

Topology selection in graphical models of autoregressive processes

Jitkomut Songsiri*[†]

Lieven Vandenberghe*

Abstract

An algorithm is presented for topology selection in graphical models of autoregressive time series. The topology of the graphical 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 maximum conditional 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: topology selection, graphical models, convex optimization, time series

1 Introduction

We consider graphical models of autoregressive (AR) Gaussian processes

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

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 independent*. 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 [Bri81, Dah00]. 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 [SDV09] 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

*Department of Electrical Engineering, University of California, Los Angeles. Email: jitkomut@ee.ucla.edu, vandenbe@ee.ucla.edu. Research supported by NSF under grants ECS-0524663 and ECCS-0824003, and by a Royal Thai government scholarship.

[†]Corresponding author

based on solving the convex optimization problem

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

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 $Y_k \in \mathbf{R}^n$, $k = 1, 2, \dots, p$. 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 [SDV09]. 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; see [Eic06, SDV09]. However this combinatorial approach is clearly limited to small graphs (say, $n \leq 6$). 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 [DES97, EDS03, SSSB05, GIF02, TLH⁺00, FMM⁺05, FD03]. A common approach is to formulate hypothesis testing problems to decide about the presence or absence of edges. Dahlhaus [Dah00] derives a statistical test for the existence of an edge in the graph, based on the maximum of a nonparametric estimate of normalized inverse spectrum; see also [DES97, EDS03, SSSB05, GIF02, TLH⁺00, FMM⁺05, FD03]. Eichler [Eic08] 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 [BJ04]. 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

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

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, *i.e.*, the problem of computing the ML estimate of the inverse covariance matrix $X = \Sigma^{-1}$ of a multivariate Gaussian variable $N(0, \Sigma)$, subject to conditional independence constraints (which, for a normal distribution, correspond to zeros in the inverse covariance) [Dem72], [Lau96, Section 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 terms, *i.e.*, on solving

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

(see [DRV04, MB06, BEd08, Lu09]). 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, El Ghaoui, and d’Aspremont [BEd08] 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 [Nes05] to a smoothed approximation of (4). The authors of [FD03] 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. In [SR09], Scheinberg and Rish also consider a coordinate ascent method applied to the primal problem. A method based on column-wise updates is given in [RBLZ08]. A related problem is explored in [YL07] where the authors make a connection between (4) and the max-det problem [VBW98], and solve the problem using interior-point methods. Lu [Lu09] observes that the dual of (4) is a smooth problem, and applies Nesterov’s method [Nes05] directly to the dual. Another closely related work is [DGK08] 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 §2 we start by reviewing the definition of conditional independence in time series, and summarize the results from [SDV09]. In section 3 we set up the topology selection problem as a regularized ML problem and discuss its properties. We describe first-order methods for solving large instances of the regularized ML estimation problem. Examples with randomly generated and real data sets are presented in section 4.

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

$$X = [X_0 \quad X_1 \quad \cdots \quad X_p]$$

with $X_0 \in \mathbf{S}^n$ and $X_1, \dots, X_p \in \mathbf{R}^{n \times n}$. The standard trace inner product $\mathbf{tr}(X^T Y)$ is used in each of 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 :

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

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

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

The adjoint of T is a mapping $D : \mathbf{S}^{n(p+1)} \rightarrow \mathbf{M}^{n,p}$ 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 $D(S) = [D_0(S) \ D_1(S) \ \cdots \ D_p(S)]$ where

$$D_0(S) = \sum_{i=0}^p S_{ii}, \quad D_k(S) = 2 \sum_{i=0}^{p-k} S_{i,i+k}, \quad k = 1, \dots, p. \quad (6)$$

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

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

The same notation is used for $P_{\mathcal{V}}$ as a mapping from $\mathbf{R}^{n \times n} \rightarrow \mathbf{R}^{n \times n}$ and as a mapping from $\mathbf{S}^n \rightarrow \mathbf{S}^n$. In both cases, $P_{\mathcal{V}}$ 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 $P_{\mathcal{V}}(X)$ denotes the $r \times s$ block matrix with i, j block $P_{\mathcal{V}}(X)_{ij} = P_{\mathcal{V}}(X_{ij})$. The subscript of $P_{\mathcal{V}}$ is omitted if the sparsity pattern \mathcal{V} is clear from the context.

2 Graphical models of autoregressive processes

2.1 Conditional independence

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

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

where $R_k = \mathbf{E} x(t+k)x(t)^T$ and $j = \sqrt{-1}$. 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 [Dah00]. If we denote by \mathcal{V} the set of index pairs i, j of conditionally independent variables, then we can use the projection operator $P = P_{\mathcal{V}}$ defined in (6) to express the conditional independence relations as

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

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

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

where $B_0 \in \mathbf{S}_{++}^n$ and $B_k \in \mathbf{R}^{n \times n}$, $k = 1, \dots, p$. 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

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

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 $k = 1, \dots, p$. If we define $B = [B_0 \ B_1 \ \dots \ B_p]$, we can use the operator D defined in (6) to express Y_k as

$$[Y_0 \ Y_1 \ \dots \ Y_p] = D(B^T B). \quad (11)$$

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

$$P(D(B^T B)) = 0. \quad (12)$$

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

2.2 Conditional maximum likelihood estimation

We now consider the problem of estimating the model parameters B from an observed sequence $\hat{x}(1), \hat{x}(2), \dots, \hat{x}(N)$ of the AR process, subject to known conditional independence constraints (12). In [SDV09] the estimation problem was formulated as the optimization problem

$$\begin{aligned} & \text{minimize} && -2 \log \det B_0 + \text{tr}(CB^T B) \\ & \text{subject to} && P(D(B^T B)) = 0. \end{aligned} \quad (13)$$

The matrix $C \in \mathbf{S}_{++}^{n(p+1)}$ is a sample estimate of the covariance matrix, *i.e.*, its blocks C_{ij} , $i \leq j$, 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*

$$C = \frac{1}{N-p} H H^T, \quad H = \begin{bmatrix} \hat{x}(p+1) & \hat{x}(p+2) & \dots & \hat{x}(N) \\ \hat{x}(p) & \hat{x}(p+1) & \dots & \hat{x}(N-1) \\ \vdots & \vdots & & \vdots \\ \hat{x}(1) & \hat{x}(2) & \dots & \hat{x}(N-p) \end{bmatrix}. \quad (14)$$

With this choice the estimation problem (13) can be interpreted as a maximum likelihood problem. From (9), the conditional density of $x(t_1), \dots, x(t_2)$, given $x(t_1-p), \dots, x(t_1-1)$, is given by

$$\left(\frac{\det B_0}{(2\pi)^{n/2}} \right)^{t_2-t_1+1} \exp \left(-\frac{1}{2} \sum_{t=t_1}^{t_2} \mathbf{x}(t)^T B^T B \mathbf{x}(t) \right), \quad (15)$$

where $\mathbf{x}(t)$ denotes the $n(p+1)$ -vector $\mathbf{x}(t) = (x(t), x(t-1), \dots, x(t-p))$. From this it can be shown that the cost function in (13) with C defined as in (14), is essentially the negative conditional log-likelihood function of the observed sequence $\hat{x}(p+1), \hat{x}(p+2), \dots, \hat{x}(N)$, given $\hat{x}(1), \dots, \hat{x}(p)$. We therefore refer to (13) 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*

$$C = \frac{1}{N} H H^T, \quad (16)$$

where

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

The windowed estimate C is block-Toeplitz, and this guarantees several useful properties of the resulting model B (for example, stability; see [SDV09]). In practice, the differences between the windowed and non-windowed estimates are small when $N \gg p$.

When there are no sparsity constraints in (13), the solution can be found from the optimality conditions

$$\begin{bmatrix} C_{00} & C_{01} & \cdots & C_{0p} \\ C_{10} & C_{11} & \cdots & C_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ C_{p0} & C_{p1} & \cdots & C_{pp} \end{bmatrix} \begin{bmatrix} I \\ A_1^T \\ \vdots \\ A_p^T \end{bmatrix} = \begin{bmatrix} \Sigma \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad (17)$$

with unknowns $\Sigma = B_0^2$, $A_k = B_0^{-1} B_k$. The bottom p equations can be interpreted as normal equations for the least-squares problem

$$\text{minimize } \mathbf{tr}(A C A^T)$$

with variable $A = [I \ A_1 \ \cdots \ A_p]$, and form a linear system from which A_1, \dots, A_p can be computed. After computing A_k we obtain Σ from the first equation. This method is known as the *covariance method* if C is the non-windowed sample covariance (14), and as the *correlation method* if C is the windowed sample covariance (16) [SM97].

