C_{1p} \\ \vdots & \vdots & & \vdots \\ C_{p0} & C_{p1} & \cdots & C_{pp} \end{bmatrix} \begin{bmatrix} I \\ A_1^T \\ \vdots \\ A_p^T \end{bmatrix} = \begin{bmatrix} \hat{\Sigma} \\ 0 \\ \vdots \\ 0 \end{bmatrix}. \tag{1.18}$$

If C is block-Toeplitz, these equations have the same form as the Yule-Walker equations (1.9), and can be solved more efficiently than when C is not block-Toeplitz. Another advantage is that the solution of (1.18) always provides a stable model if C is block Toeplitz and positive definite. This can be proved as follows (see [32]). Suppose z is a zero of $\mathbf{A}(z)$, *i.e.*, there exists a nonzero w such that $w^H\mathbf{A}(z) = 0$. Define $u_1 = w$ and $u_k = A_{k-1}^T w + zu_{k-1}$ for $k = 2, \ldots, p$. Then we have

$$u = A^T w + z\tilde{u}$$

where $u = (u_1, u_2, \ldots, u_p, 0)$, $\tilde{u} = (0, u_1, u_2, \ldots, u_p)$. From this and (1.18),

$$u^H C u = w^H \hat{\Sigma} w + |z|^2 \tilde{u}^H C \tilde{u}.$$

The first term on the righthand side is positive because $\hat{\Sigma} \succ 0$. Also, $u^H C u = \tilde{u}^H C \tilde{u}$ since C is block-Toeplitz. Therefore $|z| < 1$.

In the following two sections we give alternative interpretations of the covariance and correlation variants of the least-squares estimation method, in terms of maximum likelihood and maximum entropy estimation, respectively.

1.2.2 Maximum likelihood estimation

The exact likelihood function of an AR model (1.4), based on observations $x(1)$, $\ldots$, $x(N)$, is complicated to derive and difficult to maximize [28, 33]. A standard simplification is to treat $x(1)$, $x(2)$, $\ldots$, $x(p)$ as fixed, and to define the likelihood function in terms of the conditional distribution of a sequence $x(t)$, $x(t+1)$, $\ldots$, $x(t+N-p-1)$, given $x(t-1)$, $\ldots$, $x(t-p)$. This is called the *conditional* maximum likelihood estimation method [33, §5.1].

The conditional likelihood function of the AR process (1.4) is

$$\begin{aligned}&\frac{1}{((2\pi)^n \det\Sigma)^{(N-p)/2}} \exp\left(-\frac{1}{2}\sum_{t=p+1}^{N} \mathbf{x}(t)^T A^T \Sigma^{-1} A \mathbf{x}(t)\right)\\ &= \left(\frac{\det B_0}{(2\pi)^{n/2}}\right)^{N-p} \exp\left(-\frac{1}{2}\sum_{t=p+1}^{N} \mathbf{x}(t)^T B^T B \mathbf{x}(t)\right) \qquad (1.19)\end{aligned}$$

where $\mathbf{x}(t)$ is the $((p+1)n)$-vector $\mathbf{x}(t) = (x(t), x(t-1), \ldots, x(t-p))$ and

$$A = \begin{bmatrix} I & A_1 & \cdots & A_p \end{bmatrix}, \qquad B = \begin{bmatrix} B_0 & B_1 & \cdots & B_p \end{bmatrix},$$

with $B_0 = \Sigma^{-1/2}$, $B_k = \Sigma^{-1/2} A_k$, $k = 1, \ldots, p$. Taking the logarithm of (1.19) we obtain the conditional log-likelihood function (up to constant terms and factors)

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

where H is the matrix (1.16). If we define $C = (1/(N-p))HH^T$, we can then write the conditional ML estimation problem as

$$\text{minimize } -2\log\det B_0 + \mathbf{tr}(CB^TB) \qquad (1.20)$$

with variable $B \in \mathbf{M}^{n,p}$. This problem is easily solved by setting the gradient equal to zero: the optimal B satisfies $CB^T = (B_0^{-1}, 0, \ldots, 0)$. Written in terms of the model parameters $A_k = B_0^{-1}B_k$, $\Sigma = B_0^{-2}$, this yields

$$C\begin{bmatrix} I \\ A_1^T \\ \vdots \\ A_p^T \end{bmatrix} = \begin{bmatrix} \Sigma \\ 0 \\ \vdots \\ 0 \end{bmatrix},$$

i.e., the Yule-Walker equations with the block Toeplitz coefficient matrix replaced by C. The conditional ML estimate is therefore equal to the least-squares estimate from the covariance method.

1.2.3 Maximum entropy estimation

Consider the maximum entropy (ME) problem introduced by Burg [34]:

$$\begin{aligned}&\text{maximize } \tfrac{1}{2\pi}\textstyle\int_{-\pi}^{\pi} \log\det S(\omega)d\omega\\ &\text{subject to } \tfrac{1}{2\pi}\textstyle\int_{-\pi}^{\pi} S(\omega)e^{\mathrm{j}k\omega}d\omega = \bar{R}_k, \quad 0 \le k \le p.\end{aligned} \qquad (1.21)$$

The matrices $\bar{R}_k$ are given. The variable is the spectral density $S(\omega)$ of a real stationary Gaussian process $x(t)$, *i.e.*, the Fourier transform of the covariance function $R_k = \mathbf{E}\, x(t+k)x(t)^T$:

$$S(\omega) = R_0 + \sum_{k=0}^{\infty} \left(R_k e^{-\mathrm{j}k\omega} + R_k^T e^{\mathrm{j}k\omega} \right), \qquad R_k = \frac{1}{2\pi} \int_{-\pi}^{\pi} S(\omega) e^{\mathrm{j}k\omega} d\omega.$$

The constraints in (1.21) therefore fix the first $p+1$ covariance matrices to be equal to $\bar{R}_k$. The problem is to extend these covariances so that the entropy rate of the process is maximized. It is known that the solution of (1.21) is a Gaussian AR process of order p, and that the model parameters A_k, Σ follow from the Yule-Walker equations (1.9) with $\bar{R}_k$ substituted for R_k.

To relate the ME problem to the estimation methods of the preceding sections, we derive a dual problem. To simplify the notation later on, we multiply the two sides of the equality constraints $k = 1, \ldots, p$ by 2. We introduce a Lagrange multiplier $Y_0 \in \mathbf{S}^n$ for the first equality constraint ($k = 0$), and multipliers $Y_k \in \mathbf{R}^{n\times n}$, $k = 1, \ldots, p$, for the other p equality constraints. If we change the sign of the objective, the Lagrangian is

$$-\frac{1}{2\pi} \int_{-\pi}^{\pi} \log\det S(\omega) d\omega + \mathbf{tr}(Y_0(R_0 - \bar{R}_0) + 2\sum_{k=1}^{p} \mathbf{tr}(Y_k^T(R_k - \bar{R}_k)).$$

Differentiating with respect to R_k gives

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} S^{-1}(\omega) e^{\mathrm{j}\omega k} d\omega = Y_k, \quad 0 \le k \le p \tag{1.22}$$

and hence

$$S^{-1}(\omega) = Y_0 + \sum_{k=1}^{p} \left(Y_k e^{-\mathrm{j}k\omega} + Y_k^T e^{\mathrm{j}k\omega} \right) \triangleq Y(\omega).$$

