University of California
Los Angeles

Graphical Models of Time Series:
Parameter Estimation and Topology Selection

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Electrical Engineering

by

Jitkomut Songsiri

2010

The dissertation of Jitkomut Songsiri is approved.

Kung Yao

Vwani Roychowdhury

James S. Gibson

Lieven Vandenberghe, Committee Chair

University of California, Los Angeles

2010

To my family

Table of Contents

List of Figures

List of Tables

Acknowledgments

This thesis would not have been possible without the guidance from my advisor, Professor Lieven Vandenberghe. I am grateful for his patience from explaining to me the very basic linear algebra to deeper topics in convex optimization. Numerous discussions with him have provided me a great inspiration and encouragement on my research. His remarkable vision and dedication to work make him a distinguished role model for me to follow. I am truely honored to be his student.

I also would like to thank Professor Levan, my initial academic advisor, who had guided me during the first year through many control classes. His warm support and important advice are very much appreciated. I must also thank my committee members for valuable comments to this work. A sense of gratitude to many professors at UCLA who have strengthened my understanding of control theory, optimization and related EE courses.

The Royal Thai scholarship has provided me a financial support I needed for studying in the U.S. I must also express my gratitude to all professors at Control lab, Chulalongkorn University for motivating me to pursue a PhD degree and providing me valuable advice regarding academic and career decisions.

Many friends have helped me keep up my motivation during my school days. A big thank to John, Martin, David, Amir, Manolo, and Adolfo for sharing useful discussions in my daily work at the cubicle and cheerful support to each other. I wish to thank my Thai friends at UCLA (Anusorn, Nokku and others) for entertainment, generosity and caring they provided. My best friend at Chula, Som for being supportive for everything I do. Thanks for our friendship and for never being tired of my endlessly boring phone calls during the past five years. A special thank also goes to Oou, who had stuck with me through the difficult

times. Her support made me stronger tremendously during my years spent in grad school.

Lastly, I wish to thank my parents for their love and support throughout all levels of my education. My sister and brother (Pim and Poom) for giving a loving environment to me. I dedicate this thesis to my family, the most important thing in my life.

Vita

1978	Born, Nakhon Sawan, Thailand
1999	B.S. (Electrical Engineering), Chulalongkorn University, Thailand.
2002	M.Eng. (Electrical Engineering), Chulalongkorn University, Thailand.
2003-2004	Research Assistant, Control System laboratory, Chulalongkorn University, Thailand.
2004	Recipient of the Royal Thai government scholarship.
2008	Teaching Assistant for EE142, Electrical Engineering, UCLA.
2009-2010	Academic Graduate Researcher, Electrical Engineering, UCLA.

Publications

J. Songsiri, and L. Vandenberghe. Topology selection in graphical models of autoregressive processes. Accepted for publication in *Journal of Machine Learning Research*.

J. Songsiri, J. Dahl, and L. Vandenberghe. Graphical models of autoregressive processes, Y. Eldar and D. Palomar (Editors), *Convex Optimization in Signal*

Processing and Communications, Cambridge University Press, 2010.

J. Songsiri, J. Dahl, and L. Vandenberghe. Maximum-likelihood estimation of autoregressive models with conditional independence constraints. *Proceedings of the IEEE conference on Acoustic, Speech and Signal Processing*, 2009.

Abstract of the Dissertation

Graphical Models of Time Series: Parameter Estimation and Topology Selection

by

Jitkomut Songsiri

Doctor of Philosophy in Electrical Engineering

University of California, Los Angeles, 2010

Professor Lieven Vandenberghe, Chair

This thesis is concerned with estimation problems in graphical models of time series. The graph topology of a graphical model characterizes conditional independence relations between the variables, so estimation generally involves two problems: topology selection and parameter estimation for a given topology. We first consider the problem of fitting a Gaussian autoregressive model to a time series, subject to conditional independence constraints. This is an extension of the classical covariance selection problem to time series. The conditional independence constraints impose a sparsity pattern on the inverse of the spectral density matrix, and result in nonconvex quadratic equality constraints in the maximum likelihood formulation of the model estimation problem. We present a semidefinite relaxation, and prove via duality that the relaxation is exact when the sample covariance matrix is block-Toeplitz. We also give experimental results suggesting that the relaxation is often exact when the sample covariance matrix is not block-Toeplitz. The estimation method can be used for small topology

selection problems by enumerating all topologies, solving the estimation problem for each topology and ranking them via model selection criteria such as the Akaike or Bayes information criteria.

As a second contribution, we propose an efficient method for learning the topology of graphical models of autoregressive Gaussian time series. The method is based on an ℓ_1-type nonsmooth regularization of the conditional maximum likelihood estimation problem used to promote sparsity in the inverse of the estimated spectral density matrix. We describe a heuristic approach for choosing the regularization parameter which controls the sparsity of the esimated inverse spectrum. The estimation accuracy of the topology and AR model is illustrated by numerical examples and experiments with real data sets.

Finally, we describe a large-scale algorithm that solves a reformulation of the duals of the above two problems via the gradient projection method. Numerical results show that the method is capable of solving problems of dimensions of several hundred within a reasonable amount of time.

CHAPTER 1

Introduction

1.1 Graphical models of random variables

This thesis is concerned with estimation problems in graphical models of time series. Graphical models combine probabilistic concepts with graph theory by representing dependencies among variables as a graph. In this thesis we focus on undirected graphical models where the edges specify the conditional dependence structure of random variables. Graphical models are useful for many reasons. By exploiting the graph representation, some statistical quantities such as marginal or conditional probabilities of a subset of nodes in the graph can be calculated more efficiently [WJ08]. Graphical models also provide insight in the structure of the distribution. The conditional independence structure defined on a graph lets us associate a probabilistic model with the graph and further allows us to build a complex model out of simpler parts. This paves the way to parameter estimation methods that provide a parsimonious model for a complex system. For these reasons, graphical models have become a useful tool for many statistical applications such as modeling of complex biological systems, information extraction, speech processing, pattern recognition, communication networks, etc [Bis06, BB01].

A simple example of a graphical model is a Gaussian graphical model, associated with a multivariate Gaussian random variable. Other common examples include contingency tables, which describe conditional independence relations in

multinomial distributions, Bayesian networks, which use directed acyclic graphs to represent causal or temporal relations, and chain graphs, which are mixed graphs containing undirected and directed edges. For general introductions to graphical models, we refer the readers to [Whi90, Edw00]. A comprehensive treatment of the theory of graphical models can be found in [Lau96]. Related topics in learning and estimation problems in graphical models are presented in [Jor99, WJ08].

This work focuses on parameter and structure learning in graphical models. General estimation problems in graphical models can be divided in two groups depending on whether the topology of the graph is given or not. The topology of graphical models is defined through the notion of conditional independence and this was initially established for static multivariate random variables. For a Gaussian variable $x \sim \mathcal{N}(0, \Sigma)$, the components x_i and x_j are conditionally independent, conditional on the other components, if and only if $(\Sigma^{-1})_{ij} = 0$ (see details in section 2.1.1). The topology of a Gaussian graphical model is therefore equivalent to the zero pattern of the inverse covariance matrix. This nice characterization allows us to consider a class of estimation problems when a graph structure is given. An example of problems in this class is the maximum-likelihood (ML) estimation of a Gaussian graphical model, parametrized by a covariance matrix Σ, for a given graph topology. The ML problem can be expressed as

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

where C is the sample covariance matrix, and $\mathcal{V}$ are the pairs of nodes (i, j) that are not connected by an edge, *i.e.*, for which x_i and x_j are conditionally

independent. A change of variables $X = \Sigma^{-1}$ results in a convex problem

$$\begin{array}{ll} \text{maximize} & \log\det X - \mathbf{tr}(CX) \\ \text{subject to} & X_{ij} = 0, \quad (i,j) \in \mathcal{V}. \end{array} \tag{1.1}$$

This is known as the *covariance selection* problem introduced by [Dem72] as a technique to reduce the number of parameters in the estimation of Σ. If the problem is feasible, it provides a covariance estimate that yields a decreased variance since some coefficients are restricted to zero [Dem72].

The covariance selection problem has received a lot of attention in the machine learning community, and leads to the other class of estimation problems of graphical models, where the topology of the graph is unknown. To estimate the topology of Gaussian graphical models, a direct approach is to formulate a hypothesis test to decide about the presence or absence of edges between two nodes [Lau96, §5.3.3]. Another possibility is to enumerate different topologies, and use information-theoretic criteria (such as the Akaike or Bayes information criteria) to rank the models. 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, *i.e.*, on solving

$$\text{minimize} \quad -\log\det X + \mathbf{tr}(CX) + \gamma \textstyle\sum_{ij} |X_{ij}| \tag{1.2}$$

(see [MB06, DRV05, BEd08, RWR08a, FHT08, HLP06, YL07, Lu09, SR09]). The ℓ_1 regularization term has been used extensively in many statistical learning problems where sparseness of the solution is favored. Well-known applications include the Lasso method for subset selection in regression [Tib96] and ℓ_1-norm methods for compressed sensing [Tro06, CRT06a]. The ℓ_1 term in (1.2) helps to encourage some components of X to zero, thus yielding a sparse inverse covariance matrix. With this method, one can learn the conditional independence structure from the data automatically.

1.2 Graphical models of time series

As mentioned above, graphical models and related estimation problems have been primarily developed for static multivariate random variables. Recently there has been considerable interest in explaining relationships between components in multivariate time series as well. For example, one wishes to learn dependencies of temperatures recorded from many geographical regions [BJ04], dependencies of stock prices from major markets [BY03], or dependencies of biological signals that measure activities in the human brain [SSS05, EDS03, DES97], just to name a few. This has motivated us to explore an extension of graphical models of random variables to time series.

First of all, the notion of conditional independence for random variables can be extended to time series. This concept was first discussed in [Bri81] where it was shown that the conditional independence between components of a multivariate stationary Gaussian process can be characterized in terms of the inverse of the spectral density matrix $S(\omega)$. Two components $x_i(t)$ and $x_j(t)$ are independent, conditional on the other components of $x(t)$ if and only if

$$(S(\omega)^{-1})_{ij} = 0 \tag{1.3}$$

for all ω [Bri81, §8.3],[Dah00]. To discover a topology of graphical models, a common approach is to formulate a hypothesis test examining whether an edge is present in the graph. Dahlhaus [Dah00] derives a statistical test based on the maximum magnitude of a nonparametric estimate of the normalized inverse spectrum. The method was illustrated by the air pollution data to study interactions among polluted particles. The same approach was also applied to identification of functional neural connectivity in [DES97, EDS03]. This nonparametric approach based on a test in frequency domain has become a useful tool for many

applications. For example, [TLH00] investigated the connection between the cortical activity and tremor in patients suffering from Parkinson's disease. [SSS05] explored the correlated activities in human brain networks based on functional magnetic resonance imaging (fMRI) data. [GIF02] applied the technique to the haemodynamic system consisting of vital signs such as heart rate, or blood pressure, etc., which are crucial for detection of critical situation of patients in an intensive care unit. It can be also applied to the analysis of factors in therapy process from psychosomatic studies [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.

As opposed to the approach mentioned above, a parametric approach can be used to select a graphical model from a family of possible models by enumerating all topologies and ranking them via model selection criteria. Therefore, the identification of conditional independence structures reduces to a model selection problem in which the best model minimizes a model selection criterion.

In this thesis we consider a parametric approach: the maximum likelihood estimation for graphical models of autoregressive processes

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

where $x(t) \in \mathbf{R}^n$, and $w(t) \in \mathbf{R}^n$ is Gaussian white noise. This estimation problem is an extension of the covariance selection problem (1.1) to time series. The most relevant study of this type is the work of [Eic06a] which uses Whittle's approximation of the exact likelihood function, and imposes sparsity constraints on the inverse covariance functions. The model parameters were then obtained from the Yule-Walker equations. It was also mentioned in [DE03, §4.3] that numerical

solutions to the problem of fitting AR models with conditional independence constraints have been under exploration. This is due to the characterization in (1.3) in terms of AR parameter that results in complicated nonlinear constraints in A_k and Σ.

Our first contribution is to propose a convex framework for maximum-likelihood estimation of AR models with conditional independence constraints. We can show that the zeros in the inverse spectrum are equivalent to quadratic equality constraints on AR parameters, which are generally nonconvex. We propose a convex relaxation and prove that under some assumptions, it provides *exact* solutions for the ML problem, yielding polynomial-time algorithms. The results of this work can serve two purposes. Given a conditional independence graph of a time series, one can estimate the spectrum according to the graph structure. Furthermore, if the conditional independence is not specified *a priori*, the structure can be identified from the model selection problem using information-theoretic criteria such as AIC and BIC.

Similarly to the topology selection in Gaussian graphical models, the combinatorial approach for the model selection problem is clearly limited to small graphs. Our second contribution is to propose a convex formulation for topology selection in AR models, based on augmenting the ML estimation with a convex regularization term, similar to the ℓ_1-norm regularization in (1.2). The regularization term is chosen so that the sparsity in the estimated inverse spectrum is promoted, revealing the underlying conditional independence structure in the time series. This convex heuristic also preserves the exactness property under mild assumptions, *i.e.*, it provides the exact solutions A_k, Σ to the ML problem.

The topology selection problems in graphical models of Gaussian variables and graphical models of AR processes can be solved by interior-point methods [BV04,

§11], for example, the path-following methods developed for convex determinant maximization problems [Toh99, VBW98]. However, these methods can suffer from expensive computational cost in high dimensional problems. Therefore we are interested in less expensive first-order algorithms that can solve large problems with an acceptable accuracy within a reasonable amount of time.

The nondifferentibility of these problems due to the ℓ_1-type regularization also makes them challenging to solve in large scale. Several efficient methods have been proposed to topology selection problems in Gaussian graphical models (1.2); see [BEd08, FHT08, RWR08b, HLP06, YL07, SR09, RBL08, Lu09, Lu10, DGK08]. These algorithms are based on various techniques such as the coordinate descent method, interior-point methods, the gradient projection method, or the recent Nesterov's optimal gradient method [Nes04]. Most of these algorithms are applied to the dual of (1.2) which is a smooth problem.

These methods cannot be easily applied to topology selection in AR models because of several complications, for example the presence of extra linear equality and matrix inequality constraints that do not appear in the covariance selection problem. We use the optimality conditions to reformulate the dual of topology selection problems. We will see that the dual problem can be cast as a minimization problem with simple constraints which is suitable for the gradient projection method. Although the gradient method is known to converge slowly, we consider a variation using a special stepsize rule known as *Barzilai-Borwein or spectral steplength* [BB88] which has been shown to greatly improve the performance in practice [BMR03, FNW07, WNF09]. With this method we are able to solve problems of dimensions in the order of several hundred efficiently.

1.3 Outline of thesis

Chapter 2 provides the definition of conditional independence properties for random variables and time series. We review the properties of AR processes and derive a characterization of conditional independence which can be expressed as quadratic equalities of AR parameters. The last part of this chapter describes the least-squares method, the maximum-likelihood (ML), and the maximum-entropy (ME) estimation methods for AR processes. These are common estimation techniques used in spectral analysis. We show the connection of these three methods via the sample covariance matrix.

Chapter 3 extends the ML and ME estimation methods for AR processes to include conditional independence constraints. These become nonconvex problems due to the quadratic equalities in AR parameters from conditional independence relations. We introduce a convex relaxation to these problems, which in general is not equivalent to the original problem. Our main result is to use duality to show that the relaxation provides the exact solution to the ML and ME problems under a block-Toeplitz assumption on the sample covariance matrix. We end the chapter with some numerical results to illustrate that the relaxation is exact under a weaker condition in practice.

Chapter 4 considers the more general problem of estimating the AR model parameters and the topology of the graphical model. We start with a direct approach where we enumerate all topologies and rank these models according to information-theoretic criteria such as AIC or BIC. This combinatorial approach is feasible for small graphs only. This chapter presents another main result: an efficient method for topology selection in AR models. The method is based on an ℓ_1-type nonsmooth regularization of the ML estimation. The ℓ_1 regularization term is added to encourage the sparsity in the estimated inverse spectrum.

Results of experiments with random and real data sets are included.

In chapter 5 we investigate first-order algorithms for large-scale ML estimation with conditional independence constraints and ML estimation with the ℓ_1 regularization. The algorithm is based on the gradient projection method applied to the reformulated dual of these two problems. We compare the performance with other variants of the gradient projection method on randomly generated data.

1.4 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

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

with $X_0 \in \mathbf{S}^n$ and $X_1, \ldots, X_p \in \mathbf{R}^{n\times n}$. We denote by $(X_k)_{ij}$ the (i, j) component of X_k. $X^H = \bar{X}^T$ is the complex conjugate transpose. 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:

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

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 $X \in \mathbf{M}^{n,p}$, then

$$\mathrm{T}(X) = \begin{bmatrix} X_0 & X_1 & \cdots & X_p \\ X_1^T & X_0 & \cdots & X_{p-1} \\ \vdots & \vdots & \ddots & \vdots \\ X_p^T & X_{p-1}^T & \cdots & X_0 \end{bmatrix}. \tag{1.4}$$

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

$$\mathrm{D}_0(S) = \sum_{i=0}^{p} S_{ii}, \qquad \mathrm{D}_k(S) = 2\sum_{i=0}^{p-k} S_{i,i+k}, \quad k = 1, \ldots, p. \tag{1.5}$$

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 $i = 1, \ldots, n$, *i.e.*, the diagonal entries are not included among the zeros and that it is symmetric (if $(i, j) \in \mathcal{V}$, then $(j, i) \in \mathcal{V}$). We denote by $\mathrm{P}_{\mathcal{V}}(X)$ the projection of a matrix $X \in \mathbf{S}^n$ or $X \in \mathbf{R}^{n\times n}$ on the complement of the sparsity pattern $\mathcal{V}$:

$$\mathrm{P}_{\mathcal{V}}(X)_{ij} = \begin{cases} X_{ij} & (i, j) \in \mathcal{V} \\ 0 & \text{otherwise.} \end{cases} \tag{1.6}$$

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 $\mathrm{P}_{\mathcal{V}}(X)_{ij} = \mathrm{P}_{\mathcal{V}}(X_{ij})$. The subscript of $\mathrm{P}_{\mathcal{V}}$ is omitted if the sparsity pattern $\mathcal{V}$ is clear from the context.

CHAPTER 2

Background on graphical models and AR processes

In this chapter we describe the conditional independence property for Gaussian random variables and extend it to Gaussian time series. This property can be used to describe relations between the components of a multivariate random variable or time series. It is useful to represent these relations as a graph. This gives a characterization of a graphical model containing a set of nodes that represent the variables and a collection of edges. The absence of an edge between two nodes indicates that the corresponding two components are conditionally independent, given the other variables. We will see that this characterization can be expressed via the covariance matrix for Gaussian random variables and via the spectral density matrix for Gaussian time series.

Next we focus on graphical models of autoregressive processes by deriving the characterization of conditional independence in Gaussian autoregressive processes. To prepare for our study of estimation problems of the graphical models in chapter 3, we first review some existing techniques for estimating the parameters of AR models. These techniques include the least-squares method, the maximum-likelihood, and the maximum-entropy estimation methods. This topic is standard in estimation or spectral analysis and can be found in many textbooks [SS89, BJ76, Mar87, Kay88, SM97].

2.1 Conditional independence

2.1.1 Random variables

Let $x \sim N(0, \Sigma)$ be an n-dimensional Gaussian random variable. The conditional independence relations between x_i and x_j can be derived from the conditional distribution of the two variables, given the remaining variables. Suppose the random vector x is partitioned into component y and z with the corresponding mean and variance:

$$x = \begin{bmatrix} y \\ z \end{bmatrix}, \quad \mu = \begin{bmatrix} \mu_y \\ \mu_z \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \Sigma_{yy} & \Sigma_{yz} \\ \Sigma_{zy} & \Sigma_{zz} \end{bmatrix},$$

where $y = (x_i, x_j)$ consists of the components i and j of interest. It can be shown that the *conditional* distribution of y given z is also Gaussian with mean

$$\mu_{y|z} = \mu_y - \Sigma_{yz}\Sigma_{zz}^{-1}(z - \mu_z), \tag{2.1}$$

and covariance

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

i.e., the Schur complement of Σ_{zz}. The Schur complement also appears in Σ^{-1}:

$$\Sigma^{-1} = \begin{bmatrix} (\Sigma_{yy} - \Sigma_{yz}\Sigma_{zz}^{-1}\Sigma_{yz}^{H})^{-1} & \divideontimes \\ \divideontimes & \divideontimes \end{bmatrix}.$$

The conditional covariance matrix of size 2×2 is therefore the inverse of Schur complement of Σ_{zz} in Σ and it can be written as

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

Hence x_i and x_j are conditionally independent if and only if

$$(\Sigma^{-1})_{ij} = 0. \tag{2.2}$$

Specifying the graph topology of a Gaussian graphical model is therefore equivalent to specifying the sparsity pattern of the inverse covariance matrix.

