Journal of Machine Learning Research 11 (2010) 2671-2705 Submitted 10/09; Revised 5/10; Published 10/10

Topology Selection in Graphical Models of Autoregressive Processes

Jitkomut Songsiri JITKOMUT@EE.UCLA.EDU
Lieven Vandenberghe VANDENBE@EE.UCLA.EDU
Department of Electrical Engineering
University of California
Berkeley, CA 90095-1594, USA

Editor: Martin Wainwright

Abstract

An algorithm is presented for topology selection in graphical models of autoregressive Gaussian time series. The graph topology of the model represents the sparsity pattern of the inverse spectrum of the time series and characterizes conditional independence relations between the variables. The method proposed in the paper is based on an ℓ_1-type nonsmooth regularization of the conditional maximum likelihood estimation problem. We show that this reduces to a convex optimization problem and describe a large-scale algorithm that solves the dual problem via the gradient projection method. Results of experiments with randomly generated and real data sets are also included.

Keywords: graphical models, time series, topology selection, convex optimization

1. Introduction

We consider graphical models of autoregressive (AR) Gaussian processes

$$x(t) = -\sum_{k=1}^{p} A_k x(t-k) + w(t), \qquad w(t) \sim \mathcal{N}(0, \Sigma) \tag{1}$$

where $x(t) \in \mathbf{R}^n$, and $w(t) \in \mathbf{R}^n$ is Gaussian white noise. A graphical model of the time series is an undirected graph with n nodes, one for each component $x_i(t)$, and an edge connecting nodes i and j if the components $x_i(t)$ and $x_j(t)$ are *conditionally dependent*, given the other components of the time series. The conditional independence property has a simple characterization (which holds for general Gaussian stationary processes) in terms of the spectrum of the process: $x_i(t)$ and $x_j(t)$ are independent, conditional on the other $n-2$ components of $x(t)$, if and only if

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

for all ω, where $S(\omega)$ is the spectral density matrix (Brillinger, 1981, Chapter 8; Dahlhaus, 2000). This characterization allows us to include the conditional independence relations in an estimation problem by placing sparsity constraints on the inverse spectral density matrix.

In Songsiri et al. (2009) a convex optimization method was discussed for estimating the model parameters A_k, Σ from data, given the graph of conditional independence relations. The method is based on solving the convex optimization problem

$$\begin{array}{ll} \text{minimize} & -\log\det X_{00} + \mathbf{tr}(CX) \\ \text{subject to} & Y_k = \sum_{i=0}^{p-k} X_{i,i+k}, \quad k = 0,1,\dots,p, \\ & (Y_k)_{ij} = 0, \quad (i,j) \in \mathcal{V}, \quad k = 0,1,\dots,p, \\ & X \succeq 0. \end{array} \tag{2}$$

Here C is the sample covariance matrix and $\mathcal{V}$ is the set of conditionally independent pairs of variables. The optimization variables are $X \in \mathbf{S}^{n(p+1)}$ (the symmetric matrices of order $n(p+1)$), $Y_0 \in \mathbf{S}^n$, and $Y_k \in \mathbf{R}^{n\times n}$, $k = 1,2,\dots,p$. X_{ij} denotes the $n \times n$ subblock of X in position i, j, where the indices i and j run from 0 to p. It was shown that if the sample covariance matrix C is block-Toeplitz, then problem (2) is equivalent to the conditional maximum likelihood (ML) estimation problem, and the ML estimates for A_k and Σ are easily obtained from the optimal solution X. If C is not block-Toeplitz, the problem is a relaxation and in general not equivalent to the conditional ML problem. However in practice, the relaxation often happens to be exact (Songsiri et al., 2009). This will be discussed in more detail in §2.3.

In this paper we consider the more general problem of estimating the model parameters *and* the topology of the graphical model. The topology selection problem can be solved by enumerating all topologies, solving the ML estimation problem for each topology, and ranking them via information-theoretic criteria such as the Akaike or Bayes information criteria (Eichler, 2006; Songsiri et al., 2009). However this combinatorial approach is clearly limited to small graphs. The goal of this paper is to present an efficient alternative based on convex optimization.

Topology selection for graphical models of time series is of interest in many applications (see Dahlhaus et al., 1997; Eichler et al., 2003; Salvador et al., 2005; Gather et al., 2002; Timmer et al., 2000; Feiler et al., 2005; Friedman et al., 2008). A common approach is to formulate hypothesis testing problems to decide about the presence or absence of edges. Dahlhaus (2000) 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 (see also Dahlhaus et al., 1997; Eichler et al., 2003; Salvador et al., 2005; Gather et al., 2002; Timmer et al., 2000; Feiler et al., 2005; Fried and Didelez, 2003). Eichler (2008) presents a more general approach by introducing a hypothesis test based on the norm of some suitable function of the spectral density matrix. A related problem was studied by Bach and Jordan (2004). They use an efficient search procedure to learn the graph structure from sample estimates of the joint spectral density matrix.

If $p = 0$, the problem (2) reduces to

$$\begin{array}{ll} \text{minimize} & -\log\det X + \mathbf{tr}(CX) \\ \text{subject to} & X_{ij} = 0, \quad (i,j) \in \mathcal{V} \end{array} \tag{3}$$

with variable $X \in \mathbf{S}^n$. (Throughout the paper we take the set of positive definite matrices as the domain of the function $\log\det X$, so (3) includes an implicit constraint $X \succ 0$.) Problem (3) is known as the *covariance selection* problem, that is, the problem of computing the ML estimate of the inverse covariance matrix $X = \Sigma^{-1}$ of a multivariate Gaussian variable $\mathcal{N}(0,\Sigma)$, subject to conditional independence constraints (which, for a normal distribution, correspond to zeros in the inverse covariance); see Dempster (1972) and Lauritzen (1996, §5.2). Recently, new heuristic methods for topology selection in large Gaussian graphical models have been developed. These methods are based on augmenting the ML objective with an ℓ_1-norm regularization term, that is, on solving

$$\text{minimize} \quad -\log\det X + \mathbf{tr}(CX) + \gamma \textstyle\sum_{ij} |X_{ij}| \tag{4}$$

(see Dahl et al., 2005; Meinshausen and Bühlmann, 2006; Banerjee et al., 2008; Ravikumar et al., 2008; Friedman et al., 2008; Lu, 2009, 2010). The optimization problem (4) is convex but has $n(n+1)/2$ variables (the elements of X) and is nondifferentiable, so it can be challenging to solve when n is large. Several large-scale methods have been proposed. Banerjee et al. (2008) apply a block coordinate descent method to the dual problem. Each step of this method reduces to solving a quadratic program with box constraints. They also apply Nesterov's optimal gradient method (Nesterov, 2005) to a smooth approximation of (4). Friedman et al. (2008) observe that the dual of the subproblems in the coordinate descent algorithm can be regarded as a lasso-type problem and solved with a method called graphical lasso. Scheinberg and Rish (2009) consider a coordinate ascent method applied to the primal problem. A method based on column-wise updates is given by Rothman et al. (2008). A related problem is explored in Yuan and Lin (2007) where the authors make a connection between (4) and more general determinant maximization problems (Vandenberghe et al., 1998), and solve the problem using interior-point methods. Lu (2009) observes that the dual of (4) is a smooth problem and applies Nesterov's method (Nesterov, 2005) directly to the dual. The algorithm is further extended by Lu (2010) and compared with a projected spectral gradient method. Another closely related paper is Duchi et al. (2008) in which the gradient projection method is applied to the dual problem.

The main purpose of this paper is to develop an efficient method for topology selection in AR models, based on augmenting the estimation problem (2) with a convex regularization term, similar to the ℓ_1-norm regularization used in (4). We also discuss first-order methods for solving the resulting large-scale and nondifferentiable convex optimization problem.

The paper is organized as follows. In Section 2 we review the definition of conditional independence in time series and summarize the results from Songsiri et al. (2009). In Section 3 we set up the topology selection problem as a regularized ML problem and discuss its properties. Examples and applications are presented in Sections 4 and 5. We conclude in Section 6 with a discussion of gradient projection algorithms for solving large instances of the regularized ML estimation problem.

1.1 Notation

$\mathbf{S}^n$ is the set of real symmetric matrices of order n. $\mathbf{S}^n_+$ and $\mathbf{S}^n_{++}$ are the sets of symmetric positive semidefinite, respectively, positive definite, matrices of order n. $\mathbf{R}^{m\times n}$ is the set of $m\times n$-matrices. $\mathbf{M}^{n,p}$ 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 $\langle X,Y\rangle = \mathbf{tr}(X^TY)$ is used for the three vector spaces $\mathbf{S}^n$, $\mathbf{R}^{m\times n}$, $\mathbf{M}^{n,p}$. For a symmetric matrix X, the inequalities $X \succeq 0$ and $X \succ 0$ mean X is positive semidefinite, resp., positive definite. Row and column indices of submatrices in a block matrix start at 0. If X is a matrix with (block) entries X_{ij}, then $X_{i:j,k:l}$ will denote the submatrix formed by rows i through j and columns k through l:

$$X_{i:j,k:l} = \begin{bmatrix} X_{ik} & X_{i,k+1} & \cdots & X_{il} \\ X_{i+1,k} & X_{i+1,k+1} & \cdots & X_{i+1,l} \\ \vdots & \vdots & & \vdots \\ X_{jk} & X_{j,k+1} & \cdots & X_{j,l} \end{bmatrix}.$$

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 $X = \begin{bmatrix} X_0 & X_1 & \cdots & X_p \end{bmatrix} \in \mathbf{M}^{n,p}$, then

$$\mathrm{T}(X) = \begin{bmatrix} X_0 & X_1 & \cdots & X_p \\ X_1^T & X_0 & \cdots & X_{p-1} \\ \vdots & \vdots & \ddots & \vdots \\ X_p^T & X_{p-1}^T & \cdots & X_0 \end{bmatrix}. \tag{5}$$

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

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

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

$$\mathrm{D}_0(S) = \sum_{i=0}^{p} S_{ii}, \qquad \mathrm{D}_k(S) = 2\sum_{i=0}^{p-k} S_{i,i+k}, \quad k = 1, \ldots, p. \tag{6}$$

A symmetric sparsity pattern of a sparse matrix X of order n will be associated with the positions $\mathcal{V} \subseteq \{1, \ldots, n\} \times \{1, \ldots, n\}$ of its zero entries. We assume $(i,i) \notin \mathcal{V}$ for $i = 1, \ldots, n$, that is, the diagonal entries are not included among the zeros. $\mathrm{P}_{\mathcal{V}}(X)$ 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 $\mathcal{V}$:

$$\mathrm{P}_{\mathcal{V}}(X)_{ij} = \begin{cases} X_{ij} & (i,j) \in \mathcal{V} \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

The same notation is 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 an $r \times s$ block matrix with i,j block X_{ij}, and each block is square of order n, then $\mathrm{P}_{\mathcal{V}}(X)$ denotes the $r \times s$ block matrix with i,j block $\mathrm{P}_{\mathcal{V}}(X)_{ij} = \mathrm{P}_{\mathcal{V}}(X_{ij})$. Similarly, $\mathrm{P}_{\mathcal{V}}(X)$ with $X \in \mathbf{M}^{n,p}$ denotes

$$\begin{bmatrix} \mathrm{P}_{\mathcal{V}}(X_0) & \mathrm{P}_{\mathcal{V}}(X_1) & \cdots & \mathrm{P}_{\mathcal{V}}(X_p) \end{bmatrix}.$$

The subscript $\mathcal{V}$ in $\mathrm{P}_{\mathcal{V}}$ is omitted if the sparsity pattern is clear from the context.

2. Graphical Models of Autoregressive Gaussian Processes

In this section we describe the conditional independence property for Gaussian time series, review the maximum likelihood estimation of AR models, and provide a convex formulation for the estimation problem with conditional independence constraints.

2.1 Conditional Independence

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

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

where $R_k = \mathbf{E}\, x(t+k)x(t)^T$ and $\mathrm{j} = \sqrt{-1}$. We assume that $S(\omega)$ is invertible for all ω. Two components $x_i(t)$ and $x_j(t)$ of $x(t)$ are conditionally independent (i.e., conditional on the other components of $x(t)$) if

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

for all ω (Brillinger, 1981; Dahlhaus, 2000). If we denote by $\mathcal{V}$ the set of index pairs i, j of conditionally independent variables, then we can use the projection operator $\mathrm{P} = \mathrm{P}_{\mathcal{V}}$ defined in (7) to express the conditional independence relations as

$$\mathrm{P}(S(\omega)^{-1}) = 0. \tag{8}$$

In a graphical model of the process, the index set $\mathcal{V}$ is the set of missing edges in the graph.

To apply this result to AR processes (1) we need to express the inverse spectrum in terms of the model parameters. The notation will simplify if we first normalize the input covariance and use the model

$$B_0 x(t) = -\sum_{k=1}^{p} B_k x(t-k) + v(t), \qquad v(t) \sim \mathcal{N}(0, I), \tag{9}$$

where $B_0 \in \mathbf{S}^n_{++}$ and $B_k \in \mathbf{R}^{n\times n}$, $k = 1,\ldots,p$. If Σ is nonsingular, the two models are equivalent, and related as $B_0 = \Sigma^{-1/2}$, $B_k = \Sigma^{-1/2} A_k$ for $k \geq 1$. The inverse spectrum $S(\omega)$ of the process (9) is a trigonometric matrix polynomial

$$S(\omega)^{-1} = Y_0 + \frac{1}{2}\sum_{k=1}^{p}(e^{-\mathrm{j}k\omega} Y_k + e^{\mathrm{j}k\omega} Y_k^T) \tag{10}$$

where $Y_0 = \sum_{l=0}^{p} B_l^T B_l$, and $Y_k = 2\sum_{l=0}^{p-k} B_l^T B_{k+l}$ for $k = 1,\ldots,p$. If we define $B = \begin{bmatrix} B_0 \; B_1 \cdots B_p \end{bmatrix}$, we can use the operator D defined in (6) to express Y_k as

$$\begin{bmatrix} Y_0 & Y_1 & \cdots & Y_p \end{bmatrix} = \mathrm{D}(B^T B).$$

The expression (10) shows that $(S(\omega)^{-1})_{ij}$ is identically zero if and only if the i, j and j, i entries of Y_k are zero for $k = 0,\ldots,p$. The conditional independence condition (8) is therefore equivalent to a quadratic equation in the model parameters B_k:

$$\mathrm{P}\left(\mathrm{D}(B^T B)\right) = 0. \tag{11}$$

(Recall from the Notation section that if Y is a block matrix with square submatrices Y_k of order n, then $\mathrm{P}(Y)$ denotes the block matrix with submatrices $\mathrm{P}(Y_k)$.)

