MAXIMUM-LIKELIHOOD ESTIMATION OF AUTOREGRESSIVE MODELS WITH CONDITIONAL INDEPENDENCE CONSTRAINTS

Jitkomut Songsiri[*] *Joachim Dahl* *Lieven Vandenberghe*[*]

University of California, Los Angeles
Department of Electrical Engineering

Aalborg University
Department of Electronic Systems

University of California, Los Angeles
Department of Electrical Engineering

ABSTRACT

We propose a convex optimization method for maximum likelihood estimation of autoregressive models, subject to conditional independence constraints. This problem is an extension to times series of the classical covariance selection problem in graphical modeling. The conditional independence constraints impose quadratic equalities on the autoregressive model parameters, which makes the maximum likelihood estimation problem nonconvex and difficult to solve. We formulate a convex relaxation and prove that it is exact when the sample covariance matrix is block-Toeplitz. We also observe experimentally that in practice the relaxation is exact under much weaker conditions.

We discuss applications to topology selection in graphical models of time series, by enumerating all possible topologies, and ranking them using information-theoretic model selection criteria. The method is illustrated by an example of air pollution data.

Index Terms— graphical models, conditional independence, semidefinite programming relaxation, model selection

1. INTRODUCTION

Let X be an n-dimensional Gaussian random variable with covariance matrix Σ. Two components X_i and X_j are *conditionally independent*, given the other components, if and only if $(\Sigma^{-1})_{ij} = 0$ [6]. In a graph representation of X, the nodes represent the components X_i; two nodes are connected by an undirected edge if the corresponding variables are not conditionally independent. The problem of computing the maximum likelihood (ML) estimate of Σ subject to conditional independence constraints is known as *covariance selection problem* [6].

The notion of conditional independence can be extended to time series. Let $x(t)$, $t \in \mathbb{Z}$, be a multivariate stationary Gaussian process with spectral density matrix $S(\omega)$. The components x_i and x_j are conditionally independent given the remaining variables if and only if $(S(\omega)^{-1})_{ij} = 0$ for all ω [3, 5]. This condition allows us to consider estimation problems with conditional independence constraints by placing restrictions on the inverse of spectral density matrix.

In this paper, we consider ML estimation of autoregressive (AR) models of vector time series, subject to conditional independence constraints. The difficulty of this problem arises from the quadratic equalities on the AR parameters imposed by the zero pattern in $S(\omega)^{-1}$. This leads to a difficult nonconvex optimization problem. The main contribution of this paper is to show that under certain conditions the constrained ML estimation problem can be solved efficiently via a convex relaxation.

Related problems have been studied in [1, 7]. Bach and Jordan [1] consider the problem of learning the structure of the graphical model of a time series from sample estimates of the joint spectral density matrix. Eichler [7] uses Whittle's approximation of the exact likelihood function, and imposes conditional independence constraints via algorithms extended from covariance selection. In these two methods, a non-parametric estimate of the spectrum is first computed, taking into account the conditional independence constraints. In a second step, an AR model is obtained via the Yule-Walker equations. The method presented in this paper provides a more direct approach, and computes the AR coefficients directly from a convex reformulation of the maximum likelihood problem.

Notation. $\mathbf{S}^n$ denotes the set of symmetric matrices of order n. The sparsity pattern of a sparse matrix $X \in \mathbf{S}^n$ will be characterized by specifying the set of indices $\mathcal{V} \subseteq \{1, \ldots, n\} \times \{1, \ldots, n\}$ of its zero entries. We assume $\mathcal{V}$ is symmetric, *i.e.*, if $(i, j) \in$

[*] Research supported in part by NSF grants ECCS-0524663 and ECCS-0824003.