Substituting this in the Lagrangian gives the dual problem

$$\text{minimize } -\frac{1}{2\pi} \int_{-\pi}^{\pi} \log\det Y(\omega) + \mathbf{tr}(Y_0^T \bar{R}_0) + 2\sum_{k=1}^{p} \mathbf{tr}(Y_k^T \bar{R}_k) - n, \tag{1.23}$$

with variables Y_k. The first term in the objective can be rewritten by using Kolmogorov's formula [35]:

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} \log\det Y(\omega) d\omega = \log\det(B_0^T B_0),$$

where $Y(\omega) = \mathbf{B}(e^{\mathrm{j}\omega})^H \mathbf{B}(e^{\mathrm{j}\omega})$ and $\mathbf{B}(z) = \sum_{k=0}^{p} z^{-k} B_k$ is the minimum-phase spectral factor of Y. The second term in the objective of the dual problem (1.23) can also be expressed in terms of the coefficients B_k, using the relations $Y_k = \sum_{i=0}^{p-k} B_i^T B_{i+k}$ for $0 \le k \le p$. This gives

$$\mathbf{tr}(Y_0 \bar{R}_0) + 2\sum_{k=1}^{p} \mathbf{tr}(Y_k^T \bar{R}_k) = \mathbf{tr}(\mathrm{T}(\bar{R}) B^T B),$$

where $\bar{R} = \begin{bmatrix} \bar{R}_0 & \bar{R}_1 & \cdots & \bar{R}_p \end{bmatrix}$ and $B = \begin{bmatrix} B_0 & B_1 & \cdots & B_p \end{bmatrix}$. The dual problem (1.23) thus reduces to

$$\text{minimize } -2 \log \det B_0 + \mathbf{tr}(CB^T B) \tag{1.24}$$

where $C = \mathrm{T}(\bar{R})$. Without loss of generality, we can choose B_0 to be symmetric positive definite. The problem is then formally the same as the ML estimation problem (1.20), except for the definition of C. In (1.24) C is a block-Toeplitz matrix. If we choose for $\bar{R}_k$ the sample estimates

$$\bar{R}_k = \frac{1}{N} \sum_{t=1}^{N-k} x(t+k)x(t)^T,$$

then C is identical to the block-Toeplitz matrix (1.13) used in the autocorrelation variant of the least-squares method.

1.3 Autoregressive graphical models

In this section we first characterize conditional independence relations in multivariate Gaussian processes, and specialize the definition to AR processes. We then add the conditional independence constraints to the ML and ME estimation problems derived in the previous section, and investigate convex optimization techniques for solving the modified estimation problems.

1.3.1 Conditional independence in time series

Let $x(t)$ be an n-dimensonal stationary zero-mean Gaussian process with spectrum $S(\omega)$:

$$S(\omega) = \sum_{k=-\infty}^{\infty} R_k e^{-\mathrm{j}k\omega}, \qquad R_k = \mathbf{E}\, x(t+k)x(t)^T.$$

We assume that S is invertible for all ω. Components $x_i(t)$ and $x_j(t)$ are said to be independent, conditional on the other components of $x(t)$, if

$$(S(\omega)^{-1})_{ij} = 0$$

for all ω. This definition can be interpreted and justified as follows (see Brillinger [36, §8.1]). Let $u(t) = (x_i(t), x_j(t))$ and let $v(t)$ be the $(n-2)$-vector containing the remaining components of $x(t)$. Define $e(t)$ as the error

$$e(t) = u(t) - \sum_{k=-\infty}^{\infty} H_k v(t-k)$$

between $u(t)$ and the linear filter of $v(t)$ that minimizes $\mathbf{E}\,\|e(t)\|_2^2$. Then it can be shown that the spectrum of the error process $e(t)$ is

$$\begin{bmatrix} (S(\omega)^{-1})_{ii} & (S(\omega)^{-1})_{ij} \\ (S(\omega)^{-1})_{ji} & (S(\omega)^{-1})_{jj} \end{bmatrix}^{-1}. \tag{1.25}$$

This is the Schur complement of the submatrix in $S(\omega)$ indexed by $\{1,\ldots,n\} \setminus \{i,j\}$. The off-diagonal entry in the error spectrum (1.25) is called the *partial cross-spectrum* of x_i and x_j, after removing the effects of v. The partial cross-spectrum is zero if and only if the error covariances $\mathbf{E}\,e(t+k)e(t)^T$ are diagonal, *i.e.*, the two components of the error process $e(t)$ are independent.

We can apply this to an AR process (1.4) using the relation between the inverse spectrum $S(\omega)$ and the AR coefficients given in (1.10) and (1.11). These expressions show that $(S(\omega)^{-1})_{ij} = 0$ if and only if the i,j entries of Y_k are zero for $k = 0,\ldots,p$, where Y_k is given in (1.11). Using the notation defined in (1.6), we can write this as $(\mathrm{D}_k(A^T\Sigma^{-1}A))_{ij} = 0$, where $A = \begin{bmatrix} I & A_1 & \cdots & A_p \end{bmatrix}$, or as

$$\big(\mathrm{D}_k(B^TB)\big)_{ij} = 0, \qquad k = 0,\ldots,p, \tag{1.26}$$

where $B = \begin{bmatrix} B_0 & B_1 & \cdots & B_p \end{bmatrix}$.

1.3.2 Maximum likelihood and maximum entropy estimation

We now return to the ML and ME estimation methods for AR processes, described in sections 1.2.2 and 1.2.3, and extend the methods to include conditional independence constraints. As we have seen, the ML and ME estimation problems can be expressed as a convex optimization problem (1.20) and (1.24), with different choices of the matrix C. The distinction will turn out to be important later, but for now we make no assumptions on C, except that it is positive definite.

As for the Gaussian graphical models mentioned in the introduction, we assume that the conditional independence constraints are specified via an index set $\mathcal{V}$, with $(i,j) \in \mathcal{V}$ if the processes $x_i(t)$ and $x_j(t)$ are conditionally independent. We write the constraints (1.26) for $(i,j) \in \mathcal{V}$ as

$$\mathrm{P}_{\mathcal{V}}\left(\mathrm{D}(B^TB)\right) = 0,$$

where $\mathrm{P}_{\mathcal{V}}$ is the projection operator defined in (1.7). We assume that $\mathcal{V}$ does not contain the diagonal entries (i,i) and that it is symmetric (if $(i,j) \in \mathcal{V}$, then $(j,i) \in \mathcal{V}$). The ML and ME estimation with conditional independence constraints can therefore be expressed as

$$\begin{array}{ll} \text{minimize} & -2\log\det B_0 + \mathbf{tr}(CB^TB) \\ \text{subject to} & \mathrm{P}(\mathrm{D}(B^TB)) = 0. \end{array} \tag{1.27}$$

(Henceforth we drop the subscript of $\mathrm{P}_{\mathcal{V}}$.) The variable is $B = \begin{bmatrix} B_0 & B_1 & \cdots & B_p \end{bmatrix} \in \mathbf{M}^{n,p}$.

The problem (1.27) includes quadratic equality constraints and is therefore nonconvex. The quadratic terms in B suggest the convex relaxation

$$\begin{array}{ll} \text{minimize} & -\log\det X_{00} + \mathbf{tr}(CX) \\ \text{subject to} & \mathrm{P}(\mathrm{D}(X)) = 0 \\ & X \succeq 0 \end{array} \tag{1.28}$$

with variable $X \in \mathbf{S}^{n(p+1)}$ (X_{00} denotes the leading $n \times n$ subblock of X). The convex optimization problem (1.28) is a relaxation of (1.27) and only equivalent to (1.27) if the optimal solution X has rank n, so that it can be factored as $X = B^T B$. We will see later that this is the case if C is block-Toeplitz.