2.3 Convex formulation

The optimization problem (13) is non-convex because of the quadratic equality constraint. A convex relaxation is

$$\begin{aligned} & \text{minimize} && -\log \det X_{00} + \mathbf{tr}(C X) \\ & \text{subject to} && \mathbf{P}(D(X)) = 0 \\ & && X \succeq 0 \end{aligned} \quad (18)$$

with variable $X \in \mathbf{S}^{n(p+1)}$. The relaxation is exact, *i.e.*, the two problems (18) and (13) are equivalent, if the optimal solution X of (18) has rank n . In that case, the solution B of (18) can be calculated by factoring X as $X = B^T B$.

A condition for exactness of the relaxation follows from the dual problem of (18), which is

$$\begin{aligned} & \text{maximize} && \log \det W + n \\ & \text{subject to} && \begin{bmatrix} W & 0 \\ 0 & 0 \end{bmatrix} \preceq C + \mathbf{T}(\mathbf{P}(Z)), \end{aligned} \quad (19)$$

with variables $W \in \mathbf{S}^n$ and $Z \in \mathbf{M}^{n,p}$ [SDV09]. We can easily verify that the primal and dual problems are strictly feasible: $X = I$ is strictly feasible in the primal problem (18), 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 $C \succ 0$. The primal and dual problems are therefore solvable, and their optimal solutions are related by the optimality conditions

$$X_{00}^{-1} = W, \quad \text{tr} \left(X \left(C + \mathbf{T}(\mathbf{P}(Z)) - \begin{bmatrix} W & 0 \\ 0 & 0 \end{bmatrix} \right) \right) = 0. \quad (20)$$

From the optimality conditions it can be shown that the relaxation is exact when the trailing principal submatrix of order np in $C + \mathbf{T}(\mathbf{P}(Z))$ is positive definite at the optimum, *i.e.*,

$$(C + \mathbf{T}(\mathbf{P}(Z)))_{1:p,1:p} \succ 0. \quad (21)$$

Under this condition, the rank of

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

is at least np , and the two conditions in (20) imply that the optimal X has rank n .

The condition (21) is generally difficult to guarantee a priori. However, when C is block-Toeplitz, then (21) can be shown to hold for all dual feasible Z . This follows from an 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 [SDV09]). We therefore conclude that for positive definite block-Toeplitz C (for example, the windowed sample covariance (16) or the true covariance), the problems (13) and (18) are *equivalent*. For general non-block-Toeplitz C (for example, the non-windowed sample covariance (14)), we cannot guarantee that (21) holds at the optimum. However, we can note that the non-windowed sample covariance approaches a block-Toeplitz matrix as $N \rightarrow \infty$. 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 [SDV09].

2.4 Reformulated dual problem

The primal and dual optimization problems (18) and (19) are convex optimization problems with differentiable objectives and linear matrix inequality constraints. They can be solved by interior-point methods, for example, the path-following methods developed for convex determinant maximization problems [Toh99, VBW98]. 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 derive an equivalent formulation of the dual problem (19) which is more amenable to large-scale (first-order) algorithms.

Let $V = C + \text{T}(\text{P}(Z))$. The inequality in (19),

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

is equivalent to

$$V_{1:p,1:p} \succeq 0, \quad \text{range}(V_{1:p,0}) \subseteq \text{range}(V_{1:p,1:p}), \quad V_{00} - V_{1:p,0}^T V_{1:p,1:p}^\dagger V_{1:p,0} \succeq W, \quad (22)$$

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 (22) 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 (19) and write the problem as an unconstrained problem

$$\text{maximize} \quad -\phi(C + \text{T}(\text{P}(Z))) \quad (23)$$

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

$$\phi(V) = -\log \det \left(V_{00} - V_{1:p,0}^T V_{1:p,1:p}^\dagger V_{1:p,0} \right) - n, \quad (24)$$

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

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

and convexity of this expression follows from results in convex analysis [BV04, §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

$$\nabla \phi(V) = -V^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & V_{1:p,1:p}^{-1} \end{bmatrix}. \quad (25)$$

This can be seen, for example, from the identity $\det V = \det V_{1:p,1:p} \det \left(V_{00} - V_{1:p,0}^T V_{1:p,1:p}^{-1} V_{1:p,0} \right)$, which gives $\phi(V) = -\log \det V + \log \det V_{1:p,1:p} - n$.

If $C + \text{T}(\text{P}(Z)) \succ 0$ at the optimum of (23) then the primal optimal solution can be computed from Z via the expressions

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

where $V = C + \text{T}(\text{P}(Z))$ 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 (20) and the identities

$$V = \begin{bmatrix} V_{00} - V_{1:p,0}^T V_{1:p,1:p}^{-1} V_{1:p,0} & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} V_{1:p,0}^T V_{1:p,1:p}^{-1} \\ I \end{bmatrix} V_{1:p,1:p} \begin{bmatrix} V_{1:p,0}^T V_{1:p,1:p}^{-1} \\ I \end{bmatrix}^T, \quad (26)$$

$$V^{-1} = \begin{bmatrix} 0 & 0 \\ 0 & V_{1:p,1:p}^{-1} \end{bmatrix} + \begin{bmatrix} -I & \\ V_{1:p,1:p}^{-1} V_{1:p,0} & \end{bmatrix} (V_{00} - V_{1:p,0}^T V_{1:p,1:p}^{-1} V_{1:p,0})^{-1} \begin{bmatrix} -I & \\ V_{1:p,1:p}^{-1} V_{1:p,0} & \end{bmatrix}^T. \quad (27)$$

The formula for V^{-1} also provides an alternative form of the gradient (25).

The reformulated dual (23) is interesting because it can often be solved by *unconstrained* optimization algorithms. To explain this, we again distinguish between Toeplitz and non-Toeplitz C . If C is block-Toeplitz, then it can be shown that the function $\phi(C + T(P(Z)))$ is a *closed* convex function (*i.e.*, a function with closed sublevel sets) and that its domain is open. To see this, consider the function ϕ restricted to the set of block-Toeplitz matrices, *i.e.*, $\phi(T(R))$, where $R \in \mathbf{M}^{n,p}$. By definition, R is in the domain of $\phi(T(R))$ if $T(R) \succeq 0$ and there exists a positive definite W with

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

From the property of block-Toeplitz matrices mentioned in section 2.3, this implies $T(R) \succ 0$. In other words, the domain of $\phi(T(R))$ is the open set $\{R \mid T(R) \succ 0\}$. By a similar argument, if a sequence of matrices R in the domain of $\phi(T(R))$ converges to a point \bar{R} in the boundary of the domain, then the Schur complement of $T(\bar{R})_{1:p,1:p}$ in $T(\bar{R})$ must be singular, and hence $\phi(T(R)) \rightarrow \infty$. For a continuous function with an open domain this is equivalent to closedness [BV04, p.639].

The closedness property of $\phi(C + T(Z))$ implies that problem (23) can be solved by unconstrained convex optimization algorithms, for example, gradient descent or the conjugate gradient method.

If C is not block-Toeplitz, then the function $\phi(C + T(P(Z)))$ is not necessarily closed, and its domain is not necessarily open. One implication is that it is possible that the optimal solution of (23) is at a point in the boundary, *i.e.*, a point where $C + T(P(Z))$ is singular. However in practice, C is usually approximately block Toeplitz and one can expect that $\phi(C + T(P(Z)))$ is 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 $\{Z \mid \phi(C + T(P(Z))) \leq \phi(C + T(P(Z^{(0)})))\}$ is closed. This condition is considerably weaker than the requirement that all sublevel sets are closed.

We will discuss first-order algorithms in more detail in the next section, where we will encounter a constrained generalization of (23).