2.1.2 Time series

Gaussian stationary processes We consider an n-dimensional real-valued process $x(t)$. We assume that $x(t)$ is a zero-mean Gaussian process, which means its marginal distributions are jointly Gaussian. For Gaussian processes, it is known that the process is strictly stationary if it is wide-sense stationary, *i.e.*, $\mathbf{E}\, x(t_1)x(t_2)^T$ depends only on the difference of $t_1 - t_2$. Given a stationary process, we define the autocovariance function $R_k : \mathbf{Z} \to \mathbf{R}^{n\times n}$ as

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

Since $x(t)$ is real, we must have $R_{-k} = R_k^T$. In addition, the autocovariance function is always non-negative; that is for any $a_i, a_j \in \mathbf{R}^n$, with $i, j = 1, \ldots N$, we have

$$\sum_i^N \sum_j^N a_i^T R_{i-j} a_j \geq 0,$$

which follows from

$$\sum_i^N \sum_j^N a_i^T R_{i-j} a_j = \sum_i^N \sum_j^N \mathbf{E}[a_i^T x(i)x(j)^T a_j] = \mathbf{E}\left[\left(\sum_i^N a_i^T x(i)\right)^2\right] \geq 0.$$

This condition is equivalent to the non-negativity of the covariance matrix of any successive variables $x(t), x(t+1), \ldots, x(t+N)$:

$$C = \begin{bmatrix} R_0 & R_1 & \cdots & R_N \\ R_1^T & R_0 & \cdots & R_{N-1} \\ \vdots & \vdots & \ddots & \vdots \\ R_N^T & R_{N-1}^T & \cdots & R_0 \end{bmatrix} \succeq 0. \tag{2.3}$$

Suppose $\sum_{k=-\infty}^{\infty} \|R_k\| < \infty$ where $\|\cdot\|$ denotes an operator norm of a matrix. Then the sequence R_k is absolutely summable element-wise and therefore the spectral density matrix is well-defined, as the Fourier transform of the autocovariance sequence,

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

(where $\mathrm{j} = \sqrt{-1}$). For each ω, $S(\cdot)$ is a Hermittian matrix of size $n \times n$. From the property $R_{-k} = R_k^T$, we have $S(-\omega) = S(\omega)^T$. Moreover, $S(\omega)$ is a periodic function of period 2π. Therefore in spectral analysis, we often consider the spectrum only in the interval $\omega \in [0, \pi]$.

Since the autocovariance function is nonnegative, we can show that $S(\omega) \succeq 0$ for all ω. Consider an $(n(N+1))$-vector $x = (y, ye^{-\mathrm{j}\omega}, ye^{-\mathrm{j}2\omega}, \ldots, ye^{-\mathrm{j}N\omega})$ with any $y \in \mathbf{C}^n$. From the block-Toeplitz C in (2.3), we have

$$0 \leq \frac{1}{N+1} x^H C x = \sum_{k=-N}^{N} \left(1 - \frac{|k|}{N+1}\right) y^H R_k y \; e^{-\mathrm{j}k\omega},$$

where the first inequality follows from the nonnegativity of the autocovariance function. Then we can take the limit as $N \to \infty$ to conclude that $S(\omega)$ is nonnegative for all ω.

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

$$(S(\omega)^{-1})_{ij} = 0 \tag{2.4}$$

This definition can be interpreted and justified from Brillinger [Bri81, §8.3]. The idea is motivated from the interpretation of the conditional mean (2.1) as the optimal linear least-mean-square estimate of y given z, for Gaussian random variables [KSH00]. The problem is to find a linear function $h(z)$ that minimizes

$\mathbf{E}\,\|y - h(z)\|_2^2$. Moreover, the conditional covariance matrix $\Sigma_{y|z}$ is essentially the minimized error covariance matrix $\mathbf{E}[(y - h(z))(y - h(z))^T]$. This interpretation can be extended to time series as follows.

Let $y(t) = (x_i(t), x_j(t))$ and let $z(t)$ be the $(n-2)$-vector containing the remaining components of $x(t)$. Define $e(t)$ as the error

$$e(t) = y(t) - \sum_{k=-\infty}^{\infty} H_k z(t-k)$$

between $y(t)$ and the linear filter of $z(t)$ that minimizes $\mathbf{E}\,\|e(t)\|_2^2$. Then it can be shown in [Bri81, §8.3] that the cross spectrum of the residual error $e(t)$ is given by

$$S_{ee}(\omega) = S_{yy}(\omega) - S_{yz}(\omega) S_{zz}^{-1}(\omega) S_{zy}(\omega),$$

where S_{yy}, S_{yz}, S_{zz} are the submatrices of cross spectra between the corresponding variables in $S(\omega)$. This is again the Schur complement of $S_{zz}(\omega)$ in $S(\omega)$. Therefore the spectrum of the error process can be also written as

$$\begin{bmatrix} (S(\omega)^{-1})_{ii} & (S(\omega)^{-1})_{ij} \\ (S(\omega)^{-1})_{ji} & (S(\omega)^{-1})_{jj} \end{bmatrix}^{-1}. \tag{2.5}$$

The off-diagonal entry in the error spectrum (2.5) is called the *partial cross-spectrum* of x_i and x_j, after removing the effects of z. The partial cross-spectrum is zero if and only if the error covariances $\mathbf{E}\, e(t+k)e(t)^T$ are diagonal, *i.e.*, the two components of the error process $e(t)$ are independent.

It is interesting to see that the conditional independence characterization of stationary Gaussian time series can be extended from Gaussian random variables by replacing the covariance matrix with the spectral density matrix. The topology of a graphical model of time series can then be read out from the sparsity pattern in the inverse of spectral density matrix.

2.2 Autoregressive processes

A multivariate autoregressive model of order p is defined as

$$x(t) = -\sum_{k=1}^{p} A_k x(t-k) + w(t), \tag{2.6}$$

where $x(t) \in \mathbf{R}^n$ and $w(t) \sim \mathcal{N}(0, \Sigma)$ is Gaussian white noise. With this assumption, $x(t)$ is also a Gaussian process with zero mean.

The random process (2.6) can be viewed as a response of a linear system to a random input. In this case the process is asymptotically stationary if the linear system of (2.6) is stable; see [Bal95, §3] or [KSH00, §5.3]. The transfer function from w to x is $\mathbf{A}(z)^{-1}$ where

$$\mathbf{A}(z) = I + z^{-1}A_1 + \cdots + z^{-p}A_p. \tag{2.7}$$

Therefore the AR process (2.6) is stationary if the poles of $\mathbf{A}$ are inside the unit circle. From the expression of $\mathbf{A}(z)^{-1}$, the autoregressive model is sometimes called an all-pole system.

It is easily shown that the AR model parameters A_k, Σ, and the first $p+1$ covariance matrices R_k are related by the linear equations

$$\begin{bmatrix} R_0 & R_1 & \cdots & R_p \\ R_1^T & R_0 & \cdots & R_{p-1} \\ \vdots & \vdots & \ddots & \vdots \\ R_p^T & R_{p-1}^T & \cdots & R_0 \end{bmatrix} \begin{bmatrix} I \\ A_1^T \\ \vdots \\ A_p^T \end{bmatrix} = \begin{bmatrix} \Sigma \\ 0 \\ \vdots \\ 0 \end{bmatrix}. \tag{2.8}$$

These equations are called the *Yule-Walker equations* or *normal equations*. If Σ and A_k, $k = 1, \ldots, p$ are given, this equation is used to compute the autocovariance sequences R_k for $k \geq p$ recursively.

The spectral density matrix of a stationary AR process can be expressed as

$$S(\omega) = \mathbf{A}(e^{\mathrm{j}\omega})^{-1} \Sigma \mathbf{A}(e^{\mathrm{j}\omega})^{-H},$$

in which the canonical factorization directly implies the nonnegativity of $S(\omega)$. The stability of AR model rules out the possibility that $\mathbf{A}(e^{\mathrm{j}\omega})$ can have a zero on the unit circle. In addition we have $S(\omega) \succ 0$ for all ω.

2.2.1 Conditional independence

We can readily apply the conditional dependence relation (2.4) to a Gaussian AR process (2.6). The notation will simplify if we first normalize the input covariance and use an equivalent model:

$$B_0 x(t) = -\sum_{k=1}^{p} B_k x(t-k) + v(t), \tag{2.9}$$

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

From (2.7), the inverse spectrum of an AR process is a trigonometric matrix polynomial

$$S(\omega)^{-1} = \mathbf{A}(e^{\mathrm{j}\omega})^H \Sigma^{-1} \mathbf{A}(e^{\mathrm{j}\omega}) = Y_0 + \frac{1}{2}\sum_{k=1}^{p} (e^{-\mathrm{j}k\omega} Y_k + e^{\mathrm{j}k\omega} Y_k^T) \tag{2.10}$$

where

$$Y_k = \begin{cases} \sum_{l=0}^{p} A_l^T \Sigma^{-1} A_l \quad = \sum_{l=0}^{p} B_l^T B_l, & k = 0 \\ 2\sum_{l=0}^{p-k} A_l^T \Sigma^{-1} A_{l+k} = 2\sum_{l=0}^{p-k} B_l^T B_{l+k}, & k = 1, \ldots, p \end{cases} \tag{2.11}$$

with $A_0 = I$. These expressions show that $(S(\omega)^{-1})_{ij} = 0$ if and only if

$$(Y_k)_{ij} = 0, \quad \text{and} \quad (Y_k)_{ji} = 0, \tag{2.12}$$

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

This connection allows us to parametrize conditional independence relations in terms of AR coefficients and include the zero constraint of the inverse spectrum

in AR estimation methods. From (2.11), a zero in the inverse spectrum becomes quadratic equality constraints on the AR parameters. These are nonconvex constraints and generally difficult to overcome, even though the cost objectives in many estimation problems are already convex, as we will see in the next chapter.

2.2.2 Estimation methods

In this section we describe three time-domain estimation techniques for autoregressive models. The first technique, the linear least-squares method, is probably the most standard approach to many approximation problems. When applied to AR estimation, we show that the method has two variants, depending on the choice of sample covariance matrix. The other two techniques are the maximum-likelihood and the maximum entropy estimation methods. As an optimization problem, these two methods share the same expression of the cost objective. We provide a connection between these three methods in terms of the choice of sample covariance matrix used in each estimation problem.

2.2.2.1 Least-squares linear prediction

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

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

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

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

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

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

$R_k = \mathbf{E}\,x(t+k)x(t)^T$, and $\mathrm{T}(R)$ is the block-Toeplitz matrix with R as its first block row (see the Notation section at the end of chapter 1). Minimizing the prediction error is therefore equivalent to the quadratic optimization problem

$$\text{minimize} \quad \mathbf{tr}(A\,\mathrm{T}(R)A^T) \tag{2.13}$$

with variables $A_1, \ldots, A_p$.

In practice, the covariance matrix $\mathrm{T}(R)$ in (2.13) is replaced by an estimate C computed from samples of $x(t)$. Two common choices are as follows. Suppose samples $x(1), x(2), \ldots, x(N)$ are available.

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

$$C = \frac{1}{N} HH^T, \tag{2.14}$$

  where

$$H = \begin{bmatrix} x(1) & x(2) & \cdots & x(p+1) & \cdots & x(N) & 0 & \cdots & 0 \\ 0 & x(1) & \cdots & x(p) & \cdots & x(N-1) & x(N) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & x(1) & \cdots & x(N-p) & x(N-p+1) & \cdots & x(N) \end{bmatrix}. \tag{2.15}$$

  Note that the matrix C is block-Toeplitz.

- The *covariance method* uses the *non-windowed* estimate

$$C = \frac{1}{N-p} HH^T, \tag{2.16}$$

where

$$H = \begin{bmatrix} x(p+1) & x(p+2) & \cdots & x(N) \\ x(p) & x(p+1) & \cdots & x(N-1) \\ \vdots & \vdots & & \vdots \\ x(1) & x(2) & \cdots & x(N-p) \end{bmatrix}. \tag{2.17}$$

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

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

$$\text{minimize} \quad \mathbf{tr}(ACA^T). \tag{2.18}$$

Here, C is the exact covariance matrix, if available, or one of the two sample estimates (2.14) and (2.16). The first of these estimates is a block-Toeplitz matrix, while the second one is in general not block-Toeplitz. The solutions from both methods share the same asymptotic properties; they are consistent estimates when the AR process is stable with a white Gaussian noise [Lut05]. The covariance method is known to be slightly more accurate in practice if N is small [SM97, page 94]. The correlation method on the other hand has some important theoretical and practical properties, that are easily explained from the optimality conditions of (2.18). If we define $\hat{\Sigma} = ACA^T$ (*i.e.*, the estimate of the prediction error $\mathbf{E}\,\|e(t)\|_2^2$ obtained by substituting C for $\mathrm{T}(R)$), then the optimality conditions can be expressed as

$$\begin{bmatrix} C_{00} & C_{01} & \cdots & C_{pp} \\ C_{10} & C_{11} & \cdots & C_{1p} \\ \vdots & \vdots & & \vdots \\ C_{p0} & C_{p1} & \cdots & C_{pp} \end{bmatrix} \begin{bmatrix} I \\ A_1^T \\ \vdots \\ A_p^T \end{bmatrix} = \begin{bmatrix} \hat{\Sigma} \\ 0 \\ \vdots \\ 0 \end{bmatrix}. \tag{2.19}$$

If C is block-Toeplitz, these equations have the same form as the Yule-Walker equations (2.8), and can be solved more efficiently than when C is not block-

Toeplitz. Another advantage is that the solution of (2.19) always provides a stable model if C is block Toeplitz and positive definite. This can be proved as follows (see [SN87]). Suppose z is a zero of $\mathbf{A}(z)$, *i.e.*, there exists a nonzero w such that $w^H\mathbf{A}(z) = 0$. Define $u_1 = w$ and $u_k = A_{k-1}^T w + \bar{z}u_{k-1}$ for $k = 2, \ldots, p$. Then we have

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

where $u = (u_1, u_2, \ldots, u_p, 0)$, $\tilde{u} = (0, u_1, u_2, \ldots, u_p)$. From this and (2.19),

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

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

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

2.2.2.2 Maximum-likelihood estimation

The exact likelihood function of an AR model (2.6), based on observations $x(1)$, $\ldots$, $x(N)$, is complicated to derive and difficult to maximize [BJ76, Rei07]. A standard simplification is to treat $x(1)$, $x(2)$, $\ldots$, $x(p)$ as fixed, and to define the likelihood function in terms of the conditional distribution of a sequence $x(t)$, $x(t+1)$, $\ldots$, $x(t+N-p-1)$, given $x(t-1)$, $\ldots$, $x(t-p)$. This is called the *conditional* maximum likelihood estimation method [Rei07, §5.1].

The conditional likelihood function of the AR process (2.6) or (2.9) is

$$\begin{aligned}
&\frac{1}{((2\pi)^n \det \Sigma)^{(N-p)/2}} \exp\left(-\frac{1}{2}\sum_{t=p+1}^{N} \mathbf{x}(t)^T A^T \Sigma^{-1} A \mathbf{x}(t)\right) \\
&= \left(\frac{\det B_0}{(2\pi)^{n/2}}\right)^{N-p} \exp\left(-\frac{1}{2}\sum_{t=p+1}^{N} \mathbf{x}(t)^T B^T B \mathbf{x}(t)\right) \qquad (2.20)
\end{aligned}$$

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

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

with $B_0 = \Sigma^{-1/2}$, $B_k = \Sigma^{-1/2} A_k$, $k = 1, \ldots, p$. Taking the logarithm of (2.20) we obtain the conditional log-likelihood function (up to constant terms and factors)

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

where H is the matrix (2.17). If we define $C = (1/(N-p))HH^T$, we can then write the conditional ML estimation problem as

$$\text{minimize} \quad -2 \log \det B_0 + \mathbf{tr}(CB^T B) \qquad (2.21)$$

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

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

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

2.2.2.3 Maximum-entropy estimation

Consider the maximum entropy (ME) problem introduced by Burg [Bur75]:

$$\begin{array}{ll} \text{maximize} & \frac{1}{2\pi}\int_{-\pi}^{\pi} \log\det S(\omega)d\omega \\ \text{subject to} & \frac{1}{2\pi}\int_{-\pi}^{\pi} S(\omega)e^{\mathrm{j}k\omega}d\omega = \bar{R}_k, \quad 0 \leq k \leq p, \end{array} \tag{2.22}$$

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

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

The constraints in (2.22) therefore fix the first $p+1$ covariance matrices to be equal to $\bar{R}_k$. The problem is to extend these covariances so that the entropy rate of the process is maximized. It is known that the solution of (2.22) is a Gaussian AR process of order p, and that the model parameters A_k, Σ follow from the Yule-Walker equations (2.8) with $\bar{R}_k$ substituted for R_k.

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

$$-\frac{1}{2\pi}\int_{-\pi}^{\pi} \log\det S(\omega)d\omega + \sum_{k=0}^{p} \mathbf{tr}(Y_k^T(R_k - \bar{R}_k)).$$

Differentiating with respect to R_k gives

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} S^{-1}(\omega)e^{\mathrm{j}\omega k}d\omega = \begin{cases} Y_k, & k = 0 \\ Y_k/2, & k = 1, \ldots, p \end{cases} \tag{2.23}$$

and hence

$$S^{-1}(\omega) = Y_0 + \frac{1}{2}\sum_{k=1}^{p}\left(Y_k e^{-\mathrm{j}k\omega} + Y_k^T e^{\mathrm{j}k\omega}\right) \;\triangleq\; Y(\omega).$$

Substituting this in the Lagrangian gives the dual problem

$$\text{minimize} \quad -\frac{1}{2\pi}\int_{-\pi}^{\pi} \log\det Y(\omega) + \sum_{k=0}^{p} \mathbf{tr}(Y_k^T \bar{R}_k) - n, \tag{2.24}$$

with variables Y_k. The first term in the objective can be rewritten by using Kolmogorov's formula [Han70]:

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

where $Y(\omega) = \mathbf{B}(e^{\mathrm{j}\omega})^H \mathbf{B}(e^{\mathrm{j}\omega})$ and $\mathbf{B}(z) = \sum_{k=0}^{p} z^{-k} B_k$ is the minimum-phase spectral factor of Y. The second term in the objective of the dual problem (2.24) can also be expressed in terms of the coefficients B_k, using the relations

$$Y_k = \begin{cases} \sum_{i=0}^{p} B_i^T B_i, & k = 0 \\ 2\sum_{i=0}^{p-k} B_i^T B_{i+k}, & k = 1, \ldots, p. \end{cases}$$

This gives

$$\sum_{k=0}^{p} \mathbf{tr}(Y_k^T \bar{R}_k) = \mathbf{tr}(\mathrm{T}(\bar{R}) B^T B),$$

where $\bar{R} = \begin{bmatrix} \bar{R}_0 & \bar{R}_1 \cdots & \bar{R}_p \end{bmatrix}$ and $B = \begin{bmatrix} B_0 & B_1 & \cdots & B_p \end{bmatrix}$. The dual problem (2.24) thus reduces to

$$\text{minimize} \quad -2\log\det B_0 + \mathbf{tr}(C B^T B) \tag{2.25}$$

where $C = \mathrm{T}(\bar{R})$. Without loss of generality, we can choose B_0 to be symmetric positive definite. The problem is then formally the same as the ML estimation problem (2.21), except for the definition of C. In (2.25) C is a block-Toeplitz matrix. If we choose for $\bar{R}_k$ the sample estimates

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

then C is identical to the block-Toeplitz matrix (2.14) used in the autocorrelation variant of the least-squares method.

2.3 Summary

We have described conditional independence relations for Gaussian random variables and time series. The condition can be expressed as a zero pattern in the inverse covariance matrix for Gaussian random variables, and as a zero pattern in the inverse spectral density for Gaussian time series. We have also derived the conditional independence for Gaussian AR processes. The condition can be expressed as quadratic equalities in AR parameters, motivating us to consider a system identification problem that takes this constraint into account.

We have reviewed the least-squares, maximum likelihood, and maximum entropy estimation methods for AR models. These three techniques share some interesting connections via a covariance matrix C. The least-squares method uses a quadratic loss function while the ML and ME estimation use the log-determinant cost objective. Despite this difference, they all lead to optimality conditions that are formally the same as the normal equations, with the covariance matrix replaced by a sample estimate C. The conditional ML estimate is equivalent to the least-squares estimate using the non-block Toeplitz C while the ME estimate is equal to the least-squares estimate using the block-Toeplitz C.

In the next chapter we will combine the maximum-likelihood and maximum-entropy estimation with the conditional independence constraints (2.12). The problem is nonconvex due to the nonconvex constraint from the quadratic equalities (2.11). We will show that under some mild condition, the estimation problem can be formulated in a convex framework.

CHAPTER 3

Estimation of graphical models of AR processes

In this chapter we extend the ML and ME estimation methods for AR processes, described in sections 2.2.2.2 and 2.2.2.3, to include conditional independence constraints. As we have seen, the ML and ME estimation share the same form of a convex optimization problem (2.21) and (2.25), with different choices of the matrix C. The distinction will turn out to be important later, but for now we make no assumptions on C, except that it is positive definite.

As we mentioned in chapter 2, the conditional independence relation imposes a nonconvex constraint to the problem. In section 3.1 we introduce a convex relaxation to the ML and ME estimation with conditional independence constraint. The relaxation, in general, does not provide an exact solution to the original problem. Using the optimality conditions derived in section 3.2, we show the first main result of this thesis in section 3.3. We prove that the relaxation is equivalent to the original problem under a block-Toeplitz assumption on C. This condition is not necessary to guarantee the exactness of the relaxation. As we illustrate by examples in section 3.4, the relaxation is exact under a weaker condition, in practice.