The proof of exactness of the relaxation under assumption of block-Toeplitz structure will follow from the dual of (1.28). We introduce a Lagrange multiplier $Z = \begin{bmatrix} Z_0 & Z_1 & \cdots & Z_p \end{bmatrix} \in \mathbf{M}^{n,p}$ for the equality constraints and a multiplier $U \in \mathbf{S}^{n(p+1)}$ for the inequality constraint. The Lagrangian is

$$\begin{aligned} L(X, Z, U) &= -\log\det X_{00} + \mathbf{tr}(CX) + \mathbf{tr}(Z^T \mathrm{P}(\mathrm{D}(X))) - \mathbf{tr}(UX) \\ &= -\log\det X_{00} + \mathbf{tr}\left((C + \mathrm{T}(\mathrm{P}(Z)) - U)X\right). \end{aligned}$$

Here we made use of the fact that the mappings T and D are adjoints, and that P is self-adjoint. The dual function is the infimum of L over all X with $X_{00} \succ 0$. Setting the gradient with respect to X equal to zero gives

$$C + \mathrm{T}(\mathrm{P}(Z)) - U = \begin{bmatrix} X_{00}^{-1} & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix}.$$

This shows that Z, U are dual feasible if $C + \mathrm{T}(\mathrm{P}(Z)) - U$ is zero, except for the $0, 0$ block, which must be positive definite. If U and Z satisfy these conditions, the Lagrangian is minimized by any X with $X_{00} = (C_{00} + \mathrm{P}(Z_0) - U_{00})^{-1}$ (where C_{00} and U_{00} denote the leading $n \times n$ blocks of C and U). Hence we arrive at the dual problem

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

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

$$\begin{array}{ll} \text{maximize} & \log\det W + n \\ \text{subject to} & \begin{bmatrix} W & 0 \\ 0 & 0 \end{bmatrix} \preceq C + \mathrm{T}(\mathrm{P}(Z)). \end{array} \tag{1.29}$$

Note that for $p = 0$ problem (1.28) reduces to the covariance selection problem (1.2), and the dual problem reduces to the maximum determinant comple-

tion problem

$$\text{maximize } \log\det(C + \mathrm{P}(Z)) + n,$$

which is equivalent to (1.3).

We note the following properties of the primal problem (1.28) and the dual problem (1.29).

- 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 $C \succ 0$, 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 $C \succ 0$, we have strong duality and the primal and dual optimal values are attained. The Karush-Kuhn-Tucker (KKT) conditions are therefore necessary and sufficient for optimality of X, Z, W. The KKT conditions are:

1. *Primal feasibility.*

$$X \succeq 0, \qquad X_{00} \succ 0, \qquad \mathrm{P}(\mathrm{D}(X)) = 0, \tag{1.30}$$

2. *Dual feasibility.*

$$W \succ 0, \qquad C + \mathrm{T}(\mathrm{P}(Z)) \succeq \begin{bmatrix} W & 0 \\ 0 & 0 \end{bmatrix}. \tag{1.31}$$

3. *Zero duality gap.*

$$X_{00}^{-1} = W, \qquad \mathbf{tr}\left(X\left(C + \mathrm{T}(\mathrm{P}(Z)) - \begin{bmatrix} W & 0 \\ 0 & 0 \end{bmatrix}\right)\right) = 0. \tag{1.32}$$

The last condition can also be written as

$$X\left(C + \mathrm{T}(\mathrm{P}(Z)) - \begin{bmatrix} W & 0 \\ 0 & 0 \end{bmatrix}\right) = 0. \tag{1.33}$$

1.3.3 Properties of block-Toeplitz sample covariances

In this section we study in more detail the solution of the primal and dual problems (1.28) and (1.29) if C is block-Toeplitz. The results can be derived from connections between spectral factorization, semidefinite programming, and orthogonal matrix polynomials discussed in [37, §6.1.1]. In this section, we provide alternative and self-contained proofs.

Assume $C = \mathrm{T}(R)$ for some $R \in \mathbf{M}^{n,p}$ and that C is positive definite.

Exactness of the relaxation

We first show that the relaxation (1.28) is exact when C is block-Toeplitz, *i.e.*, the optimal X^* has rank n and the optimal B can be computed by factoring X^* as $X^* = B^T B$. We prove this result from the optimality conditions (1.30)–(1.33).

Assume X^*, W^*, Z^* are optimal. Clearly $\mathbf{rank}\, X^* \geq n$, since its $0,0$ block is nonsingular. We will show that $C + \mathrm{T}(\mathrm{P}(Z^*)) \succ 0$. Therefore the rank of

$$C + \mathrm{T}(\mathrm{P}(Z^*)) - \begin{bmatrix} W^* & 0 \\ 0 & 0 \end{bmatrix}$$

is at least np, and the complementary slackness condition (1.33) implies that X^* has rank at most n, so we can conclude that

$$\mathbf{rank}\, X^* = n.$$

The positive definitess of $C + \mathrm{T}(\mathrm{P}(Z^*))$ follows from the dual feasibility condition (1.31) and the following basic property of block-Toeplitz matrices: If $\mathrm{T}(S)$ is a symmetric block-Toeplitz matrix, with $S \in \mathbf{M}^{n,p}$, and

$$\mathrm{T}(S) \succeq \begin{bmatrix} Q & 0 \\ 0 & 0 \end{bmatrix} \tag{1.34}$$

for some $Q \in \mathbf{S}^n_{++}$, then $\mathrm{T}(S) \succ 0$. We can verify this by induction on p. The property is obviously true for $p = 0$, since the inequality (1.34) then reduces to $S = S_0 \succeq Q$. Suppose the property holds for $p - 1$. Then (1.34) implies that the leading $np \times np$ submatrix of $\mathrm{T}(S)$, which is a block Toeplitz matrix with first row $\begin{bmatrix} S_0 & \cdots & S_{p-1} \end{bmatrix}$, is positive definite. Let us denote this matrix by V. Using the Toeplitz structure, we can partition $T(S)$ as

$$\mathrm{T}(S) = \begin{bmatrix} S_0 & U^T \\ U & V \end{bmatrix},$$

where $V \succ 0$. The inequality (1.34) implies that the Schur complement of V in the matrix $\mathrm{T}(S)$ satisfies

$$S_0 - U^T V^{-1} U \succeq Q \succ 0$$

Combined with $V \succ 0$ this shows that $\mathrm{T}(S) \succ 0$.

Stability of estimated models

It follows from (1.30)–(1.33) and the factorization $X^* = B^T B$, that

$$(C + \mathrm{T}(\mathrm{P}(Z))) \begin{bmatrix} I \\ A_1^T \\ \vdots \\ A_p^T \end{bmatrix} = \begin{bmatrix} \Sigma \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \tag{1.35}$$

if we define $\Sigma = B_0^{-2}$, $A_k = B_0^{-1} B_k$. These equations are Yule-Walker equations with a positive definite block-Toeplitz coefficient matrix. As mentioned at the

end of section 1.2.1, this imples that the zeros of $\mathbf{A}(z) = I + z^{-1}A_1 + \cdots + z^{-p}A_p$ are inside the unit circle. Therefore the solution to the convex problem (1.28) provides a stable AR model.