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, *i.e.*, 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

$$\text{AIC} = -2\mathcal{L} + 2k, \quad \text{AIC}_c = -2\mathcal{L} + \frac{2Nk}{N - k - 1}, \quad \text{BIC} = -2\mathcal{L} + k \log N \quad (28)$$

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

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

where X is the optimal solution of (18), and the effective number of parameters is

$$k = \frac{n(n+1)}{2} - |\mathcal{V}| + p(n^2 - 2|\mathcal{V}|), \quad (29)$$

where $|\mathcal{V}|$ is the number of conditionally independent pairs of variables. The information-theoretic topology selection method 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:

$$\begin{aligned} & \text{minimize} && -\log \det X_{00} + \mathbf{tr}(CX) + \gamma h(D(X)) \\ & \text{subject to} && X \succeq 0, \end{aligned} \quad (30)$$

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

$$h(Y) = \sum_{i>j} \max_{k=0,\dots,p} \max \{|Y_{k,ij}|, |Y_{k,ji}|\}. \quad (31)$$

When $p = 0$, the penalty term reduces to $h(Y) = \sum_{i>j} |Y_{ij}|$ for $Y \in \mathbf{S}^n$, and we obtain the formulation (4), considered in [BEd08, Lu09, FHT08] (with the minor difference that we do not penalize the diagonal entries of X).

The dual problem of (30), which will be important in section 3.3, can be derived as follows. To simplify the derivation we introduce a variable $Y = D(X)$ and write the problem as

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

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

$$\begin{aligned} L(X, Y, Z, U) &= -\log \det X_{00} + \mathbf{tr}(CX) + \gamma h(Y) - \mathbf{tr}(UX) + \mathbf{tr}(Z^T(D(X) - Y)) \quad (32) \\ &= -\log \det X_{00} + \mathbf{tr}((C + T(Z) - U)X) + \gamma h(Y) - \mathbf{tr}(Z^T Y). \quad (33) \end{aligned}$$

(Recall that the mappings T and D defined in (5) and (6) are adjoints, *i.e.*, $\mathbf{tr}(Z^T D(X)) = \mathbf{tr}(T(Z)X)$.) The dual function is the infimum of the Lagrangian over X and Y .

We first minimize over Y . The term $h(Y)$ 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

$$\mathbf{diag}(Z_k) = 0, \quad k = 0, 1, \dots, p. \quad (34)$$

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

$$2Z_{0,ij}Y_{0,ij} + \sum_{k=1}^p (Z_{k,ij}Y_{k,ij} + Z_{k,ji}Y_{k,ji}) + \gamma \max \left\{ |Y_{0,ij}|, \max_{k=1,\dots,p} |Y_{k,ij}|, \max_{k=1,\dots,p} |Y_{k,ji}| \right\}$$

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

$$2|Z_{0,ij}| + \sum_{k=1}^p (|Z_{k,ij}| + |Z_{k,ji}|) \leq \gamma, \quad i \neq j, \quad (35)$$

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

$$\inf_Y L(X, Y, Z, U) = \begin{cases} -\log \det X_{00} + \mathbf{tr}((C + \mathbf{T}(Z) - U)X) & (34), (35) \\ -\infty & \text{otherwise.} \end{cases}$$

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

$$(C + \mathbf{T}(Z) - U)_{00} \succ 0, \quad (C + \mathbf{T}(Z) - U)_{ij} = 0, \quad (i, j) \neq 0, \quad (36)$$

and an expression for the dual function

$$g(Z, U) = \inf_{X, Y} L(X, Y, Z, U) = \begin{cases} \log \det(C + \mathbf{T}(Z) - U)_{00} + n & (34), (35), (36) \\ -\infty & \text{otherwise.} \end{cases}$$

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

$$\begin{aligned} & \text{maximize} && \log \det W + n \\ & \text{subject to} && \begin{bmatrix} W & 0 \\ 0 & 0 \end{bmatrix} \preceq C + \mathbf{T}(Z) \\ & && \sum_{k=0}^p (|Z_{k,ij}| + |Z_{k,ji}|) \leq \gamma, \quad i \neq j \\ & && \mathbf{diag}(Z_k) = 0, \quad k = 0, \dots, p. \end{aligned} \quad (37)$$

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

$$\begin{aligned} & \text{maximize} && \log \det(C + Z) + n \\ & \text{subject to} && |Z_{ij}| \leq \gamma, \quad i \neq j \\ & && \mathbf{diag}(Z) = 0. \end{aligned}$$

Except for the equality constraint, this is the problem considered in [Lu09, DGK08].

3.2 Optimality conditions

The primal problem (30) is always strictly feasible ($X = I$ is strictly feasible). The dual problem (30) 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

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

Dual feasibility. W and Z satisfy

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

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

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 (32), we have equality between the Lagrangian and the primal objective if $\mathbf{tr}(UX) = 0$. Therefore

$$\mathbf{tr} \left(X \left(C + \mathsf{T}(Z) - \begin{bmatrix} W & 0 \\ 0 & 0 \end{bmatrix} \right) \right) = 0 \quad (38)$$

at the optimum. This condition is known as *complementary slackness*.

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

$$X_{00}^{-1} = W. \quad (39)$$

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

$$t_{ij} = \max \left\{ |Y_{0,ij}|, \max_{k=1,\dots,p} |Y_{k,ij}|, \max_{k=1,\dots,p} |Y_{k,ji}| \right\}.$$

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

$$\sum_{k=0}^p (|Z_{k,ij}| + |Z_{k,ji}|) < \gamma, \quad t_{ij} = 0,$$

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 $k = 0, \dots, p$.

The conditions (38)–(39) show that the optimal X has rank n under the same conditions as for the problem with given sparsity pattern (18). If

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

then the optimal X has rank n , and this is always the case if C is block-Toeplitz. Under these conditions, the optimization problem (30) is equivalent to a regularized (conditional) ML estimation problem for the model parameters B :

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

3.3 First-order algorithms

As in section 2.4 we can eliminate W and reformulate the dual problem as

$$\begin{aligned} & \text{maximize} && -\phi(C + \mathbf{T}(Z)) \\ & \text{subject to} && \sum_{k=0}^p (|Z_{k,ij}| + |Z_{k,ji}|) \leq \gamma, \quad i \neq j \\ & && \mathbf{diag}(Z_k) = 0, \quad k = 0, \dots, p, \end{aligned} \quad (40)$$

with ϕ defined in (24). If $C + \mathbf{T}(Z) \succ 0$ at the optimum, the primal optimal X can be expressed in terms of Z as

$$X = (C + \mathbf{T}(Z))^{-1} - \begin{bmatrix} 0 & 0 \\ 0 & (C + \mathbf{T}(Z))_{1:p,1:p}^{-1} \end{bmatrix}. \quad (41)$$

In this section we present numerical results for a first-order algorithm applied to the reformulated dual problem. The algorithm is the classical gradient projection algorithm [Ber99, §2.3] with a backtracking line search. It starts at a feasible $Z^{(0)}$ (e.g., $Z^{(0)} = 0$) and repeats the projected gradient update

$$Z^{(k)} = \mathcal{P}(Z^{(k-1)} - t_k \nabla f(Z^{(k-1)})), \quad k = 0, 1, \dots,$$

where $f(Z) = \phi(C + \mathbf{T}(Z))$, until a stopping criterion is satisfied. Here, \mathcal{P} denotes the projection on the set

$$\mathcal{C} = \left\{ Z \in \mathbf{M}^{n,p} \mid \sum_{k=0}^p (|Z_{k,ij}| + |Z_{k,ji}|) \leq \gamma \text{ for } i \neq j, \mathbf{diag}(Z_k) = 0 \text{ for } k = 0, \dots, p \right\}, \quad (42)$$

and t_k is a positive step size determined by a backtracking line search (see the appendix). The most important steps in the computation are evaluations of the gradient of the objective function $f(Z) = \phi(C + \mathbf{T}(Z))$, which is given by

$$\nabla f(Z) = \mathbf{D} \left((C + \mathbf{T}(Z))^{-1} - \begin{bmatrix} 0 & 0 \\ 0 & (C + \mathbf{T}(Z))_{1:p,1:p}^{-1} \end{bmatrix} \right), \quad (43)$$

and the projections on \mathcal{C} . We now explain these two steps and the stopping criterion in more detail. More background and a convergence analysis of the gradient projection algorithm are provided in the appendix.