3.1 Convex formulation

Using the notation defined in (1.5) and the conditional independence constraint (2.11)-(2.12), we can write this as

$$(\mathrm{D}_k(A^T\Sigma^{-1}A))_{ij} = 0,$$

where $A = [\ I \quad A_1 \quad \cdots \quad A_p\ ]$, or as

$$\big(\mathrm{D}_k(B^TB)\big)_{ij} = 0, \qquad k = 0, \ldots, p, \tag{3.1}$$

where $B = [\ B_0 \quad B_1 \quad \cdots \quad B_p\ ]$. To simplify the formulation later, we write the constraint (3.1) by using the projection operator defined in (1.6). We assume that the conditional independence constraints are specified via an index set $\mathcal{V}$, with $(i,j) \in \mathcal{V}$ if the processes $x_i(t)$ and $x_j(t)$ are conditionally independent (see the assumptions on $\mathcal{V}$ in the notation section). The constraints (3.1) for $(i,j) \in \mathcal{V}$ can be written as

$$\mathrm{P}_{\mathcal{V}}\left(\mathrm{D}(B^TB)\right) = 0 \tag{3.2}$$

where $\mathrm{P}_{\mathcal{V}}$ is defined in (1.6).

The ML and ME estimation with conditional independence constraints (3.2) can be expressed as

$$\begin{array}{ll} \text{minimize} & -2\log\det B_0 + \mathbf{tr}(CB^TB) \\ \text{subject to} & \mathrm{P}(\mathrm{D}(B^TB)) = 0 \end{array} \tag{3.3}$$

with variable $B = [\ B_0 \quad B_1 \quad \cdots \quad B_p\ ] \in \mathbf{M}^{n,p}$. (Henceforth we drop the subscript of $\mathrm{P}_{\mathcal{V}}$.) The problem (3.3) includes quadratic equality constraints and is therefore nonconvex. The quadratic terms in B suggest the convex relaxation

$$\begin{array}{ll} \text{minimize} & -\log\det X_{00} + \mathbf{tr}(CX) \\ \text{subject to} & \mathrm{P}(\mathrm{D}(X)) = 0 \\ & X \succeq 0 \end{array} \tag{3.4}$$

with variable $X \in \mathbf{S}^{n(p+1)}$ (X_{00} denotes the leading $n \times n$ subblock of X). The convex optimization problem (3.4) is a relaxation of (3.3) and only equivalent to (3.3) if the optimal solution X has rank n, so that it can be factored as $X = B^T B$. In section 3.3 we provide the main result; we show that X has rank n if C is block-Toeplitz. The proof of exactness of the relaxation under assumption of block-Toeplitz structure will be justified from the dual of (3.4) and the optimality conditions derived in section 3.2. To verify the theoretical result, we show some numerical results on the exactness of the relaxation in section 3.4.

3.2 Duality and optimality conditions

Suppose that $C \succ 0$. The goal of this section is to prove that the relaxation (3.4) is equivalent to the original estimation problem (3.3) under a condition on block-Toeplitz structure of C. This can be shown from the dual problem as follows.

The derivation of the dual problem starts with the Lagrangian defined as the cost function plus a weighted sum of the constraints. As the weights of the constraints, we introduce a Lagrange multiplier $Z = [\ Z_0 \quad Z_1 \quad \cdots \quad Z_p\ ] \in \mathbf{M}^{n,p}$ for the equality constraints and a multiplier $U \in \mathbf{S}^{n(p+1)}$ for the inequality constraint. The Lagrangian is

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

Here we made use of the fact that the mappings T and D are adjoints, and that P is self-adjoint (see the definitions from section 1.4). The Lagrange dual function

$g(Z,U)$ is defined as

$$g(Z,U) = \inf_{X_{00} \succ 0} L(X,Z,U) = \inf_{X_{00} \succ 0} \left(-\log\det X_{00} + \mathbf{tr}\left((C + \mathrm{T}(\mathrm{P}(Z)) - U)X \right) \right).$$

Setting the gradient of $L(X,Z,U)$ with respect to X equal to zero gives

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

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

The dual function is therefore given by

$$g(Z,U) = \begin{cases} \log\det(C_{00} + \mathrm{P}(Z_0) - U_{00}) + n, & C_{i,i+k} + \mathrm{P}(Z_k) - U_{i,i+k} = 0, \\ & \qquad k = 1, \ldots, p, \quad i = 0, \ldots, p-k \\ -\infty, & \text{otherwise} \end{cases}$$

($C_{i,j}$ and $U_{i,j}$ denote the (i,j) blocks of size $n \times n$ of C and U). The Lagrange dual problem is to maximize $g(Z,U)$ over $U \succeq 0$ (where U is the Lagrange multiplier corresponding to the inequality constraint). Hence we arrive at the dual problem

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

If we define $W = C_{00} + \mathrm{P}(Z_0) - U_{00}$ we can eliminate the slack variable U, and write the dual problem of (3.4) more simply as

$$\begin{array}{ll} \text{maximize} & \log\det W + n \\ \text{subject to} & \begin{bmatrix} W & 0 \\ 0 & 0 \end{bmatrix} \preceq C + \mathrm{T}(\mathrm{P}(Z)). \end{array} \tag{3.5}$$

Note that for $p = 0$ problem (3.4) reduces to the covariance selection problem (1.1), and the dual problem reduces to the maximum determinant completion problem

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

This completion problem is to determine whether a completion of C, specified by the nonzero entries in $\mathrm{P}(Z)$, is positive definite or not (if a completion exists). Among all the positive definite completions, there is the unique solution that maximizes the determinant [GHJ99].

The optimal duality gap defined as the difference between the optimal values of the primal and the dual problems, (3.4), (3.5), is given by

$$\eta = \log\det X_{00}^* + \mathbf{tr}(CX^*) - \log\det W^* - n = \mathbf{tr}(CX^*) - n$$

where X^* is the optimal solution of (3.4) and W^*, Z^* are the optimal solutions of (3.5). The optimality gap is always nonnegative and when it is zero, we say strong duality holds. For convex optimization problems, strong duality holds if Slater's condition is satisfied [BV04, §5.2.3], *i.e.*, either there exists a strictly primal feasible X in (3.4) or strictly dual feasible Z, U in (3.5).

We note the following properties of the primal problem (3.4) and the dual problem (3.5).

- The primal problem is strictly feasible ($X = I$ is strictly feasible), so Slater's

condition holds. This implies strong duality, and also that the dual optimum is attained if the optimal value is finite.

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

By the definitions of L and g

$$-\log\det X_{00}^* + \mathbf{tr}(CX^*) \geq L(X^*, Z^*, U^*) \geq g(Z^*, U^*).$$

The first inequality holds since $\mathbf{tr}(Z^{*T}\,\mathrm{P}(\mathrm{D}(X^*))) = 0$ and $\mathbf{tr}(U^*X^*) \geq 0$. The second inequality follows from the definition of the dual function g. The zero duality gap implies that the two inequalities hold with equality. Therefore, in order to have the first inequality tight, we must have $\mathbf{tr}(U^*X^*) = 0$ or equivalently $U^*X^* = 0$ since $U^* \succeq 0, X^* \succeq 0$. This condition is known as complementary slackness.

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

1. *Primal feasibility.*

$$X \succeq 0, \qquad X_{00} \succ 0, \qquad \mathrm{P}(\mathrm{D}(X)) = 0, \tag{3.6}$$

2. *Dual feasibility.*

$$W \succ 0, \qquad C + \mathrm{T}(\mathrm{P}(Z)) \succeq \begin{bmatrix} W & 0 \\ 0 & 0 \end{bmatrix}. \tag{3.7}$$

3. *Complementary slackness.*

$$X_{00}^{-1} = W, \qquad \mathbf{tr}\left(X\left(C + \mathrm{T}(\mathrm{P}(Z)) - \begin{bmatrix} W & 0 \\ 0 & 0 \end{bmatrix}\right)\right) = 0. \tag{3.8}$$

The last condition can also be written as

$$X\left(C + \mathrm{T}(\mathrm{P}(Z)) - \begin{bmatrix} W & 0 \\ 0 & 0 \end{bmatrix}\right) = 0. \tag{3.9}$$

(if A, B are positive semidefinite matrices then $\mathbf{tr}(AB) = 0$ if and only if $AB = 0$.) These conditions are called the Karush-Kuhn-Tucker (KKT) conditions which will be the basis of our main result in the following section.

3.3 Properties of block-Toeplitz sample covariances

In this section we study in more detail the solution of the primal and dual problems (3.4) and (3.5) if C is block-Toeplitz. The results can be derived from connections between spectral factorization, semidefinite programming, and orthogonal matrix polynomials discussed in [Hac03, §6.1.1]. In this section, we provide alternative and self-contained proofs.

Assume C has a block Toeplitz structure, *i.e.*, $C = \mathrm{T}(R)$ for some $R \in \mathbf{M}^{n,p}$ and that C is positive definite.

Exactness of the relaxation We first show that the relaxation (3.4) is exact when C is block-Toeplitz, *i.e.*, the optimal X^* has rank n and the optimal B can be computed by factoring X^* as $X^* = B^T B$. We prove this result from the optimality conditions (3.6)–(3.9).

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

rank of

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

is at least np, and the complementary slackness condition (3.9) implies that X^* has rank at most n, so we can conclude that

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

(recall that $C, X \in \mathbf{S}^{n(p+1)}$). The positive definiteness of $C + \mathrm{T}(\mathrm{P}(Z^*))$ follows from the dual feasibility condition (3.7) and the following basic property of block-Toeplitz matrices: If $\mathrm{T}(S)$ is a symmetric block-Toeplitz matrix, with $S \in \mathbf{M}^{n,p}$, and

$$\mathrm{T}(S) \succeq \begin{bmatrix} Q & 0 \\ 0 & 0 \end{bmatrix} \tag{3.10}$$

for some $Q \in \mathbf{S}^n_{++}$, then $\mathrm{T}(S) \succ 0$. We can verify this by induction on p. The property is obviously true for $p = 0$, since the inequality (3.10) then reduces to $S = S_0 \succeq Q$. Suppose the property holds for $p - 1$. Then (3.10) implies that the leading $np \times np$ submatrix of $\mathrm{T}(S)$, which is a block Toeplitz matrix with first row $\begin{bmatrix} S_0 & \cdots & S_{p-1} \end{bmatrix}$, is positive definite. Let us denote this matrix by V. Using the Toeplitz structure, we can partition $T(S)$ as

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

where $V \succ 0$. The inequality (3.10) implies that the Schur complement of V in the matrix $\mathrm{T}(S)$ satisfies

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

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

Stability of estimated models It follows from (3.6)–(3.9) and the factorization $X^* = B^T B$, that

$$(C + \mathrm{T}(\mathrm{P}(Z))) \begin{bmatrix} I \\ A_1^T \\ \vdots \\ A_p^T \end{bmatrix} = \begin{bmatrix} \Sigma \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \tag{3.11}$$

if we define $\Sigma = B_0^{-2}$, $A_k = B_0^{-1} B_k$. These equations are Yule-Walker equations with a positive definite block-Toeplitz coefficient matrix. As mentioned at the end of section 2.2.2.1, this implies that the zeros of $\mathbf{A}(z) = I + z^{-1}A_1 + \cdots + z^{-p}A_p$ are inside the unit circle. Therefore the solution to the convex problem (3.4) provides a stable AR model.

3.4 Examples with randomly generated data

In this section we evaluate the ML and ME estimation methods on several data sets. The convex optimization package `CVX` [GB08a, GB08b] was used to solve the ML and ME estimation problems with small dimensions. We will further investigate large-scale algorithms in chapter 5.

The first set of experiments uses data randomly generated from AR models with sparse inverse spectra. The purpose is to examine the quality of the semidefinite relaxation (3.4) of the ML estimation problem for finite N. We generated 50 sets of time series from four AR models of different dimensions. We solved (3.4) for different N. Figure 3.1 shows the percentage of the 50 data sets for which the relaxation was exact (the optimal X in (3.4) had rank n.) The results illustrate that the relaxation is often exact for moderate values of N, even when the matrix C is not block-Toeplitz.

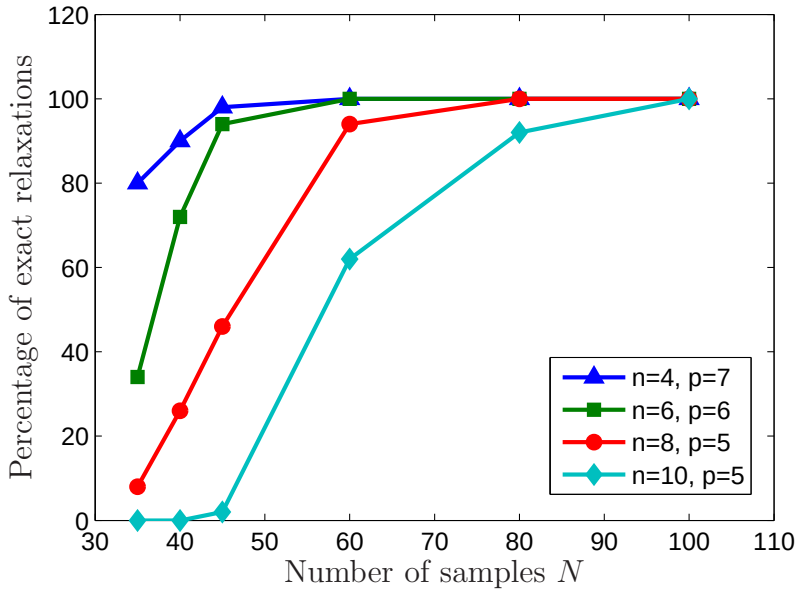

Figure 3.1: Number of cases where the convex relaxation of the ML problem is exact, versus the number of samples.

The next figure shows the convergence rate of the ML and ME estimates, with and without imposed conditional independence constraints, to the true model, as a function of the number of samples. The data were generated from an AR model of dimension $n = p = 6$ with nine zeros in the inverse spectrum. We use the Kullback-Leibler (KL) divergence between two zero-mean Gaussian processes [BJ04]

$$J(S(\omega)\|G(\omega)) = \frac{1}{2\pi}\int_0^{2\pi} I(S(\omega)\|G(\omega))d\omega$$

where $I(S\|G) = -(1/2)\log\det(SG^{-1}) - (1/2)\,\mathbf{tr}(I - SG^{-1})$, as a measure of estimation accuracy. Figure 3.2 shows the KL divergence between the estimated and the true spectra as a function of N, for four estimation methods: the ML and ME estimation methods without conditional independence constraints, and the ML and ME estimation methods with the correct conditional independence constraints. We notice that the KL divergences decrease at the same rate for the four

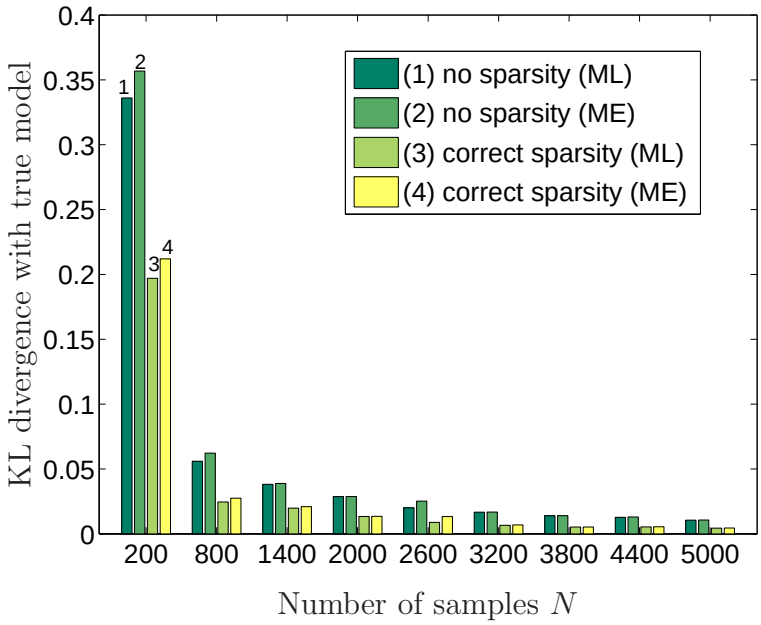

Figure 3.2: KL divergence between estimated AR models and the true model ($n = 6$, $p = 6$) versus the number of samples.

estimates. However, the ML and ME estimates without the sparsity constraints give models with substantially larger values of KL divergence when N is small. For sample size under 3000, the ME estimates (with and without the sparsity constraints) are also found to be less accurate than their ML counterparts. This effect is well known in spectral analysis (see, for example, [SM97, page 94]). As N increases, the difference between the ME and ML methods disappears.

3.5 Summary

We have proposed convex relaxations for the problems of conditional ML and ME estimation of AR models with conditional independent constraints. The two problems have the same form with different choices for the sample covariance matrix C. For the ME problem, C is given by (2.14), while for the conditional

ML problem, it is given by (2.16). In both cases, C is positive definite if the information matrix H has full rank. This is sufficient to guarantee that the relaxed problem (3.4) is bounded below.

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

CHAPTER 4

Topology selection

In chapter 3 we have introduced a convex formulation of the problem of estimating AR models when a conditional independence structure is given. This section considers the other class of estimation problem in graphical modeling, where the topology of the graph is unknown. We focus on topology selection in graphical models of AR processes.

Discovering a sparsity pattern of a graph is closely related to model selection problems. Each graph topology corresponds to a model with a different number of parameters. When using these models to explain given data, a model with more parameters certainly improves the fitting error; however, it also results in increased variance. Therefore, one should pick a model with the smallest possible number of parameters that can adequately explain the data. This concept is known as *the principle of parsimony* [BA02], and the general tradeoff between a bias versus variance in statistics.

Section 4.1 presents a direct approach to the topology selection problem. The method is to enumerate all topologies, and rank the models using information-theoretic criteria, each of which introduces a penalty term for model complexity. As we will see from examples in section 4.2, this approach is limited to small graphs only since the number of all possible topologies grows exponentially as a function of n.

Therefore, section 4.3 considers a more efficient framework to discover the topology of large graphical models. A common approach to find sparse models is typically based on ℓ_1-norm regularization. Examples are the ℓ_1-regularized least-squares problem (or *the Lasso*) considered in [Tro06, CRT06a, CRT06b, FNW07] and the covariance selection problem (1.2). In section 4.3.2 we propose a convex formulation based on ℓ_1-type regularization to encourage sparsity in the inverse spectrum for topology selection in AR models. Examples with random and real data sets are illustrated in section 4.4 and 4.5.

4.1 Model selection via information criteria

In model selection, we make a statistical inference from the data to select a good approximate model from a set of candidates. A theoretical basis for model selection includes model selection criteria as a measure of goodness of fit of an estimated model. Three popular model selection criteria are the *Akaike Information Criterion* (AIC), the second-order variant of AIC (AIC_c), and the *Bayes information criterion* (BIC) [BA02]. These criteria are used to make a fair comparison between models of different complexity. They assign to an estimated model a score equal to $-2\mathcal{L}$, where $\mathcal{L}$ is the likelihood of the model, augmented with a term that depends on the effective number of parameters k in the model:

$$\begin{aligned} \text{AIC} &= -2\mathcal{L} + 2k, && (4.1)\\ \text{AIC}_c &= -2\mathcal{L} + \frac{2kN}{N-k-1}, && (4.2)\\ \text{BIC} &= -2\mathcal{L} + k\log N, && (4.3) \end{aligned}$$

where N is the sample size. The second term places a penalty on models with high complexity. When comparing different models, we rank them according to one of the criteria and select the model with the lowest score. Of these three criteria,

the AIC is known to perform poorly if N is small compared to the number of parameters k. The $\mathrm{AIC_c}$ was developed as a correction to the AIC for small N. For large N the BIC favors simpler models than the AIC or $\mathrm{AIC_c}$.

To select a suitable graphical AR model for observed samples of an n-dimensional time series, we can enumerate models of different lengths p and with different graphs. For each model, we solve the ML estimation problem with conditional independence constraints (3.4), calculate the AIC, $\mathrm{AIC_c}$, or BIC score, and select the model with the best (lowest) score. Obviously, an exhaustive search of all sparsity patterns is only feasible for small n (say, $n \leq 6$), since there are

$$\sum_{m=0}^{n(n-1)/2} \binom{n(n-1)/2}{m} = 2^{n(n-1)/2} \tag{4.4}$$

different graphs with n nodes.