1.3.4 Summary

We have proposed convex relaxations for the problems of conditional ML and ME estimation of AR models with conditional independent constraints. The two problems have the same form with different choices for the sample covariance matrix C. For the ME problem, C is given by (1.13), while for the conditional ML problem, it is given by (1.15). In both cases, C is positive definite if the information matrix H has full rank. This is sufficient to guarantee that the relaxed problem (1.28) is bouned below.

The relaxation is exact if the matrix C is block-Toeplitz, *i.e.*, for the ME problem. The Toeplitz structure also ensures stability of the estimated AR model. In the conditional ML problem, C is in general not block-Toeplitz, but approaches a block-Toeplitz matrix as N goes to infinity. We conjecture that the relaxation of the ML problem is exact with high probability even for moderate values of N. This will be illustrated by the experimental results in the next section.

1.4 Numerical examples

In this section we evaluate the ML and ME estimation methods on several data sets. The convex optimization package `CVX` [38, 39] was used to solve the ML and ME estimation problems.

1.4.1 Randomly generated data

The first set of experiments uses data randomly generated from AR models with sparse inverse spectra. The purpose is to examine the quality of the semidefinite relaxation (1.28) of the ML estimation problem for finite N. We generated 50 sets of time series from four AR models of different dimensions. We solved (1.28) for different N. Figure 1.1 shows the percentage of the 50 data sets for which the relaxation was exact (the optimal X in (1.28) had rank n.) The results illustrate that the relaxation is often exact for moderate values of N, even when the matrix C is not block-Toeplitz.

The next figure shows the convergence rate of the ML and ME estimates, with and without imposed conditional independence constraints, to the true model, as a function of the number of samples. The data were generated from an AR model of dimension $n = p = 6$ with nine zeros in the inverse spectrum. Figure 1.2 shows the Kullback-Leibler (KL) divergence [24] between the estimated and the true spectra as a function of N, for four estimation methods: the ML and ME estima-

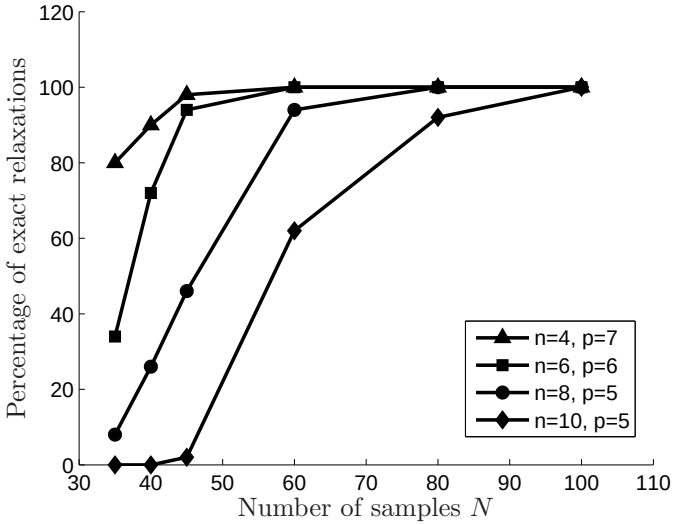

Figure 1.1 Number of cases where the convex relaxation of the ML problem is exact, versus the number of samples.

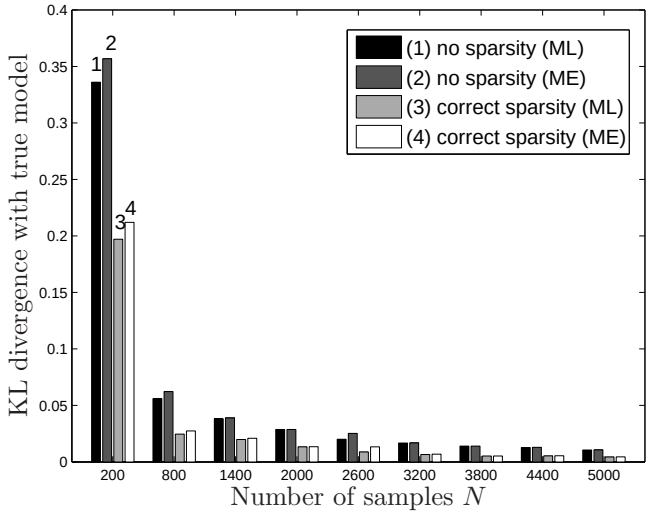

Figure 1.2 KL divergence between estimated AR models and the true model ($n = 6$, $p = 6$) versus the number of samples.

tion methods without conditional independence constraints, and the ML and ME estimation methods with the correct conditional independence constraints. We notice that the KL divergences decrease at the same rate for the four estimates. However, the ML and ME estimates without the sparsity constraints give models

with substantially larger values of KL divergence when N is small. For sample size under 3000, the ME estimates (with and without the sparsity constraints) are also found to be less accurate than their ML counterparts. This effect is well known in spectral analysis (see, for example, [31, page 94]). As N increases, the difference between the ME and ML methods disappears.

1.4.2 Model selection

The next experiment is concerned with the problem of topology selection in graphical AR models.

Three popular model selection criteria are the *Akaike Information Criterion* (AIC), the second-order variant of AIC (AIC_c), and the *Bayes information criterion* (BIC) [40]. These criteria are used to make a fair comparison between models of different complexity. They assign to an estimated model a score equal to $-2L$, where L is the likelihood of the model, augmented with a term that depends on the effective number of parameters k in the model:

$$\mathrm{AIC} = -2L + 2k, \qquad \mathrm{AIC}_c = -2L + \frac{2kN}{N-k-1}, \qquad \mathrm{BIC} = -2L + k\log N.$$

The second term places a penalty on models with high complexity. When comparing different models, we rank them according to one of the criteria and select the model with the lowest score. Of these three criteria, the AIC is known to perform poorly if N is small compared to the number of parameters k. The AIC_c was developed as a correction to the AIC for small N. For large N the BIC favors simpler models than the AIC or AIC_c.

To select a suitable graphical AR model for observed samples of an n-dimensional time series, we can enumerate models of different lengths p and with different graphs. For each model, we solve the ML estimation problem, calculate the AIC, AIC_c, or BIC score, and select the model with the best (lowest) score. Obviously, an exhaustive search of all sparsity patterns is only feasible for small n (say, $n \leq 6$), since there are

$$\sum_{m=0}^{n(n-1)/2} \binom{n(n-1)/2}{m} = 2^{n(n-1)/2} \tag{1.36}$$

different graphs with n nodes.

In the experiment we generate $N = 1000$ samples from an AR model of dimension $n = 5$, $p = 4$, and zeros in positions $(1,2)$, $(1,3)$, $(1,4)$, $(2,4)$, $(2,5)$, $(4,5)$ of the inverse spectrum. We show only results for the BIC. In the BIC we substitute the conditional likelihood discussed in section 1.2.2 for the exact likelihood L. (For sufficiently large N the difference is negligible.) As effective number of parameters we take

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

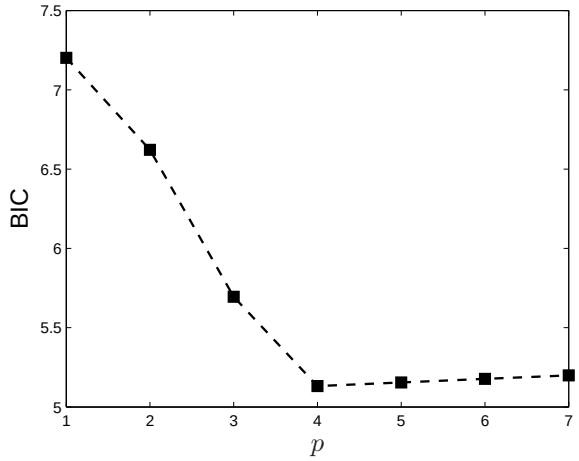

Figure 1.3 BIC score scaled by $1/N$ of AR models of order p.