The gradient (43) can be evaluated using a Cholesky factorization $C + \mathbf{T}(Z) = L^T L$ with L a lower triangular matrix. If we partition L as

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

then the gradient of $-f(C + \mathbf{T}(Z))$ at Z simplifies to

$$-\mathbf{D} \left(\begin{bmatrix} I \\ -L_{1:p,1:p}^{-1} L_{1:p,0} \end{bmatrix} L_{00}^{-1} L_{00}^{-T} \begin{bmatrix} I \\ -L_{1:p,1:p}^{-1} L_{1:p,0} \end{bmatrix}^T \right).$$

The Euclidean projection $\mathcal{P}(U)$ of a matrix $U \in \mathbf{M}^{p,n}$ on the set \mathcal{C} is defined as

$$\mathcal{P}(U) = \underset{Z \in \mathcal{C}}{\operatorname{argmin}} \|Z - U\|_F^2.$$

Clearly, the diagonal entries of $\mathcal{P}(U)_k$ are zero for $k = 0, \dots, p$. To find the off-diagonal entries we can solve an independent problem

$$\begin{aligned} \text{minimize} \quad & 2(Z_{0,ij} - U_{0,ij})^2 + \sum_{k=1}^p ((Z_{k,ij} - U_{k,ij})^2 + (Z_{k,ji} - U_{k,ji})^2) \\ \text{subject to} \quad & 2|Z_{0,ij}| + \sum_{k=1}^p (|Z_{k,ij}| + |Z_{k,ji}|) \leq \gamma \end{aligned}$$

for each i, j with $i > j$. This is essentially the problem of projecting a vector on a ball in ℓ_1 -norm, and the solution is readily obtained via Lagrange duality. We have

$$\mathcal{P}(U)_{k,ij} = \begin{cases} U_{k,ij} + \lambda & U_{k,ij} \leq -\lambda \\ 0 & |U_{k,ij}| \leq \lambda \\ U_{k,ij} - \lambda & U_{k,ij} \geq \lambda, \end{cases}$$

where the dual multiplier λ is zero if $2|U_{0,ij}| + \sum_{k=1}^p (|U_{k,ij}| + |U_{k,ji}|) \leq \gamma$, and equal to the unique solution of the piecewise linear equation

$$2 \max\{|U_{0,ij}| - \lambda, 0\} + \sum_{k=1}^p (\max\{|U_{k,ij}| - \lambda, 0\} + \max\{|U_{k,ji}| - \lambda, 0\}) = \gamma$$

otherwise.

We will use the following stopping criterion in the experiments. At each iteration, we compute X in (41) from the current iterate Z . This matrix X is primal feasible, as can be seen from the identity (27) and the fact that $C + \mathbf{T}(Z) \succ 0$. By taking the Schur complement of $(C + \mathbf{T}(Z))_{1:p,1:p}$ we also find a dual feasible W in (37). The duality gap between this primal feasible X and the dual feasible Z, W is

$$\begin{aligned} \eta &= -\log \det X_{00} + \mathbf{tr}(CX) + \gamma h(\mathbf{D}(X)) - \log \det W - n \\ &= \mathbf{tr}(CX) - n + \gamma h(\mathbf{D}(X)) \\ &= \mathbf{tr}((C + \mathbf{T}(Z))X) - n - \mathbf{tr}(X \mathbf{T}(Z)) + \gamma h(\mathbf{D}(X)) \\ &= -\mathbf{tr}(X \mathbf{T}(Z)) + \gamma h(\mathbf{D}(X)). \end{aligned} \tag{44}$$

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

Numerical example We generate an AR model with a sparse inverse spectrum by setting $B_0 = I$ and randomly generated sparse lower triangular matrices B_k with entries ± 0.5 . This procedure is repeated until a stable AR model is found. The AR process is then used to generate N samples of the time series. In this experiment, the model dimensions are $n = 300$, $p = 4$, $N = 2n(p + 1)$. The true inverse spectrum has 857 non-zero entries in the upper triangular part or about 2% of the number of all entries. The penalty parameter γ is set at $\gamma = 0.1$.

The variable Z in the reformulated dual problem (40) is a matrix in $\mathbf{M}^{300,4}$, so the problem has $n(n + 1)/2 + pn^2 = 405,150$ optimization variables. We start the gradient projection algorithm at a strictly feasible $Z^{(0)} = 0$, and terminate when the duality gap is below 10^{-1} (the optimal value is on the order of hundreds). The backtracking line search of iteration k is initialized at $t = \lambda_{\min}^2(C + \mathbf{T}(Z^{(k-1)}))$. With this initialization, the line searches required at most five backtracking steps to find an acceptable step size.

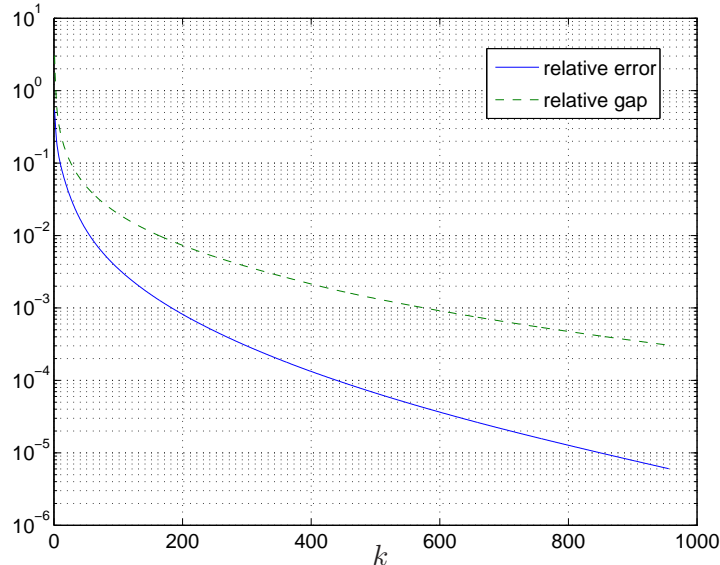


Figure 1: Plot of relative error $(f(Z^{(k)}) - f(Z^*)) / |f^*|$ (solid curve) and relative duality gap $\eta^{(k)} / |f^*|$ (dashed curve) versus the number of iterations.

Figure 1 shows the relative error $(f(Z^{(k)}) - f^*) / |f^*|$ where f^* is the optimal value, and the relative duality gap $\eta^{(k)} / |f^*|$ versus the iteration number for a typical instance. As can be seen, a solution with a moderate accuracy (relative error in the range of 10^{-4} – 10^{-3}) is obtained in a number of iterations that is only a fraction of the problem size.

We have also experimented with gradient projection algorithms from the class of optimal first-order methods originated by Nesterov (in particular, algorithm 1 in [Tse08]; see also [Nes04, BT09]). For functions with a Lipschitz continuous gradient, 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 a sequence of non-increasing step sizes ($t_{k+1} \leq t_k$) determined by a backtracking line search [BT09, Tse08]. In the problems considered here, the gradient of f is not Lipschitz continuous on \mathcal{C} (in fact \mathcal{C} is not even contained in $\text{dom } f$). Nevertheless, an implementation with a backtracking line search worked well in our experiments. If monotonicity of the step sizes is imposed (as required by the theory), the convergence was not significantly faster than the classical gradient projection method. However, an implementation in which the step sizes are not forced to be decreasing, was often about five times faster.

4 Examples

The regularized ML estimation problem (30) is motivated by the fact that the resulting AR model typically has a sparse inverse spectrum $S(\omega)^{-1}$. Since the regularized problem is also convex, it can be solved as an efficient heuristic for topology selection. In this section we present some examples that illustrate the effectiveness of this approach.

The sparsity of the inverse spectrum 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, [FHT08, HLPL06, BEd08]. Meinshausen and Bühlmann [MB06] give explicit formulas for γ based on a statistical analysis of the probability of errors in the topology (see also [YL07, BEd08]). Asadi *et al.* [ARS⁺09] 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 (30), *i.e.*, between the log-likelihood and the penalty function $h(D(X))$. The trade-off curve can be computed by solving (30) 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 (18) 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.

Before we discuss the results, we elaborate on two key steps: the computation of the trade-off curves, and the thresholding used to estimate the topology of a given process. The trade-off curves are computed by solving (30) for a sequence of values of γ . To obtain an accurate curve with a small number of values γ , we choose the sequence as follows. We first take two values of γ near the opposing ends of the trade-off curve. As the next value of γ we take the slope of the line between the first two points on the trade-off curve. This procedure is continued recursively. At each step we take as the next value for γ the slope of the line between two previously computed adjacent points on the trade-off curve. We will see in the examples that a small number of points is usually sufficient to get an accurate approximation of the curve.