2.2 Conditional Maximum Likelihood Estimation

We now consider the problem of estimating the model parameters B from an observed sequence $\tilde{x}(1), \tilde{x}(2), \ldots, \tilde{x}(N)$ of the AR process, subject to known conditional independence constraints (11). In Songsiri et al. (2009) the estimation problem was formulated as the optimization problem

$$\begin{array}{ll} \text{minimize} & -2\log\det B_0 + \mathbf{tr}(CB^T B) \\ \text{subject to} & \mathrm{P}(\mathrm{D}(B^T B)) = 0. \end{array} \tag{12}$$

The matrix $C \in \mathbf{S}_+^{n(p+1)}$ is a sample estimate of the covariance matrix, that is, its blocks C_{ij}, $i \le j$, are estimates of the covariances $R_{j-i} = \mathbf{E}\, x(t+j-i)x(t)^T$, calculated from the observed sequence. Two choices of C are common. The first choice is the *non-windowed estimate*

$$C = \frac{1}{N-p} H H^T, \qquad H = \begin{bmatrix} \tilde{x}(p+1) & \tilde{x}(p+2) & \cdots & \tilde{x}(N) \\ \tilde{x}(p) & \tilde{x}(p+1) & \cdots & \tilde{x}(N-1) \\ \vdots & \vdots & & \vdots \\ \tilde{x}(1) & \tilde{x}(2) & \cdots & \tilde{x}(N-p) \end{bmatrix}. \tag{13}$$

With this choice the estimation problem (12) can be interpreted as a maximum likelihood problem. Indeed, from (9), the conditional density of a sequence $x(t_1)$, $x(t_1+1)$, $\ldots x(t_2)$, given $x(t_1-p)$, $\ldots$, $x(t_1-1)$, is given by

$$\left(\frac{\det B_0}{(2\pi)^{n/2}}\right)^{t_2-t_1+1} \exp\left(-\frac{1}{2}\sum_{t=t_1}^{t_2} \mathbf{x}(t)^T B^T B \mathbf{x}(t)\right),$$

where $\mathbf{x}(t)$ denotes the $n(p+1)$-vector $\mathbf{x}(t) = (x(t), x(t-1), \ldots, x(t-p))$. From this it can be shown that the cost function in (12) with C defined as in (13), is essentially the negative conditional log-likelihood function of the observed sequence $\tilde{x}(p+1)$, $\tilde{x}(p+2)$, $\ldots$, $\tilde{x}(N)$, given $\tilde{x}(1)$, $\ldots$, $\tilde{x}(p)$. We therefore refer to (12) as the *conditional maximum likelihood problem*. For AR processes, the conditional ML formulation is substantially simpler and more often used than the exact ML formulation. Moreover, when the data length N is sufficiently large compared to p, the difference between the exact and conditional ML formulations is small.

The second choice for C is the *windowed estimate*

$$C = \frac{1}{N} H H^T, \tag{14}$$

where

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

The windowed estimate C is block-Toeplitz, and this guarantees several useful properties of the resulting model B (for example, stability; see Songsiri et al., 2009). In practice, the differences between the windowed and non-windowed estimates are small when $N \gg p$.

We will assume that C is positive definite. If n is small compared to N, this is a reasonable assumption but not guaranteed to be true. (As a counterexample, assume $\tilde{x}(1)$, $\ldots$, $\tilde{x}(n)$ are the first n unit vectors and the remainder of the sequence is zero. The matrix C in (14) then has rank $n+p$.) If C is not positive definite, it may be necessary to add a small multiple of the identity. This is equivalent to a quadratic regularization term proportional to $\|B\|_F^2$ in the objective of (12).

When there are no sparsity constraints in (12), the solution can be found by setting the gradient of the cost function equal to zero, which gives

$$\begin{bmatrix} C_{00} & C_{01} & \cdots & C_{0p} \\ C_{10} & C_{11} & \cdots & C_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ C_{p0} & C_{p1} & \cdots & C_{pp} \end{bmatrix} \begin{bmatrix} B_0 \\ B_1^T \\ \vdots \\ B_p^T \end{bmatrix} = \begin{bmatrix} B_0^{-1} \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

Written in terms of the original variables $\Sigma = B_0^{-2}$, $A_k = B_0^{-1}B_k$, this gives

$$\begin{bmatrix} C_{00} & C_{01} & \cdots & C_{0p} \\ C_{10} & C_{11} & \cdots & C_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ C_{p0} & C_{p1} & \cdots & C_{pp} \end{bmatrix} \begin{bmatrix} I \\ A_1^T \\ \vdots \\ A_p^T \end{bmatrix} = \begin{bmatrix} \Sigma \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \tag{15}$$

with unknowns $\Sigma = B_0^{-2}$, $A_k = B_0^{-1}B_k$. The bottom p equations form a set of linear equations from which $A_1, \ldots, A_p$ can be determined. Plugging in the solution in the first equation gives Σ. Later in the paper we will refer to the solution as the *least-squares estimate* because the bottom p equations can be interpreted as normal equations for the least-squares problem

$$\text{minimize} \quad \mathbf{tr}(ACA^T)$$

with variable $A = \begin{bmatrix} I & A_1 & \cdots & A_p \end{bmatrix}$. This method is also known as the *covariance method* if C is the non-windowed sample covariance (13), and as the *correlation method* if C is the windowed sample covariance (14) (see Stoica and Moses, 1997).

2.3 Convex Formulation

The optimization problem (12) is non-convex because of the quadratic equality constraint. A convex relaxation is

$$\begin{array}{ll} \text{minimize} & -\log\det X_{00} + \mathbf{tr}(CX) \\ \text{subject to} & \mathrm{P}(\mathrm{D}(X)) = 0 \\ & X \succeq 0 \end{array} \tag{16}$$

with variable $X \in \mathbf{S}^{n(p+1)}$. The relaxation is exact, that is, the two problems (16) and (12) are equivalent, if the optimal solution X of (16) has rank n. In that case, the solution B of (16) can be calculated by factoring X as $X = B^TB$.

A condition for exactness of the relaxation follows from the dual problem of (16), which is

$$\begin{array}{ll} \text{maximize} & \log\det W + n \\ \text{subject to} & \begin{bmatrix} W & 0 \\ 0 & 0 \end{bmatrix} \preceq C + \mathrm{T}(\mathrm{P}(Z)), \end{array} \tag{17}$$

with variables $W \in \mathbf{S}^n$ and $Z \in \mathbf{M}^{n,p}$ (for the derivation, see Songsiri et al., 2009). The variable Z is the Lagrange multiplier associated with the equality constraint in (16); the slack matrix in the inequality in (17) is the multiplier associated with the primal inequality $X \succeq 0$. To find the relation between primal and dual solutions, we first note that the primal and dual problems are strictly feasible: $X = I$ is strictly feasible in the primal problem (16), since by assumption $\mathcal{V}$ does not contain any diagonal entries; in the dual problem $Z = 0$ and a sufficiently small positive definite W are strictly feasible, because $C \succ 0$ by assumption. From convex duality, strict primal and dual feasibility imply that the primal and dual problems are solvable, and that their optimal solutions are related by the optimality conditions

$$X_{00}^{-1} = W, \qquad \mathbf{tr}\left(X\left(C + \mathrm{T}(\mathrm{P}(Z)) - \begin{bmatrix} W & 0 \\ 0 & 0 \end{bmatrix}\right)\right) = 0 \tag{18}$$

(Boyd and Vandenberghe, 2004, Chapter 5). The second condition is known as *complementary slackness* between the optimal X and the dual variable associated with the inequality $X \succeq 0$. From these optimality conditions, it can be shown that the relaxation is exact when the trailing principal submatrix of order np in the matrix $C + \mathrm{T}(\mathrm{P}(Z)) \in \mathbf{S}^{n(p+1)}$ is positive definite at the optimum, that is,

$$(C + \mathrm{T}(\mathrm{P}(Z)))_{1:p,1:p} \succ 0. \tag{19}$$

Under this condition, the rank of

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

is at least np. Since X has order $n(p+1)$, the two conditions in (18) imply that the optimal X has rank n.

In general it is difficult to guarantee a priori that the condition (19) holds at optimum. However, when C is block-Toeplitz, then (19) can be shown to hold for all dual feasible Z. This follows from the following easily established property of block-Toeplitz matrices: if $V \in \mathbf{S}^{n(p+1)}$ is a symmetric block-Toeplitz matrix with $n \times n$ blocks V_{ij}, and

$$V = \begin{bmatrix} V_{00} & V_{0,1:p} \\ V_{1:p,0} & V_{1:p,1:p} \end{bmatrix} \succeq \begin{bmatrix} W & 0 \\ 0 & 0 \end{bmatrix}$$

for some $W \succ 0$, then V is positive definite (see Songsiri et al., 2009, §3.3.3). We therefore conclude that for positive definite block-Toeplitz C (for example, the windowed sample covariance (14) or the true covariance), the problems (12) and (16) are *equivalent*. For general non-block-Toeplitz C (for example, the non-windowed sample covariance (13)), we cannot guarantee that (19) holds at the optimum. However, we can note that the non-windowed sample covariance approaches a block-Toeplitz matrix as $N \to \infty$. It is therefore not surprising that even for the non-windowed estimate, the relaxation is often exact, as was observed in the experimental results in Songsiri et al. (2009).

3. Topology Selection Via Nonsmooth Regularization

In the previous section we have described a convex formulation of the (conditional) ML estimation problem with given conditional independence constraints, that is, a given graph topology. In many applications the topology is not known, and needs to be discovered from the data. Information theoretic model selection criteria such as the Akaike, second-order Akaike, or Bayes information criteria can be used for this purpose. They require enumerating all possible topologies, solving the ML problem for each topology, and ranking the ML estimates according to their information criterion score. These scores are defined as

$$\mathrm{AIC} = -2\mathcal{L} + 2k, \qquad \mathrm{AIC_c} = -2\mathcal{L} + \frac{2Nk}{N-k-1}, \qquad \mathrm{BIC} = -2\mathcal{L} + k\log N \tag{20}$$

where $\mathcal{L}$ is the log-likelihood of the ML estimate, N is the sample size, and k is the effective number of parameters. In our application, $\mathcal{L}$ is given by

$$\mathcal{L} = \frac{N-p}{2}\left(\log\det X_{00} - \mathbf{tr}(CX)\right)$$

where X is the optimal solution of (16), and we use for k the total number of parameters in the estimation problem,

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

where $|\mathcal{V}|$ is the number of conditionally independent pairs of variables. This topology selection method based on information-theoretic criteria is feasible if the number of possible topologies is not too large, but quickly becomes intractable even for small values of n. In this section and the next we describe a more scalable approach based on a convex optimization problem that extends the ℓ_1-norm heuristic (4) for sparse covariance selection.

3.1 Regularized ML Problem

In analogy with the convex heuristic for covariance selection (4), we can formulate a regularized ML problem by adding a nonsmooth ℓ_1-type penalty:

$$\begin{array}{ll} \text{minimize} & -\log\det X_{00} + \mathbf{tr}(CX) + \gamma h(\mathrm{D}(X)) \\ \text{subject to} & X \succeq 0, \end{array} \tag{21}$$

where $\gamma > 0$ is a weighting parameter. The penalty $h : \mathbf{M}^{n,p} \to \mathbf{R}$ is a convex function, chosen to encourage a sparse solution X with a common, symmetric sparsity pattern for the $p+1$ blocks of $\mathrm{D}(X)$. We will use the penalty function

$$h_\infty(Y) = \sum_{j>i} \max\left\{ \left|(Y_0)_{ij}\right|, \max_{k=1,\ldots,p} \left|(Y_k)_{ij}\right|, \max_{k=1,\ldots,p} \left|(Y_k)_{ji}\right| \right\} \tag{22}$$

that is, a sum of the ℓ_∞-norms of vectors of i,j and j,i-entries of the coefficients Y_k. In the examples (Section 4) we will also discuss penalty functions defined as sums of ℓ_α-norms, with $\alpha = 1,2$.

Regularization with a convex sum-of-norms penalty is a popular technique for achieving sparsity of groups of variables. Examples from statistics are the *composite absolute penalties* (CAP) (Zhao et al., 2009) and the *group lasso* (Yuan and Lin, 2006; Kim et al., 2006). When $p = 0$ and $X \in \mathbf{S}^n$ in (21) the penalty term reduces to $\sum_{i>j} |X_{ij}|$ and we obtain problem (4), studied in Banerjee et al. (2008), Lu (2009) and Friedman et al. (2008), with the minor difference that we do not penalize the diagonal entries of X.