$\mathcal{V}$ then $(j,i) \in \mathcal{V}$, and that it does not contain any diagonal entries, *i.e.*, $(i,i) \notin \mathcal{V}$ for $i = 1,\ldots,n$. $P(X)$ denotes the projection of a square symmetric or non-symmetric matrix X on $\mathcal{V}$:

$$P(X)_{ij} = \begin{cases} X_{ij} & (i,j) \in \mathcal{V} \\ 0 & \text{otherwise.} \end{cases} \qquad (1)$$

If X is a $p \times p$ block-matrix with i,j block X_{ij}, then we define $P(X)$ as the $p \times p$ block matrix with i,j block $P(X)_{ij} = P(X_{ij})$.

2. CONDITIONAL INDEPENDENCE

Let $x(t)$, $t \in \mathbb{Z}$, be an n-dimensional Gaussian time series with positive spectral density matrix $S(\omega)$. As mentioned in the introduction, the components x_i and x_j are conditionally independent given all other variables if and only if

$$(S^{-1}(\omega))_{ij} = 0 \quad \forall\omega \qquad (2)$$

(see [3, 5]). We now apply this characterization to autoregressive models. Consider an autoregressive model of order p,

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

where $v(t) \sim \mathcal{N}(0,I)$, and B_0 is nonsingular. (Without loss of generality we can assume that B_0 is symmetric positive definite.) It is easily shown that (2) reduces to

$$(Y_k)_{ij} = (Y_k)_{ji} = 0,$$

where $Y_k = \sum_{l=0}^{p-k} B_l^T B_{l+k}$ for $k = 0,\ldots,p$.

3. MAXIMUM-LIKELIHOOD ESTIMATION

We are interested in ML estimation based on $N+p$ observations, $x(1)$,..., $x(N+p)$. The exact ML problem is difficult, even without sparsity constraints. A standard approach is to condition on the first p observations $x(1)$, ..., $x(p)$ and to use the conditional density function of the last N observations [9]. If we define $B = \begin{bmatrix} B_0 & B_1 & \cdots & B_p \end{bmatrix}$, then it can be verified that the log-likelihood function is

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

where $R = (1/N)HH^T$ and

$$H = \begin{bmatrix} x(p+1) & x(p+2) & \ldots & x(N+p) \\ x(p) & x(p+1) & \ldots & x(N+p-1) \\ \vdots & \vdots & & \vdots \\ x(1) & x(2) & \ldots & x(N) \end{bmatrix}. \qquad (3)$$

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

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

The variables are $Y_0, B_0 \in \mathbf{S}^n$ and $Y_k, B_k \in \mathbf{R}^{n\times n}$, $k = 1,\ldots,p$. By using the projection notation defined in (1), the ML problem can be written more clearly as

$$\begin{array}{ll} \text{min.} & -\log\det B_0 + \frac{1}{2}\,\mathbf{tr}(RB^TB) \\ \text{s.t.} & P\left(\sum_{i=0}^{p-k} B_i^T B_{i+k}\right) = 0, \quad k = 0,\ldots p. \end{array} \qquad (4)$$

This problem includes quadratic equality constraints and is therefore nonconvex.

4. CONVEX RELAXATION

A change of variables $X = B^TB$ transforms (4) in the equivalent problem

$$\begin{array}{ll} \text{min.} & -\log\det X_{00} + \mathbf{tr}(RX) \\ \text{s.t.} & P\left(\sum_{i=0}^{p-k} X_{i,i+k}\right) = 0, \quad k = 0,\ldots,p \\ & X \succeq 0, \quad \mathbf{rank}(X) = n, \end{array}$$

with variable $X \in \mathbf{S}^{n(p+1)}$, a symmetric block matrix with block entries X_{ij} of size $n \times n$. Deleting the rank constraint results in a convex relaxation:

$$\begin{array}{ll} \text{min.} & -\log\det X_{00} + \mathbf{tr}(RX) \\ \text{s.t.} & P\left(\sum_{i=0}^{p-k} X_{i,i+k}\right) = 0, \quad k = 0,\ldots,p \\ & X \succeq 0. \end{array} \qquad (5)$$

The convex optimization problem (5) is equivalent to (4) if it can be guaranteed that the optimal solution X has rank n. This is the case under certain assumptions, as we now show. We assume that R is block-Toeplitz and positive definite, and partitioned as

$$R = \begin{bmatrix} R_0 & R_1 & \cdots & R_p \\ R_1^T & R_0 & \cdots & R_{p-1} \\ \vdots & \vdots & \ddots & \vdots \\ R_p^T & R_{p-1}^T & \cdots & R_0 \end{bmatrix} \qquad (6)$$

with $R_0 \in \mathbf{S}^n$, $R_1,\ldots,R_p \in \mathbf{R}^{n\times n}$.