The topology of a given AR model will be estimated by applying a threshold to the maximum magnitude

$$\rho_{ij} = \sup_{\omega} |R(\omega)_{ij}| \quad (45)$$

of the entries of the normalized inverse spectrum $R(\omega)$, defined as

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

($R(\omega)$ is the inverse spectrum $S(\omega)^{-1}$ with rows and columns scaled to make the diagonal one.) The normalized inverse spectrum $R(\omega)$ is known as *partial coherence* [Bri81, Dah00]. 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. In the experiments we use a threshold of 10^{-1} , *i.e.*, we set entries with $\rho_{ij} \leq 10^{-1}$ equal to zero.

4.1 Randomly generated data

We start with a few experiments with randomly generated models. The random AR model coefficients B_k were generated using the method described at the end of section 3.3. The model dimensions in the two experiments are $n = 20$ and $p = 2$. The sample covariance matrix C was calculated from a sequence of $N = 1000$ samples of the process.

Experiment 1 In the first experiment we assume that the model order p is known. We first calculate the trade-off curve between the penalty $h(D(X))$ and the log-likelihood $\mathcal{L}(X)$ (figure 2, top). Next 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 $\rho_{ij} \leq 10^{-1}$

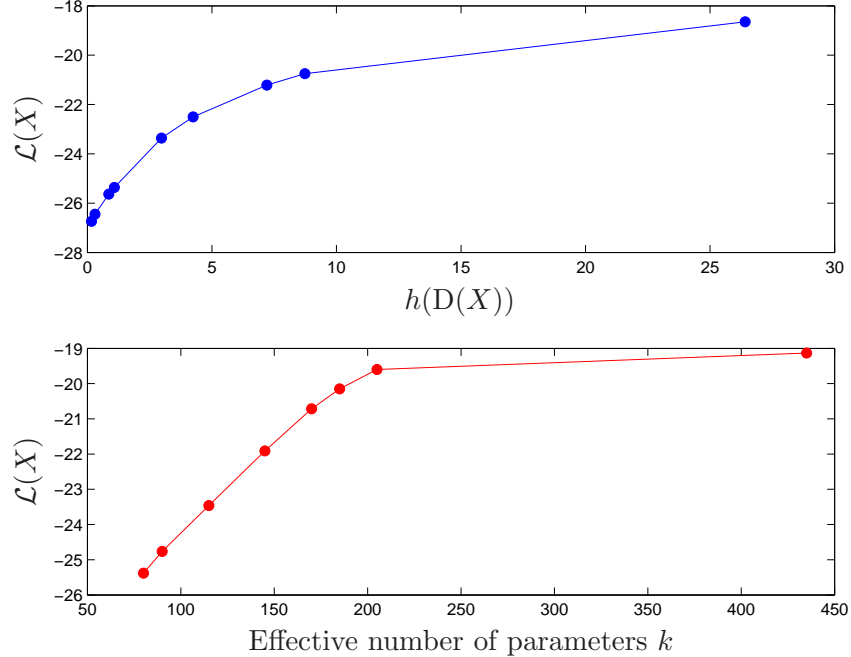


Figure 2: *Top.* Trade-off curve between the log-likelihood $\mathcal{L}(X)$ and $h(D(X))$. *Bottom.* Log-likelihood $\mathcal{L}(X)$ as a function of the effective number of parameters for points on the trade-off curve.

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 (17) with C given in (14) or, equivalently, the ML solution of (13) without the sparsity constraints). For each of the nine sparsity patterns, we solve the ML problem subject to sparsity constraints (18). We rank the nine solutions using the AIC_c and BIC scores defined in (28). Figure 4 shows the two scores and the maximized log-likelihood as functions of γ . The models that minimize the AIC/BIC scores turn out to be the same in this example (the models for $\gamma = 0.13$) and the corresponding topology is shown in figure 5 (right). The sparsity pattern on the left in figure 5 is the topology estimated by thresholding the partial coherence spectrum of the least-squares solution. We see that the convex heuristic correctly identified all except five entries, and is much more accurate than the least-squares estimate.

Experiment 2 The purpose of the second experiment is to examine how fast the error in the topology selection decreases with increasing sample length N . We vary the sample size N , and for each N select a value of γ as in the previous experiment. We solve the regularized ML problem (30) for 50 different sample sequences (*i.e.*, 50 different C).

Figure 6 shows the errors as a function of N . “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 6 show the fraction of false positives and false negatives versus the sample size. The bottom graphs show the total fraction of misclassified entries. The experiment illustrates that the errors in the estimated topology decrease much more rapidly than in an estimation based on the least-squares method.

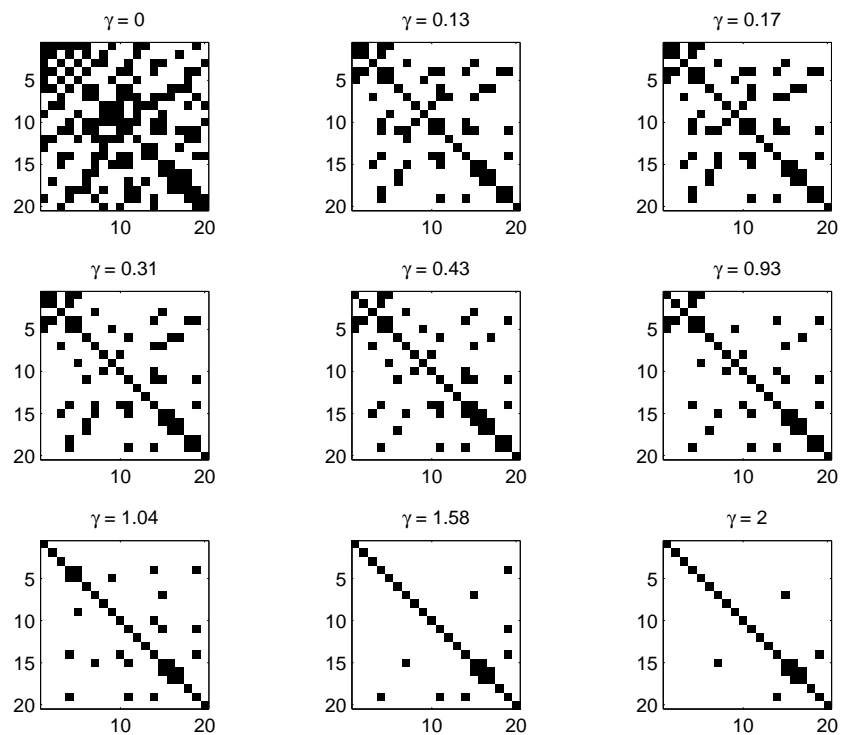


Figure 3: Topologies of solutions along the tradeoff curve in figure 2 (ordered from right to left on the tradeoff curve).

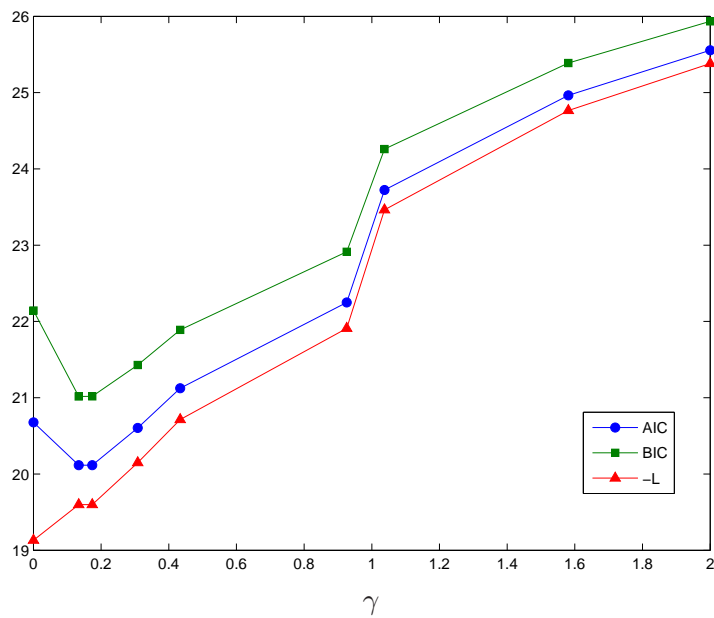


Figure 4: AIC and BIC scores, and maximized log-likelihood for solutions on the trade-off curve in figure 2.