This approach can be more clearly illustrated by the following example. We generate $N = 1000$ samples from an AR model of dimension $n = 5$, $p = 4$, and zeros in positions $(1,2)$, $(1,3)$, $(1,4)$, $(2,4)$, $(2,5)$, $(4,5)$ of the inverse spectrum. We show only results for the BIC. In the BIC we substitute the conditional likelihood discussed in section 2.2.2.2 for the exact likelihood $\mathcal{L}$. (For sufficiently large N the difference is negligible.) As effective number of parameters we take the total number of optimization variables

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

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

Figure 4.1 shows the scores of the estimated models as a function of p. For each p the score shown is the best score among all graph topologies. The BIC selects the correct model order $p = 4$. Figure 4.2 shows the seven best models

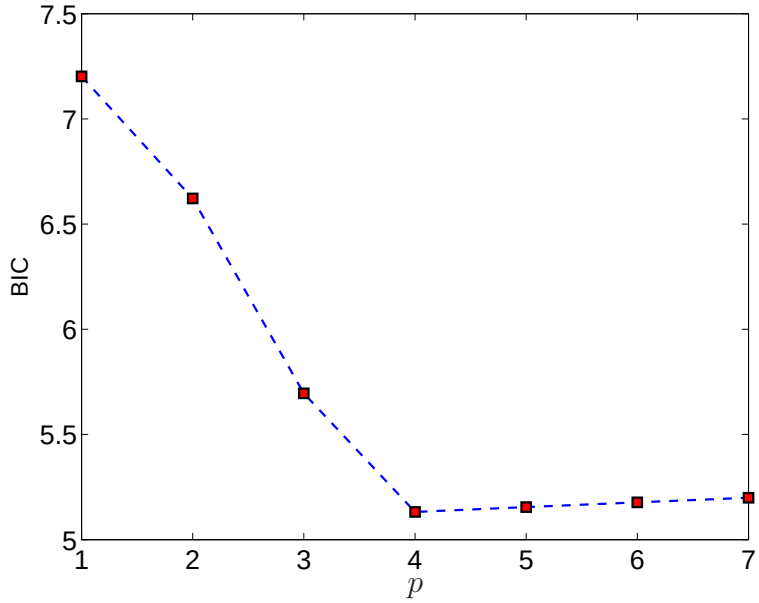

Figure 4.1: BIC score scaled by $1/N$ of AR models of order p.

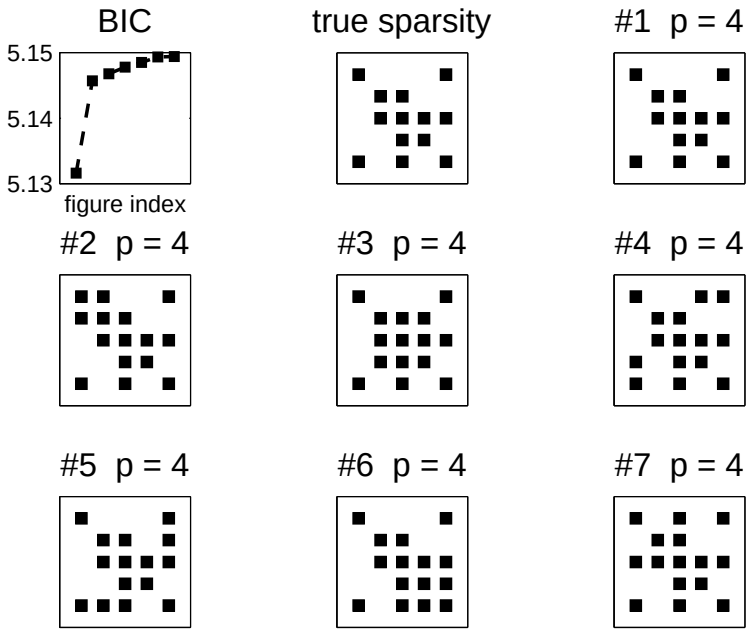

Figure 4.2: Seven best ranked topologies according to the BIC.

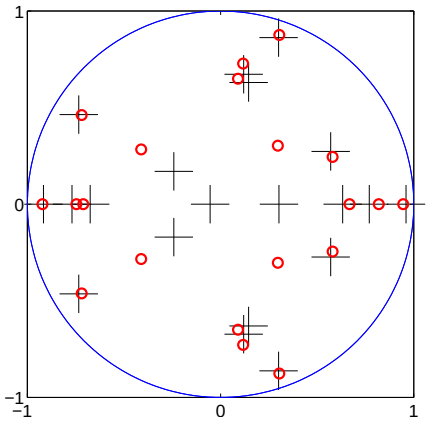

Figure 4.3: Poles of the true model (plus signs) and the estimated model (circles).

according to the BIC. The subgraphs labeled #1 to #7 show the estimated model order p, and the selected sparsity pattern. The corresponding scores are shown in the first subgraph, and the true sparsity pattern is shown in the second subgraph. The BIC identified the correct sparsity pattern. Figure 4.3 shows the location of the poles of the true AR model and the model selected by the BIC.

In figures 4.4 and 4.5 we compare the spectrum of the model selected by the BIC with the spectrum of the true model and with a nonparametric estimate of the spectrum. The lower half of the figures shows the *coherence spectrum*, *i.e.*, the spectrum normalized to have diagonal one:

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

where $\mathbf{diag}(S)$ is the diagonal part of S. The upper half shows the *partial coherence spectrum* $R(\omega)$ [Bri81, Dah00], *i.e.*, the inverse spectrum normalized to have diagonal one:

$$R(\omega) = \mathbf{diag}(S(\omega)^{-1})^{-1/2} S(\omega)^{-1}\, \mathbf{diag}(S(\omega)^{-1})^{-1/2}. \tag{4.5}$$

The i, j entry of the coherence spectrum is a measure of how dependent components i and j of the time series are. The i, j entry of the partial coherence

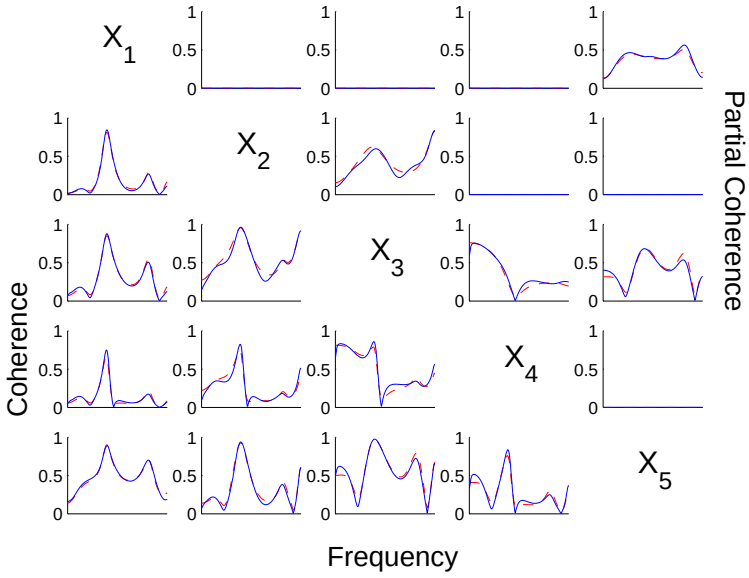

Figure 4.4: Partial coherence and coherence spectra of the AR model: true spectrum (dashed lines) and ML estimates (solid lines).

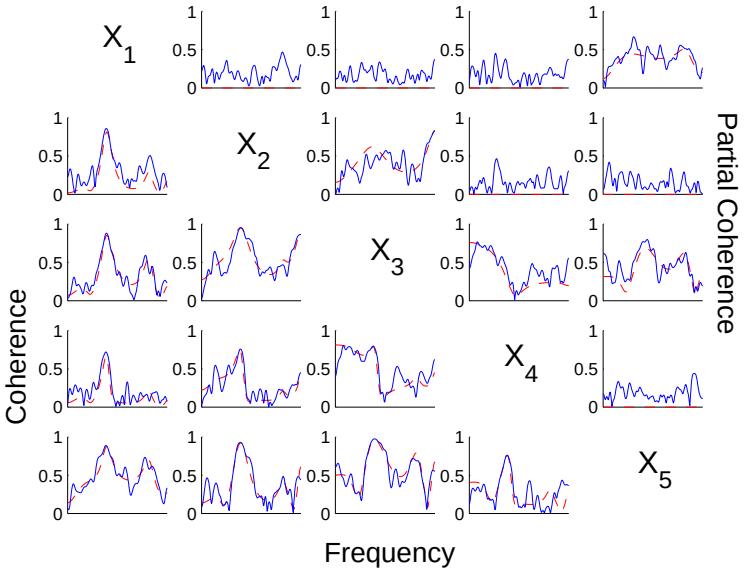

Figure 4.5: Partial coherence and coherence spectra of the AR model: true spectrum (dashed lines) and nonparametric estimates (solid lines).

spectrum on the other hand measures the *conditional* dependence between the components i and j, after removing the linear effects from the other variables. The dashed lines show the spectra of the true model. The solid lines in figure 4.4 are the spectra of the ML estimates. The solid lines in figure 4.5 are nonparametric estimates of the spectrum, obtained with Welch's method (see [Pro01, §12.2.2]) using a Hamming window of length 40 (see [Pro01, page 642]). The nonparametric estimate in figure 4.5 of the partial coherence spectrum clearly gives a poor indication of the correct sparsity pattern.

4.2 Examples with small real data sets

Air pollution data

This data set consists of a time series of dimension $n = 5$. The components are four air pollutants, CO, NO, NO_2, O_3, and the solar radiation intensity R, recorded hourly during 2006 at Azusa, California. The entire data set consists of $N = 8370$ observations, and was obtained from Air Quality and Meteorological Information System (AQMIS) (`www.arb.ca.gov/aqd/aqdcd/aqdcd.htm`). The daily averages over one year are shown in figure 4.6. A similar data set was studied previously in [Dah00], using a nonparametric approach.

We use the BIC to compare models with orders ranging from $p = 1$ to $p = 8$. Table 4.1 lists the models with the best ten BIC scores (which differ by only 0.84%). Figure 4.7 shows the coherence and partial coherence spectra obtained from a nonparametric estimation (solid red lines), and the ML model with the best BIC score (dashed blue lines).

From table 4.1, the lowest BIC scores of each model of order $p = 4$, 5, 6 correspond to the missing edge between NO and the solar radiation. This agrees

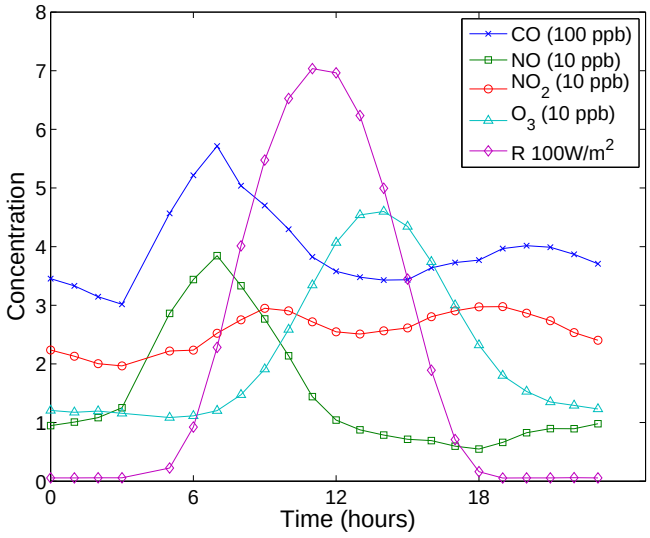

Figure 4.6: Average of daily concentration of CO, NO, NO_2, and O_3, and the solar radiation (R).

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

Table 4.1: Models with the lowest BIC scores for the air pollution data, determined by an exhaustive search of all models of orders $p = 1, \ldots, 8$. $\mathcal{V}$ is the set of conditionally independent pairs in the model.

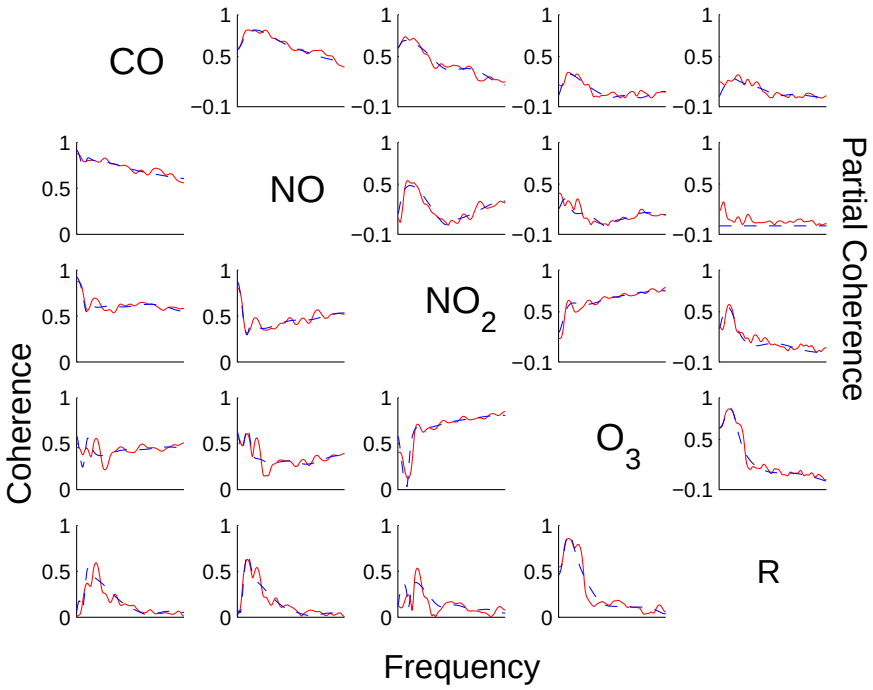

Figure 4.7: Coherence (lower half) and partial coherence spectra (upper half) for the first model in table 4.1. Nonparametric estimates are in solid red lines, and ML estimates in dashed blue lines.

with the empirical partial coherence in figure 4.7 where the pair NO-R is weakest. Table 4.1 also suggests that other weak links are $(\mathrm{CO}, \mathrm{O}_3)$ and $(\mathrm{CO}, \mathrm{R})$. The partial coherence spectra of these pairs are not identically zero, but are relatively small compared to the other pairs.

The presence of the stronger components in the partial coherence spectra are consistent with the discussion in [Dah00]. For example, the solar radiation plays a role in the photolysis of NO_2 and the generation of O_3. The concentration of CO and NO are highly correlated because both are generated by traffic.

International stock markets

We consider a multivariate time series of five stock market indices: the S&P 500 composite index (U.S.), Nikkei 225 share index (Japan), the Hang Seng stock composite index (Hong Kong), the FTSE 100 share index (United Kingdom), and the Frankfurt DAX 30 composite index (Germany). The data were downloaded from `www.globalfinancialdata.com` with the record from June 4, 1997 to June 15, 1999. (The data were converted to US dollars to take the volatility of exchange rates into account. We also replaced missing data due to national holidays by the most recent values.) For each market we use as variable the return between trading day $k-1$ and k, defined as

$$r_k = 100 \log(p_k/p_{k-1}), \tag{4.6}$$

where p_k is the closing price on day k. The resulting five-dimensional time series of length 528 is shown in figure 4.8. This data set is a subset of the data set used in [BY03].

We enumerate all graphical models of orders ranging from $p = 1$ to $p = 9$. Because of the relatively small number of samples, the $\mathrm{AIC_c}$ criterion will be used to compare the models. Figure 4.9 shows the optimal $\mathrm{AIC_c}$ (optimized over all models of a given lag p) versus p. Table 4.2 shows the model order and topology of the five models with the best $\mathrm{AIC_c}$ scores. The column labeled $\mathcal{V}$ shows the list of conditionally independent pairs of variables.

Figure 4.10 shows the coherence (bottom half) and partial coherence (upper half) spectra for the model selected by the $\mathrm{AIC_c}$, and for a nonparametric estimate.

It is interesting to compare the results with the conclusions in [BY03]. For example, the authors of [BY03] mention a strong connection between the German

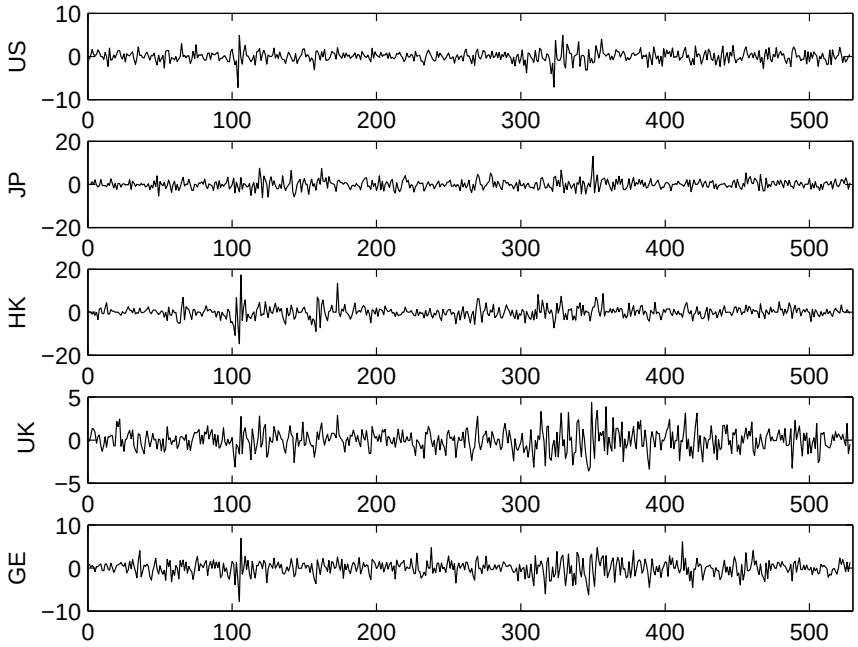

Figure 4.8: Detrended daily returns for five stock market indices between June 4, 1997 and June 15, 1999.

Table 4.2: Five best AR models, ranked according to $\mathrm{AIC_c}$ scores, for the international stock market data.

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

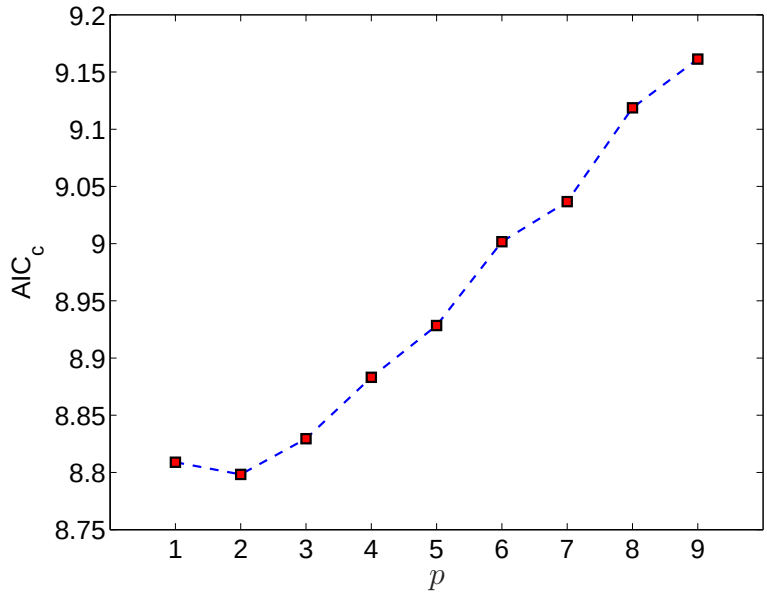

Figure 4.9: Minimized $\mathrm{AIC_c}$ scores (scaled by $1/N$) of p^{th}-order models for the stock market return data.

and the other European stock markets, in particular, the UK. This agrees with the high value of the UK-GE component of the partial coherence spectrum in figure 4.10. The lower strength of the connections between the Japanese and the other stock markets is also consistent with the findings in [BY03]. Another conclusion from [BY03] is that the volatility in the US stock markets transmits to the world through the German and Hong Kong markets. As far as the German market is concerned, this seems to be confirmed by the strength of the US-GE component in the partial coherence spectrum.

European stock markets

This data set is similar to the previous one. We consider a five-dimensional time series consisting of the following stock market indices: the FTSE 100 share index (United Kingdom), CAC 40 (France), the Frankfurt DAX 30 composite

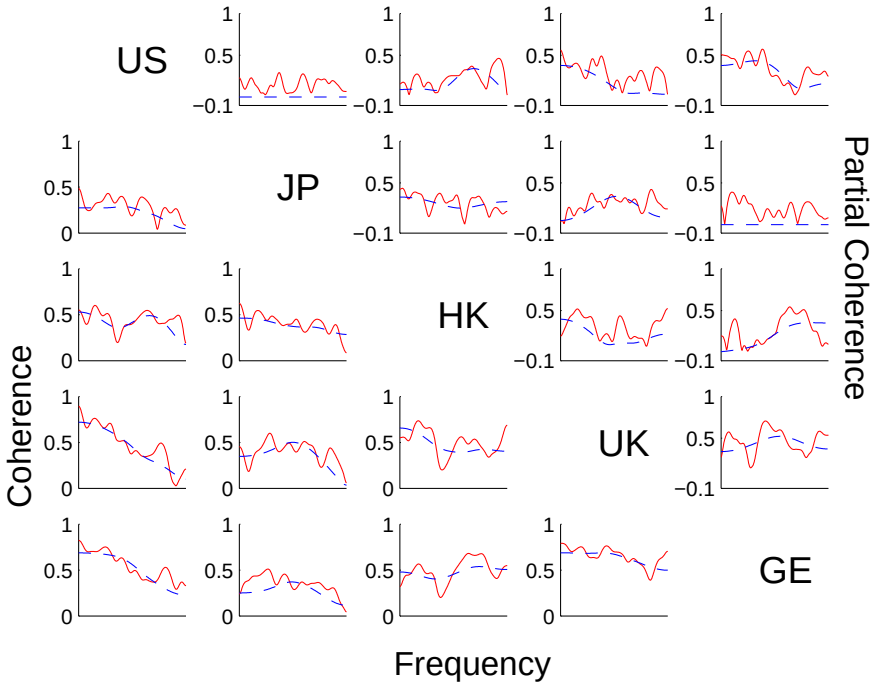

Figure 4.10: Coherence and partial coherence spectra of international stock market data, for the first model in table 4.2. Nonparametric estimates are shown in solid red lines and ML estimates are shown in dashed blue lines.