We now derive the dual problem of (21) which will be important in Section 6. To simplify the derivation we introduce a variable $Y = \mathrm{D}(X)$ and write the problem as

$$\begin{array}{ll} \text{minimize} & -\log\det X_{00} + \mathbf{tr}(CX) + \gamma h_\infty(Y) \\ \text{subject to} & Y = \mathrm{D}(X) \\ & X \succeq 0. \end{array}$$

If we use a multiplier $Z \in \mathbf{M}^{n,p}$ for the equality constraint $Y = \mathrm{D}(X)$ and a multiplier $U \in \mathbf{S}^{n(p+1)}$ for the inequality $X \succeq 0$, the Lagrangian of the problem is

$$\begin{aligned} L(X,Y,Z,U) &= -\log\det X_{00} + \mathbf{tr}(CX) + \gamma h_\infty(Y) - \mathbf{tr}(UX) + \mathbf{tr}(Z^T(\mathrm{D}(X) - Y)) \\ &= -\log\det X_{00} + \mathbf{tr}((C + \mathrm{T}(Z) - U)X) + \gamma h_\infty(Y) - \mathbf{tr}(Z^T Y). \end{aligned} \tag{23}$$

(Recall that the mappings T and D defined in (5) and (6) are adjoints, that is, $\mathbf{tr}(Z^T\,\mathrm{D}(X)) = \mathbf{tr}(\mathrm{T}(Z)X)$.) The dual function is the infimum of the Lagrangian over X and Y. We first minimize over Y. The nonlinear penalty term does not depend on the diagonal entries of the blocks Y_k. The minimization over the diagonal entries of Y_k is therefore unbounded below unless

$$\mathbf{diag}(Z_k) = 0, \quad k = 0, 1, \dots, p. \tag{24}$$

The minimization over the off-diagonal part of the blocks Y_k decomposes into independent minimizations of the functions

$$-\sum_{k=0}^{p}((Z_k)_{ij}(Y_k)_{ij} + (Z_k)_{ji}(Y_k)_{ji}) + \gamma \max\left\{|(Y_0)_{ij}|, \max_{k=1,\dots,p}|(Y_k)_{ij}|, \max_{k=1,\dots,p}|(Y_k)_{ji}|\right\}$$

for each element i, j with $i > j$. This expression is unbounded below unless

$$2|(Z_0)_{ij}| + \sum_{k=1}^{p}(|(Z_k)_{ij}| + |(Z_k)_{ji}|) \le \gamma, \quad i \ne j, \tag{25}$$

and, if this condition holds, the infimum over Y is zero. The result of the partial minimization of the Lagrangian over Y can be summarized as

$$\inf_Y L(X,Y,Z,U) = \begin{cases} -\log\det X_{00} + \mathbf{tr}((C + \mathrm{T}(Z) - U)X) & \text{(24), (25)} \\ -\infty & \text{otherwise.} \end{cases}$$

Next, we carry out the minimization over X. The terms in X_{00} are bounded below if and only if $(C + \mathrm{T}(Z) - U)_{00} \succ 0$, and if this holds, they are minimized by $X_{00} = (C + \mathrm{T}(Z) - U)_{00}^{-1}$. The Lagrangian is linear in the other blocks X_{ij}, and therefore bounded below (and identically zero) only if $(C + \mathrm{T}(Z) - U)_{ij} = 0$ for blocks $(i, j) \ne (0, 0)$. This gives a third set of dual feasibility conditions:

$$(C + \mathrm{T}(Z) - U)_{00} \succ 0, \qquad (C + \mathrm{T}(Z) - U)_{ij} = 0, \quad (i, j) \ne 0, \tag{26}$$

and an expression for the dual function

$$g(Z,U) = \inf_{X,Y} L(X,Y,Z,U) = \begin{cases} \log\det(C + \mathrm{T}(Z) - U)_{00} + n & \text{(24), (25), (26)} \\ -\infty & \text{otherwise.} \end{cases}$$

The dual problem is to maximize $g(Z,U)$ subject to $U \succeq 0$. If we add a variable $W = C_{00} + Z_0 - U_{00}$ and eliminate the slack variable U, we can express the dual problem as

$$\begin{array}{ll} \text{maximize} & \log\det W + n \\ \text{subject to} & \begin{bmatrix} W & 0 \\ 0 & 0 \end{bmatrix} \preceq C + \mathrm{T}(Z) \\ & \sum_{k=0}^{p}(|(Z_k)_{ij}| + |(Z_k)_{ji}|) \le \gamma, \quad i \ne j \\ & \mathbf{diag}(Z_k) = 0, \quad k = 0, \dots, p. \end{array} \tag{27}$$

The variables are $W \in \mathbf{S}^n$ and $Z \in \mathbf{M}^{n,p}$. When $p = 0$, the problem reduces to

$$\begin{array}{ll} \text{maximize} & \log\det(C + Z) + n \\ \text{subject to} & |Z_{ij}| \le \gamma/2, \quad i \ne j \\ & \mathbf{diag}(Z) = 0, \end{array}$$

Except for the equality constraint, this is the problem considered in Lu (2009) and Duchi et al. (2008).

If a sum of ℓ_α-norms

$$h_\alpha(Y) = \sum_{j>i} \left(\sum_{k=0}^{p} (|(Y_k)_{ij}|^\alpha + |(Y_k)_{ji}|^\alpha) \right)^{1/\alpha} \tag{28}$$

is used as penalty function in (21), the second constraint in the corresponding dual problem (27) is replaced by

$$\left(\sum_{k=0}^{p} \left(|(Z_k)_{ij}|^\beta + |(Z_k)_{ji}|^\beta \right) \right)^{1/\beta} \leq \gamma, \quad i \neq j$$

with $\beta = \alpha/(\alpha-1)$.

3.2 Optimality Conditions

The primal problem (21) is always strictly feasible ($X = I$ is strictly feasible). The dual problem (21) is strictly feasible if $C \succ 0$ (we can take $Z = 0$ and W positive definite and sufficiently small). It follows that the primal and dual problems are solvable, have equal optimal values, and that their solutions are characterized by the following set of necessary and sufficient optimality (or KKT) conditions.

Primal feasibility. X and Y satisfy

$$X \succeq 0, \qquad X_{00} \succ 0, \qquad Y = \mathrm{D}(X).$$

Dual feasibility. W and Z satisfy

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

$$\sum_{k=0}^{p} (|(Z_k)_{ij}| + |(Z_k)_{ji}|) \leq \gamma, \quad i \neq j, \qquad \mathbf{diag}(Z_k) = 0, \quad k = 0, 1, \ldots p.$$

Zero duality gap. The Lagrangian evaluated at the primal and dual optimal solutions is equal to the primal objective at the optimal X, Y, and equal to the dual objective evaluated at the optimal W, Z. From (23), we have equality between the Lagrangian and the primal objective if $\mathbf{tr}(UX) = 0$. Therefore the complementary slackness condition

$$\mathbf{tr}\left(X \left(C + \mathrm{T}(Z) - \begin{bmatrix} W & 0 \\ 0 & 0 \end{bmatrix} \right) \right) = 0 \tag{29}$$

holds at the optimum. Equality between the Lagrangian and the dual objective requires that the primal optimal X, Y minimize the Lagrangian evaluated at the dual optimal W, Z. Reviewing the derivation of the dual problem, we see that X_{00} minimizes the Lagrangian if

$$X_{00}^{-1} = W. \tag{30}$$

To express the conditions from the minimization over Y, we define

$$t_{ij} = \max \left\{ |(Y_0)_{ij}|, \max_{k=1,\ldots,p} |(Y_k)_{ij}|, \max_{k=1,\ldots,p} |(Y_k)_{ji}| \right\}.$$

Then we see that Y minimizes the Lagrangian if for all $i \neq j$, we either have

$$\sum_{k=0}^{p} (|(Z_k)_{ij}| + |(Z_k)_{ji}|) < \gamma,$$

or we have $\sum_{k=0}^{p}(|(Z_k)_{ij}| + |(Z_k)_{ji}|) = \gamma$ and

$$(Z_k)_{ij} = 0,\ |(Y_k)_{ij}| \leq t_{ij} \quad \text{or} \quad (Z_k)_{ij} < 0,\ (Y_k)_{ij} = -t_{ij} \quad \text{or} \quad (Z_k)_{ij} > 0,\ (Y_k)_{ij} = t_{ij}$$

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

The conditions (29)–(30) show that the optimal X has rank n under the same conditions as for the problem with given sparsity pattern (16). If

$$(C + \mathrm{T}(Z))_{1:p,1:p} \succ 0$$

then the optimal X has rank n, and this is always the case if C is block-Toeplitz. Under these conditions, the optimization problem (21) is equivalent to a regularized (conditional) ML estimation problem for the model parameters B:

$$\text{minimize} \quad -2\log\det B_0 + \mathbf{tr}(CB^TB) + \gamma h_\infty(\mathrm{D}(B^TB)).$$

4. Examples with Randomly Generated Data

Our interest in the regularized ML formulation (21) is motivated by the fact that the resulting AR model typically has a sparse inverse spectrum $S(\omega)^{-1}$. Since the regularized problem is convex, it is interesting as an efficient heuristic for topology selection. In this section we illustrate several aspects of this approach using experiments with randomly generated data. In Section 5 we will apply the method to real data sets. Numerical algorithms for solving the regularized problem (21) are discussed in Section 6.

4.1 Method

We first explain in greater detail how we will use the results of the regularized ML problem for model selection.

4.1.1 Choice of Regularization Parameter γ

The sparsity in the inverse spectrum of the solution of the regularized ML problem is controlled by the weighting coefficient γ. As γ varies, the sparsity pattern varies from dense (γ small) to diagonal (γ large). Several authors have discussed the choice of γ in the context of covariance selection (i.e., heuristics based on solving problem (4) or closely related problems). A common approach is to select γ via cross-validation; see, for example, Friedman et al. (2008), Huang et al. (2006) and Banerjee et al. (2008). Meinshausen and Bühlmann (2006) give explicit formulas for γ based

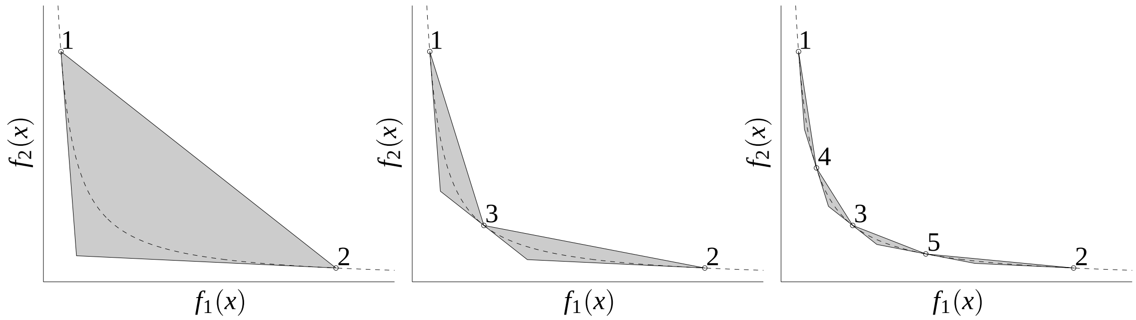


Figure 1: Method for approximating the trade-off curve between two convex objectives.

on a statistical analysis of the probability of errors in the topology (see also Yuan and Lin, 2007; Banerjee et al., 2008). Asadi et al. (2009) consider γ as a random variable and use a maximum a posteriori probability (MAP) estimation to choose γ and the covariance matrix.

In the examples of this section we will use the following method for selecting γ. We first compute the entire trade-off curve between the two terms in the objective of (21), that is, between the log-likelihood and the penalty function $h_\infty(\mathrm{D}(X))$. The trade-off curve can be computed by solving (21) for a number of different values of γ (see below). We collect the topologies of the solutions along the trade-off curve, and solve the ML problem (16) for each of these topologies. We then rank the models using the Bayes information criterion (BIC), as discussed at the beginning of Section 3, and select the model with the lowest score. In this approach, the convex heuristic is used as a preprocessing step to reduce the number of topologies that are examined using the BIC, and to filter out topologies that are unlikely to be competitive.

4.1.2 TRACING TRADE-OFF CURVES

The trade-off curves are computed by solving (21) for a sequence of values of γ. To obtain an accurate estimate of the curve with only a small number of values γ we use a method which is illustrated in Figure 1 for a generic trade-off between two convex cost functions f_1 and f_2. We first solve the scalarized problem

$$\text{minimize} \quad f_1(x) + \gamma f_2(x) \tag{31}$$

for two positive values γ_1, γ_2 near the opposite ends of the trade-off curve. This gives the points labeled 1 and 2 on the trade-off curve. The values of γ_1 and γ_2 also define the slopes of straight lines that support the trade-off curve at points 1 and 2. Since the trade-off curve is convex, we can conclude that the curve between 1 and 2 lies somewhere in the shaded triangular region. As γ_3, we choose the value that corresponds to the slope of the straight line between 1 and 2. Solving problem (31) with $\gamma = \gamma_3$ gives point 3 on the trade-off curve and a straight line that supports the curve at point 3. The trade-off curve between points 1 and 2 is now known to lie in the union of the two shaded triangles. Next, we solve the problem (31) for a value γ_4 corresponding to the slope of the straight line between points 1 and 3, and a value γ_5 corresponding to the slope of the straight line between 3 and 2. In this example, we obtain fairly accurate upper and lower bounds of the actual trade-off curve after solving five scalarized problems (31).

4.1.3 THRESHOLDING

With a proper value of γ, the regularized ML problem (21) has a sparse solution Y, resulting in a sparse inverse spectrum $S(\omega)^{-1}$. When solved with a limited accuracy, the entries of Y are not exactly zero. We will use the following method to determine the topology from the computed solution.

We calculate the inverse spectrum $S(\omega)^{-1}$ and normalize it by scaling its rows and columns so that the diagonal is one:

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

The normalized inverse spectrum $R(\omega)$ is known as the *partial coherence* (Brillinger, 1981; Dahlhaus, 2000). Its entries are between 0 and 1 in magnitude, and measure the conditional dependence between the corresponding variables, after removing the linear effects from the other variables. In the static case ($p = 0$), $R(\omega)$ reduces to the normalized concentration matrix. To estimate the graph topology we compare the L_∞-norms of the entries of $R(\omega)$,

$$\rho_{ij} = \sup_{\omega} |R(\omega)_{ij}|$$

with a given threshold. This thresholding step is similar to thresholding in other sparse methods, for example the thresholded lasso and Dantzig estimators in Lounici (2008).

To simplify the interpretation we will use the same threshold value (10^{-1}) in all the experiments, that is, we remove edge (i, j) from the graph if $\rho_{ij} \le 10^{-1}$.

4.2 Experiment 1

In the first series of experiments we generate AR models with sparse inverse spectra by setting $B_0 = I$ and randomly choosing sparse lower triangular matrices B_k with entries ± 0.5. The random trials are continued until a stable AR model is found. The AR process is then used to generate N samples of the time series. The model dimensions are $n = 20$ and $p = 2$.