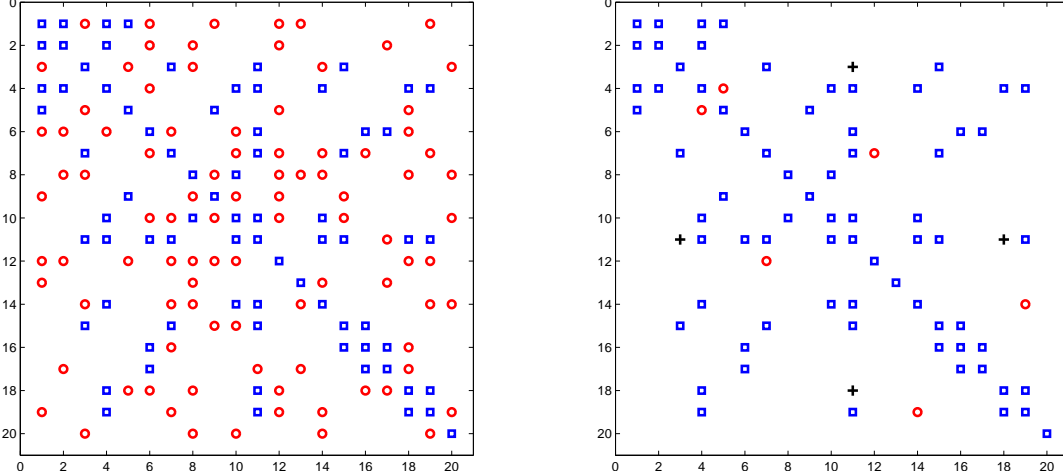


Figure 5: *Left.* The sparsity pattern from the least-squares solution. *Right.* The sparsity pattern from the regularized ML problem with $\gamma = 0.134$. The blue squares are the correctly identified nonzero entries (true positives). The red circles are the entries that are misclassified as nonzero (false positives). The black crosses are entries that are misclassified as zeros (false negatives).

4.2 International stock markets

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 shown in figure 7. Similar time series for a smaller number of markets were analyzed in [BY03, AAA08].

We solve the regularized ML problem with model orders ranging from $p = 1$ 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 $\gamma = 0.18$). Figure 8 (right) shows ρ_{ij} , the maximum magnitude of the partial coherence of the model, and compares it with a nonparametric estimate and the thresholded 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.

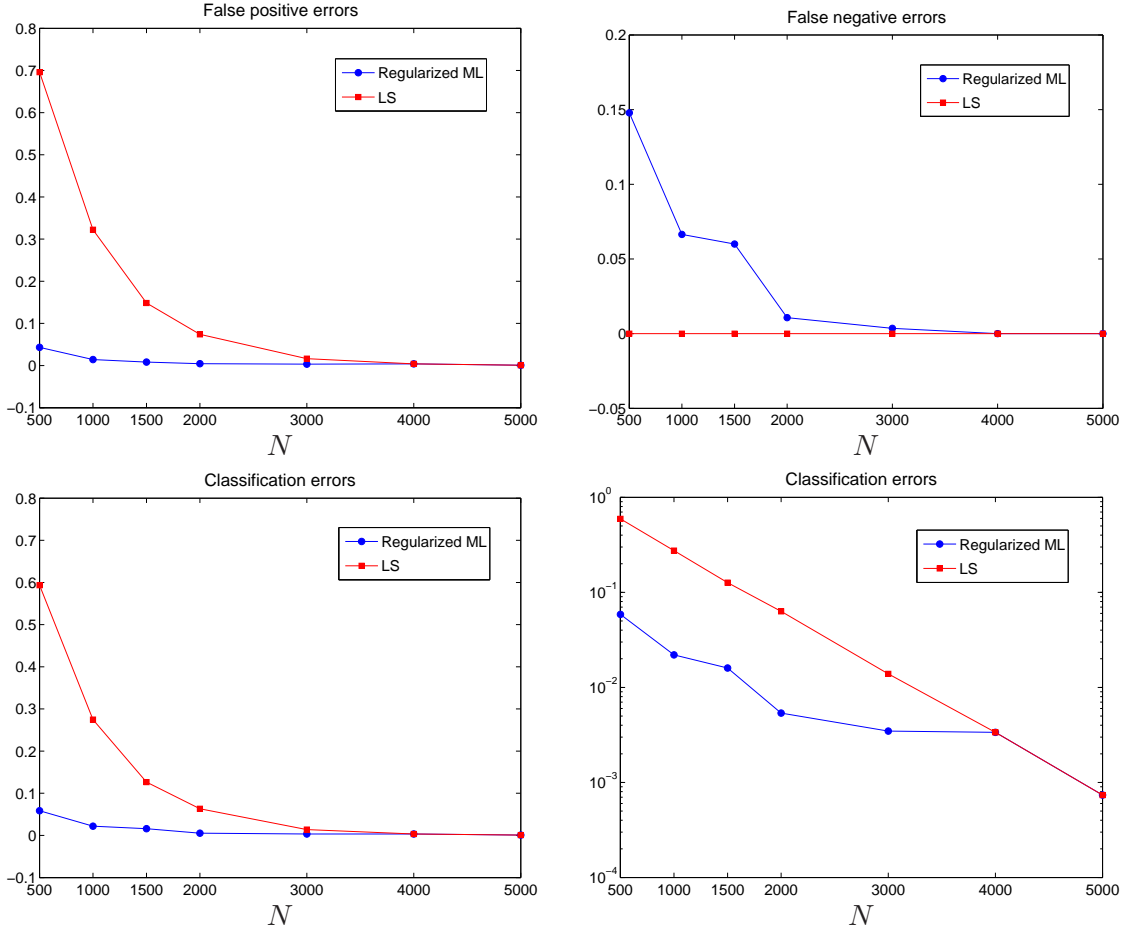


Figure 6: *Top left.* Fraction of incorrectly added edges in the estimated graph (number of upper triangular nonzeros in the estimated pattern that are incorrect, divided by the number of upper triangular zeros in the correct pattern). *Top right.* Fraction of incorrectly removed edges in the estimated 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. The left plot uses a linear scale for the x-axis, the right plot is the same curve on a logarithmic scale. The results for the least-squares model estimates are in red squares and the results from the convex heuristic are in blue circles.

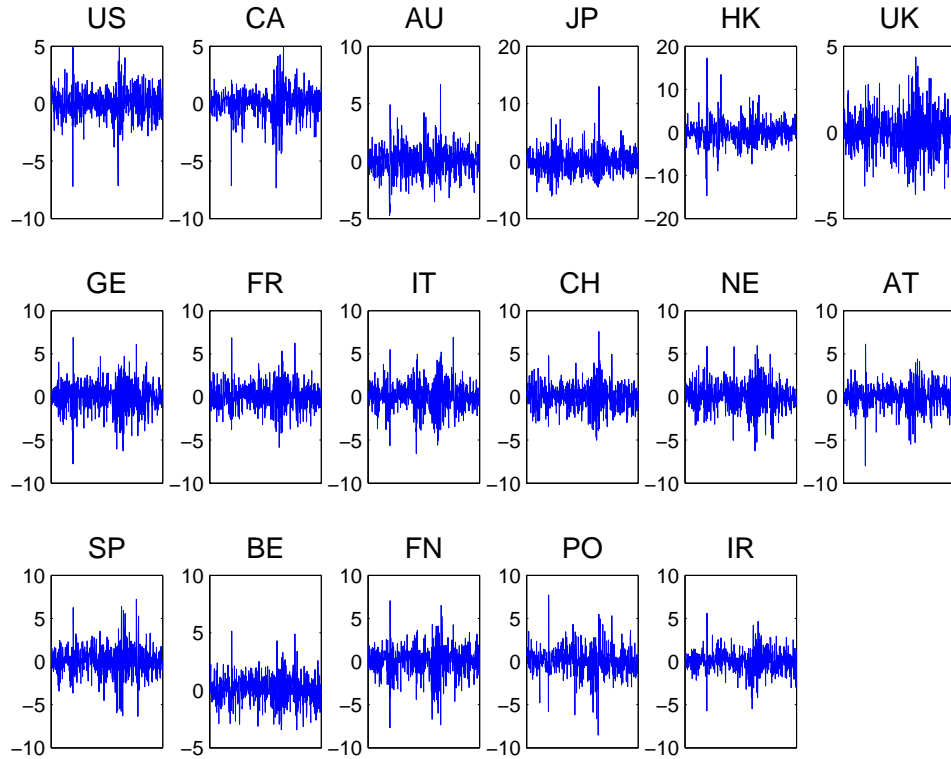


Figure 7: Time series of daily returns for stock market indices of 17 countries.