4.1. The dual problem

The dual problem of (5) can be expressed as

$$\begin{array}{ll} \text{maximize} & \log\det W + n \\ \text{subject to} & \begin{bmatrix} W & 0 \\ 0 & 0 \end{bmatrix} \preceq R + P(Z), \end{array} \qquad (7)$$

with variables $W \in \mathbf{S}^n$ and $Z \in \mathbf{S}^{n(p+1)}$. The matrix Z is block-Toeplitz and partitioned as in (6). It follows from standard results in convex duality that the optimal values of (5) and (7) are equal, and that the optimal solutions X, W, Z are related by

$$\begin{aligned} X_{00}^{-1} &= W, \\ X\left(R + P(Z) - \begin{bmatrix} W & 0 \\ 0 & 0 \end{bmatrix}\right) &= 0. \end{aligned} \qquad (8)$$

(See [2, chapter 5].)

4.2. Exactness of the relaxation

Assume $X^\star$, $W^\star$, $Z^\star$ are optimal. We will use the following result to show that $X^\star$ has rank n.

Let R be a symmetric block-Toeplitz matrix with block-dimensions as in (6). If R satisfies

$$R \succeq \begin{bmatrix} W & 0 \\ 0 & 0 \end{bmatrix},$$

for some positive definite $W \in \mathbf{S}^n$, then $R \succ 0$. This is easily proved by induction on p, and the proof is omitted here.

Using this result we see that the constraints in (7) imply that $R + P(Z^\star) \succ 0$ if $Z^\star$ is dual optimal. Thus the rank of the matrix

$$R + P(Z^\star) - \begin{bmatrix} W^\star & 0 \\ 0 & 0 \end{bmatrix}$$

is at least np, and its nullspace has dimension at most n. It follows from (8) that $\mathbf{rank}(X^*) \leq n$, and since $X_{00}^* \succ 0$, we have $\mathbf{rank}(X^*) = n$.

This result shows that if R is block-Toeplitz and positive definite, the optimal solution $X^\star$ of the convex problem (5) can be factorized as $X = B^T B$, with $B_0 = X_{00}^{1/2}$, and that B is the globally optimal solution of (4).

The matrix R in the ML problem (3) approaches a block-Toeplitz matrix as N increases. We conjecture that the convex relaxation remains exact if R is almost Toeplitz. This is confirmed by experimental results in the next section.

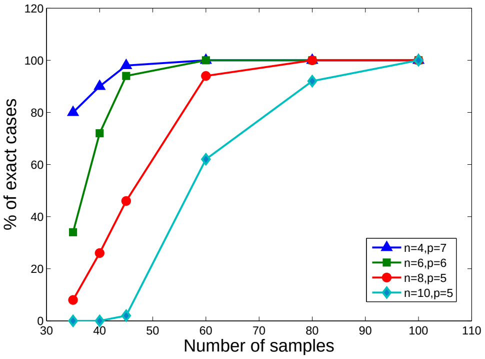

Fig. 1: Percentage of rank-n solutions versus the number of samples.

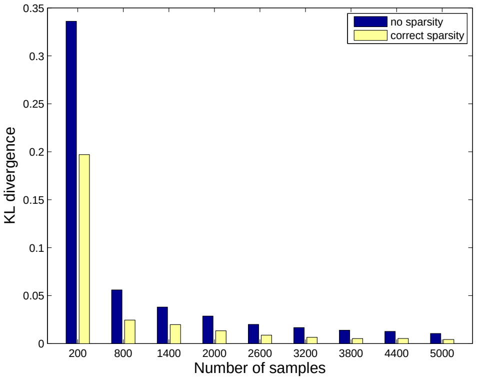

Fig. 2: KL divergence from the true model versus the number of samples.