4.2.1 TOPOLOGY SELECTION

We first illustrate the basic topology selection method outlined above using the correct model order ($p = 2$). The sample size is $N = 512$.

Figure 2 shows the trade-off curve between the penalty $h_\infty(\mathrm{D}(X))$ and the log-likelihood $\mathcal{L}(X)$. We calculate the inverse spectra (10) for the computed points on the trade-off curve, and apply a threshold to them (as explained above, by setting entries with $\rho_{ij} \le 10^{-1}$ to zero). The resulting topologies are shown in Figure 3. The patterns range from quite dense (small γ) to very sparse (large γ). The sparsity of the densest solution ($\gamma = 10^{-5}$) is identical to the sparsity of the least-squares estimate (i.e., the solution of the equations (15) with C given in (13) or, equivalently, the ML solution of (12) without the sparsity constraints). For each of the nine sparsity patterns, we solve the ML problem subject to sparsity constraints (16). We rank the nine solutions using the $\mathrm{AIC_c}$ and BIC scores defined in (20). Figure 4 shows the two scores and the negative log-likelihood as functions of γ. The models that minimize the $\mathrm{AIC_c}$/BIC scores turn out to be the same in this example (the models for $\gamma = 0.15$) and the corresponding topology is shown in Figure 5 (left). Only seven entries are misclassified (six entries are misclassified as zeros; one as nonzero). The sparsity pattern in the middle is the topology estimated by thresholding the partial coherence spectrum of

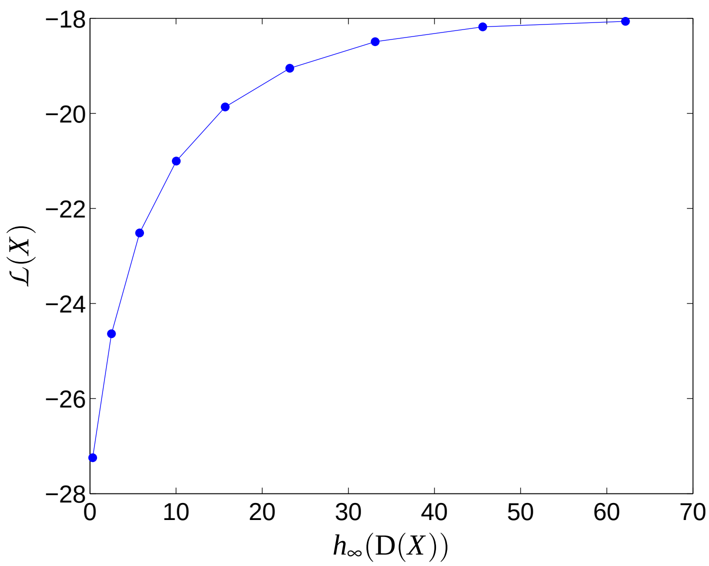


Figure 2: Trade-off curve between the log-likelihood $\mathcal{L}(X)$ and $h_\infty(\mathrm{D}(X))$.

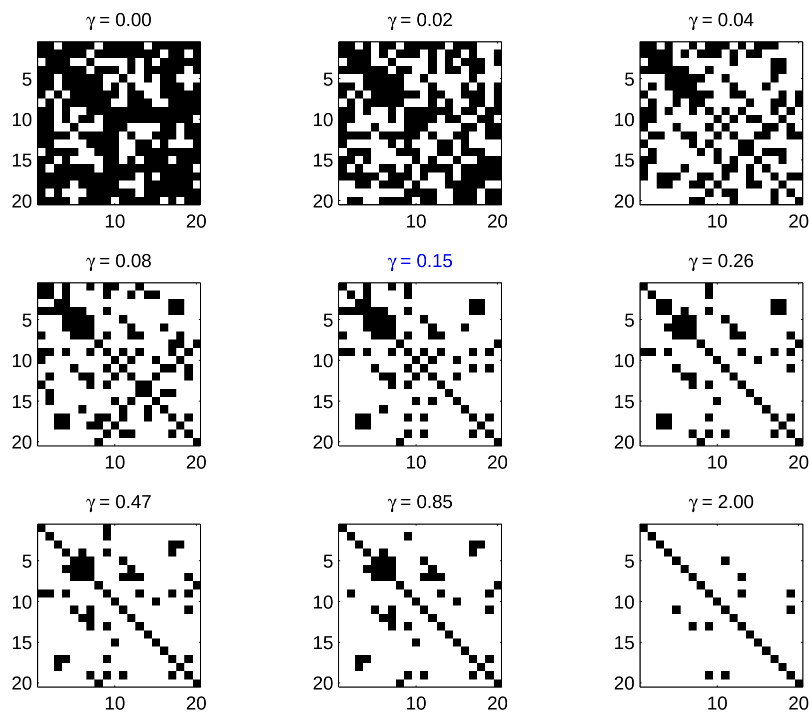


Figure 3: Topologies of solutions along the tradeoff curve in Figure 2 (ordered from right to left on the tradeoff curve).

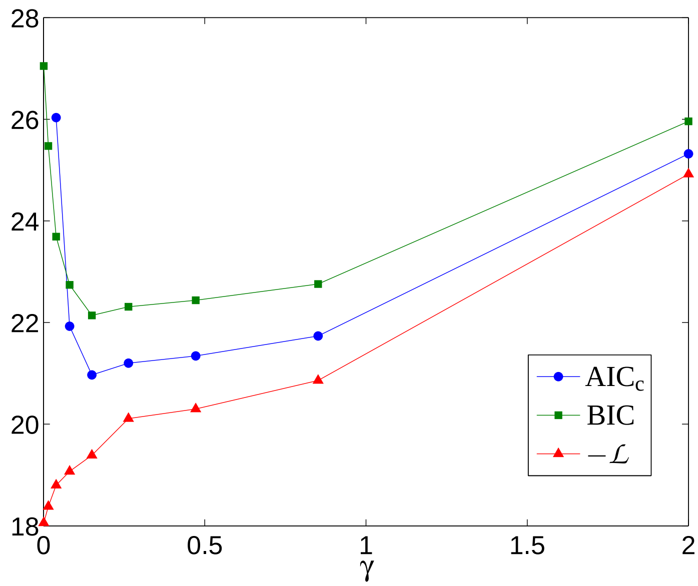


Figure 4: $\mathrm{AIC_c}$ and BIC scores, and maximized log-likelihood for solutions on the trade-off curve in Figure 2.

the least-squares solution with the correct model order ($p = 2$). This pattern is computed by solving the ML problem (12) without constraints, and then thresholding the partial coherence (using the same threshold value 0.1 as in the other experiments). The difference between the two patterns clearly shows the benefits of the nonsmooth regularization for estimating a sparse topology. The sparsity pattern on the right of Figure 5 is obtained from the covariance selection method with ℓ_1-norm regularization (i.e., by setting $p = 0$ in the regularized ML problem (21)) and thresholding the partial coherence. Ignoring the model dynamics substantially increased the error in the topology selection.

4.2.2 COMPARISON WITH OTHER TYPES OF REGULARIZATION

To compare the quality of the sparse models with the models obtained from other estimation methods we evaluate the Kullback-Leibler (KL) divergence (Bach and Jordan, 2004) between the true and the estimated spectra as a function of the sample size N for the following six methods.

1. ML estimation without conditional independence constraints (or least-squares estimate). This is the solution of (12) without the constraints, and it can be computed by solving the normal equations (15).

2. ML estimation with conditional independence constraints determined by thresholding the partial coherence matrix of the least-squares estimate (solution 1).

3. ML estimation with Tikhonov regularization and without conditional independence constraints. Tikhonov regularization (also known as *ridge regression* or ℓ_2-regularization) is widely used in statistics and estimation (Hastie et al., 2009, §3.4). A Tikhonov-regularized ML estimate

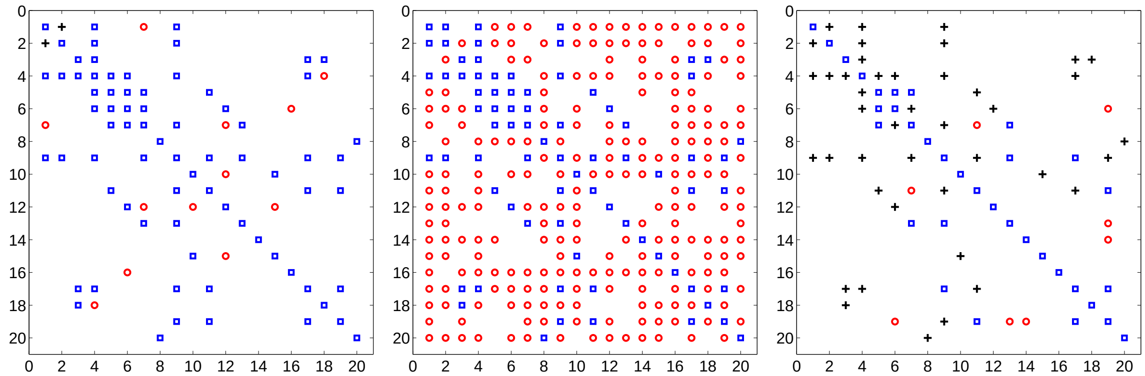


Figure 5: *Left.* The sparsity pattern from the regularized ML problem with $\gamma = 0.15$. *Middle.* The sparsity pattern estimated from the least-squares solution. *Right.* The sparsity pattern from the regularized ML problem for a static model ($p = 0$). The blue squares are the correctly identified nonzero entries (true positives). The red circles are the entries that are misclassified as nonzero (false positives). The black crosses are entries that are misclassified as zeros (false negatives).

is the solution of

$$\text{minimize} \quad -2\log\det B_0 + \mathbf{tr}(CB^T B) + \gamma\|B\|_F^2.$$

The solution can be computed from the normal equations (15) with C replaced by $C+\gamma I$. The solution of this problem can therefore also be viewed as a ML estimate using a perturbed sample covariance matrix $C+\gamma I$. In the experiment, the value of γ is determined by performing a five-fold cross-validation (Hastie et al., 2009, §7.10).

4. ML estimation with conditional independence constraints determined by thresholding the inverse spectral density for the Tikhonov estimate (solution 3).

5. Regularized ML estimation with h_∞-penalty. This is the solution of problem (21) with penalty function (22).

6. ML estimation with conditional independence constraints determined by thresholding the inverse spectral density for the h_∞-regularized ML estimate (solution 5).

The total number of variables in this example is $n(n+1)/2 + pn^2 = 1010$ variables. We show the results in Figure 6 in two different settings: with small sample sizes ($N < 1010$) and with moderate to large sample sizes ($N \geq 1010$). We can note that for small sample sizes N the constrained ML estimates (models 2,4,6) are not better than the unconstrained estimates (models 1,3,5), and much worse in the case of the Tikhonov-regularized estimates. This can be explained by large errors in the estimated topology. For larger N the constrained estimates are consistently better than the unconstrained models, and for very large N the three constrained ML estimates give the same accuracy. For small and moderate N we also see that model 6 (ML estimate for the topology selected via nonsmooth regularization) is much more accurate than the other methods.

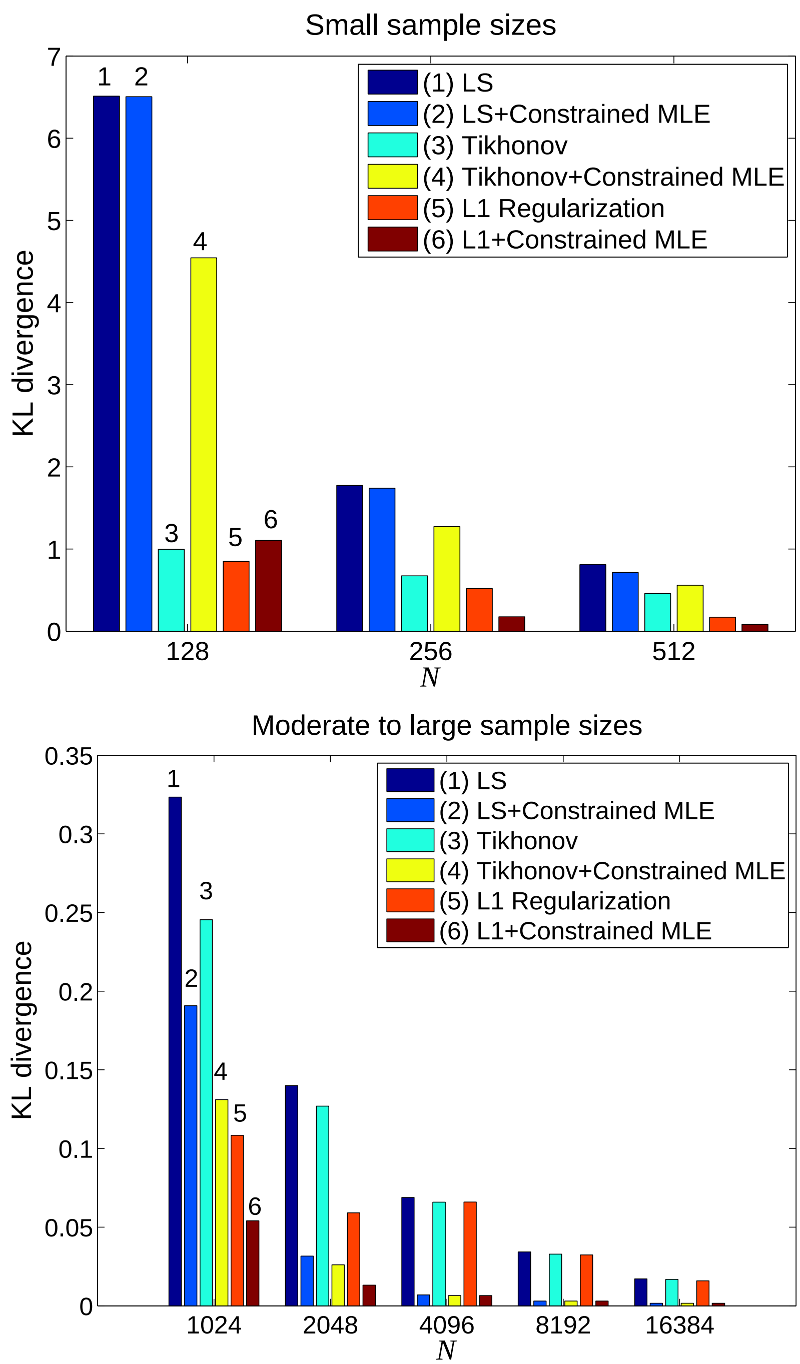


Figure 6: KL divergence between estimated AR models and the true model $(n = 20, p = 2)$ versus the number of samples N. We compare six methods: (1) least-squares estimate, (2) constrained ML estimate with topology estimated by thresholding solution 1, (3) ML estimate with Tikhonov regularization, (4) constrained ML estimate with topology estimated by thresholding solution 3, (5) regularized ML estimate with h_∞-penalty, (6) constrained ML estimate with topology estimated by thresholding solution 5.