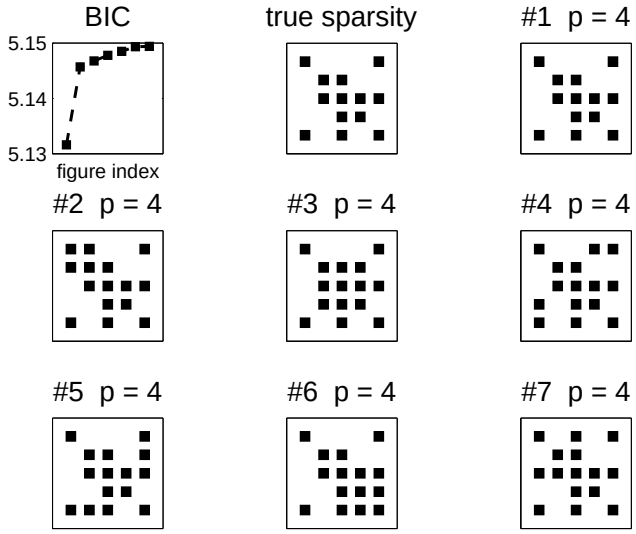

Figure 1.4 Seven best ranked topologies according to the BIC.

where $|\mathcal{V}|$ is the number of conditional independence constraints, *i.e.*, the number of zeros in the lower triangular part of the inverse spectrum.

Figure 1.3 shows the scores of the estimated models as a function of p. For each p the score shown is the best score among all graph topologies. The BIC selects the correct model order $p = 4$. Figure 1.4 shows the seven best models according to the BIC. The subgraphs labeled #1 to #7 show the estimated model order

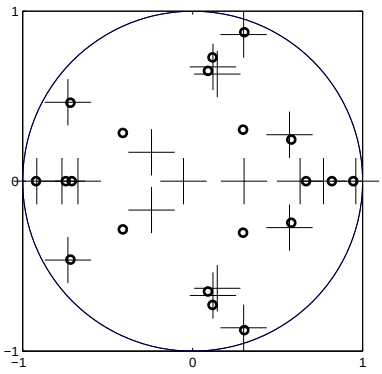

Figure 1.5 Poles of the true model (plus signs) and the estimated model (circles).

p, and the selected sparsity pattern. The corresponding scores are shown in the first subgraph, and the true sparsity pattern is shown in the second subgraph. The BIC identified the correct sparsity pattern. Figure 1.5 shows the location of the poles of the true AR model and the model selected by the BIC.

In figures 1.6 and 1.7 we compare the spectrum of the model selected by the BIC with the spectrum of the true model and with a nonparametric estimate of the spectrum. The lower half of the figures show the *coherence spectrum*, *i.e.*, the spectrum normalized to have diagonal one:

$$\mathbf{diag}(S(\omega))^{-1/2} S(\omega)\, \mathbf{diag}(S(\omega))^{-1/2},$$

where $\mathbf{diag}(S)$ is the diagonal part of S. The upper half shows the *partial coherence spectrum*, *i.e.*, the inverse spectrum normalized to have diagonal one:

$$\mathbf{diag}(S(\omega)^{-1})^{-1/2} S(\omega)^{-1}\, \mathbf{diag}(S(\omega)^{-1})^{-1/2}.$$

The i, j entry of the coherence spectrum is a measure of how dependent components i and j of the time series are. The i, j entry of the partial coherence spectrum on the other hand is a measure of *conditional* dependence. The dashed lines show the spectra of the true model. The solid lines in figure 1.6 are the spectra of the ML estimates. The solid lines in figure 1.7 are nonparametric estimates of the spectrum, obtained with Welch's method (see [41, §12.2.2]) using a Hamming window of length 40 (see [41, page 642]). The nonparametric estimate of the partial coherence spectrum clearly gives a poor indication of the correct sparsity pattern.

1.4.3 Air pollution data

The data set used in this section consists of a time series of dimension $n = 5$. The components are four air pollutants, CO, NO, NO_2, O_3, and the solar radiation intensity R, recorded hourly during 2006 at Azusa, Cal-

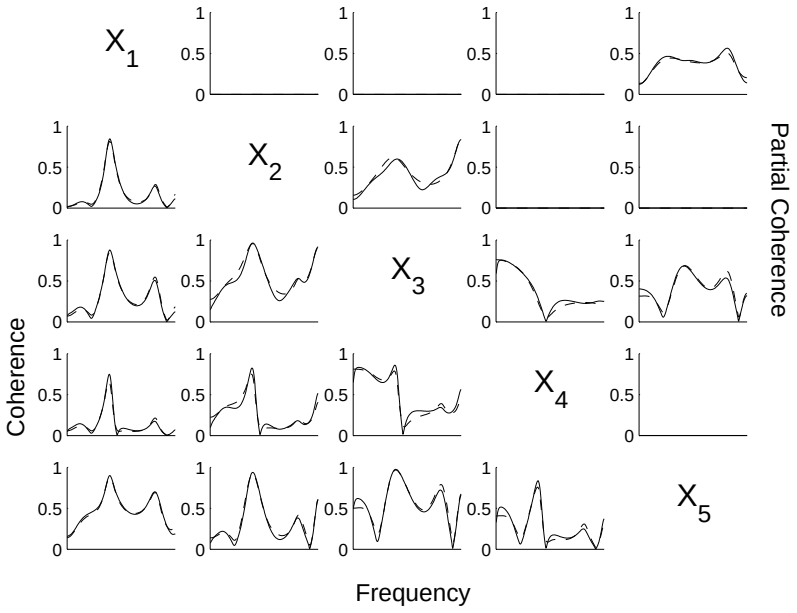

Figure 1.6 Partial coherence and coherence spectra of the AR model: true spectrum (dashed lines) and ML estimates (solid lines).

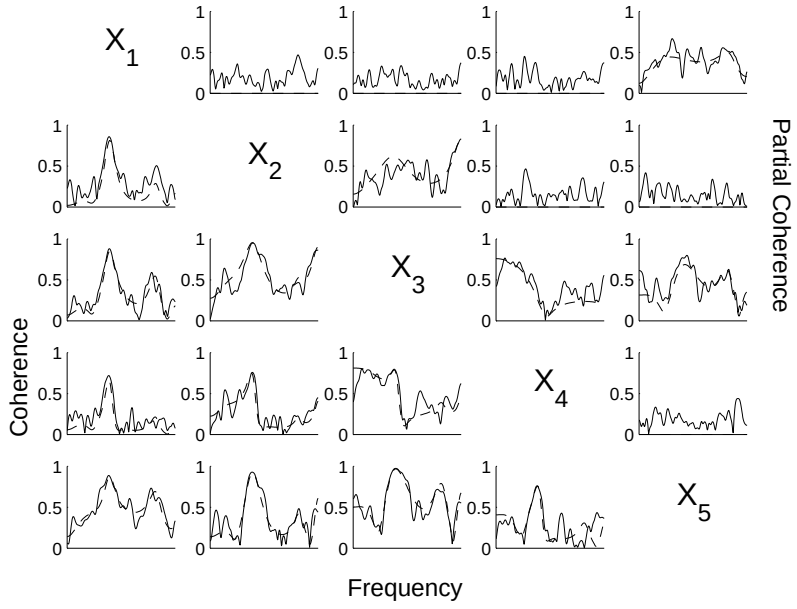

Figure 1.7 Partial coherence and coherence spectra of the AR model: true spectrum (dashed lines) and nonparametric estimates (solid line).