5. EXAMPLES

5.1. Randomly generated data

We generated 50 sets of time series from four AR models with sparse spectral densities. We solved (5) with different numbers of samples (N) and show the percentage of rank-n solutions in Figure 1. The figure illustrates that the relaxation is exact for moderate values of N, even though the matrix R is not block-Toeplitz.

Figure 2 shows the convergence rate of the ML estimates (with and without sparsity constraints) to the true model, as a function of the number of samples. We use a model has dimension $n = p = 6$. The Kullback-Leibler (KL) divergence [1] is used to measure the difference between the estimated and the true spectrum. The figure illustrates that the ML estimate without the sparsity constraint gives the model with substantially larger values of KL divergence when N is small.

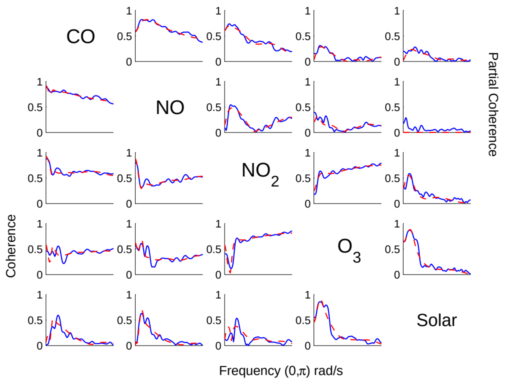

Fig. 3: Partial coherence spectrum $S(\omega)^{-1}$ (upper triangular part) and coherence spectrum $S(\omega)$ (lower triangular part) for the air pollution data. Nonparametric estimates are in solid blue lines and ML estimates are in dashed red lines.

5.2. Real data set

We illustrate the proposed method by 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. The data were observed from Jan 1, 2006 to Dec 31, 2006 in Azusa, California, and were obtained from Air Quality and Meteorological Information System (AQMIS) (`www.arb.ca.gov/aqd/aqdcd/aqdcd.htm`). This application was discussed previously in [5] by using a nonparametric approach (and with a different data set).

In order to learn the conditional independence graph, we enumerate the AR models of orders $p = 1$ to $p = 8$ with all possible sparsity constraints. For each sparsity pattern and each p, we constructed R from (3) and solved (5) using `CVX` [8], and then decomposed the optimal rank-n X to obtain AR parameters A_k. For each fitted model, we computed the BIC (Bayesian information criterion) score $\mathrm{BIC} = k \log N - 2L$, where L is the maximized log-likelihood, N is the sample size, and k is the effective number of parameters [4]. The best BIC score is a model of lag $p = 4$, and a conditional independence pattern in which only the pair $(\mathrm{NO}, \mathrm{R})$ is conditionally independent. (Several other topologies gave BIC scores that were only slightly worse.) Figure 3 shows the estimates of the partial coherence spectrum (normalized $S(\omega)^{-1}$) and coherence spectrum (normalized $S(\omega)$), obtained from a nonparametric estimation, and for the ML model with the best BIC score. These results are consistent with the discussion in Dahlhaus [5].

6. CONCLUSIONS

We have considered a parametric approach for maximum-likelihood estimation of autoregressive models with conditional independence constraints. The zero constraints on the inverse of spectral density matrix result in nonconvex constraints in the maximum likelihood estimation problem. We have formulated a relaxation which can be solved efficiently by convex optimization, and derived conditions that guarantee that the relaxation is exact. This allows us to solve a graphical inference problem by fitting autoregressive models to different topologies, and comparing the topologies with information-theoretic model selection criteria. The approach was illustrated with randomly generated and real data.

7. REFERENCES

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

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

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

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

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

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

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

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

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