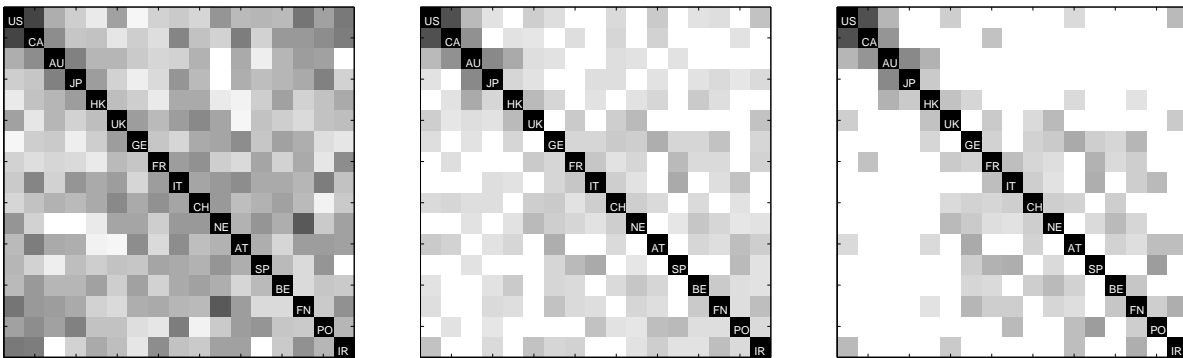


Figure 8: The maximum magnitude of the partial coherence ρ_{ij} for three models of the stock exchange data. The partial coherence left is a nonparametric sample estimate. The plot in the middle is the sparsity pattern of the thresholded least-squares estimate. The plot on the right is the result of the convex heuristic based on the regularized ML problem.

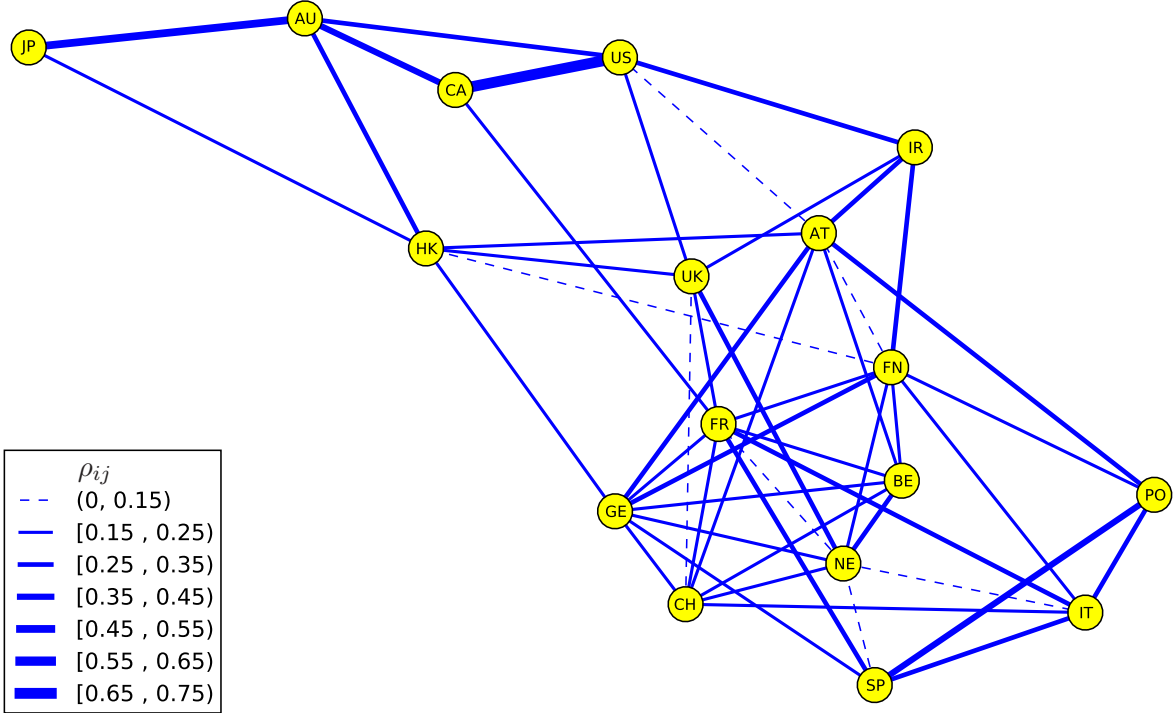


Figure 9: A graphical model of stock market data. The strength of connections is represented by the width of the blue links, which is proportional to $\rho_{ij} = \max_{\omega} |R(\omega)_{ij}|$ if it is greater than 0.15.

Figure 9 shows the graphical model estimated by the 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.

5 Conclusion

We have presented a convex optimization method for topology selection in graphical models of Gaussian autoregressive 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 a time series of international stock market indices were included to confirm the effectiveness of this approach.

References

- [AAA08] 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.
- [ARS⁺09] 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.
- [BEd08] 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.
- [Ber99] D. P. Bertsekas. *Nonlinear Programming*. Athena Scientific, second edition, 1999.
- [BJ04] F. R. Bach and M. I. Jordan. Learning graphical models for stationary time series. *IEEE Transactions on Signal Processing*, 32:2189–2199, 2004.
- [Bri81] D. R. Brillinger. *Time series: Data analysis and theory*. Holden-Day, San Francisco, CA, expanded edition, 1981.
- [BT09] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2009. To appear.
- [BV04] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004. www.stanford.edu/~boyd/cvxbook.
- [BY03] D.A. Bessler and J. Yang. The structure of interdependence in international stock markets. *Journal of International Money and Finance*, 22(2):261–287, 2003.
- [Dah00] R. Dahlhaus. Graphical interaction models for multivariate time series. *Metrika*, 51(2):157–172, 2000.
- [Dem72] A. P. Dempster. Covariance selection. *Biometrics*, 28:157–175, 1972.
- [DES97] 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.
- [DGK08] J. Duchi, S. Gould, and D. Koller. Projected subgradient methods for learning sparse Gaussians. *Proceeding of the Conference on Uncertainty in AI*, 2008.
- [DRV04] 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, 2004.
- [EDS03] M. Eichler, R. Dahlhaus, and J. Sandkühler. Partial correlation analysis for the identification of synaptic connections. *Biological Cybernetics*, 89(4):289–302, 2003.
- [Eic06] M. Eichler. Fitting graphical interaction models to multivariate time serie. *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence*, 2006.

- [Eic08] M. Eichler. Testing nonparametric and semiparametric hypotheses in vector stationary processes. *Journal of Multivariate Analysis*, 99(5):968–1009, 2008.
- [FD03] R. Fried and V. Didelez. Decomposability and selection of graphical models for multivariate time series. *Biometrika*, 90(2):251–267, 2003.
- [FHT08] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432, 2008.
- [FMM⁺05] 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.
- [FNW07] 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.
- [GIF02] U. Gather, M. Imhoff, and R. Fried. Graphical models for multivariate time series from intensive care monitoring. *Statistics in Medicine*, 21(18):2685–2701, 2002.
- [HLPL06] 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.
- [Lau96] S. L. Lauritzen. *Graphical Models*. Oxford University Press, 1996.
- [Lu09] Z. Lu. Smooth optimization approach for sparse covariance selection. *SIAM Journal on Optimization*, 19:1807, 2009.
- [MB06] N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the Lasso. *Annals of Statistics*, 34(3):1436–1462, 2006.
- [Nes04] Y. Nesterov. *Introductory Lectures on Convex Optimization*. Kluwer Academic Publishers, 2004.
- [Nes05] Yu. Nesterov. Smooth minimization of non-smooth functions. *Mathematical Programming Series A*, 103:127–152, 2005.
- [Pol87] B. T. Polyak. *Introduction to Optimization*. Optimization Software, Inc., 1987.
- [RBLZ08] A. J. Rothman, P. J. Bickel, E. Levina, and J. Zhu. Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515, 2008.
- [SDV09] 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, 2009.
- [SM97] P. Stoica and R. L. Moses. *Introduction to Spectral Analysis*. Prentice Hall, 1997.
- [SR09] K. Scheinberg and I. Rish. SINCO - a greedy coordinate ascent method for sparse inverse covariance selection problem. 2009. Submitted.
- [SSSB05] R. Salvador, J. Suckling, C. Schwarzbauer, and E. Bullmore. Undirected graphs of frequency-dependent functional connectivity in whole brain networks. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1457):937–946, 2005.