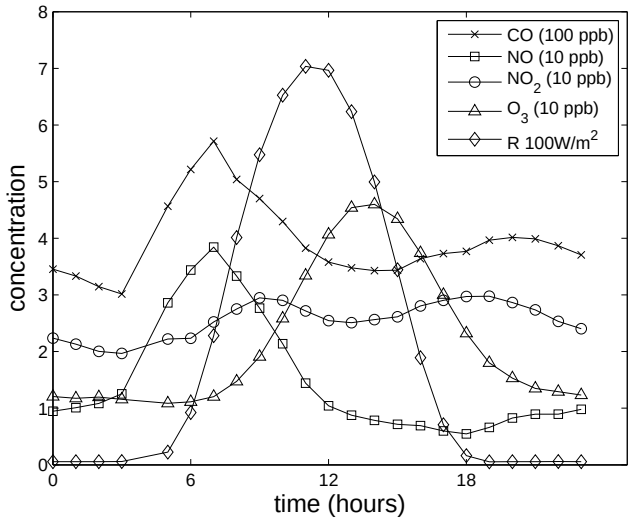

Figure 1.8 Average of daily concentration of CO, NO, NO_2, and O_3, and the solar radiation (R).

Rank	p	BIC score	$\mathcal{V}$
1	4	15414	$(\mathrm{NO}, \mathrm{R})$
2	5	15455	$(\mathrm{NO}, \mathrm{R})$
3	4	15461	
4	4	15494	$(\mathrm{CO}, \mathrm{O_3})$, $(\mathrm{CO}, \mathrm{R})$
5	4	15502	$(\mathrm{CO}, \mathrm{R})$
6	5	15509	$(\mathrm{CO}, \mathrm{O_3})$, $(\mathrm{CO}, \mathrm{R})$
7	5	15512	
8	4	15527	$(\mathrm{CO}, \mathrm{O_3})$
9	6	15532	$(\mathrm{NO}, \mathrm{R})$
10	5	15544	$(\mathrm{CO}, \mathrm{R})$

Table 1.1. Models with the lowest BIC scores for the air pollution data, determined by an exhaustive search of all models of orders $p = 1, \ldots, 8$. $\mathcal{V}$ is the set of conditionally independent pairs in the model.

ifornia. The entire data set consists of $N = 8370$ observations, and was obtained from Air Quality and Meteorological Information System (AQMIS) (`www.arb.ca.gov/aqd/aqdcd/aqdcd.htm`). The daily averages over one year are shown in figure 1.8. A similar data set was studied previously in [15], using a nonparametric approach.

We use the BIC to compare models with orders ranging from $p = 1$ to $p = 8$. Table 1.1 lists the models with the best ten BIC scores (which differ by only 0.84%). Figure 1.9 shows the coherence and partial coherence spectra obtained

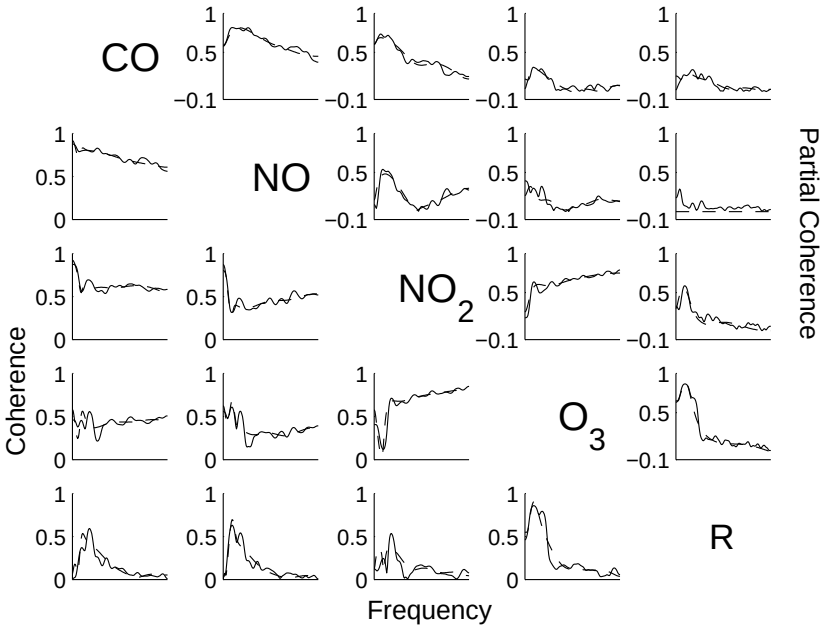

Figure 1.9 Coherence (lower half) and partial coherence spectra (upper half) for the first model in table 1.1. Nonparametric estimates are in solid lines, and ML estimates in dashed lines.

from a nonparametric estimation (solid lines), and the ML model with the best BIC score (dashed lines).

From table 1.1, 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 figure 1.9 where the pair NO-R is weakest. Table 1.1 also suggests that other weak links are $(\mathrm{CO}, \mathrm{O}_3)$ and $(\mathrm{CO}, \mathrm{R})$. The partial coherence spectra of these pairs are not identically zero, but are relatively small compared to the other pairs.

The presence of the stronger components in the partial coherence spectra are consistent with the discussion in [15]. For example, the solar radiation plays a role in the photolysis of NO_2 and the generation of O_3. The concentration of CO and NO are highly correlated because both are generated by traffic.

1.4.4 International stock markets

We consider a multivariate time series of five stock market indices: the S&P 500 composite index (U.S.), Nikkei 225 share index (Japan), the Hang Seng stock composite index (Hong Kong), the FTSE 100 share index (United Kingdom), and the Frankfurt DAX 30 composite index (Germany). The data were

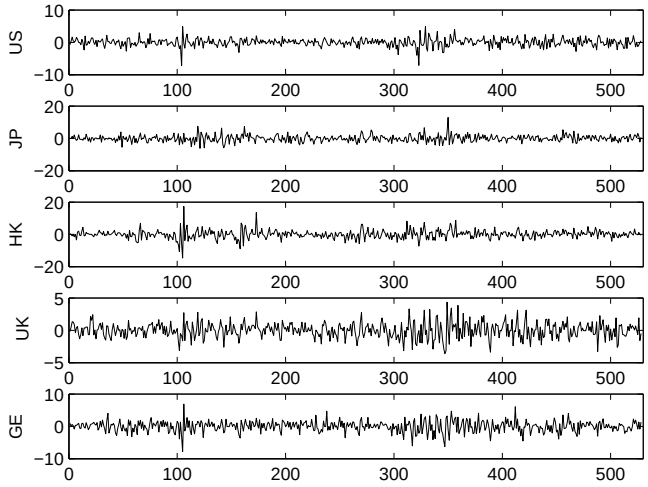

Figure 1.10 Detrended daily returns for five stock market indices between June 4, 1997 and June 15, 1999.

recorded from June 4, 1997 to June 15, 1999, and were downloaded from `www.globalfinancial.com`. (The data were converted to US dollars to take the volatility of exchange rates into account. We also replaced missing data due to national holidays by the most recent values.) For each market we use as variable the return between trading day $k-1$ and k, defined as

$$r_k = 100 \log(p_k/p_{k-1}), \tag{1.37}$$

where p_k is the closing price on day k. The resulting five-dimensional time series of length 528 is shown in figure 1.10. This data set is a subset of the data set used in [42].