4.2.3 ERRORS IN TOPOLOGY AS A FUNCTION OF SAMPLE SIZE

In the last figure (Figure 7) we examine how fast the error in the topology selection decreases with increasing sample length N for three topology selection methods: LS estimation followed by thresholding, ML estimation with Tikhonov regularization followed by thresholding, and ML estimation with nonsmooth regularization followed by thresholding. For each sample size N we show the errors averaged over 50 sample sequences (i.e., 50 different sample covariance matrices C). “False positives” refers to entries that are incorrectly classified as nonzeros (i.e., incorrectly added edges in the graphical model). “False negatives” are entries that are incorrectly classified as zeros (i.e., incorrectly deleted edges). The top graphs in Figure 7 show the fraction of false positives and false negatives versus the sample size. The bottom graphs show the total fraction of misclassified entries. We compare the three methods listed above. As can be seen, the total error in the estimated topology is reduced in the regularized estimates, and the errors decrease more rapidly when we regularize with the sum-of-norms penalty h_∞.

4.3 Experiment 2

In the second experiment we compare different penalty functions h for the regularized ML problem (21): the ‘sum-of-ℓ_∞-norms’ penalty h_∞ defined in (22), the ‘sum-of-ℓ_2-norms’ penalty h_2 defined in (28) with $\alpha = 2$, and the ‘sum-of-ℓ_1-norms’ penalty h_1 defined in (28) with $\alpha = 1$. These penalty functions all yield models with a sparse inverse spectrum

$$S(\omega)^{-1} = Y_0 + \frac{1}{2}\sum_{k=1}^{p}(e^{-\mathrm{j}k\omega}Y_k + e^{\mathrm{j}k\omega}Y_k^T),$$

but have different degrees of sparsity for the entries $(Y_k)_{ij}$ within each group i, j.

The data are generated by randomly choosing sparse coefficients Y_k of an inverse spectrum (10). For each (i, j) of nonzero locations in $S(\omega)^{-1}$, we select random values $(Y_k)_{ij}$ with about the same magnitude for all k. If necessary, a multiple of the identity matrix is added to Y_0 to guarantee the positiveness of the spectrum. An AR realization of the spectrum is then computed by spectral factorization and used to generate sample time series. The model dimensions are $n = 5$, $p = 7$.

Figure 8 shows typical values for the estimated coefficients $(Y_k)_{ij}$. The three penalty functions all give the same topology, but a different sparsity with the same group i, j of coefficients. The sparsity within each group is largest for the h_1-penalty and smallest for the h_∞-penalty.

Table 1 shows the results of topology selection with the three penalties, for sample size $N = 512$ and averaged over 50 sample sequences. The h_∞-penalty gives the models with the smallest KL divergence and smallest error in topology. This is to be expected, given the distribution of the nonzero coefficients $(Y_k)_{ij}$ in the AR models that were used to generate the data. The results also agree with a comparison of different norms in a composite penalty function (Zhao et al., 2009). In general the best choice of norm will depend on how the coefficients are distributed within each group.

5. Applications

This section presents two examples of real data sets to demonstrate how topology selection can facilitate studies of relationship in multivariate time series.

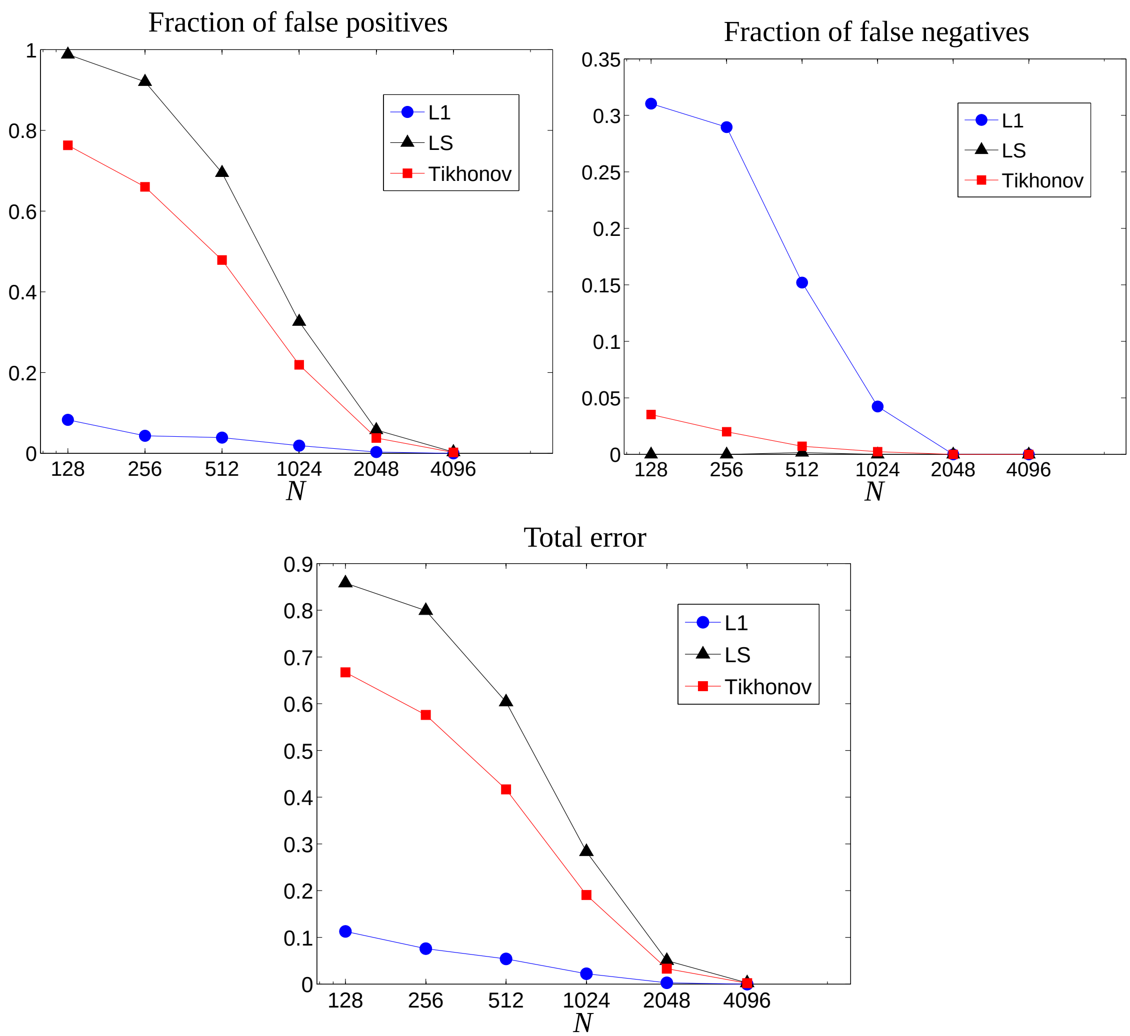


Figure 7: *Top left.* Fraction of incorrectly added edges in the estimated graph (number of upper triangular nonzeros in the estimated pattern that are incorrect, divided by the number of upper triangular zeros in the correct pattern). *Top right.* Fraction of incorrectly removed edges in the estimated graph (number of upper triangular zeros in the estimated pattern that are incorrect, divided by the number of upper triangular nonzeros in the correct pattern). *Bottom.* The combined classification error computed as the sum of the false positives and false negatives divided by the number of upper triangular entries in the pattern.

5.1 Functional Magnetic Resonance Imaging (fMRI) Data

In this section we apply the topology selection method to a functional magnetic resonance imaging (fMRI) time series. We use the StarPlus fMRI data set[1] (Mitchell et al., 2004), which was analyzed

1. StarPlus data can be found at `www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-81/www/`.

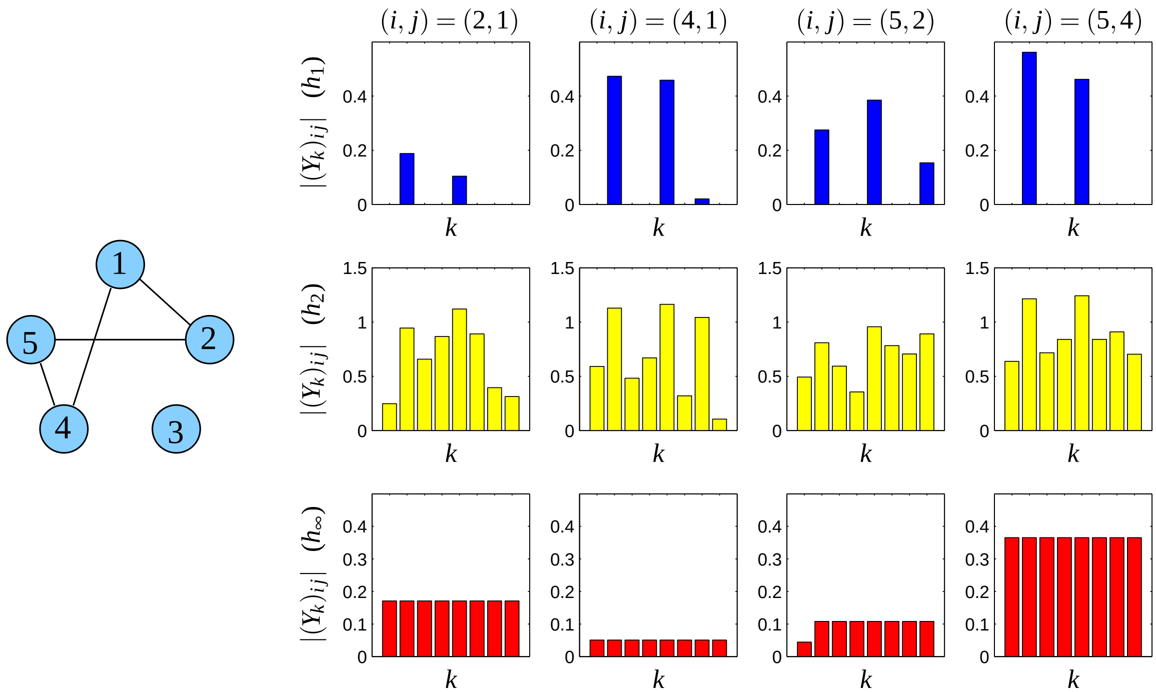


Figure 8: Nonzero coefficients $|(Y_k)_{ij}|$ for regularized ML estimates with penalty h_α, for $\alpha = 1, 2, \infty$.

Dimensions	KL divergence			Error in topology (%)		
	h_1	h_2	h_∞	h_1	h_2	h_∞
$n = 20,\ p = 2$	0.24	0.22	**0.21**	11.8	11.9	11.6
$n = 20,\ p = 4$	0.33	0.24	**0.19**	1.65	1.19	0.51
$n = 30,\ p = 2$	0.40	0.35	**0.30**	9.95	8.83	7.96
$n = 30,\ p = 4$	0.59	0.46	**0.40**	5.18	3.97	3.53

Table 1: Accuracy of topology selection methods with penalty h_α for $\alpha = 1, 2, \infty$. The table shows the average KL divergence with respect to the true model and the average percentage error in the estimated topology (defined as the sum of the false positives and false negatives divided by the number of upper triangular entries in the pattern), averaged over 50 instances.

using covariance selection in Scheinberg and Rish (2009). The data consists of 80 time series (runs) of brain image scans. In half of the 80 runs the input stimulus shown to the subject is a picture; in the other half it is a sentence. Each run contains 16 images, resulting in 640 images for each input. Mitchell et al. (2004) suggest a region of interest (ROI) of 1718 voxels. To reduce the dimension we took averages over groups of voxels in the ROI and considered four reduced graphs with $n = 7$, 50, 100, and 190 nodes, respectively.

We fit two different AR models, one for each input. The AR model orders selected by the BIC are shown in Table 2. As the problem size (n) becomes larger, the BIC tends to pick a static model

Input	$n=7$	$n=50$	$n=100$	$n=190$
Picture	$p=1$	$p=1$	$p=0$	$p=0$
Sentence	$p=1$	$p=1$	$p=0$	$p=0$

Table 2: AR model orders for the fMRI data set.

Input	Static models ($p=0$)			Time series models ($p=1$)		
	ℓ_1	Tikhonov	LS	ℓ_1	Tikhonov	LS
Picture	991	4116	4203	**0**	13467	13465
Sentence	922	4021	4131	**0**	13240	13238

Table 3: Relative BIC scores of six models fitted to two fMRI time series of size $n=50$. The 'static' models are Gaussian graphical models (i.e., AR models of order $p=0$), the time series models are AR models of order $p=1$. The models are constrained ML estimates with topologies estimated using three different methods: Regularized ML estimate with h_α-penalty, Tikhonov-regularized ML estimate, and the least-squares estimate. The BIC scores are relative to the score of the best model (time series models of regularized ML estimate with h_α-penalty).