- [TLH⁺00] 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. *International Journal of Bifurcation and Chaos in applied Sciences and Engineering*, 10(11):2595–2610, 2000.
- [Toh99] K.-C. Toh. Primal-dual path-following algorithms for determinant maximization problems with linear matrix inequalities. *Computational Optimization and Applications*, 14:309–330, 1999.
- [Tse08] P. Tseng. On accelerated proximal gradient methods for convex-concave optimization. *SIAM Journal on Optimization*, 2008. submitted.
- [VBW98] 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.
- [YL07] M. Yuan and Y. Lin. Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1):19, 2007.

A Gradient projection

In this appendix we describe the gradient projection algorithm used in the paper. To simplify the notation, we use a generic problem format

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

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

$$S = \{x \in \mathcal{C} \cap \mathbf{dom} f \mid f(x) \leq f(x^{(0)})\} \quad (46)$$

is closed and bounded.

The projection of a point y on \mathcal{C} will be denoted $\mathcal{P}(y) = \operatorname{argmin}_{z \in \mathcal{C}} \|z - y\|_2$. The projection satisfies

$$(y - \mathcal{P}(y))^T (z - \mathcal{P}(y)) \leq 0 \quad \forall z \in \mathcal{C}. \quad (47)$$

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

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

for $t > 0$ [Nes04, §2.2.3]. It can be shown that a point $x \in \mathcal{C} \cap \mathbf{dom} f$ is optimal if and only if $G_t(x) = 0$ for any $t > 0$. A useful property of the gradient map follows by applying (47) to $y = x - t\nabla f(x)$ and $\mathcal{P}(y) = x - tG_t(x)$: we have

$$(G_t(x) - \nabla f(x))^T (z - x + tG_t(x)) \leq 0 \quad \forall z \in \mathcal{C}. \quad (48)$$

for all $x \in \mathbf{dom} f$ and $z \in \mathcal{C}$. For $z = x$, this further reduces to the inequality

$$\nabla f(x)^T G_t(x) \geq \|G_t(x)\|_2^2 \quad (49)$$

for all $x \in \mathcal{C} \cap \mathbf{dom} f$.

Algorithm The gradient projection algorithm [Ber99, Pol87] is a classical algorithm for optimization problems with simple constraints, and is especially popular for large-scale problems (see, for example, [FNW07]). The algorithm starts at a feasible $x^{(0)}$ and computes

$$x^{(k)} = \mathcal{P}(x^{(k-1)} - t_k \nabla f(x^{(k-1)})), \quad k = 1, 2, \dots,$$

until a stopping criterion is satisfied. A backtracking line search can be used to determine the step size t_k . The line search starts at a given $t = \hat{t} > 0$, and backtracks ($t := \beta t$, where $\beta \in (0, 1)$ is an algorithm parameter) until the condition

$$f(x - tG_t(x)) \leq f(x) - t\nabla f(x)^T G_t(x) + \frac{t}{2} \|G_t(x)\|_2^2 \quad (50)$$

holds, where $x = x^{(k-1)}$. The final value of t is used as step size t_k . If $x - tG_t(x) \notin \mathbf{dom} f$, then the lefthand side of the inequality (50) is interpreted as plus infinity, and the inequality as false.

Other types of line searches are described in [Ber99, FNW07].

Analysis A common assumption in the literature on the gradient projection algorithm is that $\mathcal{C} \subseteq \mathbf{dom} f$ and that the gradient ∇f is Lipschitz continuous on \mathcal{C} . Under this assumption it is known that the error $f(x^{(k)}) - f^*$ decreases as $1/k$ [BT09, Nes04]. These assumptions are not valid for the applications in this paper: here, $\mathcal{C} \not\subseteq \mathbf{dom} f$ and the gradient of f is not Lipschitz continuous on $\mathcal{C} \cap \mathbf{dom} f$. For completeness we therefore include a convergence analysis. The proof is adapted from Beck and Teboulle [BT09].

The assumption that the sublevel set (46) associated with the initial point is closed and bounded implies that there exists an optimal x^* , and that the gradient ∇f satisfies a Lipschitz condition on S : there exists an $L > 0$ such that

$$\|\nabla f(u) - \nabla f(v)\|_2 \leq L\|u - v\|_2 \quad \forall u, v \in S. \quad (51)$$

Define $x = x^{(i-1)}$ and assume that $x \in S$. From the Lipschitz property (51),

$$f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{L}{2} \|y - x\|_2^2$$

for all $y \in S$. Applying this to $y = \mathcal{P}(x - t\nabla f(x)) = x - tG_t(x)$ and using (49) gives

$$\begin{aligned} f(\mathcal{P}(x - t\nabla f(x))) &\leq f(x) - t\nabla f(x)^T G_t(x) + \frac{Lt^2}{2} \|G_t(x)\|_2^2 \\ &\leq f(x) - t\left(1 - \frac{Lt}{2}\right) \|G_t(x)\|_2^2 \end{aligned} \quad (52)$$

if $\mathcal{P}(x - t\nabla f(x)) \in S$. Define

$$\tau = \sup\{\tau \geq 0 \mid \mathcal{P}(x - t\nabla f(x)) \in S \text{ for } t \in [0, \tau]\}.$$

We have $\tau > 0$ because for small positive t ,

$$f(\mathcal{P}(x - t\nabla f(x))) \approx f(x) - t\nabla f(x)^T G_t(x) \leq f(x) - t\|G_t(x)\|_2^2 < f(x),$$

from (49) and the fact that $G_t(x) \neq 0$. Therefore $\mathcal{P}(x - t\nabla f(x)) \in S$ for small positive t . Since $\mathcal{P}(x - t\nabla f(x))$ is continuous in t , and S is a closed set, we either have $\tau = \infty$, or τ is

finite and $\mathcal{P}(x - \tau \nabla f(x))$ is in the boundary of S , *i.e.*, $f(\mathcal{P}(x - \tau \nabla f(x))) = f(x^{(0)})$. From the bound (52) we can then note that $\tau \geq 2/L$, because otherwise the inequality evaluated at $t = \tau$ would imply that $f(\mathcal{P}(x - \tau \nabla f(x))) < f(x)$, a contradiction. Evaluating (52) at $t = 1/L$, we see that $t = 1/L$ satisfies (50). We conclude that if $x \in S$, then the line search terminates with a value $t \geq t_{\min} = \min\{\hat{t}, \beta/L\}$.

Next, we note that if (50) holds, then for all $y \in \mathcal{C} \cap \mathbf{dom} f$,

$$\begin{aligned} f(x - tG_t(x)) &\leq f(x) - t\nabla f(x)^T G_t(x) + \frac{t}{2} \|G_t(x)\|_2^2 \\ &\leq f(y) + \nabla f(x)^T (x - y) + t(G_t(x) - \nabla f(x))^T G_t(x) - \frac{t}{2} \|G_t(x)\|_2^2 \\ &\leq f(y) + G_t(x)^T (x - y) - \frac{t}{2} \|G_t(x)\|_2^2. \end{aligned}$$

The last step follows from (48) with $z = y$. Taking $y = x$ shows that $f(x - tG_t(x)) < f(x)$, so the algorithm is a descent method, and if $x^{(i-1)} \in S$ then $x^{(i)} \in S$. Taking $y = x^*$ gives

$$\begin{aligned} f(x - tG_t(x)) &\leq f(x^*) + G_t(x)^T (x - x^*) - \frac{t}{2} \|G_t(x)\|_2^2 \\ &= f(x^*) + \frac{1}{2t} (\|x - x^*\|_2^2 - \|x - tG_t(x) - x^*\|_2^2) \\ &\leq f(x^*) + \frac{1}{2t_{\min}} (\|x - x^*\|_2^2 - \|x - tG_t(x) - x^*\|_2^2), \end{aligned}$$

i.e.,

$$f(x^{(i)}) - f(x^*) \leq \frac{1}{2t_{\min}} (\|x^{(i-1)} - x^*\|_2^2 - \|x^{(i)} - x^*\|_2^2).$$

Combining these bounds for $i = 1, \dots, k$ gives

$$f(x^{(k)}) - f(x^*) \leq \frac{1}{k} \sum_{i=1}^k (f(x^{(i)}) - f(x^*)) \leq \frac{1}{2kt_{\min}} \|x^{(0)} - x^*\|_2^2.$$