We enumerate all graphical models of orders ranging from $p = 1$ to $p = 9$. Because of the relatively small number of samples, the AIC_c criterion will be used to compare the models. Figure 1.11 shows the optimal AIC_c (optimized over all models of a given lag p) versus p. Table 1.2 shows the model order and topology of the five models with the best AIC_c scores. The column labeled $\mathcal{V}$ shows the list of conditionally independent pairs of variables.

Figure 1.12 shows the coherence (bottom half) and partial coherence (upper half) spectra for the model selected by the AIC_c, and for a nonparametric estimate.

It is interesting to compare the results with the conclusions in [42]. For example, the authors of [42] mention a strong connection between the German and the other European stock markets, in particular, the UK. This agrees with the high value of the UK-GE component of the partial coherence spectrum in figure 1.12. The lower strength of the connections between the Japanese and the other stock

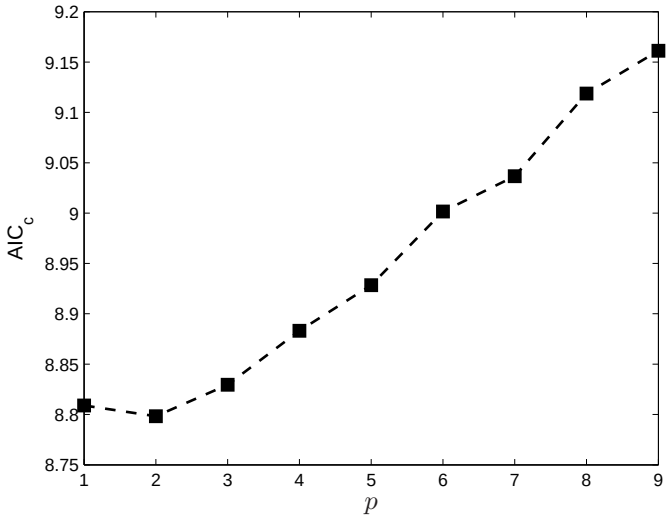

Figure 1.11 Minimized AIC_c scores (scaled by $1/N$) of pth-order models for the stock market return data.

Rank	p	AIC_c score	$\mathcal{V}$
1	2	4645.5	(US,JP), (JP,GE)
2	2	4648.0	(US,JP)
3	1	4651.1	(US,JP), (JP,GE)
4	1	4651.6	(US,JP)
5	2	4653.1	(JP,GE)

Table 1.2. Five best AR models, ranked according to AIC_c scores, for the international stock market data.

markets is also consistent with the findings in [42]. Another conclusion from [42] is that the volatility in the US stock markets transmits to the world through the German and Hong Kong markets. As far as the German market is concerned, this seems to be confirmed by the strength of the US-GE component in the partial coherence spectrum.

1.4.5 European stock markets

This data set is similar to the previous one. We consider a five-dimensional time series consisting of the following stock market indices: the FTSE 100 share index (United Kingdom), CAC 40 (France), the Frankfurt DAX 30 composite index (Germany), MIBTEL (Italy), Austrian Traded Index ATX (Austria). The data were stock index closing prices recorded from January 1, 1999 to July 31, 2008, and obtained from `www.globalfinancial.com`. The stock market daily returns

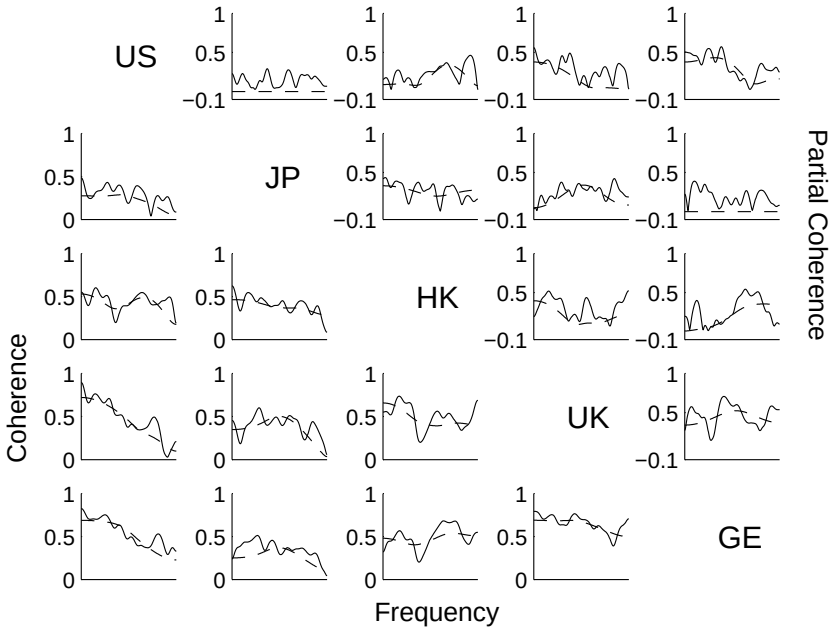

Figure 1.12 Coherence and partial coherence spectra of international stock market data, for the first model in table 1.2. Nonparametric estimates are shown in solid lines and ML estimates are shown in dashed lines.

were computed from (1.37), resulting in a five-dimensional time series of length $N = 2458$.

The BIC selects a model with lag $p = 1$, and with (UK,IT), (FR,AU), and (GE, AU) as the conditionally independent pairs. The coherence and partial coherence spectra for this model are shown in figure 1.13. The partial coherence spectrum suggests that the French stock market is the market on the Continent most strongly connected to the UK market. The French, German, and Italian stock markets are highly inter-dependent, while the Austrian market is more weakly connected to the other markets. These results agree with conclusions from the analysis in [43].

1.5 Conclusion

We have considered a parametric approach for maximum likelihood estimation of autoregressive models with conditional independence constraints. These constraints impose a sparsity pattern on the inverse of the spectral density matrix, and result in nonconvex equalities in the estimation problem. We have formulated a convex relaxation of the ML estimation problem and shown that the