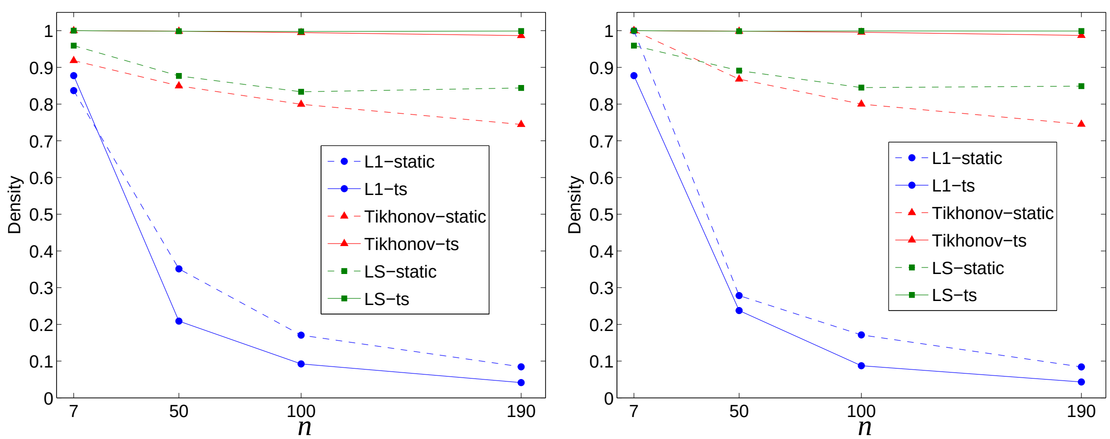


Figure 9: Density of the graphical models of fMRI data for 'picture' stimulus (*left*) and for 'sentence' stimulus (*right*). The density is computed as the number of nonzero entries in the estimated inverse spectrum divided by n^2.

($p=0$). Table 3 shows the BIC scores of different models for the experiment with size $n=50$.

The topologies selected by the BIC are the regularized ML estimates with h_∞-penalty. Figure 9 shows the sparsity of the estimated graphs from the least-squares, Tikhonov-regularized ML, and h_∞-regularized ML methods. The plots show that the h_∞-regularization produces much sparser graphs than the other two methods.

To get an idea of the accuracy of the estimated network structure, we validated the result with a simple classification experiment. For each input we keep one fMRI run as a test problem and use

model order	$n = 7$	$n = 50$	$n = 100$	$n = 190$
$p = 0$	0.21	0.16	**0.11**	**0.06**
$p = 1$	**0.20**	**0.16**	0.16	0.11

Table 4: Classification error of fMRI data versus model size. The error is the number of runs for which the stimulus input is correctly identified divided by the total number of runs (40).

the 39 remaining runs to estimate a sparse AR model. The two models are then used to guess the inputs shown to the subject during the test run. The classification algorithm computes the likelihood of each input, based on the two models, and selects the input with the highest likelihood. We repeat this for each of the 40 choices of test run. Table 4 shows the classification error versus the number of nodes in the graph. We see that the classification is quite successful and achieves an error in the range 6–20%. The error tends to be smaller if we use less averaging (larger n). We also note that for each n, the AR model of order p chosen in Table 2 also performs slightly better in the classification experiment.

5.2 International Stock Market Data

We consider a multivariate time series of 17 stock market indices: the S&P 5000 composite index (U.S.), Toronto stock exchange 300 index (Canada), the All ordinary composite stock index (Australia), the Nikkei 225 stock index (Japan), the Hang Seng stock composite index (Hong Kong), the FTSE 100 share index (United Kingdom), the Frankfurt DAX 30 composite index (German), the CAC 40 stock composite index (France), MIBTEL index (Italy), the Zurich Swiss Market composite index (Switzerland), the Amsterdam exchange index (Netherlands), the Austrian traded index (Austria), IBEX 35 (Spain), BEL 20 (Belgium), the OMX Helsinki 25 index (Finland), the Portugese stock index (Portugal), the Irish stock exchange index (Ireland). The data were stock index closing prices recorded from June 3, 1997 to June 30, 1999 and obtained from `www.globalfinancialdata.com`. The data were converted to US dollars. Missing data due to national holidays were replaced by the most recent values. For each market we use as variable the return between trading day $k-1$ and k, defined as

$$r_k = 100 \log(\pi_k / \pi_{k-1}),$$

where π_k is the closing price on day k. This results in 17-dimensional time series of length 540. Similar time series for a smaller number of markets were analyzed in Bessler and Yang (2003) and Abdelwahab et al. (2008).

We solve the h_∞-regularized ML problem with model orders ranging from $p = 0$ to $p = 3$, and for each value collect the topologies along the trade-off curve, as in the previous examples. The AIC$_c$ and BIC criteria were then used to select a model. Both criteria selected a model of order $p = 1$ and the same sparsity pattern (corresponding to a value $\gamma = 0.34$). Figure 10 (right) shows ρ_{ij}, the maximum magnitude of the partial coherence of the model, and compares it with a thresholded nonparametric estimate obtained with Welch's method (Proakis, 2001) and the constrained ML model with topology obtained by thresholding the least-squares estimate. We note that the graph topologies suggested by the nonparametric and least-squares estimates are much denser than the regularized ML estimate.

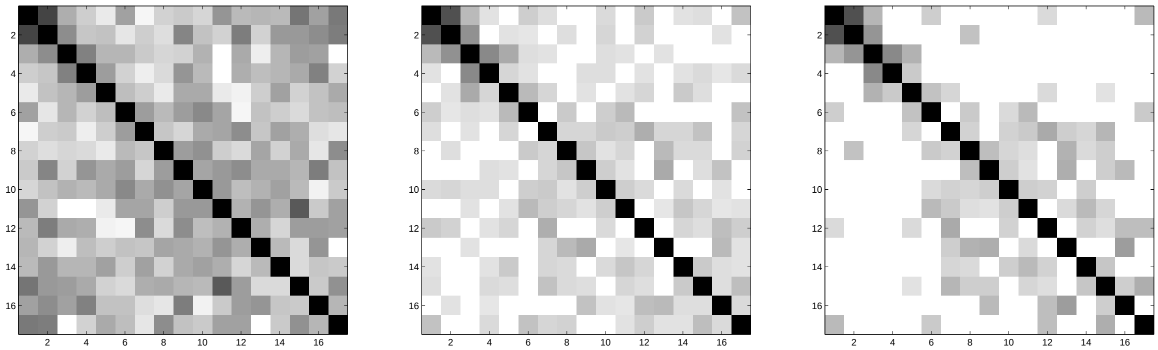


Figure 10: The maximum magnitude ρ_{ij} of the partial coherence for three models of the stock exchange data. *Left:* Thresholded nonparametric sample estimate using Welch's method. *Middle:* Constrained ML estimate with topology determined from the LS solution. *Right:* Constrained ML estimate with topology determined from the h_∞-regularized ML estimate.

Figure 11 shows the graphical model estimated by the h_∞-regularized ML problem. The thickness of the edges is proportional to ρ_{ij}. We recognize many connections that can be explained from geographic proximity or economic ties between the countries. For example, we see strong connections between the U.S. and Canada, between Australia, Japan, and Hong Kong, between Hong Kong and U.K., between the southern European countries, et cetera. Overall the graphical model seems plausible, and the experiment suggests that the topology selection method is quite effective.

6. First-order Optimization Algorithms

In the preceding sections we have considered four convex optimization problems. The constrained ML estimation problem (16) and its dual (17) have differentiable objectives and linear equality and matrix inequality constraints. The regularized ML problem (21) also includes a nondifferentiable term in the objective, and its dual (27) has a differentiable objective but constraints that involve nondifferentiable functions. These optimization problems can be solved by interior-point methods, for example, the path-following methods developed for convex determinant maximization problems (Toh, 1999; Vandenberghe et al., 1998). In practice, however, the problems are often too large for interior-point methods because they involve matrix variables (X or Z) of high dimension. In this section we therefore investigate less expensive first-order algorithms applied to a reformulation of the dual problems (17) and (27). The dual approach avoids several difficulties that arise in first-order methods applied to the primal problems: the complicated constraints in the constrained ML problem (16), the fact that its objective, which is also the first term in the objective of the regularized ML problem (21), is not strictly convex, the nondifferentiability of the penalty term in (21), and, most important, the fact the solution X has low rank and therefore lies on the boundary of the feasible set. (For the regularized ML problem (21), these difficulties could be addressed as in the covariance selection method of Banerjee et al. (2008), by applying Nesterov's fast gradient method to an approximation of the primal problem with a smoothed objective and a closed bounded constraint set (Nesterov, 2005). In our limited experience, with a fixed and conservative choice of

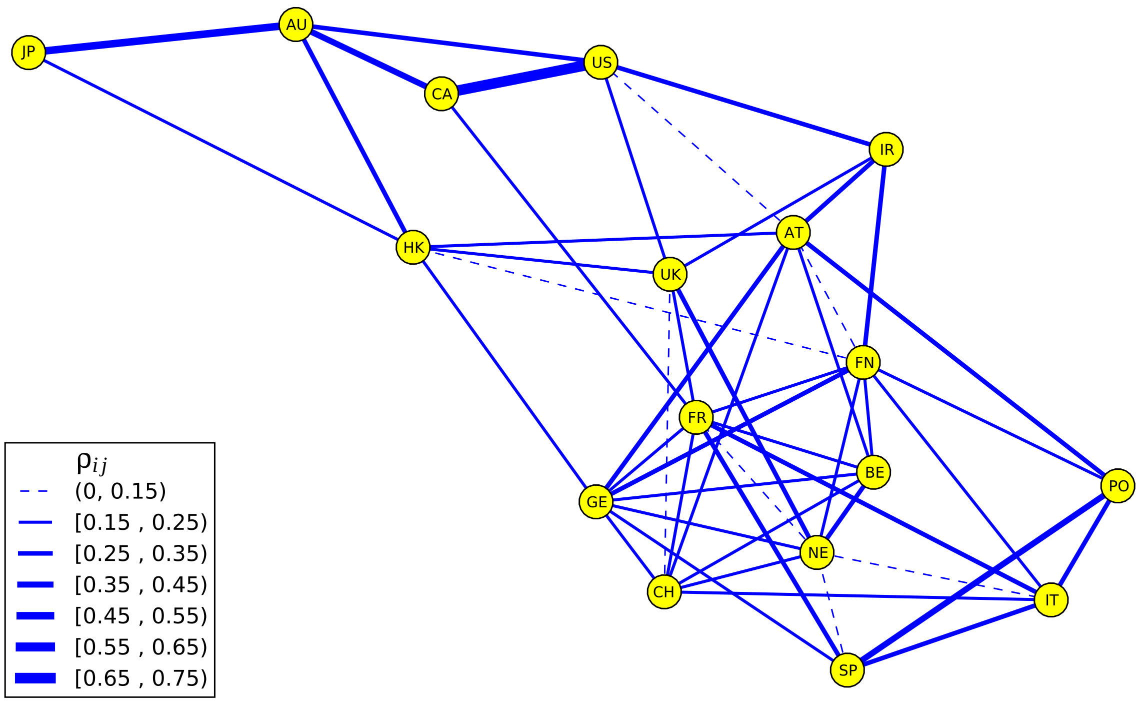


Figure 11: A graphical model of stock market data. The strength of connections is represented by the width of the blue links, which is proportional to $\rho_{ij} = \sup_\omega |R(\omega)_{ij}|$ if it is greater than 0.15.

the smoothing and bounding parameters, this algorithm was slower than the dual gradient projection method described in this section, so we will not pursue it here.)

6.1 Reformulated Dual Problems

To reformulate the dual problems we eliminate the variable W in (17) and (27). Let $V = C + \mathbf{T}(\mathbf{P}(Z))$, respectively, $V = C + \mathbf{T}(Z)$. The inequality

$$V - \begin{bmatrix} W & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} V_{00} - W & V_{1:p,0}^T \\ V_{1:p,0} & V_{1:p,1:p} \end{bmatrix} \succeq 0,$$

is equivalent to

$$V_{1:p,1:p} \succeq 0, \qquad \mathrm{range}(V_{1:p,0}) \subseteq \mathrm{range}(V_{1:p,1:p}), \qquad V_{00} - V_{1:p,0}^T V_{1:p,1:p}^\dagger V_{1:p,0} \succeq W, \tag{32}$$

where $V_{1:p,1:p}^\dagger$ is the pseudo-inverse of $V_{1:p,1:p}$. If $V \succeq 0$, then the matrix W with maximum determinant that satisfies (32) is equal to $V_{00} - V_{1:p,0}^T V_{1:p,1:p}^\dagger V_{1:p,0}$, the *Schur complement* of $V_{1:p,1:p}$ in V. This observation allows us to eliminate W from (17) and (27). Problem (17) can be written as an unconstrained problem

$$\text{maximize} \quad -\phi(C + \mathbf{T}(\mathbf{P}(Z))), \tag{33}$$

and problem (27) as a problem with simple constraints

$$\begin{array}{ll} \text{maximize} & -\phi(C+\mathrm{T}(Z)) \\ \text{subject to} & \sum_{k=0}^{p}(|(Z_k)_{ij}|+|(Z_k)_{ji}|) \leq \gamma, \quad i \neq j \\ & \mathbf{diag}(Z_k) = 0, \quad k=0,\ldots,p. \end{array} \tag{34}$$

Here $\phi : \mathbf{S}^{n(p+1)} \to \mathbf{R}$ is defined as

$$\phi(V) = -\log\det\left(V_{00} - V_{1:p,0}^T V_{1:p,1:p}^{\dagger} V_{1:p,0}\right) - n,$$

with domain $\mathbf{dom}\,\phi = \{V \in \mathbf{S}_+^{n(p+1)} \mid V_{00} - V_{1:p,0}^T V_{1:p,1:p}^{\dagger} V_{1:p,0} \succ 0\}$. This function is convex, since it can be expressed as

$$\phi(V) = \inf\left\{-\log\det W \;\middle|\; \begin{bmatrix} W & 0 \\ 0 & 0 \end{bmatrix} \preceq V \right\} - n,$$

and convexity of this expression follows from results in convex analysis (Boyd and Vandenberghe, 2004, §3.2.5). It is also a smooth function on the interior of its domain and its gradient at a positive definite V can be expressed as

$$\nabla\phi(V) = -V^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & V_{1:p,1:p}^{-1} \end{bmatrix}. \tag{35}$$

This can be seen, for example, from the identity $\det V = \det V_{1:p,1:p} \det(V_{00} - V_{1:p,0}^T V_{1:p,1:p}^{-1} V_{1:p,0})$, which gives $\phi(V) = -\log\det V + \log\det V_{1:p,1:p} - n$, and the fact that the gradient of $\log\det X$ is X^{-1}.