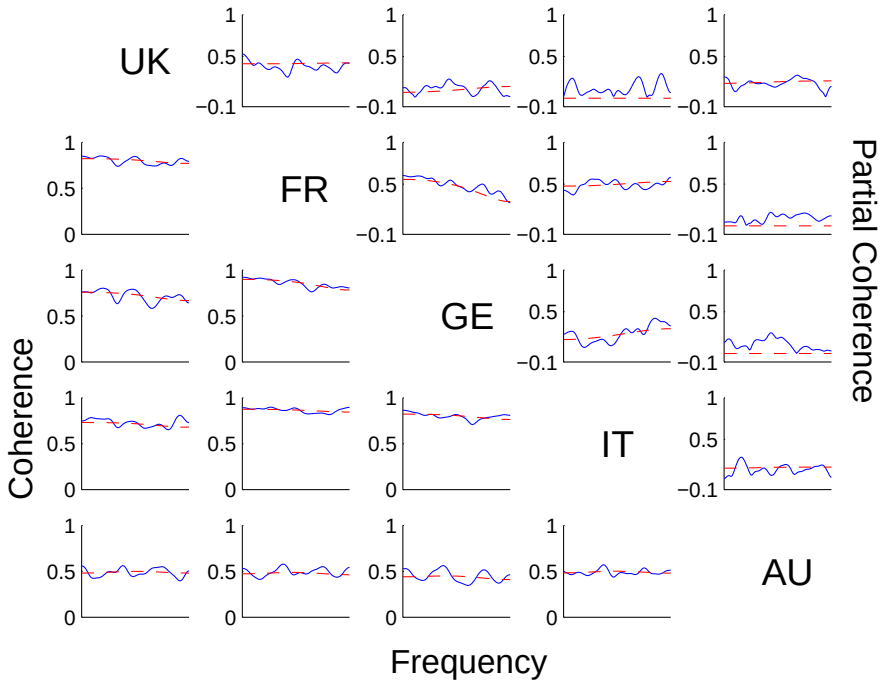

Figure 4.11: Coherence and partial coherence spectrum of the model for the European stock return data. Nonparametric estimates (solid red lines) and ML estimates (dashed blue lines) for the best model selected by the BIC.

index (Germany), MIBTEL (Italy), Austrian Traded Index ATX (Austria). The data were stock index closing prices recorded from January 1, 1999 to July 31, 2008, and obtained from `www.globalfinancialdata.com`. The stock market daily returns were computed from (4.6), resulting in a five-dimensional time series of length $N = 2458$.

The BIC selects a model with lag $p = 1$, and with (UK,IT), (FR,AU), and (GE, AU) as the conditionally independent pairs. The coherence and partial coherence spectra for this model are shown in figure 4.11. The partial coherence spectrum suggests that the French stock market is the market on the Continent most strongly connected to the UK market. The French, German, and Italian

stock markets are highly inter-dependent, while the Austrian market is more weakly connected to the other markets. These results agree with conclusions from the analysis in [YML03].

4.3 Model selection via ℓ_1-regularized ML estimation

As we have seen in section 4.1, the 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 we describe a more scalable approach based on a convex optimization problem that extends the ℓ_1-norm regularization (1.2) for sparse covariance selection.

4.3.1 Regularization methods

In statistical learning, a regularization is a technique used to add prior knowledge about the behavior of the solution. A typical example of regularization is *Ridge regression or Tikhonov regularization* [HTF09, §3.4] of a least-squares problem:

$$\text{minimize} \quad \|Ax - b\|_2^2 + \gamma \|x\|_2^2$$

with $\gamma > 0$. This regularization penalizes the norm of x and control its amount by the weighting parameter γ. In estimation, it can alleviate the problem of having high variance in x due to variations in A. The Tikhonov estimate has the analytical solution

$$x = (A^T A + \gamma I)^{-1} A^T b.$$

The regularization can also be used when A is ill-conditioned. As we see, it requires no assumption on the rank of A since $A^T A + \gamma I$ is always invertible for any $\gamma > 0$.

Another regularization that is very relevant to our study is ℓ_1-norm regularization used in the *Lasso* (in signal processing also known as *basis pursuit*):

$$\text{minimize} \quad \|Ax - b\|_2^2 + \gamma\|x\|_1.$$

By making γ large enough, some coefficients of x become zero because of the nature of ℓ_1 norm, so the Lasso is used as a heuristic for regression selection to find a sparse solution.

Regularization methods can be also interpreted as a maximum a posterior probability (MAP) estimation. From the above two examples, the ℓ_2 penalty corresponds to log-prior of the Gaussian distribution, and the ℓ_1 penalty log-prior of the Laplace distribution for each x_i.

In the next section we consider a regularization of the ML estimation that is similar to the ℓ_1 regularization in the least-squares problem.

4.3.2 ℓ_1-regularized ML estimation

From the conditional independence constraints in (3.4), our goal is to promote a sparsity in $\mathrm{D}(X)$. In analogy with the convex heuristic for covariance selection (1.2), we can formulate a regularized ML problem by adding a nonsmooth ℓ_1-type penalty:

$$\begin{array}{ll} \text{minimize} & -\log\det X_{00} + \mathbf{tr}(CX) + \gamma h(\mathrm{D}(X)) \\ \text{subject to} & X \succeq 0, \end{array} \tag{4.7}$$

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

$$h(Y) = h_\infty(Y) = \sum_{j>i} \max\left\{ |(Y_0)_{ij}|, \max_{k=1,\ldots,p} |(Y_k)_{ij}|, \max_{k=1,\ldots,p} |(Y_k)_{ji}| \right\} \tag{4.8}$$

i.e., a sum of the ℓ_∞-norms of vectors of i, j and j, i-entries of the coefficients Y_k. In the examples (section 4.4.3) 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) [ZRY09] and the *group lasso* [YL06, KKK06].

When $p = 0$ and $X \in \mathbf{S}^n$ in (4.7) the penalty term reduces to $\sum_{i>j} |X_{ij}|$ and we obtain the formulation (1.2), studied in [BEd08, Lu09, FHT08], with the minor difference that we do not penalize the diagonal entries of X.

In the following we will use the result from duality to conclude the low rank property of the optimal X, similar to the result shown in section 3.2. The dual problem will also become important later in section 5, in terms of numerical implementation. 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

$$\begin{aligned} & L(X, Y, Z, U) \\ = \; & -\log\det X_{00} + \mathbf{tr}(CX) + \gamma h_\infty(Y) - \mathbf{tr}(UX) + \mathbf{tr}(Z^T(\mathrm{D}(X) - Y)) \quad (4.9) \\ = \; & -\log\det X_{00} + \mathbf{tr}((C + \mathrm{T}(Z) - U)X) + \gamma h_\infty(Y) - \mathbf{tr}(Z^T Y). \end{aligned}$$

(Recall that the mappings T and D defined in (1.4) and (1.5) are adjoints, *i.e.*, $\mathbf{tr}(Z^T\,\mathrm{D}(X)) = \mathbf{tr}(\mathrm{T}(Z)X)$.) 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

$$\mathbf{diag}(Z_k) = 0, \quad k = 0, 1, \ldots, p. \tag{4.10}$$

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

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

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

$$2|Z_{0,ij}| + \sum_{k=1}^{p} (|(Z_k)_{ij}| + |(Z_k)_{ji}|) \leq \gamma, \quad i \neq j, \tag{4.11}$$

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

$$\inf_{Y} L(X, Y, Z, U) = \begin{cases} -\log\det X_{00} + \mathbf{tr}((C + \mathrm{T}(Z) - U)X) & (4.10),\ (4.11) \\ -\infty & \text{otherwise.} \end{cases}$$

Next, we carry out the minimization over X. The terms in X_{00} are bounded below if only if $(C + \mathrm{T}(Z) - U)_{00} \succ 0$, 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 $(i, j) \neq (0, 0)$. This gives a third set of dual feasibility conditions:

$$(C + \mathrm{T}(Z) - U)_{00} \succ 0, \qquad (C + \mathrm{T}(Z) - U)_{ij} = 0, \quad (i, j) \neq 0, \tag{4.12}$$

and an expression for the dual function

$$g(Z, U) = \inf_{X,Y} L(X, Y, Z, U) = \begin{cases} \log\det(C + \mathrm{T}(Z) - U)_{00} + n & (4.10),\ (4.11),\ (4.12) \\ -\infty & \text{otherwise.} \end{cases}$$

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

$$\begin{array}{ll} \text{maximize} & \log\det W + n \\ \text{subject to} & \begin{bmatrix} W & 0 \\ 0 & 0 \end{bmatrix} \preceq C + \mathrm{T}(Z) \\ & \sum_{k=0}^{p} (|(Z_k)_{ij}| + |(Z_k)_{ji}|) \leq \gamma, \quad i \neq j \\ & \mathbf{diag}(Z_k) = 0, \quad k = 0, \ldots, p. \end{array} \tag{4.13}$$

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

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

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

If a sum of ℓ_α-norms

$$h_\alpha(Y) = \sum_{j>i} \left(\sum_{k=0}^{p} (|(Y_k)_{ij}|^\alpha + |(Y_k)_{ji}|^\alpha) \right)^{1/\alpha} \tag{4.14}$$

is used as penalty function in (4.7), the second constraint in the corresponding dual problem (4.13) 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 $\beta = (\alpha - 1)/\alpha$.

4.3.3 Optimality conditions

The optimal duality gap between the optimal values of the primal problem (4.7) and the dual problem (4.13) is

$$\eta = -\log\det X_{00}^* + \mathbf{tr}(CX^*) + \gamma h(\mathrm{D}(X^*)) - \log\det W^* - n,$$

where X^* is the optimal solution of (4.7) and W^* is the optimal solution of (4.13). For convex optimization problems, the duality gap is zero if either the primal problem or the dual problem is strictly feasible.

The primal problem (4.7) is always strictly feasible ($X = I$ is strictly feasible). The dual problem (4.7) 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 (4.9), we have

equality between the Lagrangian and the primal objective if $\mathbf{tr}(UX) = 0$. This is the complementary slackness condition that holds at optimum and can be equivalently expressed as

$$\mathbf{tr}\left(X\left(C + \mathrm{T}(Z) - \begin{bmatrix} W & 0 \\ 0 & 0 \end{bmatrix}\right)\right) = 0. \tag{4.15}$$

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

$$X_{00}^{-1} = W. \tag{4.16}$$

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

$$(Z_k)_{ij} = 0,\ |(Y_k)_{ij}| \leq t_{ij}$$

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

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

The conditions (4.15)–(4.16) show that the optimal X has rank n under the same conditions as for the problem with given sparsity pattern (3.4). 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 (by reviewing the property of block-Toeplitz matrix in section 3.3.) Under these conditions, the optimization problem (4.7) is equivalent to a regularized (conditional) ML estimation problem for the model parameters B:

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

4.4 Examples with randomly generated data

This section illustrates the effectiveness of our method by means of a few numerical experiments on randomly generated sparse AR models. We first show how we use the topology selection problem (4.7) in practice. We discuss how to select the weighting parameter γ, how to determine the topology from the computed solution, and examine the accuracy of the selected model. The experiments also include a comparison between our method and other types of regularization and a comparison of different penalty functions h_α for the regularization problem.

4.4.1 Method

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

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 (1.2) or closely related problems). A common approach is to select γ via cross-validation; see,

for example, [FHT08, HLP06, 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.* [ARS09] consider γ as a random variable and use a maximum a posterior 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 (4.7), *i.e.*, between the log-likelihood and the penalty function $h_\infty(\mathrm{D}(X))$. The trade-off curve can be computed by solving (4.7) 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 (3.4) for each of these topologies. We then rank the models using the Bayes information criterion (BIC), as discussed at the beginning of section 4.1, 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.

Tracing trade-off curves The trade-off curves are computed by solving (4.7) 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 4.12 for a generic trade-off between two convex cost functions f_1 and f_2. We first solve the scalarized problem

$$\text{minimize} \quad f_1(x) + \gamma f_2(x) \tag{4.17}$$

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 give the slopes of straight lines that support the trade-off curve at points 1

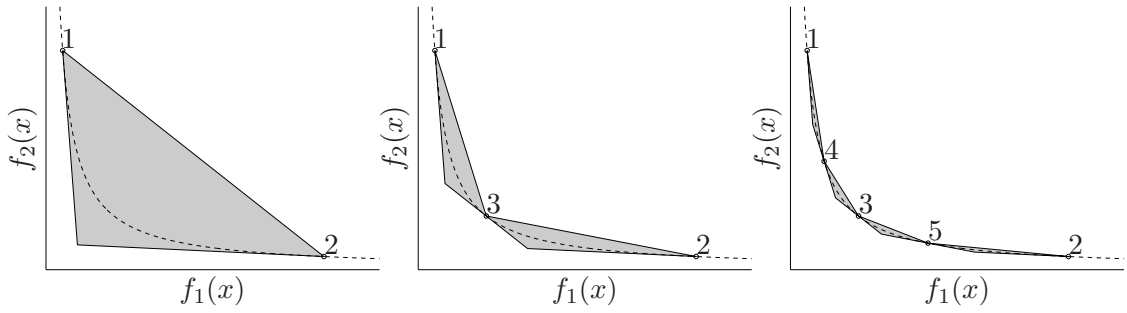

Figure 4.12: Method for approximating the trade-off curve between two convex objectives.

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 (4.17) 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 (4.17) 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 (4.17).

Thresholding With a proper value of γ, the regularized ML problem (4.7) has a sparse solution Y, resulting in a sparse inverse spectrum $S^{-1}(\omega)$. 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 partial coherence $R(\omega)$ as described in (4.5). It is essentially the inverse spectrum $S(\omega)^{-1}$ normalized to have diagonal one:

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

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 $R(\omega)$,

$$\rho_{ij} = \sup_{\omega} |R(\omega)_{ij}| \tag{4.18}$$

with a given threshold. In the experiments we use a value of 10^{-1} for the threshold, *i.e.*, we remove edge (i, j) from the graph if $\rho_{ij} \leq 10^{-1}$. Of course, other choices of norm (L_2 or L_1) can be used. They give similar results for the estimated topology.

4.4.2 Experiment 1: performance of the ℓ_1 regularization

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 $p = 2$.

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 4.13 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 (2.10) for the computed points on the trade-off curve, and apply a threshold to them (as explained above, by setting entries with $\rho_{ij} \leq 10^{-1}$ to zero). The resulting topologies are shown in figure 4.14. 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 (2.19) with C given in (2.16) or, equivalently, the ML solution of (3.3) without the sparsity constraints). For each of the nine sparsity patterns, we solve the ML problem

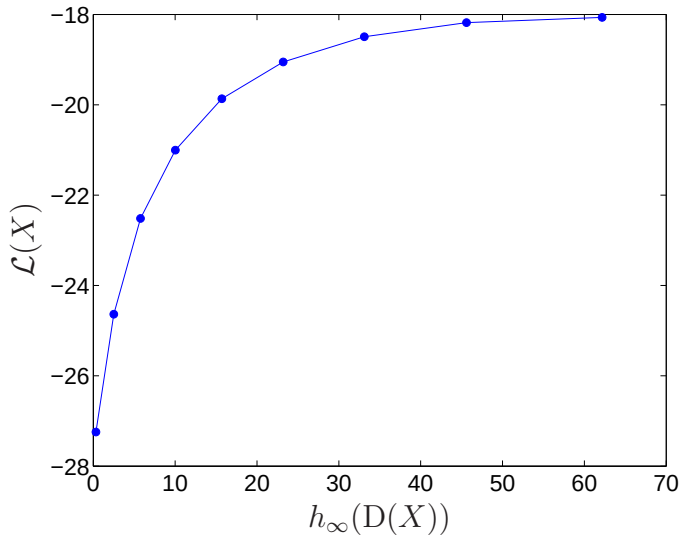

Figure 4.13: Trade-off curve between the log-likelihood $\mathcal{L}(X)$ and $h_\infty(\mathrm{D}(X))$.

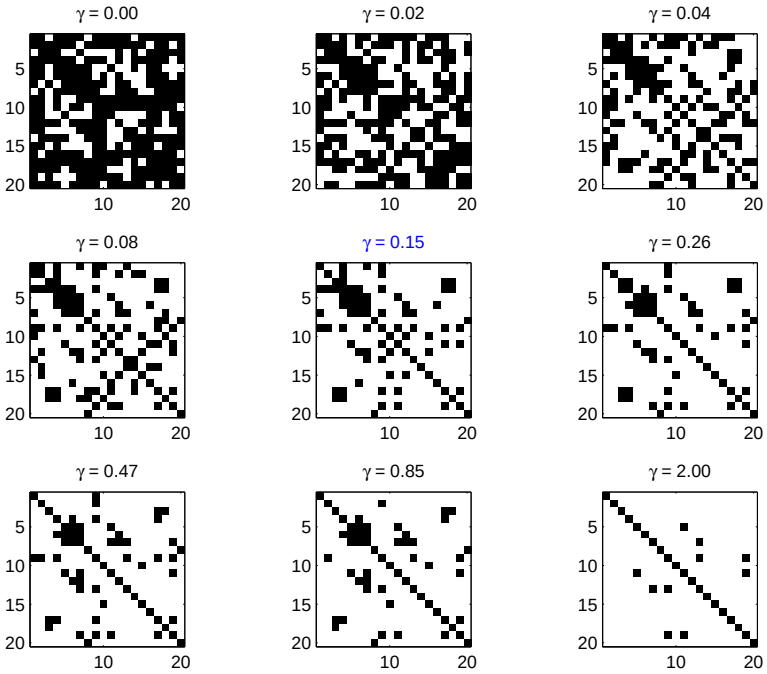

Figure 4.14: Topologies of solutions along the trade-off curve in figure 4.13 (ordered from right to left on the trade-off curve).

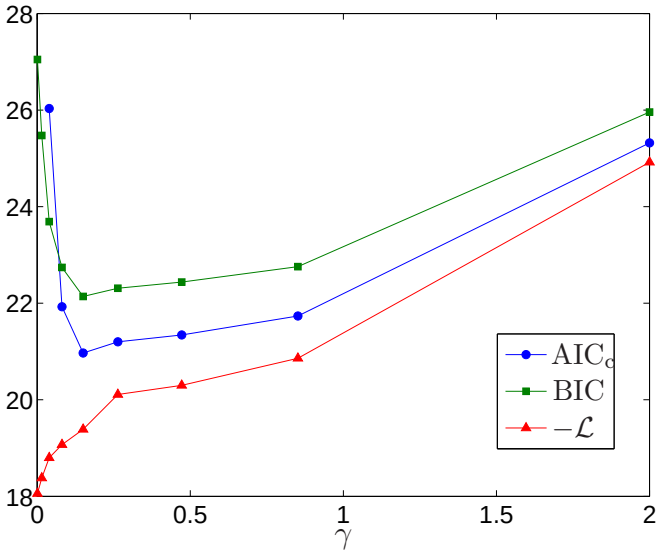

Figure 4.15: AIC_c and BIC scores, and maximized log-likelihood for solutions on the trade-off curve in figure 4.13.

subject to sparsity constraints (3.4). We rank the nine solutions using the AIC_c and BIC scores defined in (4.2)-(4.3). Figure 4.15 shows the two scores and the negative log-likelihood as functions of γ. The models that minimize the 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 4.16 (top left). Only seven entries are misclassified (six entries are misclassified as zeros; one as nonzero). The sparsity pattern in the top right is the topology estimated by thresholding the partial coherence spectrum of the least-squares solution with the correct model order ($p = 2$). This pattern is computed by solving the ML problem (3.3) 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 bottom of figure (4.16) is obtained from the covariance selection method with ℓ_1-norm regularization (*i.e.*, by setting $p = 0$

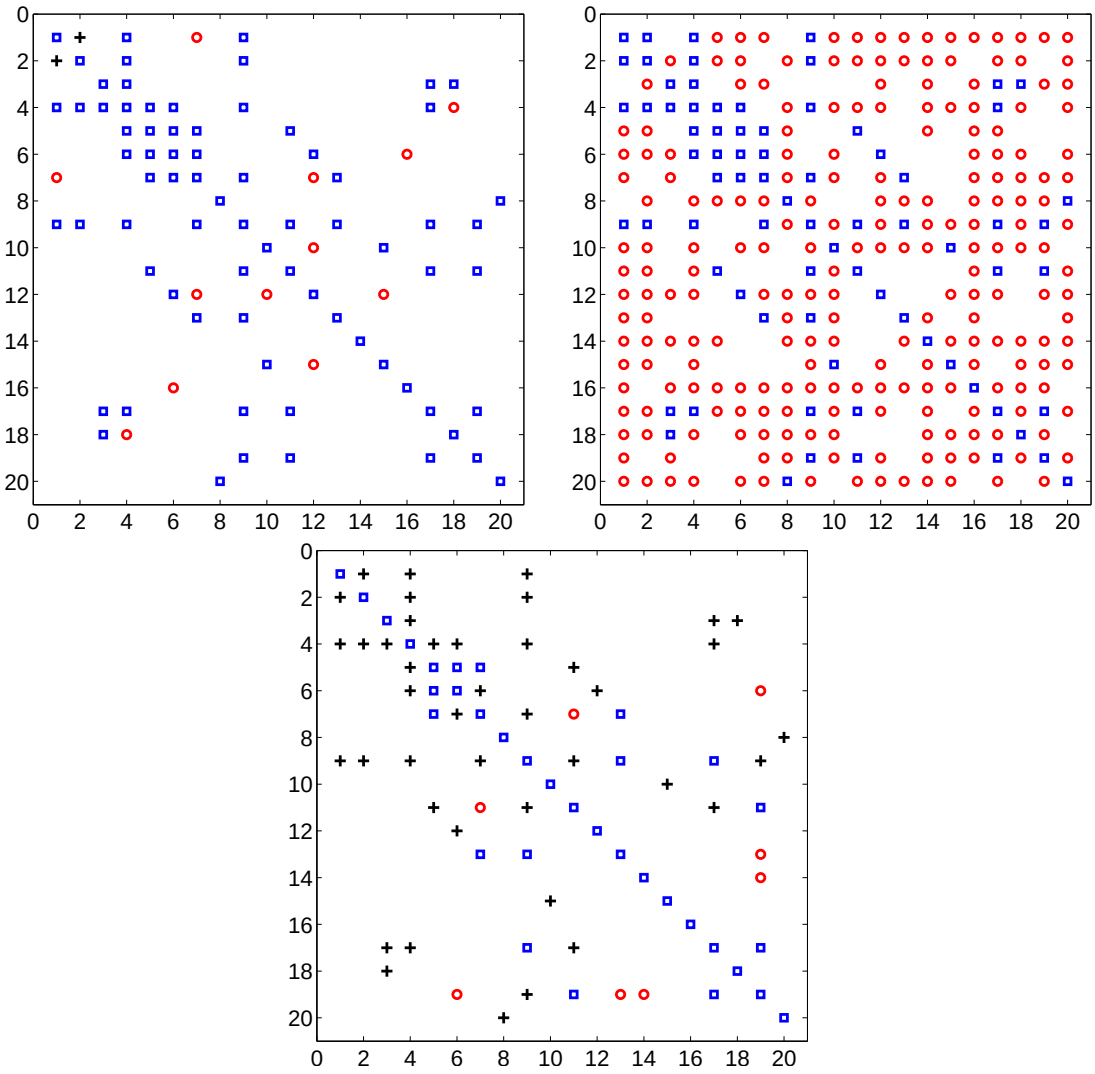

Figure 4.16: *Top Left.* The sparsity pattern from the regularized ML problem with $\gamma = 0.15$. *Top Right.* The sparsity pattern estimated from the least-squares solution. *Bottom.* The sparsity pattern from the regularized ML problem for a static model ($p = 0$). The blue squares are the correctly 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).

in the regularized ML problem (4.7)) and thresholding the partial coherence. Ignoring the model dynamics substantially increases the error in the topology selection.