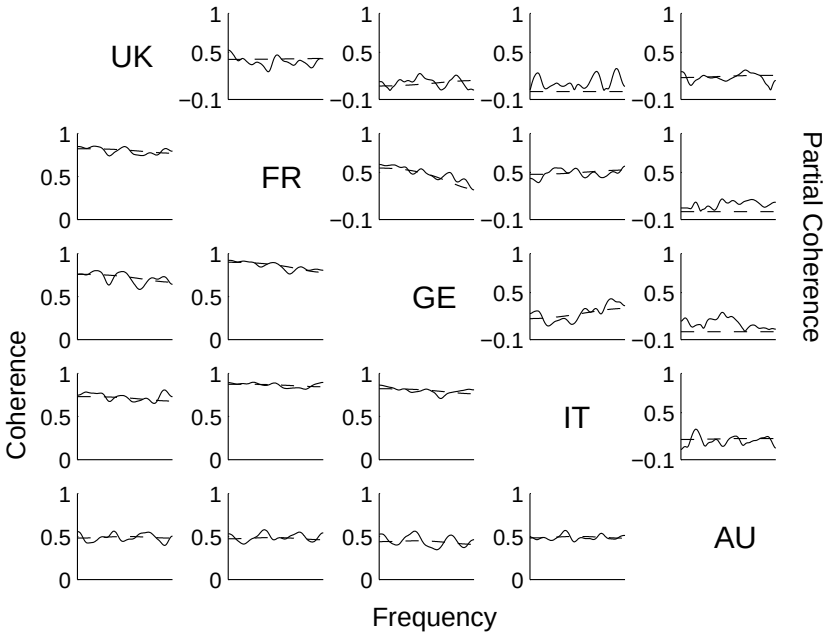

Figure 1.13 Coherence and partial coherence spectrum of the model for the European stock return data. Nonparametric estimates (solid lines) and ML estimates (dashed lines) for the best model selected by the BIC.

relaxation is exact when the sample covariance matrix in the objective of the estimation problem is block-Toeplitz. We have also noted from experiments that the relaxation is often exact for covariance matrices that are not block-Toeplitz.

The convex formulation allows us to select graphical models by fitting autoregressive models to different topologies, and ranking the topologies using information theoretic model selection criteria. The approach was illustrated with randomly generated and real data, and works well when the number of models in the comparison is small, or the number of nodes is small enough for an exhaustive search. For larger model selection problems, it will be of interest to extend recent techniques for covariance selection [12, 13] to time series.

Acknowledgments

The research was supported in part by NSF grants ECS-0524663 and ECCS-0824003, and a Royal Thai government scholarship. Part of the research by Joachim Dahl was carried out during his affiliation with Aalborg University, Denmark.

References

[1] J. Pearl, *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, 1988.

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

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

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

[5] M. I. Jordan, Ed., *Learning in Graphical Models*. MIT Press, 1999.

[6] J. Pearl, *Causality. Models, Reasoning, and Inference*. Cambridge University Press, 2000.

[7] M. I. Jordan, "Graphical models," *Statistical Science*, vol. 19, pp. 140–155, 2004.

[8] A. P. Dempster, "Covariance selection," *Biometrics*, vol. 28, pp. 157–175, 1972.

[9] R. Grone, C. R. Johnson, E. M. Sá, and H. Wolkowicz, "Positive definite completions of partial Hermitian matrices," *Linear Algebra and Applications*, vol. 58, pp. 109–124, 1984.

[10] J. Dahl, L. Vandenberghe, and V. Roychowdhury, "Covariance selection for non-chordal graphs via chordal embedding," *Optimization Methods and Software*, vol. 23, no. 4, pp. 501–520, 2008.

[11] N. Meinshausen and P. Bühlmann, "High-dimensional graphs and variable selection with the Lasso," *Annals of Statistics*, vol. 34, no. 3, pp. 1436–1462, 2006.

[12] O. Banerjee, L. El Ghaoui, and A. d'Aspremont, "Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data," *Journal of Machine Learning Research*, vol. 9, pp. 485–516, 2008.

[13] Z. Lu, "Adaptive first-order methods for general sparse inverse covariance selection," 2008, manuscript.

[14] D. Brillinger, "Remarks concerning graphical models for time series and point processes," *Revista de Econometria*, vol. 16, pp. 1–23, 1996.

[15] R. Dahlhaus, "Graphical interaction models for multivariate time series," *Metrika*, vol. 51, no. 2, pp. 157–172, 2000.

[16] R. Dahlhaus, M. Eichler, and J. Sandkühler, "Identification of synaptic connections in neural ensembles by graphical models," *Journal of Neuroscience Methods*, vol. 77, no. 1, pp. 93–107, 1997.

[17] M. Eichler, R. Dahlhaus, and J. Sandkühler, "Partial correlation analysis for the identification of synaptic connections," *Biological Cybernetics*, vol. 89, no. 4, pp. 289–302, 2003.

[18] 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*, vol. 360, no. 1457, pp. 937–946, 2005.

[19] U. Gather and M. I. nd R. Fried, "Graphical models for multivariate time series from intensive care monitoring," *Statistics in Medicine*, vol. 21, no. 18, pp. 2685–2701, 2002.

[20] J. Timmer, M. Lauk, S. Häußler, V. Radt, B. Köster, B. Hellwig, B. Guschlbauer, C. Lücking, M. Eichler, and G. Deuschl, "Cross-spectral analysis of tremor time series," *Internation Journal of Bifurcation and Chaos*, vol. 10, pp. 2595–2610, 2000.

[21] S. Feiler, K. Muller, A. Muller, R. Dahlhaus, and W. Eich, "Using interaction graphs for analysing the therapy process," *Psychother Psychosom*, vol. 74, no. 2, pp. 93–99, 2005.

[22] R. Fried and V. Didelez, "Decomposability and selection of graphical models for multivariate time series," *Biometrika*, vol. 90, no. 2, p. 251, 2003.

[23] M. Eichler, "Testing nonparametric and semiparametric hypotheses in vector stationary processes," *Journal of Multivariate Analysis*, vol. 99, no. 5, pp. 968–1009, 2008.

[24] F. Bach and M. Jordan, "Learning graphical models for stationary time series," *IEEE Transactions on Signal Processing*, vol. 52, no. 8, pp. 2189–2199, 2004.

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

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

[27] T. Soderstrom and P. Stoica, *System Identification*. Prentice Hall, 1989.

[28] G. Box and G. Jenkins, *Time Series Analysis, Forecasting and Control*. Holden-Day, Incorporated, 1976.

[29] S. Marple, *Digital spectral analysis with applications*. Prentice-Hall, Inc., 1987.

[30] S. Kay, *Modern Spectral Estimation: Theory and Application*. Prentice Hall, 1988.

[31] P. Stoica and R. Moses, *Introduction to Spectral Analysis*. Prentice Hall, Inc., 1997.

[32] P. Stoica and A. Nehorai, "On stability and root location of linear prediction models," *IEEE Transaxtions on Acoustics, Speech, and Signal Processing*, vol. 35, pp. 582–584, 1987.

[33] G. C. Reinsel, *Elements of Multivariate Time Series Analysis*, 2nd ed. Springer, 2007.

[34] J. P. Burg, "Maximum entropy spectral analysis," Ph.D. dissertation, Stanford University, 1975.

[35] E. Hannon, *Multiple Time Series*. John Wiley and Sons, Inc., 1970.

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

[37] Y. Hachez, "Convex optimization over non-negative polynomials: Structured algorithms and applications," Ph.D. dissertation, Université catholique de Louvain, Belgium, 2003.

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

[39] ——, "Graph implementations for nonsmooth convex programs," in *Recent Advances in Learning and Control (a tribute to M. Vidyasagar)*, V. Blondel, S. Boyd, and H. Kimura, Eds. Springer Verlag, 2008.

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

[41] J. Proakis, *Digital Communications*, 4th ed. McGraw-Hill, 2001.

[42] D. Bessler and J. Yang, "The structure of interdependence in international stock markets," *Journal of International Money and Finance*, vol. 22, no. 2, pp. 261–287, 2003.

[43] J. Yang, I. Min, and Q. Li, "European stock market integration: Does EMU matter?" *Journal of Business Finance & Accounting*, vol. 30, no. 9-10, pp. 1253–1276, 2003.