If $V = C+\mathrm{T}(\mathrm{P}(Z)) \succ 0$ at the optimum of (33) then the primal optimal solution can be computed from Z via the expressions

$$X = V^{-1} - \begin{bmatrix} 0 & 0 \\ 0 & V_{1:p,1:p}^{-1} \end{bmatrix} = \begin{bmatrix} -I \\ V_{1:p,1:p}^{-1} V_{1:p,0} \end{bmatrix} W^{-1} \begin{bmatrix} -I \\ V_{1:p,1:p}^{-1} V_{1:p,0} \end{bmatrix}^T \tag{36}$$

where $V = C+\mathrm{T}(\mathrm{P}(Z))$ and $W = V_{00} - V_{1:p,0}^T V_{1:p,1:p}^{-1} V_{1:p,0}$. The expression for X follows from the optimality condition (18) and the identities

$$V = \begin{bmatrix} V_{00} - V_{1:p,0}^T V_{1:p,1:p}^{-1} V_{1:p,0} & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} V_{1:p,0}^T V_{1:p,1:p}^{-1} \\ I \end{bmatrix} V_{1:p,1:p} \begin{bmatrix} V_{1:p,0}^T V_{1:p,1:p}^{-1} \\ I \end{bmatrix}^T,$$

$$V^{-1} = \begin{bmatrix} 0 & 0 \\ 0 & V_{1:p,1:p}^{-1} \end{bmatrix} + \begin{bmatrix} -I \\ V_{1:p,1:p}^{-1} V_{1:p,0} \end{bmatrix} (V_{00} - V_{1:p,0}^T V_{1:p,1:p}^{-1} V_{1:p,0})^{-1} \begin{bmatrix} -I \\ V_{1:p,1:p}^{-1} V_{1:p,0} \end{bmatrix}^T. \tag{37}$$

The formula for V^{-1} also provides an alternative form of the gradient (35).

Similarly, if $C+\mathrm{T}(Z) \succ 0$ at the optimum of (34) then the primal optimal X can be computed from (36) with $V = C+\mathrm{T}(Z)$.

The reformulated dual problems are interesting because they can often be solved by gradient algorithms for unconstrained optimization or gradient projection algorithms for problems with simple constraints. To explain this, we again distinguish between Toeplitz and non-Toeplitz C. If C is block-Toeplitz, then it can be shown that the functions $\phi(C+\mathrm{T}(\mathrm{P}(Z)))$ and $\phi(C+\mathrm{T}(Z))$ are *closed* convex functions (i.e., with closed sublevel sets) and that their domains are open. Consider the function ϕ restricted to the set of block-Toeplitz matrices, that is, $\phi(\mathrm{T}(R))$, where $R \in \mathbf{M}^{n,p}$. By definition, R is in the domain of $\phi(\mathrm{T}(R))$ if $\mathrm{T}(R) \succeq 0$ and there exists a positive definite W with

$$\mathrm{T}(R) \succeq \begin{bmatrix} W & 0 \\ 0 & 0 \end{bmatrix}.$$

From the property of block-Toeplitz matrices mentioned in Section 2.3, this implies $\mathrm{T}(R) \succ 0$. In other words, the domain of $\phi(\mathrm{T}(R))$ is the open set $\{R \mid \mathrm{T}(R) \succ 0\}$. By a similar argument, if a sequence of matrices R in the domain of $\phi(\mathrm{T}(R))$ converges to a point $\bar{R}$ in the boundary of the domain, then the Schur complement of $\mathrm{T}(\bar{R})_{1:p,1:p}$ in $\mathrm{T}(\bar{R})$ must be singular, and hence $\phi(\mathrm{T}(R)) \to \infty$. For a continuous function with an open domain this is equivalent to closedness (Boyd and Vandenberghe, 2004, p.639).

If C is not block-Toeplitz, then the functions $\phi(C+\mathrm{T}(\mathrm{P}(Z)))$ and $\phi(C+\mathrm{T}(Z))$ are not necessarily closed, and their domains not necessarily open. One implication is that it is possible that the optimal solution of (33) or (34) is at a point in the boundary of the domain of the cost function, that is, a point where $C+\mathrm{T}(\mathrm{P}(Z))$ or $C+\mathrm{T}(Z)$ are singular. However in practice, C is usually approximately block-Toeplitz and one can expect that the functions are often closed. Moreover, in order to apply unconstrained minimization algorithms it is sufficient that the algorithm is started at a point $Z^{(0)}$ for which the sublevel set $\{Z \mid \phi(C+\mathrm{T}(\mathrm{P}(Z))) \le \phi(C+\mathrm{T}(\mathrm{P}(Z^{(0)})))\}$ is closed. This condition is considerably weaker than the requirement that all sublevel sets are closed.

6.2 Gradient Projection Algorithms

We now present some details on first-order algorithms for the reformulated dual problems. We focus on the constrained problem (34) since the unconstrained problem (33) can be handled as a special case. We first describe a version of the classical gradient projection with a backtracking line search (Polyak, 1987; Bertsekas, 1999). To simplify the notation we will use a generic problem format

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in \mathcal{C} \end{array}$$

where $f : \mathbf{R}^n \to \mathbf{R}$ is convex and continuously differentiable with an open domain, and $\mathcal{C}$ is a closed convex set. We assume that a feasible point $x^{(0)}$ is known and that the sublevel set

$$\mathcal{S} = \{x \in \mathbf{dom}\, f \cap \mathcal{C} \mid f(x) \le f(x^{(0)})\}$$

is closed and bounded. The closedness assumption is satisfied if f is a closed function. (See the previous paragraph on the validity of this assumption for problems (33) and (34).) We assume that projections on $\mathcal{C}$ are inexpensive and we denote the projection operator by $\mathcal{P}$:

$$\mathcal{P}(y) = \operatorname*{argmin}_{x \in \mathcal{C}} \|x-y\|_2.$$

The *gradient map* associated with f and $\mathcal{C}$ is defined as

$$G_t(x) = \frac{1}{t}\left(x - \mathcal{P}(x - t\nabla f(x))\right)$$

for $t > 0$. For an unconstrained problem, the gradient map is $G_t(x) = \nabla f(x)$, independent of t.

6.2.1 BASIC GRADIENT PROJECTION

The basic gradient projection method starts at $x^{(0)}$ and continues the iteration

$$\begin{aligned} x^{(k)} &= \mathcal{P}\left(x^{(k-1)} - t_k \nabla f(x^{(k-1)})\right) \\ &= x^{(k-1)} - t_k G_{t_k}(x^{(k-1)}) \end{aligned} \tag{38}$$

until a stopping criterion is satisfied. A classical convergence result states that $x^{(k)}$ converges to an optimal solution if t_k is fixed and equal to $1/L$, where L is a constant that satisfies

$$\|\nabla f(u) - \nabla f(v)\|_2 \leq L\|u - v\|_2 \quad \forall u, v \in \mathcal{S}, \tag{39}$$

(Polyak, 1987, §7.2.1). Although our assumptions ($\mathcal{S}$ is closed and bounded, and **dom** f is open) imply that the Lipschitz condition (39) holds for some constant $L > 0$, its value is not known in practice, so the fixed step size rule $t_k = 1/L$ cannot be used. We therefore determine t_k using a backtracking search (Beck and Teboulle, 2009). The step size search algorithm in iteration k starts at a value $t_k := \overline{t_k}$ where

$$\overline{t_k} = \min\left\{\frac{s^T s}{s^T y}, t_{\max}\right\}, \tag{40}$$

where

$$s = x^{(k-1)} - x^{(k-2)}, \qquad y = \nabla f(x^{(k-1)}) - \nabla f(x^{(k-2)}),$$

and $t_{\max}$ is a positive constant. (In the first iteration we initialize the step size as $t_1 = t_{\max}$.) The search then repeats the update $t_k := \beta t_k$ (where $\beta \in (0,1)$ is an algorithm parameter) until $x^{(k-1)} - t_k G_{t_k}(x^{(k-1)}) \in$ **dom** f and

$$f(x^{(k-1)} - t_k G_{t_k}(x^{(k-1)})) \leq f(x^{(k-1)}) - t_k \nabla f(x^{(k-1)})^T G_{t_k}(x^{(k-1)}) + \frac{t_k}{2}\|G_{t_k}(x^{(k-1)})\|_2^2. \tag{41}$$

The resulting step size t_k is used in the update to $x^{(k)}$ in (38). Note that the trial points

$$x^{(k-1)} - t_k G_{t_k}(x^{(k-1)}) = \mathcal{P}\left(x^{(k-1)} - t_k \nabla f(x^{(k-1)})\right)$$

generated during the step size search are not necessarily on a straight line. The trajectory is sometimes referred to as the *projection arc* (Bertsekas, 1999, §2.3).

The step length $\|s\|_2^2/s^T y$ is known as the *Barzilai-Borwein* step size and forms the basis of *spectral gradient* methods (Barzilai and Borwein, 1988; Birgin et al., 2003; Figueiredo et al., 2007; Wright et al., 2009). It can be motivated by the easily established fact that $\|s\|_2^2/s^T y \geq 1/L$ if f satisfies (39), so it is a readily computed upper bound for $1/L$.

6.2.2 VARIATIONS

The basic gradient projection method can be varied in several ways, some of which will be compared in the numerical experiments below. To avoid computing a projection for each trial step size t_k in the step size search, we can replace the gradient update with

$$x^{(k)} = x^{(k-1)} - t_k G_{\overline{t_k}}(x^{(k-1)}) \tag{42}$$

where $\bar{t_k}$ is held fixed at the value (40) and t_k is determined by a backtracking search: we take $t_k := \bar{t_k}$ and then backtrack ($t_k := \beta t_k$) until $x^{(k-1)} - t_k G_{\bar{t_k}}(x^{(k-1)}) \in \mathbf{dom}\, f$ and

$$f(x^{(k-1)} - t_k G_{\bar{t_k}}(x^{(k-1)})) \leq f(x^{(k-1)}) - t_k \nabla f(x^{(k-1)})^T G_{\bar{t_k}}(x^{(k-1)}) + \frac{t_k}{2} \|G_{\bar{t_k}}(x^{(k-1)})\|_2^2. \tag{43}$$

In this method the trial points during the step size selection follow a straight line, and each step only requires a function evaluation.

Many alternatives to the step size rules (38) and (42) are available in the literature, for example, the Armijo rule (Bertsekas, 1999, §2.3), and conditions that allow non-monotone convergence (Birgin et al., 2000; Lu and Zhang, 2009). In our experiments these variations gave similar results as the step size rules outlined above.

Another attractive class of gradient projection algorithms are the optimal first-order methods originated by Nesterov (Nesterov, 2004; Tseng, 2008; Beck and Teboulle, 2009). For functions whose gradient is Lipschitz continuous on $\mathcal{C}$, these algorithms have a better complexity than the classical gradient projection method (at most $O(\sqrt{1/\epsilon})$ iterations are needed to reach an accuracy ϵ, as opposed to $O(1/\epsilon)$ for the gradient projection method). These theoretical complexity results are valid if a constant step size $t_k = 1/L$ is used where L is the Lipschitz constant for the gradient, or if the step sizes form an nonincreasing sequence ($t_{k+1} \leq t_k$) determined by a backtracking line search (Beck and Teboulle, 2009; Tseng, 2008). The assumption that the gradient is Lipschitz continuous on $\mathcal{C}$ does not hold for the problem considered here, and it is not clear if the convergence analysis can be extended to the case when the gradient is Lipschitz continuous only on the initial sublevel set. Nevertheless, an implementation with a backtracking line search worked well in our experiments (see next section).

6.2.3 Implementation Details

The most important steps in the gradient projection algorithms applied to (33) are the evaluations of the gradient of the objective function and the projections on the set defined by the constraints. We now explain these two steps and the stopping criterion in more detail.

The gradient (35) of ϕ at a point V can be evaluated from a Cholesky factorization $V = L^T L$ with L lower triangular. If we partition L as

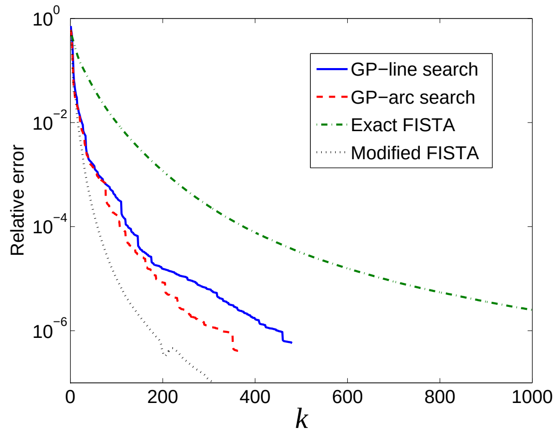
$$L = \begin{bmatrix} L_{00} & 0 \\ L_{1:p,0} & L_{1:p,1:p} \end{bmatrix}$$

then

$$\nabla\phi(V) = \begin{bmatrix} I \\ -L_{1:p,1:p}^{-1} L_{1:p,0} \end{bmatrix} L_{00}^{-1} L_{00}^{-T} \begin{bmatrix} I \\ -L_{1:p,1:p}^{-1} L_{1:p,0} \end{bmatrix}^T.$$

The projection $\mathcal{P}(U)$ of a matrix $U \in \mathbf{M}^{p,n}$ on the set defined by the constraints in (34) can be efficiently computed as follows. Clearly, the diagonal entries of $\mathcal{P}(U)_k$ are zero for $k = 0, \ldots, p$. To find the off-diagonal entries we can solve an independent problem

$$\begin{array}{ll} \text{minimize} & 2((Z_0)_{ij} - (U_0)_{ij})^2 + \sum_{k=1}^{p} \left(((Z_k)_{ij} - (U_k)_{ij})^2 + ((Z_k)_{ji} - (U_k)_{ji})^2 \right) \\ \text{subject to} & \sum_{k=0}^{p} (|(Z_k)_{ij}| + |(Z_k)_{ji}|) \leq \gamma \end{array}$$



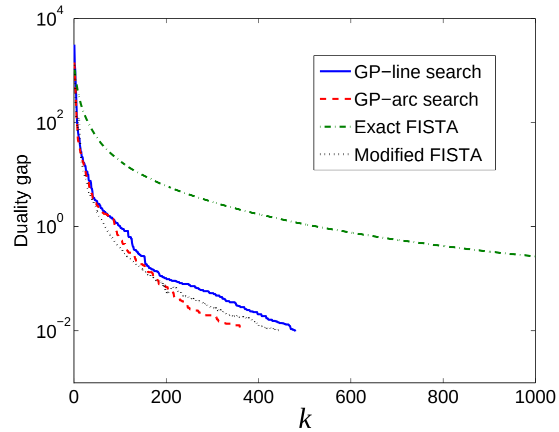


Figure 12: Convergence of gradient projection algorithms. *Left:* Relative error $(f(\mathrm{Z}^{(k)}) - f^\star)/|f^\star|$ versus the number of iterations. *Right:* Duality gap versus the number of iterations.

for each i, j with $j > i$. This is the problem of projecting a vector on the ℓ_1-norm ball. The solution is easily derived from duality and can be calculated by applying to the entries $(U_k)_{ij}$ the shrinkage operation familiar in sparse optimization (see, for example, Tibshirani, 1996).