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 [BJ04] 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 (3.3) without the constraints, and it can be computed by solving the normal equations (2.19).

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. Similar to ℓ_2-regularized least-squares problem in section 4.3.1, Tikhonov regularization (or ℓ_2 regularization) for the ML estimation can be obtained by adding the Frobenious norm of B to the unconstrained ML problem (2.21):

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

   The solution can be computed from the normal equations (2.19) 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 [HTF09, §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 (4.7) with penalty (4.8).

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 4.17 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 gave 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.

Errors in topology as a function of sample size In the last figure (Figure 4.18) we examine how fast the error in the topology selection decreases with increasing sample length N for three topology selection methods: LS estimation

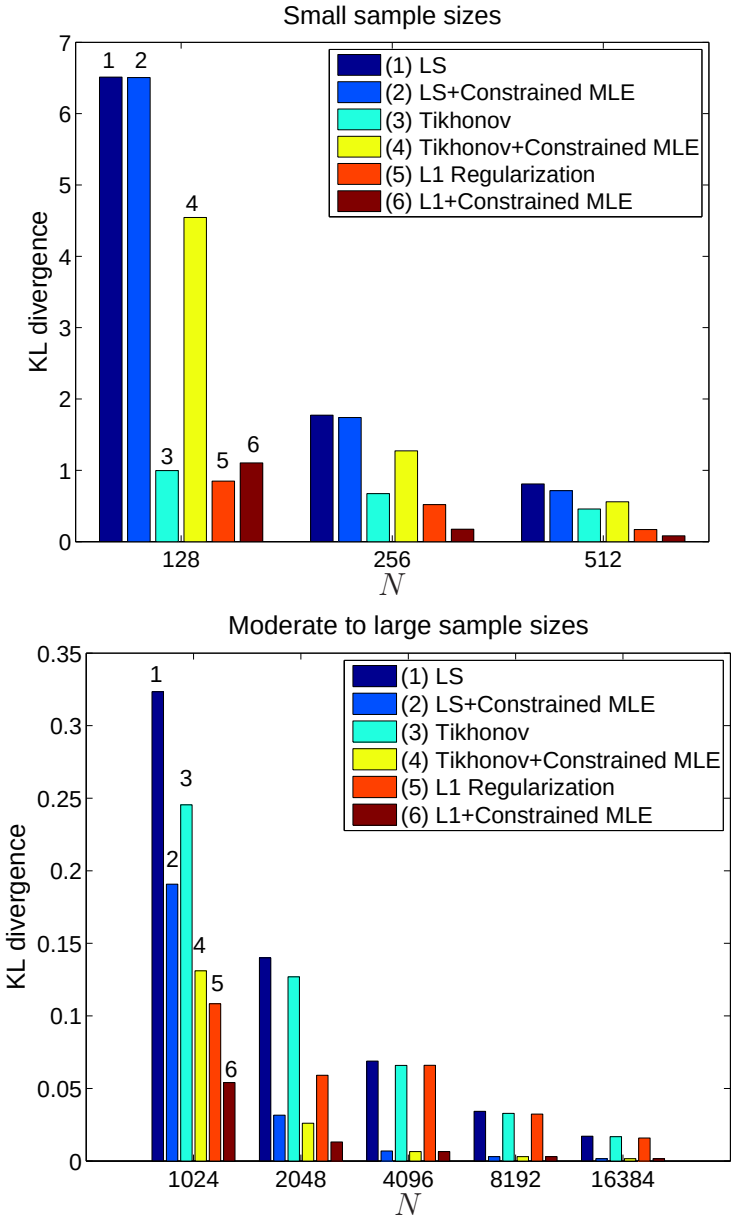

Figure 4.17: KL divergence between estimated AR models and the true model ($n = 20, p = 2$) versus the number of samples N. We compare six methods: (1) least-squares estimate, (2) constrained ML estimate with topology estimated by thresholding solution 1, (3) ML estimate with Tikhonov regularization, (4) constrained ML estimate with toplogy estimated by thresholding solution 3, (5) regularized ML estimate with h_∞-penalty, (6) constrained ML estimate with topology estimated by thresholding solution 5.

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 4.18 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. The total error in the estimated topology is reduced by regularizing, and the errors decrease more rapidly when we regularize with the sum-of-norms penalty h_∞.

4.4.3 Experiment 2: sum-of-ℓ_α-norms penalties

In the second experiment we compare different penalty functions h for the regularized ML problem (4.7): the 'sum-of-ℓ_∞-norms' penalty h_∞ defined in (4.8), the 'sum-of-ℓ_2-norms' penalty h_2 defined in (4.14) with $\alpha = 2$, and the 'sum-of-ℓ_1-norms' penalty h_1 defined in (4.14) 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 (2.10). 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

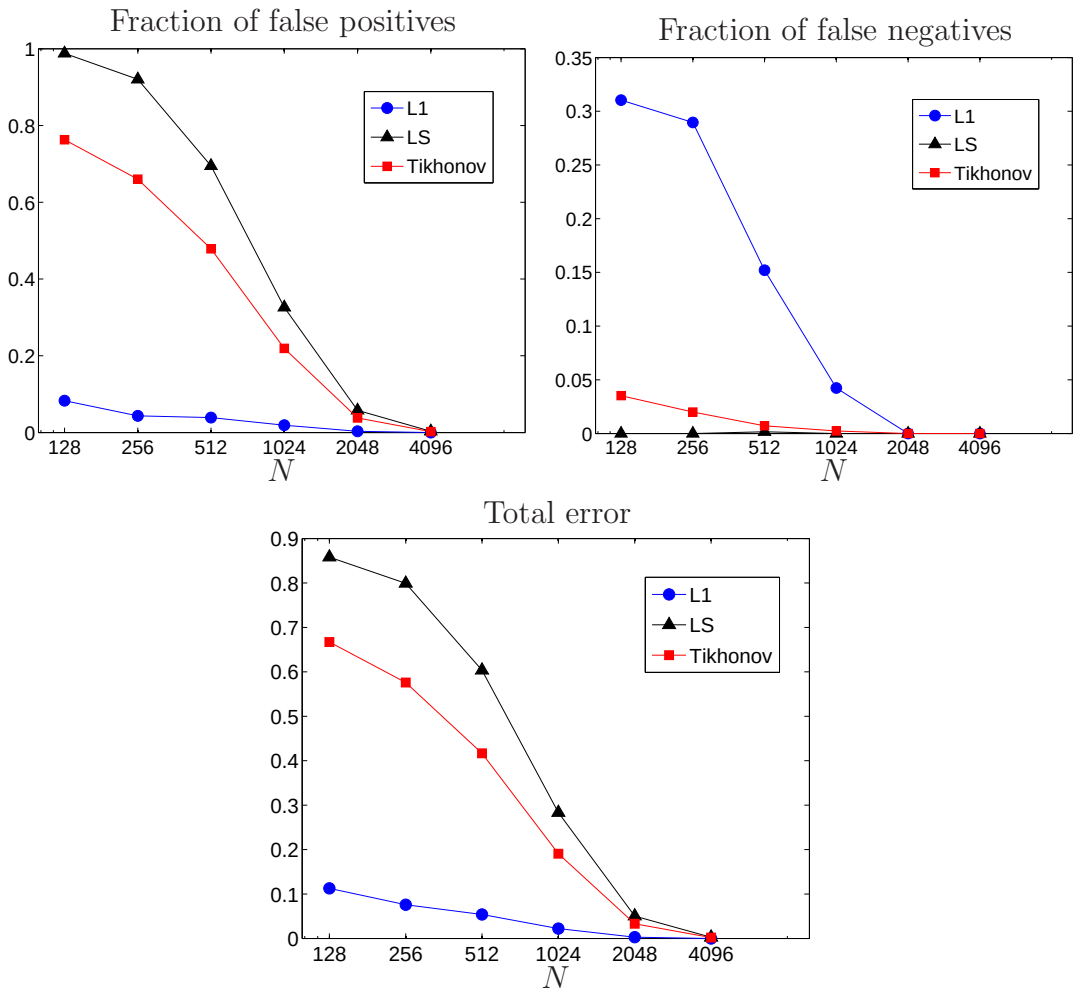

Figure 4.18: *Top left.* Fraction of incorrectly added edges in the estimated graph (number of upper triangular nonzeros in the estimated pattern that are incorrect, divided by the number of upper triangular zeros in the correct pattern). *Top right.* Fraction of incorrectly removed edges in the estimated 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.

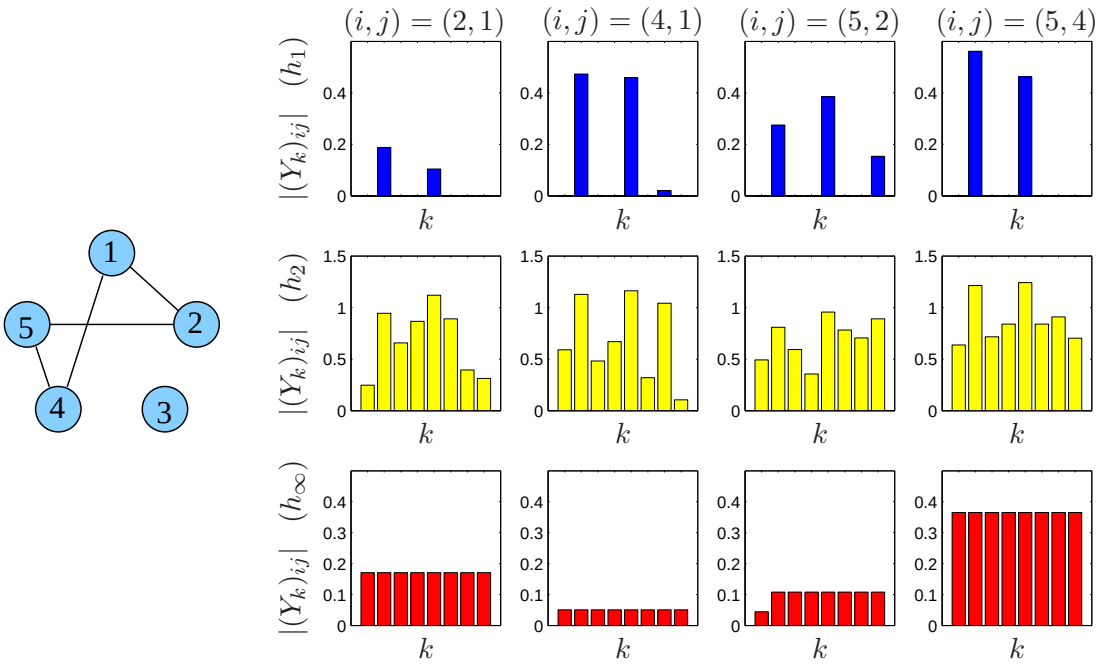

Figure 4.19: Nonzero coefficients $|(Y_k)_{ij}|$ for regularized ML estimates with penalty h_α, for $\alpha = 1, 2, \infty$.

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 4.19 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 4.3 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

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 4.3: 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.

with a comparison of different norms in a composite penalty function [ZRY09]. In general the best choice of norm will depend on how the coefficients are distributed within each group.

4.5 Examples with moderate and large real data sets

In this section we present two examples of real data sets to demonstrate how topology selection can facilitate studies of relationships in multivariate time series. We discuss fMRI data from neuroscience and stock return data from economics.

4.5.1 Functional magnetic resonance imaging (fMRI) data

There is great interest in using fMRI measurement to analyze interactions between active brain regions that are either stimulated by certain tasks or in resting states. Analyzing associations between interested regions could bring some

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 4.4: AR model orders for the fMRI data set.

insight understanding on the brain function to neuroscientists. It is widely accepted that the functional activity of each subregion can be demonstrated by human functional magnetic resonance imaging (fMRI) time series in which most cases, depicts blood oxygenation level. It is based on the assumption that the more activities the brain has, the higher level of oxygen will be used. Inference about the functional connectivity can be explained from the underlying dependence structure of the system. A graph-theoretical approach has been suggested to accommodate such analysis (see [SSS05, Eic05, RFG05, HPF03], and [Eic06b, §14]) by applying the concept of Granger-causality graphs [Eic07] instead of conditional independence graphs considered in this thesis.

In this section we apply the topology selection method to a functional magnetic resonance imaging (fMRI) time series. The data set is from [MHN04] and was analyzed in [SR09] using covariance selection. 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. The authors of [MHN04] 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 4.4. As the problem size (n) becomes larger, the BIC tends to pick a static model ($p = 0$). Table 4.5 shows the BIC

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 4.5: 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).

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 4.20 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 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 4.6 shows

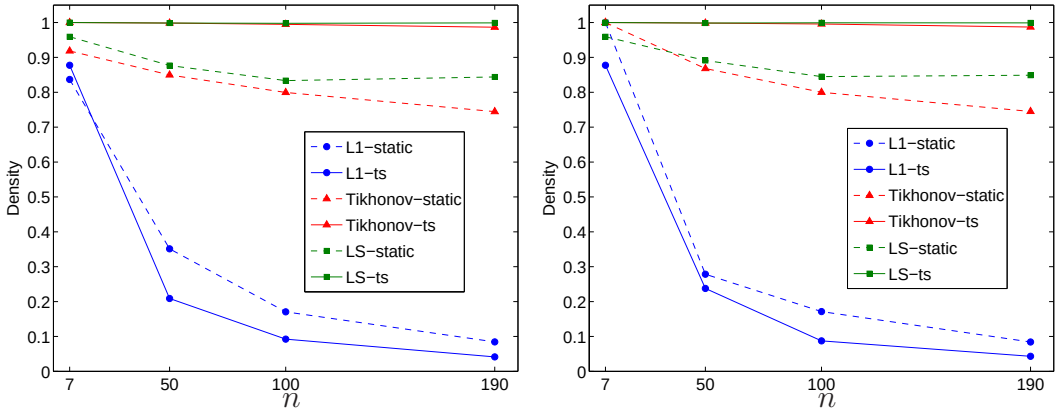

Figure 4.20: Density of the graphical models of fMRI data for 'picture' stimulus (*Left*) and for 'sentence' stimulus (*Right*). The density is computed as the number of nonzero entries in the estimated inverse spectrum divided by n^2.

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 4.6: 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 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 4.4 also performs slightly better in the classification experiment.

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

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 $\mathrm{AIC_c}$ 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 4.21 (right) shows ρ_{ij}, the maximum magnitude of the partial coherence of the model, and compares it with

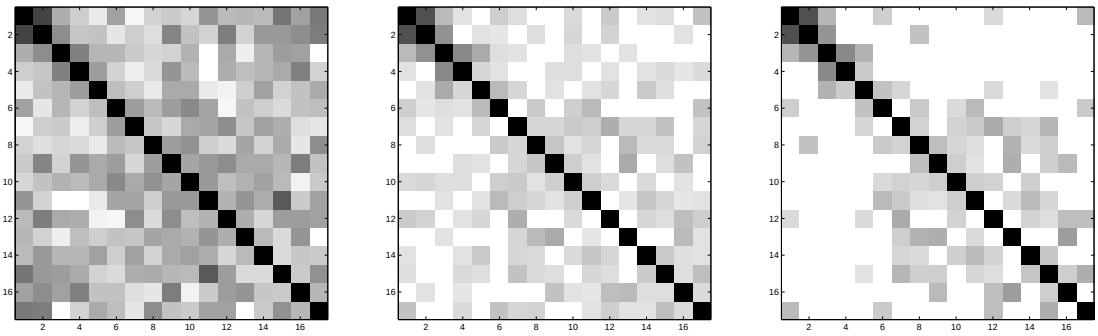

Figure 4.21: The maximum magnitude of the partial coherence ρ_{ij} for three models of the stock exchange data after applying a threshold. *(Left.)* A nonparametric sample estimate using Welch's method. *(Middle.)* Thresholded least-squares estimate. *(Right.)* Result of the h_∞-regularized ML problem.

a nonparametric estimate obtained with Welch's method [Pro01] 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 4.22 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.

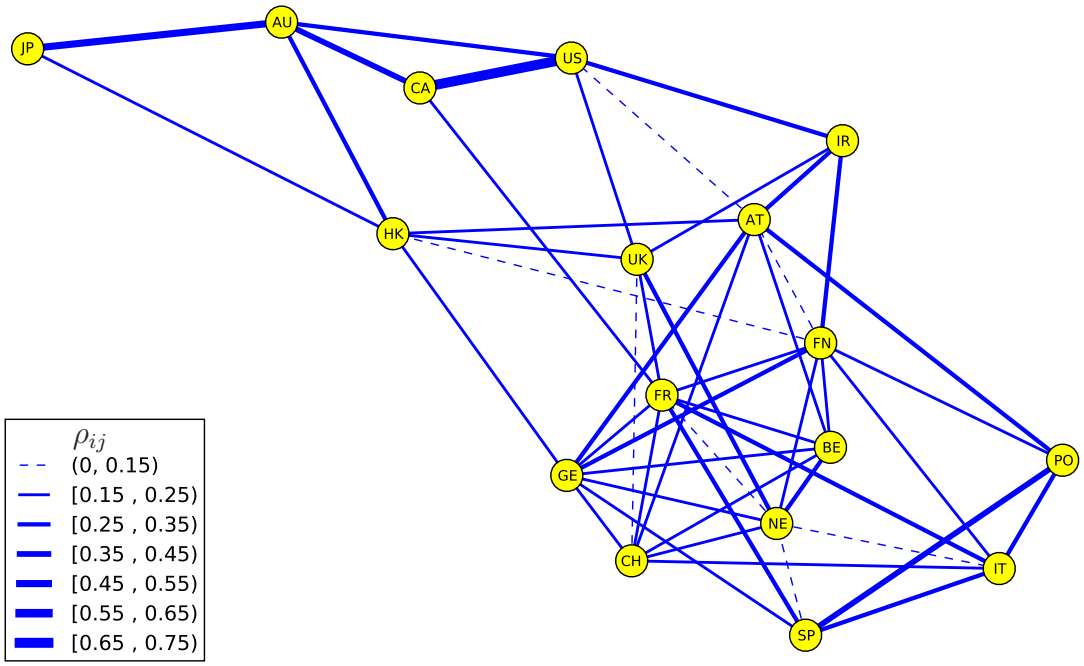

Figure 4.22: A graphical model of stock market data. The strength of connections is represented by the width of the blue links, which is proportional to $\rho_{ij} = \sup_\omega |R(\omega)_{ij}|$ if it is greater than 0.15.

CHAPTER 5

Algorithms

In this thesis we have encountered several matrix optimization problems. In chapter 2 we have shown that the ML estimation with conditional independence constraints can be solved via a convex problem

$$\begin{array}{ll} \text{minimize} & -\log\det X_{00} + \mathbf{tr}(CX) \\ \text{subject to} & \mathrm{P}(\mathrm{D}(X)) = 0 \\ & X \succeq 0 \end{array} \tag{5.1}$$

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

$$\begin{array}{ll} \text{maximize} & \log\det W + n \\ \text{subject to} & \left[\begin{array}{cc} W & 0 \\ 0 & 0 \end{array}\right] \preceq C + \mathrm{T}(\mathrm{P}(Z)) \end{array} \tag{5.2}$$

(with variables $W \in \mathbf{S}^n$ and $Z \in \mathbf{M}^{n,p}$). These two problems have differentiable objectives and linear equality and matrix inequality constraints. In chapter 3 the topology selection problem in graphical models was solved via an ℓ_1-regularized convex problem

$$\begin{array}{ll} \text{minimize} & -\log\det X_{00} + \mathbf{tr}(CX) + \gamma h(Y) \\ \text{subject to} & Y = \mathrm{D}(X), \quad X \succeq 0 \end{array} \tag{5.3}$$

(with variable $X \in \mathbf{S}^{n(p+1)}$ and $Y \in \mathbf{M}^{n,p}$). The objective function of this problem contains a nondifferentiable term ($h(Y)$ given by (4.8)) while its dual

$$
\begin{array}{ll}
\text{maximize} & \log\det W + n \\
\text{subject to} & \begin{bmatrix} W & 0 \\ 0 & 0 \end{bmatrix} \preceq C + \mathrm{T}(Z) \\
& \sum_{k=0}^{p} (|(Z_k)_{ij}| + |(Z_k)_{ji}|) \leq \gamma, \quad i \neq j \\
& \mathbf{diag}(Z_k) = 0, \quad k = 0, \ldots, p
\end{array}
\tag{5.4}
$$

(with variables $W \in \mathbf{S}^n$ and $Z \in \mathbf{M}^{n,p}$) has a differentiable objective, but the constraints involve a nondifferentiable function. As we see, the additional complications that do not appear in the covariance selection problems (1.1) and (1.2) are the complicated equality constraints and the fact that the smooth term $\log\det X_{00} + \mathbf{tr}(CX)$ in the primal (5.1) and (5.3) is *not* strictly convex. Moreover, there are extra matrix inequalities in the dual (5.2) and (5.4).

Nevertheless, these four convex problems can be solved by interior-point methods [BV04, §11]. For example, the path-following method was applied to related problems of determinant maximization in [Toh99, VBW98]. The number of iterations of an interior-point method to achieve a desired accuracy is known to grow slowly compared to the problem size. However, the most significant part is the Newton's step which is equivalent to solving sets of linear equations that involve the Hessian matrix of the objective. This step becomes prohibitive when the optimization variables have high dimension. In this thesis we therefore investigate less expensive first-order algorithms (methods that require only a knowledge of the gradient) for the two estimation problems.

In section 5.1 we review first-order methods recently used in related sparse optimization problems. Section 5.2 describes the gradient projection method, later shown to be suitable for a reformulation of the dual problems (5.2) and (5.4) in section 5.3. The analysis of the convergence and some numerical results are

included in section 5.4 and 5.5 respectively.

5.1 First-order methods for sparse optimization

There is currently great interest in efficient methods for solving large-scale optimization problems involving ℓ_1-norm. The simplest example is the ℓ_1-norm regularized least-squares problem

$$\text{minimize} \quad \frac{1}{2}\|y - Ax\|_2^2 + \gamma\|x\|_1, \tag{5.5}$$

with variable $x \in \mathbf{R}^n$, for given $y \in \mathbf{R}^m, A \in \mathbf{R}^{m\times n}$ and $\gamma > 0$. This is motivated by the fact that the ℓ_1-norm term promotes a sparsity to the solution. This problem, early introduced in [CDS01] as *basis pursuit* and in [Tib96] as *Lasso*, is now widely used in many applications; including signal processing [Tro06], wavelet-based image reconstruction [FN03], and compressed sensing [Don06, CRT06a, CRT06b, FNW07]. The problem (5.5) can be immediately solved by general solvers based on interior-point methods [BV04, §11] such as SDPT3 [TTT99]. However, these methods can suffer from expensive computational cost in the high dimensions setting of real-world applications. This has recently led to research seeking inexpensive first-order algorithms, for example gradient projection [FNW07], proximal gradient [WNF09, BT09], optimal first-order methods (or Nesterov's method) [Nes07, BBC09], and other competitive algorithms reviewed in [YGS10]. Despite the nonsmoothness of (5.5), two important properties that make the problem suitable for efficient methods are that the domain of (5.5) is the whole space and that the ℓ_1 term is *separable*.

Solving the covariance selection problem in large scale is more involved. As

we have seen, the primal problem:

$$\begin{array}{ll} \text{minimize} & -\log\det X + \mathbf{tr}(CX) + \gamma \sum_{j>i} |X_{ij}| \\ \text{subject to} & X \succeq 0 \end{array} \tag{5.6}$$

has a nonsmooth penalty term similar to the ℓ_1 term in (5.5), but it is *nonseparable* because of the cone constraint. The dual problem:

$$\begin{array}{ll} \text{maximize} & \log\det(C+Z) + n \\ \text{subject to} & |Z_{ij}| \leq \gamma/2, \quad i \neq j \end{array} \tag{5.7}$$

has a differentiable cost objective and simple box constraints, but its domain is not all the feasible set. Several large-scale methods have been proposed in the literature. Banerjee *et al.* [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 smooth approximation of (5.6). The authors of [FHT08] 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 consider a coordinate ascent method applied to the primal problem. A method based on column-wise updates is given in [RBL08]. A related problem is explored in [YL07] where the authors make a connection between (5.6) and more general determinant maximization problems [VBW98], and solve the problem using interior-point methods. Lu [Lu09] observes that the dual (5.7) is a smooth problem, and applies Nesterov's method [Nes05] directly to the dual. The algorithm is further extended in [Lu10] and compared with a projected spectral gradient method. Another closely related paper is [DGK08] in which the gradient projection method is applied to the dual problem.

These methods are efficient in practice and can solve problems of dimensions in the order of thousand. However, it is not obvious how to extend these methods

to solve the problems (5.1)-(5.4). A reformulation of the dual problems (5.2) and (5.4) as minimization problems over a simple set makes it applicable to the gradient projection method described in the next section.

5.2 First-order algorithms

This section presents some details on first-order algorithms. To simplify the notation we use a generic problem format

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in \mathcal{C} \end{array} \tag{5.8}$$

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. It is interesting to consider the classical gradient projection method [Pol87, Ber99] when $\mathcal{C}$ has a relatively simple structure that makes a projection on $\mathcal{C}$ inexpensive. Examples are the probability simplex, the nonnegative orthant, box constraints, or the positive semidefinite cone. We will show in section 5.3 that $\mathcal{C}$ in our problem is related to a ℓ_1-norm ball, for which it is also simple to compute a projection.

To give more detail on the gradient projection algorithm, we assume that a feasible point $x^{(0)}$ is known and that the sublevel set

$$\mathcal{S} = \{x \in \mathbf{dom}\, f \cap \mathcal{C} \mid f(x) \leq f(x^{(0)})\} \tag{5.9}$$

is closed and bounded. The closedness assumption is satisfied if f is a closed function. (We will show in section 5.4.1 that this is the case for problems (5.2) and (5.4).) 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. Moreover, 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 $t > 0$.

5.2.1 Basic gradient projection

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

$$\begin{aligned} x^{(k)} &= \mathcal{P}\left(x^{(k-1)} - t_k \nabla f(x^{(k-1)})\right) \\ &= x^{(k-1)} - t_k G_{t_k}(x^{(k-1)}) \end{aligned} \tag{5.10}$$

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

$$\|\nabla f(u) - \nabla f(v)\|_2 \leq L\|u - v\|_2 \quad \forall u, v \in \mathcal{S}, \tag{5.11}$$

[Pol87, §7.2.1]. Although our assumptions ($\mathcal{S}$ is closed and bounded, and $\mathbf{dom}\, f$ is open) imply that the Lipschitz condition (5.11) 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 [BT09]. The step size search algorithm in iteration k starts at a value $t_k := \bar{t}_k$ where

$$\bar{t}_k = \min\left\{\frac{s^T s}{s^T y}, t_{\max}\right\}, \tag{5.12}$$

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 $x^{(k-1)} - t_k G_{t_k}(x^{(k-1)}) \in \mathbf{dom}\, f$ and

$$f(x^{(k-1)} - t_k G_{t_k}(x^{(k-1)})) \leq f(x^{(k-1)}) - t_k \nabla f(x^{(k-1)})^T G_{t_k}(x^{(k-1)}) + \frac{t_k}{2} \|G_{t_k}(x^{(k-1)})\|_2^2. \tag{5.13}$$

The resulting step size t_k is used in the update to $x^{(k)}$ in (5.10). 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* [Ber99, §8.3].

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 [BB88, BMR03, SZZ05, FNW07, WNF09] that have been observed to greatly improve the convergence in practice. It can be motivated by the easily established fact that $\|s\|_2^2 / s^T y \geq 1/L$ if f satisfies (5.11), so it is a readily computed upper bound for $1/L$.

This fact is related to the motivation suggested by [BB88] that $a = \|s\|_2^2 / s^T y$ is chosen such that aI is a good approximation of the inverse Hessian over the last step. It requires that $s \approx ay$ in the least-squares sense, *i.e.*,

$$a = \operatorname*{argmin}_a \|s - ay\|_2^2 = \frac{s^T s}{s^T y}.$$

5.2.2 Step size rules

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

$$x^{(k)} = x^{(k-1)} - t_k G_{\bar{t}_k}(x^{(k-1)}) \tag{5.14}$$

where $\bar{t}_k$ is held fixed at the value (5.12) 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

$$f(x^{(k-1)} - t_k G_{\bar{t}_k}(x^{(k-1)})) \leq f(x^{(k-1)}) - t_k \nabla f(x^{(k-1)})^T G_{\bar{t}_k}(x^{(k-1)}) + \frac{t_k}{2} \|G_{\bar{t}_k}(x^{(k-1)})\|_2^2. \tag{5.15}$$

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 (5.10) and (5.14) are available in the literature, for example, the Armijo rule [Ber99, §2.3], and conditions that allow non-monotone convergence [BMR00, LZ09]. In our experiments these variations gave similar results as the step size rules outlined above.

5.2.3 Optimal first-order methods

Another attractive class of gradient projection algorithms is the optimal first-order methods originated by Nesterov [Nes04, Tse08, BT09]. 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 [BT09, Tse08]. 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 section 5.5).

For nondifferentiable functions, if the objective can be written as $f(x) = g(x) + h(x)$ where $g(x)$ has Lipschitz continuous gradient, and $h(x)$ is the nonsmooth term, such as ℓ_1-regularized problems (5.5), the optimal methods [Nes04, Tse08, BT09] also yield a better complexity of $O(1/\epsilon)$, as opposed to $O(1/\epsilon^2)$ for the subgradient method. A nonsmooth problem that has a special saddle format can also be solved by a recently famous smoothing technique [Nes05] with complexity $O(1/\epsilon)$. The idea is to compute a smooth approximation of the objective and then apply the optimal first-order method to the approximated problem. This approach has been used in sparse approximation problems [GJL07, BEd08, BBC09, KY09, DHJ10] by formulating the ℓ_1 term into a saddle format.

5.3 Reformulated dual problems

In this section we reformulate the dual problems (5.2) and (5.4) to make them suitable for gradient projection algorithms, *i.e.*, as in (5.8). As we have seen the primal problems (5.1) and (5.3) poses several difficulties for the gradient methods. In addition to the nonsmoothness in (5.3), the primal problems contain complicated equality constraints and matrix inequalities. The term $-\log\det X_{00} + \mathbf{tr}(CX)$ is not strictly convex and does not increase to infinity near the boundary of its domain. Moreover, the solution always lies on the boundary of the domain (the optimal X has low rank), so an algorithm must take into account the cone constraint. We will see shortly that these difficulties can be avoided in the dual approach.

To reformulate the dual problems we eliminate the variable W in (5.2) and (5.4). Let $V = C + \mathrm{T}(\mathrm{P}(Z))$, respectively, $V = C + \mathrm{T}(Z)$. 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

$$V_{1:p,1:p} \succeq 0, \quad \mathrm{range}(V_{1:p,0}) \subseteq \mathrm{range}(V_{1:p,1:p}), \quad V_{00} - V_{1:p,0}^T V_{1:p,1:p}^\dagger V_{1:p,0} \succeq W, \tag{5.16}$$

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 (5.16) 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 (5.2) and (5.4). Problem (5.2) can be written as an unconstrained problem

$$\text{maximize} \quad -\phi(C + \mathrm{T}(\mathrm{P}(Z))), \tag{5.17}$$

and problem (5.4) as a problem with simple constraints

$$\begin{array}{ll} \text{maximize} & -\phi(C + \mathrm{T}(Z)) \\ \text{subject to} & \sum_{k=0}^{p} (|(Z_k)_{ij}| + |(Z_k)_{ji}|) \leq \gamma, \quad i \neq j \\ & \mathbf{diag}(Z_k) = 0, \quad k = 0, \ldots, p. \end{array} \tag{5.18}$$

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

$$\phi(V) = -\log\det\left(V_{00} - V_{1:p,0}^T V_{1:p,1:p}^\dagger V_{1:p,0}\right) - n, \tag{5.19}$$

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 [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

$$\nabla\phi(V) = -V^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & V_{1:p,1:p}^{-1} \end{bmatrix}. \tag{5.20}$$

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 (5.17) then the primal optimal solution can be computed from Z via the expressions

$$X = V^{-1} - \begin{bmatrix} 0 & 0 \\ 0 & V_{1:p,1:p}^{-1} \end{bmatrix} = \begin{bmatrix} -I \\ V_{1:p,1:p}^{-1} V_{1:p,0} \end{bmatrix} W^{-1} \begin{bmatrix} -I \\ V_{1:p,1:p}^{-1} V_{1:p,0} \end{bmatrix}^T \tag{5.21}$$

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 (3.9) and the identities

$$V = \begin{bmatrix} V_{00} - V_{1:p,0}^T V_{1:p,1:p}^{-1} V_{1:p,0} & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} V_{1:p,0}^T V_{1:p,1:p}^{-1} \\ I \end{bmatrix} V_{1:p,1:p} \begin{bmatrix} V_{1:p,0}^T V_{1:p,1:p}^{-1} \\ I \end{bmatrix}^T, \tag{5.22}$$

$$V^{-1} = \begin{bmatrix} 0 & 0 \\ 0 & V_{1:p,1:p}^{-1} \end{bmatrix} + \begin{bmatrix} -I \\ V_{1:p,1:p}^{-1} V_{1:p,0} \end{bmatrix} (V_{00} - V_{1:p,0}^T V_{1:p,1:p}^{-1} V_{1:p,0})^{-1} \begin{bmatrix} -I \\ V_{1:p,1:p}^{-1} V_{1:p,0} \end{bmatrix}^T. \tag{5.23}$$

The formula for V^{-1} also provides an alternative form of the gradient (5.20).

Similarly, if $C + \mathrm{T}(Z) \succ 0$ at the optimum of (5.18) then the primal optimal X can be computed from (5.21) 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. The method requires an important assumption on the closedness of the sublevet set (5.9). This property will be discussed in the next section.

5.4 Analysis of gradient projection

A common assumption in the literature on the gradient projection algorithm for the problem (5.8) 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 $f(x^{(k)}) - f^\star$ decreases as $1/k$ [BT09, Nes04]. These assumptions are *not* valid for the applications in this thesis: here from (5.17)-(5.18), $\mathcal{C} \not\subseteq \mathbf{dom}\, f$ and the gradient of f is not Lipschitz continuous on $\mathcal{C} \cap \mathbf{dom}\, f$. For completeness we therefore include a convergence analysis in section 5.4.2. The proof is adapted from Beck and Teboulle [BT09] and requires the closedness property of the reformulated dual objective ϕ.

5.4.1 Closedness property

The closedness property of ϕ can be concluded by the structure of 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, *i.e.*, $\phi(\mathrm{T}(R))$, 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 3.3, 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 [BV04, 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 (5.17) or (5.18) is at a point in the boundary of the domain of the cost function, *i.e.*, 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 $\{Z \mid \phi(C + \mathrm{T}(\mathrm{P}(Z))) \leq \phi(C + \mathrm{T}(\mathrm{P}(Z^{(0)})))\}$ is closed. This condition is considerably weaker than the requirement that all sublevel sets are closed.

5.4.2 Convergence analysis

In this section we provide a convergence proof of the gradient projection method for the problem (5.8). The proof requires an important property of the projection. The projection satisfies

$$(y - \mathcal{P}(y))^T (z - \mathcal{P}(y)) \leq 0 \quad \forall z \in \mathcal{C}. \tag{5.24}$$

A useful property of the gradient map follows by applying (5.24) to $y = x - t\nabla f(x)$ and $\mathcal{P}(y) = x - tG_t(x)$: we have

$$(G_t(x) - \nabla f(x))^T(z - x + tG_t(x)) \leq 0 \quad \forall z \in \mathcal{C} \tag{5.25}$$

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

$$\nabla f(x)^T G_t(x) \geq \|G_t(x)\|_2^2 \tag{5.26}$$

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

Before we give the convergence proof, we first show that the backtracking line search (5.13) is well-defined; it terminates after a finite number of steps. The assumption that the sublevel set (5.9) associated with the initial point is closed and bounded implies that there exists an optimal $x^\star$, and that the gradient ∇f satisfies a Lipschitz condition on S: there exists an $L > 0$ such that

$$\|\nabla f(u) - \nabla f(v)\|_2 \leq L\|u - v\|_2 \quad \forall u, v \in S. \tag{5.27}$$

Define $x = x^{(i-1)}$ and assume that $x \in S$. From the Lipschitz property (5.27),

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

for all $y \in S$. Applying this to $y = \mathcal{P}(x - t\nabla f(x)) = x - tG_t(x)$ and using (5.26) gives

$$\begin{aligned} f(\mathcal{P}(x - t\nabla f(x))) &\leq f(x) - t\nabla f(x)^T G_t(x) + \frac{Lt^2}{2}\|G_t(x)\|_2^2 \\ &\leq f(x) - t(1 - \frac{Lt}{2})\|G_t(x)\|_2^2 \end{aligned} \tag{5.28}$$

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 (5.26) 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 (5.28) 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 (5.28) at $t = 1/L$, we see that $t = 1/L$ satisfies (5.13). 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 give a convergence proof of the gradient projection. We note that if (5.13) holds, then for all $y \in \mathcal{C} \cap \mathbf{dom}\, f$,

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

The last step follows from (5.25) 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 $y = x^\star$ gives

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

i.e.,

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

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

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

This shows that the number of iterations to reach accuracy $f(x) - f^* \leq \epsilon$ is bounded by $O(1/\epsilon)$.

5.5 Numerical examples

We generate AR models as in the experiment described in section 4.4.2. 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 (5.18) 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 $Z^{(0)} = 0$, and terminate when the duality gap is below 10^{-2} (the optimal value is on the order of hundreds).

Figure 5.1 shows the relative error $(f(Z^{(k)}) - f^\star)/|f^\star|$ where $f(Z) = \phi(C + \mathrm{T}(Z))$ 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 (5.10) with step size rule (5.13). 'GP with line search' refers to the gradient projection method (5.14) with step size rule (5.15). 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 back-

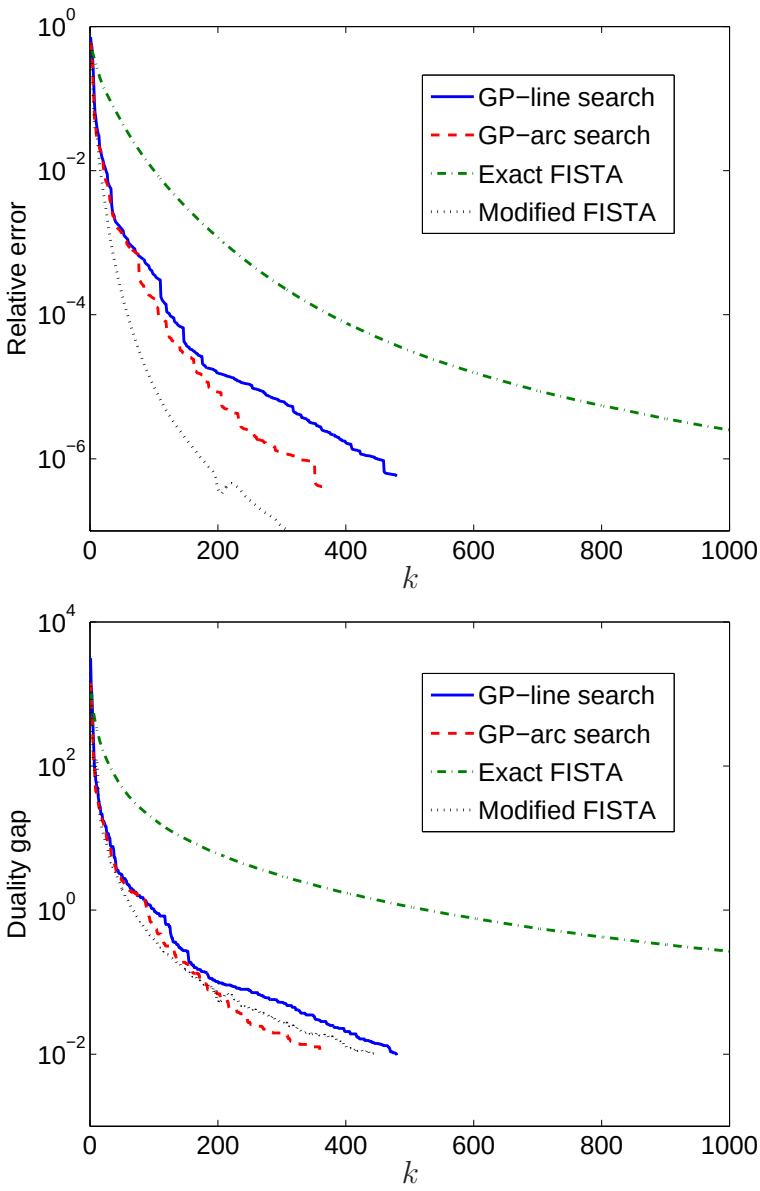

Figure 5.1: Convergence of gradient projection algorithms. *Left:* Relative error $(f(Z^{(k)}) - f^\star)/|f^\star|$ versus the number of iterations. *Right:* Duality gap versus the number of iterations.

tracking line search from [BT09] using monotonically decreasing step sizes ($t_k \leq t_{k-1}$, as required by the theory in [BT09]). 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 searches is initialized at the Barzilai and Borwein steplength (5.12), was often about five times faster. This algorithm is referred to as 'Modified FISTA' in the figure.