The following stopping criterion will be used in the experiments. At each iteration, we compute X in (36) from the current iterate Z. This matrix X is primal feasible, as can be seen from the identity (37) and the fact that $C + \mathrm{T}(Z) \succ 0$. By taking the Schur complement of $(C + \mathrm{T}(Z))_{1:p,1:p}$ we also find a dual feasible W in (27). The duality gap between this primal feasible X and the dual feasible Z, W is

$$\begin{aligned}
\eta &= -\log\det X_{00} + \mathbf{tr}(CX) + \gamma h(\mathrm{D}(X)) - \log\det W - n \\
&= \mathbf{tr}(CX) - n + \gamma h(\mathrm{D}(X)) \\
&= \mathbf{tr}((C + \mathrm{T}(Z))X) - n - \mathbf{tr}(X\,\mathrm{T}(Z)) + \gamma h(\mathrm{D}(X)) \\
&= -\mathbf{tr}(X\,\mathrm{T}(Z)) + \gamma h(\mathrm{D}(X)).
\end{aligned}$$

We terminate when the duality gap is below a given tolerance.

6.3 Numerical Example

We generate AR models as in the experiment described in Section 4.2. In the first experiment, the model dimensions are $n = 300$, $p = 2$, $N = 2n(p+1)$. The true inverse spectrum has 10428 non-zero entries in the upper triangular part (a density of about 12%). The penalty parameter γ is set at $\gamma = 0.1$. The variable Z in the reformulated dual problem (34) is a matrix in $\mathbf{M}^{300,2}$, so the problem has $n(n+1)/2 + pn^2 = 225150$ optimization variables. We start the gradient projection algorithm at a strictly feasible $\mathrm{Z}^{(0)} = 0$, and terminate when the duality gap is below 10^{-2} (the optimal value is on the order of hundreds).

Figure 12 shows the relative error $(f(\mathrm{Z}^{(k)}) - f^\star)/|f^\star|$ where $f(\mathrm{Z}) = \phi(C + \mathrm{T}(\mathrm{P}(\mathrm{Z})))$ and $f^\star$ is the optimal value. It also shows the duality gap $\eta^{(k)}$ versus the iteration number for a typical instance. 'GP with arc search' refers to the gradient projection method (38) with step size rule (41).

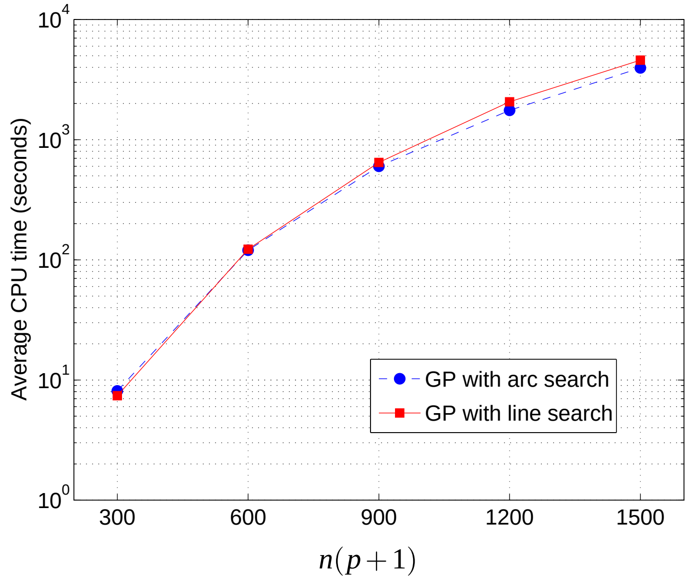


Figure 13: Average CPU times (averaged over 10 runs) of the gradient projection algorithm versus the problem size. The algorithm stops when the duality gap is less than 10^{-1}. The red squares correspond to ‘GP with line search’ and the blue squares correspond to ‘GP with arc search’.

‘GP with line search’ refers to the gradient projection method (42) with step size rule (43). The step size searches required at most 15 backtracking steps to find an acceptable step size. As can be seen, a solution with a moderate accuracy (relative error in the range 10^{-4}–10^{-3}) is obtained after a number of iterations that is only a fraction of the problem size. The convergence of the ‘arc search’ method is slightly faster, but it should be kept in mind that this method is more expensive than the ‘line search’.

The ‘Exact FISTA’ method is the gradient projection algorithm with backtracking line search from Beck and Teboulle (2009) using monotonically decreasing step sizes ($t_k \leq t_{k-1}$, as required by the theory in Beck and Teboulle 2009). As can be seen the convergence was not faster than the classical gradient projection method. A heuristic modification in which the step sizes are not forced to be nonincreasing, but at each iteration the line searche is initialized at the Barzilai and Borwein steplength (40), was often about five times faster. This algorithm is referred to as ‘Modified FISTA’ in the figure.

Figure 13 shows the CPU time versus problem size on a 3GHz Intel Pentium(R) 4 processor with 2.94 GB of RAM, for the ‘GP with arc search’ and ‘GP with line search’ algorithms. The test problems are generated as in the previous experiment, with $p = 2$ and varying n. The algorithms stop when it achieves a duality gap less than $\varepsilon = 0.1$. This yields a solution with a moderate accuracy (relative gap in the range 10^{-4}–10^{-3}). The plot shows that the times needed to solving the regularized ML estimation using both algorithms are fairly comparable with a slight advantage for ‘GP with arc search’ when n is large. Although the backtracking steps in the arc search method are more expensive, the gradient projection method with this step size selection required fewer iterations in most cases.

7. Conclusion

We have presented a convex optimization method for topology selection in graphical models of autoregressive Gaussian processes. The method is based on augmenting the maximum likelihood estimation problem with an ℓ_1-type penalty function, chosen to promote sparsity in the inverse spectrum. By tracing the trade-off curve between the log-likelihood and the penalty function, we obtain a small set of sparse graph topologies, that can then be ranked according to information-theoretic criteria such as the AIC or BIC. This procedure avoids the combinatorial complexity of enumerating all possible topologies, and produces accurate results for smaller sample sizes than methods based on empirical or least-squares estimates. To solve the large, nonsmooth convex optimization problems that result from this formulation, we have investigated a gradient projection method applied to a reformulated dual problem. Experiments with randomly generated examples, and an analysis of an fMRI time series and a time series of international stock market indices were included to confirm the effectiveness of this approach.

Acknowledgments

The authors thank Zhaosong Lu for interesting discussions on algorithms for the penalized ML problem. This research was supported by NSF under grant ECCS-0824003 and by a Royal Thai government scholarship.

References

A. Abdelwahab, O. Amor, and T. Abdelwahed. The analysis of the interdependence structure in international financial markets by graphical models. *International Research Journal of Finance and Economics*, 15:291–306, 2008.

N. Bani Asadi, I. Rish, K. Scheinberg, D. Kanevsky, and B. Ramabhadran. A MAP approach to learning sparse Gaussian markov networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1721–1724, 2009.

F. R. Bach and M. I. Jordan. Learning graphical models for stationary time series. *IEEE Transactions on Signal Processing*, 32:2189–2199, 2004.

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*, 9:485–516, 2008.

J. Barzilai and J. M. Borwein. Two-point step size gradient methods. *IMA Journal of Numerical Analysis*, 8:141–148, 1988.

A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

D. P. Bertsekas. *Nonlinear Programming*. Athena Scientific, Nashua, New Hampshire, second edition, 1999.

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.

E. G. Birgin, J. M. Martínez, and M. Raydan. Nonmonotone spectral projected gradient methods on convex sets. *SIAM J. on Optimization*, 10(4):1196–1211, 2000.

E. G. Birgin, J. M. Martínez, and M. Raydan. Inexact spectral projected gradient methods on convex sets. *IMA Journal of Numerical Analysis*, 23:539–559, 2003.

S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, 2004. `www.stanford.edu/~boyd/cvxbook`.

D. R. Brillinger. *Time series: Data analysis and theory*. Holden-Day, San Francisco, California, expanded edition, 1981.

J. Dahl, V. Roychowdhury, and L. Vandenberghe. Maximum-likelihood estimation of multivariate normal graphical models: large-scale numerical implementation and topology selection. Technical report, Electrical Engineering Department, UCLA, 2005.

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

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

A. P. Dempster. Covariance selection. *Biometrics*, 28:157–175, 1972.

J. Duchi, S. Gould, and D. Koller. Projected subgradient methods for learning sparse Gaussians. In *Proceeding of the Conference on Uncertainty in AI*, 2008.

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

M. Eichler. Testing nonparametric and semiparametric hypotheses in vector stationary processes. *Journal of Multivariate Analysis*, 99(5):968–1009, 2008.

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

S. Feiler, K.G. Müller, A. Müller, R. Dahlhaus, and W. Eich. Using interaction graphs for analysing the therapy process. *Psychother Psychosom*, 74(2):93–99, 2005.

M. A. T. Figueiredo, R. D. Nowak, and S. J. Wright. Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems. *IEEE Journal of Selected Topics in Signal Processing*, 1(4):586–597, 2007.

R. Fried and V. Didelez. Decomposability and selection of graphical models for multivariate time series. *Biometrika*, 90(2):251–267, 2003.

J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432, 2008.

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.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer, New York, 2nd edition, 2009.

J. Z. Huang, N. Liu, M. Pourahmadi, and L. Liu. Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika*, 93(1):85–98, 2006.

Y. Kim, J. Kim, and Y. Kim. Blockwise sparse regression. *Statistica Sinica*, 16(2):375–390, 2006.

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

K. Lounici. Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators. *Electronic Journal of Statistics*, 2:90–102, 2008.

Z. Lu. Smooth optimization approach for sparse covariance selection. *SIAM Journal on Optimization*, 19:1807, 2009.

Z. Lu. Adaptive first-order methods for general sparse inverse covariance selection. *SIAM Journal on Matrix Analysis and Applications*, 2010. forthcoming.

Z. Lu and Y. Zhang. An augmented lagrangian approach for sparse principal component analysis. 2009. Submitted.

N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the Lasso. *Annals of Statistics*, 34(3):1436–1462, 2006.

T.M. Mitchell, R. Hutchinson, R.S. Niculescu, F. Pereira, and X. Wang. Learning to decode cognitive states from brain images. *Machine Learning*, 57(1):145–175, 2004.

Y. Nesterov. *Introductory Lectures on Convex Optimization*. Kluwer Academic Publishers, Dordrecht, The Netherlands, 2004.

Y. Nesterov. Smooth minimization of non-smooth functions. *Mathematical Programming Series A*, 103:127–152, 2005.

B. T. Polyak. *Introduction to Optimization*. Optimization Software, Inc., New York, 1987.

J. Proakis. *Digital Communications*. McGraw-Hill, New York, fourth edition, 2001.

R. Ravikumar, M. J. Wainwright, G. Raskutti, and B. Yu. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence, 2008. URL `http://www.citebase.org/abstract?id=oai:arXiv.org:0811.3628`. Available from `arxiv.org/abs/0811.3628`.

A. J. Rothman, P. J. Bickel, E. Levina, and J. Zhu. Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515, 2008.

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.

K. Scheinberg and I. Rish. SINCO - a greedy coordinate ascent method for sparse inverse covariance selection problem. 2009. Submitted.

J. Songsiri, J. Dahl, and L. Vandenberghe. Graphical models of autoregressive processes. In Y. Eldar and D. Palomar, editors, *Convex Optimization in Signal Processing and Communications*. Cambridge University Press, Cambridge, 2009.

P. Stoica and R. L. Moses. *Introduction to Spectral Analysis*. Prentice Hall, London, 1997.

R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

J. Timmer, M. Lauk, S. Häußler, V. Radt, B. Köster, B. Hellwig, B. Guschlbauer, C.H. Lücking, M. Eichler, and G. Deuschl. Cross-spectral analysis of tremor time series. *Internation Journal of Bifurcation and Chaos in applied Sciences and Engineering*, 10(11):2595–2610, 2000.

K.-C. Toh. Primal-dual path-following algorithms for determinant maximization problems with linear matrix inequalities. *Computational Optimization and Applications*, 14:309–330, 1999.

P. Tseng. On accelerated proximal gradient methods for convex-concave optimization. *SIAM Journal on Optimization*, 2008. Submitted.

L. Vandenberghe, S. Boyd, and S.-P. Wu. Determinant maximization with linear matrix inequality constraints. *SIAM J. on Matrix Analysis and Applications*, 19(2):499–533, April 1998.

S.J. Wright, R.D. Nowak, and M.A.T. Figueiredo. Sparse reconstruction by separable approximation. *IEEE Transactions on Signal Processing*, 57(7):2479–2493, 2009.

M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B Statistical Methodology*, 68(1):49–67, 2006.

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

P. Zhao, G. Rocha, and B. Yu. The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics*, 37(6A):3468–3497, 2009.