Figure 5.2 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 $\epsilon = 0.1$. 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.

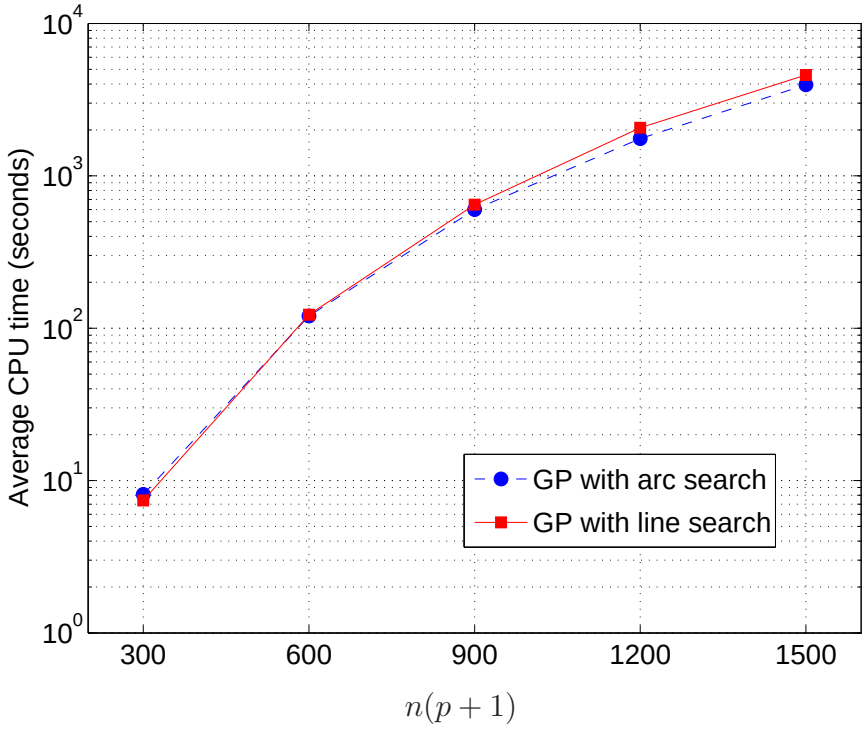

Figure 5.2: Average CPU times (averaged over 10 runs) of the gradient projection algorithm versus the problem size. The algorithm stops when the duality gap is less than 10^{-1}. The red squares correspond to 'GP with line search' and the blue squares correspond to 'GP with arc search'.

CHAPTER 6

Conclusions

6.1 Contributions

In the first part of the thesis, we have considered a parametric approach for maximum likelihood estimation of autoregressive models with conditional independence constraints. These constraints impose a sparsity pattern on the inverse of the spectral density matrix, and result in nonconvex equalities in the estimation problem. We have formulated a convex relaxation of the ML estimation problem and shown that the relaxation is exact when the sample covariance matrix in the objective of the estimation problem is block-Toeplitz. We have also noted from experiments that the relaxation is often exact for covariance matrices that are not block-Toeplitz.

The convex formulation allows us to select graphical models by fitting autoregressive models to different topologies, and ranking the topologies using information theoretic model selection criteria. The approach was illustrated with randomly generated and real data, and works well when the number of models in the comparison is small, or the number of nodes is small enough for an exhaustive search.

For larger model selection problems, 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 esti-

mation 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 more 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.

6.2 Suggestions for future research

In chapter 5, the gradient projection method with a special stepsize rule is proposed as a method to solve the convex optimization problems considered in this thesis. One may consider other types of algorithms for these problems. Recently the Nesterov's optimal method [Nes04, Tse08, BT09] has been applied to many large-scale convex optimization problems because of its low computational cost and a better convergence rate than the classical gradient method. This method seems to be promising for the reformulated duals (5.17) and (5.18) with a more careful analysis on the convergence. One may also investigate a possibility to apply the block coordinate descent method [Ber99, §2.7] to the dual problems (5.2) and (5.4).

Chapter 2 has described the integration between graph theory and the concept of conditional independence for identifying the interactions among variables. One

can investigate other definitions for interaction. For example one can discuss graphical models in which the direction of connections demonstrates causality. We can apply the concept of Granger causality which has been extensively used for economic time series and follow a graphical framework from [Eic07, Eic06b].

A Granger causality graphical model of AR processes (2.6) is a mixed graph with directed and undirected edges. The absence of a directed edge from node j to i indicates that x_i is *not Granger-caused* by x_j (knowing x_j does not help to improve the prediction of x_i), and this can be characterized by [Lut05]

$$(A_k)_{ij} = 0, \quad k = 1, 2, \ldots, p.$$

The absence of an undirected edge between nodes i and j implies that there is *no instantaneous causality* between x_i and x_j (in period t, knowing $x_i(t+1)$ does not help to improve the forecast of $x_j(t+1)$ and vice versa). This condition can be characterized via the noise covariance matrix [Lut05] as

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

Suppose the sets of the pairs of nodes (i, j) that are not connected by directed edges and undirected edges are characterized by $\mathcal{V}_1$ and $\mathcal{V}_2$ respectively. We can formulate the maximum-likelihood estimation of Granger causality graphical models of AR processes as

$$\begin{array}{ll} \text{minimize} & \log\det\Sigma + \mathbf{tr}(\Sigma^{-1}\bar{\Sigma}) + \mathbf{tr}(\Sigma^{-1}(A-\bar{A})C(A-\bar{A})^T) \\ \text{subject to} & 0 \preceq \Sigma \preceq 2\bar{\Sigma} \\ & (A_k)_{ij} = 0\ ,\ k = 1, \ldots p, \quad (i,j) \in \mathcal{V}_1 \\ & \Sigma_{ij} = 0, \quad (i,j) \in \mathcal{V}_2 \end{array} \tag{6.1}$$

with variables $A = [A_1\ A_2\ \ldots\ A_p]$ with $A_k \in \mathbf{R}^{n\times n}$ and $\Sigma \in \mathbf{S}^n$. The matrices $\bar{\Sigma}, \bar{A}$ are given by the least-squares estimates and C is the sample covariance

matrix. It can be shown that the problem (6.1) is convex under the first inequality constraint. We can also consider ℓ_1 regularization for topology selection in Granger graphical models by adding ℓ_1-type penalties to A and Σ, similar to the problem of conditional independence graphs in (4.7).

A convex framework of estimation problems for graphical models is motivated by recent applications in neuroscience and in particular, by applications of fMRI time series [GRK03, SSS05, VSL05, Eic06b, DCB06, VBV06, DRD08, DHS08, DLJ08]. There are several aspects of this application that require further study. Firstly, the most suitable definition of brain connectivity and the best models for making causal inferences are still being debated [Fri09, RFG09]. Secondly, in most experiments fMRI time series are recorded while a subject is responding to a sequence of stimuli, so it is important to use a model that includes input dynamics. Examples of such models are dynamic causal modeling introduced by Friston *et al.* [FHP03], autoregressive exogenous (ARX) models, or hidden Markov AR models [CWM08]. Lastly, it is of interest to develop methods for making group inferences, *i.e.*, determining graphical models for different subjects that have the same topology.

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.

[ARS09] 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*, pp. 1721–1724, 2009.

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

[Bal95] A.V. Balakrishnan. *Introduction to Random Processes in Engineering*. Wiley New York, 1995.

[BB88] J. Barzilai and J. M. Borwein. "Two-point step size gradient methods." *IMA Journal of Numerical Analysis*, **8**:141–148, 1988.

[BB01] P. Baldi and S. Brunak. *Bioinformatics: the Machine Learning Approach*. The MIT Press, 2001.

[BBC09] S. Becker, J. Bobin, and E. Candès. "NESTA: a fast and accurate first-order method for sparse recovery." submitted, 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.

[Bis06] C.M. Bishop. *Pattern Recognition and Machine Learning*. Springer New York, 2006.

[BJ76] G.E.P. Box and G. Jenkins. *Time Series Analysis, Forecasting and Control*. Holden-Day, San Francisco, CA, 1976.

[BJ04] F. R. Bach and M. I. Jordan. "Learning graphical models for stationary time series." *IEEE Transactions on Signal Processing*, **52**(8):2189–2199, 2004.

[BMR00] E. G. Birgin, J. M. Martínez, and M. Raydan. "Nonmonotone spectral projected gradient methods on convex sets." *SIAM Journal on Optimization*, **10**(4):1196–1211, 2000.

[BMR03] 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.

[Bri81] D. R. Brillinger. *Time Series Analysis: Data Analysis and Theory.* Holden-Day, Inc., 1981.

[BT09] 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.

[Bur75] J. P. Burg. *Maximum entropy spectral analysis.* PhD thesis, Stanford University, 1975.

[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.

[CDS01] S.S. Chen, D. L. Donoho, and M.A. Saunders. "Atomic decomposition by basis pursuit." *SIAM review*, **43**(1):129–159, 2001.

[CRT06a] E. J. Candès, J. Romberg, and T. Tao. "Stable signal recovery from incomplete and inaccurate measurements." *Communications on Pure and Applied Mathematics*, **59**(8):1207–1223, 2006.

[CRT06b] E.J. Candès, J.K. Romberg, and T. Tao. "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information." *IEEE Transactions on information theory*, **52**(2):489–509, 2006.

[CWM08] J. Chiang, Z.J. Wang, and M.J. McKeown. "A Hidden Markov, Multivariate Autoregressive (HMM-mAR) Network Framework for Analysis of Surface EMG (sEMG) Data." *IEEE Transactions on Signal Processing*, **56**(8):4069–4081, 2008.

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

[DCB06] M. Ding, Y. Chen, and S.L. Bressler. "Granger causality: Basic theory and application to neuroscience." In B. Schelter, M. Winterhalder, and J. Timmer, editors, *Handbook of time series analysis: recent theoretical developments and applications*. Wiley, 2006.

[DE03] R. Dahlaus and M. Eichler. "Causality and graphical models in time series analysis." In P.J. Green, N.L. Hjort, and S. Richardson, editors, *Highly Structured Stochastic Systems*. Oxford University Press, 2003.

[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.

[DHJ10] J. Dahl, P.C. Hansen, S.H. Jensen, and T.L. Jensen. "Algorithms and software for total variation image reconstruction via first-order methods." *Numerical Algorithms*, **53**(1):67–92, 2010.

[DHS08] G. Deshpande, X. Hu, R. Stilla, and K. Sathian. "Effective connectivity during haptic perception: A study using Granger causality analysis of functional magnetic resonance imaging data." *Neuroimage*, **40**(4):1807–1814, 2008.

[DLJ08] G. Deshpande, S. LaConte, G.A. James, S. Peltier, and X. Hu. "Multivariate Granger causality analysis of fMRI data." *Human brain mapping*, **30**(4):1361–1373, 2008.

[Don06] D. Donoho. "Compressed sensing." *IEEE Transactions on Information Theory*, **52**(4):1289–1306, 2006.

[DRD08] M. Dhamala, G. Rangarajan, and M. Ding. "Analyzing information flow in brain networks with nonparametric Granger causality." *NeuroImage*, **41**(2):354–362, 2008.

[DRV05] 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.

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

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

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

[Eic06a] M. Eichler. "Fitting graphical interaction models to multivariate time serie." *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence*, 2006.

[Eic06b] M. Eichler. "Graphical Modeling of Dynamic Relationships in Multivariate Time Series." In B. Schelter, M. Winterhalder, and J. Timmer, editors, *Handbook of Time Series Analysis: Recent Theoretical Developments and Applications*. Wiley, 2006.

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

[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.

[FHP03] K.J. Friston, L. Harrison, and W. Penny. "Dynamic causal modelling." *Neuroimage*, **19**(4):1273–1302, 2003.

[FHT08] J. Friedman, T. Hastie, and R. Tibshirani. "Sparse inverse covariance estimation with the graphical lasso." *Biostatistics*, **9**(3):432, 2008.

[FN03] M. Figueiredo and R. Nowak. "An EM algorithm for wavelet-based image restoration." *IEEE Transactions on Image Processing*, **12**(8):906–916, 2003.

[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.

[Fri09] K. Friston. "Causal modelling and brain connectivity in functional magnetic resonance imaging." *PLoS Biology*, **7**(2), 2009.

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

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

[GHJ99] W. Glunt, T.L. Hayden, C.R. Johnson, and P. Tarazaga. "Positive definite completions and determinant maximization." *Linear Algebra and its Applications*, **288**:1–10, 1999.

[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.

[GJL07] L.E. Ghaoui, M.I Jordan, and G.R.G. Lanckriet. "A direct formulation for sparse PCA using semidefinite programming." *SIAM Review*, **49**:434, 2007.

[GRK03] R. Goebel, A. Roebroeck, D.S. Kim, and E. Formisano. "Investigating directed cortical interactions in time-resolved fMRI data using vector autoregressive modeling and Granger causality mapping." *Magnetic Resonance Imaging*, **21**(10):1251–1261, 2003.

[Hac03] Y. Hachez. *Convex Optimization over Non-Negative Polynomials: Structured Algorithms and Applications*. PhD thesis, Université catholique de Louvain, Belgium, 2003.

[Han70] E. J. Hannon. *Multiple Time Series*. Wiley, New York, 1970.

[HLP06] 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.

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

[HTF09] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer, 2nd edition, 2009.

[Jor99] M. I. Jordan, editor. *Learning in Graphical Models*. MIT Press, Cambridge, MA, 1999.

[Kay88] S.M. Kay. *Modern Spectral Estimation: Theory and Application*. Prentice Hall, Englewood Cliffs, NJ, 1988.

[KKK06] Y. Kim, J. Kim, and Y. Kim. "Blockwise sparse regression." *Statistica Sinica*, **16**(2):375–390, 2006.

[KSH00] T. Kailath, A. H. Sayed, and B. Hassibi. *Linear estimation*. Prentice Hall, New Jersey, 2000.

[KY09] K.C. K.C. Toh and S. Yun. "An accelerated proximal gradient algorithm for nuclear norm regularized least squares problems." *To appear in Pacific J. of Optimization*, 2009.

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

[Lu09] Z. Lu. "Smooth optimization approach for sparse covariance selection." *SIAM Journal on Optimization*, **19**:1807, 2009.

[Lu10] Z. Lu. "Adaptive first-order methods for general sparse inverse covariance selection." *SIAM Journal on Matrix Analysis and Applications*, 2010. To appear.

[Lut05] H. Lütkepohl. *New Introduction to Multiple Time Series Analysis*. Springer, 2005.

[LZ09] Z. Lu and Y. Zhang. "An Augmented Lagrangian Approach for Sparse Principal Component Analysis." Submitted., 2009.

[Mar87] S.L. Marple. *Digital spectral analysis with applications*. Prentice-Hall, Englewood Cliffs, NJ, 1987.

[MB06] N. Meinshausen and P. Bühlmann. "High-dimensional graphs and variable selection with the Lasso." *Annals of Statistics*, **34**(3):1436–1462, 2006.

[MHN04] 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.

[Nes04] Yu. 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.

[Nes07] Y. Nesterov. "Gradient methods for minimizing composite objective function." Technical report, Center for Operations Research and Econometrics (CORE), Catholic University of Louvain, Louvain-la-Neuve, Belguim, 2007.

[Pol87] B. T. Polyak. *Introduction to Optimization*. Optimization Software, Inc., 1987.

[Pro01] J. Proakis. *Digital Communications*. McGraw-Hill, Boston, MA, 4 edition, 2001.

[RBL08] A. J. Rothman, P. J. Bickel, E. Levina, and J. Zhu. "Sparse permutation invariant covariance estimation." *Electronic Journal of Statistics*, **2**:494–515, 2008.

[Rei07] G. C. Reinsel. *Elements of Multivariate Time Series Analysis*. Springer, second edition, 2007.

[RFG05] A. Roebroeck, E. Formisano, and R. Goebel. "Mapping directed influence over the brain using Granger causality and fMRI." *Neuroimage*, **25**(1):230–242, 2005.

[RFG09] A. Roebroeck, E. Formisano, and R. Goebel. "The identification of interacting networks in the brain using fMRI: Model selection, causality and deconvolution." *NeuroImage*, 2009.

[RWR08a] R. Ravikumar, M. J. Wainwright, G. Raskutti, and B. Yu. "High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence.", 2008. Available from `arxiv.org/abs/0811.3628`.

[RWR08b] R. Ravikumar, M. J. Wainwright, G. Raskutti, and B. Yu. "Model selection in Gaussian graphical models: High-dimensional consistency of ℓ_1-regularized MLE." In *Advances in Neural Information Processing (NIPS)*, volume 21, 2008.

[SM97] P. Stoica and R.L. Moses. *Introduction to Spectral Analysis*. Prentice Hall, Upper Saddle River, NJ, 1997.

[SN87] P. Stoica and A. Nehorai. "On stability and root location of linear prediction models." *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **35**:582–584, 1987.

[SR09] K. Scheinberg and I. Rish. "SINCO - a greedy coordinate ascent method for sparse inverse covariance selection problem." 2009. Submitted.

[SS89] T. Söderström and P. Stoica. *System Identification*. Prentice Hall International, London, 1989.

[SSS05] 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.

[SZZ05] T. Serafini, G. Zanghirati, and L. Zanni. "Gradient projection methods for quadratic programs and applications in training support vector machines." *Optimization Methods and Software*, **20**(2):353–378, 2005.

[Tib96] R. Tibshirani. "Regression shrinkage and selection via the Lasso." *Journal of the Royal Statistical Society. Series B (Methodological)*, **58**(1):267–288, 1996.

[TLH00] 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.

[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.

[Tro06] J. Tropp. "Just relax: Convex programming methods for identifying sparse signals in noise." *IEEE Transactions on Information Theory*, **52**(3):1030–1051, 2006.

[Tse08] P. Tseng. "On accelerated proximal gradient methods for convex-concave optimization." *SIAM Journal on Optimization*, 2008. submitted.

[TTT99] K.C. Toh, M.J. Todd, and R. H. Tütüncü. "SDPT3a Matlab software package for semidefinite programming." *Optimization Methods and Software*, **11**(12):545–581, 1999.

[VBV06] P.A. Valdés-Sosa, J.M. Bornot-Sánchez, M. Vega-Hernández, L. Melie-García, A. Lage-Castellanos, and E. Canales-Rodríguez. "Granger Causality on Spatial Manifolds: Applications to Neuroimaging." In B. Schelter, M. Winterhalder, and J. Timmer, editors, *Handbook of Time Series Analysis: Recent Theoretical Developments and Applications*. Wiley, 2006.

[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.

[VSL05] P.A. Valdés-Sosa, J. M. Sánchez-Bornot, A. Lage-Castellanos, M. Vega-Hernández, J. Bosch-Bayard, L. Melie-García, and E. Canales-Rodríguez. "Estimating brain functional connectivity with sparse multivariate autoregression." *Philosophical Transactions of the Royal Society B: Biological Sciences*, **360**(1457):969–981, 2005.

[Whi90] J. Whittaker. *Graphical Models in Applied Multivariate Statistics*. Wiley, New York, 1990.

[WJ08] M.J. Wainwright and M.I. Jordan. "Graphical models, exponential families, and variational inference." *Foundations and Trends® in Machine Learning*, **1**(1-2):1–305, 2008.

[WNF09] 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.

[YGS10] A.Y. Yang, Arvind Ganesh, S. Shankar Sastry, and Y. Ma. "Fast ℓ_1-Minimization Algorithms and An Application in Robust Face Recognition: A Review." Technical report, Electrical Engineering and Computer Sciences, University of California, Berkeley, 2010.

[YL06] 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.

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

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

[ZRY09] 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.