ARMA Identification of Graphical Models

Enrico Avventi, Anders G. Lindquist, *Life Fellow, IEEE*, and Bo Wahlberg, *Fellow, IEEE*

***Abstract*—Consider a Gaussian stationary stochastic vector process with the property that designated pairs of components are conditionally independent given the rest of the components. Such processes can be represented on a graph where the components are nodes and the lack of a connecting link between two nodes signifies conditional independence. This leads to a sparsity pattern in the inverse of the matrix-valued spectral density. Such graphical models find applications in speech, bioinformatics, image processing, econometrics and many other fields, where the problem to fit an autoregressive (AR) model to such a process has been considered. In this paper we take this problem one step further, namely to fit an autoregressive moving-average (ARMA) model to the same data. We develop a theoretical framework and an optimization procedure which also spreads further light on previous approaches and results. This procedure is then applied to the identification problem of estimating the ARMA parameters as well as the topology of the graph from statistical data.**

***Index Terms*—Autoregressive moving-average (ARMA) modeling, conditional independence, graphical models, system identification.**

I. Introduction

GRAPHICAL models represent families of probability distributions in the form of graphs which expose relative conditional independences. In this paper we consider a real, m-dimensional, zero-mean, Gaussian, stationary stochastic vector process $\{x(t)\}_{t\in\mathbb{Z}}$ with the property that designated pairs of components are conditionally independent given the rest of the components. In fact, such processes can be represented on a graph where the components are nodes and the lack of a connecting link between two nodes signifies conditional independence [4]. As was shown in [3], this is manifested by a sparsity pattern in the inverse of the $m \times m$ matrix-valued spectral density

$$\Phi(e^{i\theta}) = \sum_{k=-\infty}^{\infty} R_k e^{ik\theta} \tag{I.1}$$

where

$$R_k := E\{x(k)x(0)^{\mathsf{T}}\} \tag{I.2}$$

and where we assume that $\Phi(e^{i\theta}) > 0$ for all $\theta \in [-\pi, \pi]$. Let $\mathcal{S}_+^m$ denote the class of such spectral densities that are integrable on $[-\pi, \pi]$. In fact, as we shall demonstrate in more detail in Section III, it can be shown that

$$\left[\Phi(e^{i\theta})^{-1}\right]_{k\ell} = 0, \quad -\pi \le \theta \le \pi \tag{I.3}$$

for pairs (k, ℓ) such that x_k and x_j are conditionally independent given the rest of the components of the process x; also see [3], [4]. Such graphical models find applications in speech, bioinformatics, image processing, econometrics and many other fields [1], [2], [6], [7].

More precisely, given $V := \{1, 2, \dots m\}$ and

$$X_I = \operatorname{span}\{x_j(t) : j \in I, t \in \mathbb{Z}\}$$

for an arbitrary index set $I \subset V$, (I.3) holds for all pairs (k, ℓ) such that $X_{\{k\}}$ and $X_{\{\ell\}}$ are *conditionally independent* given $X_{V\setminus\{k,\ell\}}$, which we write as

$$X_{\{k\}} \perp X_{\{\ell\}} \mid X_{V\setminus\{k,\ell\}}.$$

The set of all such conditional independence relations constitutes a graph $G = (V, E)$ where V, defined as above, is the set of vertices and $E \subseteq V \times V$ is a set of edges defined in the following way:

$$(k, \ell) \notin E \iff k \neq \ell, \;\; X_{\{k\}} \perp X_{\{\ell\}} | X_{V\setminus\{k,\ell\}}.$$

A typical such graph is depicted in Fig. 1, where the lack of an arch between nodes k and ℓ signifies conditional independence between the processes $\{x_k(t)\}_{\mathbb{Z}}$ and $\{x_\ell(t)\}_{\mathbb{Z}}$ given the rest of the component processes. Graphs of this type are referred to as a *partial/conditional independence graph* or, more simply, as an *interaction graph* in the literature. A model of the process x which takes conditional independence relations into consideration is commonly referred as a *graphical model*. In Section II we present some motivating examples of applications exhibiting such graphical models.

The problem to fit an autoregressive (AR) model to such a process has been considered in [4], [6] as a means for assessing conditional independence. The basic idea is to use a maximum likelihood and ask for consistency of the AR model with the data together with conditional independence between particular nodes. More precisely, given the (estimates of) autocovariances $R_0, R_1, \dots, R_n$, the problem in these papers is to find a multivariate autoregressive model

$$\sum_{j=0}^{n} A_j x(t-j) = e(t) \tag{I.4}$$

Manuscript received August 12, 2011; revised September 24, 2012; accepted November 27, 2012. Date of publication November 30, 2012; date of current version April 18, 2013. This work was supported by the Swedish Research Foundation through ACCESS and other VR grants and by the Swedish Foundation for Strategic Research through CIAM. Recommended by Associate Editor E. Weyer.

E. Avventi is with the Division of Optimization and Systems Theory, Department of Mathematics, Center for Industrial and Applied Mathematics (CIAM), ACCESS, KTH Royal Institute of Technology, Stockholm 100 44, Sweden (e-mail: avventi@kth.se).

A. G. Lindquist is with the Department of Automation, Shanghai Jiao Tong University, Shanghai 100 44, China and also with the Division of Optimization and Systems Theory, Department of Mathematics, Center for Industrial and Applied Mathematics (CIAM), and ACCESS, KTH Royal Institute of Technology, Stockholm 100 44, Sweden . (e-mail address: alq@kth.se).

B. Wahlberg is with the Automatic Control Lab and ACCESS, KTH Royal Institute of Technology, Stockholm 100 44, Sweden (e-mail address: bo.wahlberg@ee.kth.se).

Digital Object Identifier 10.1109/TAC.2012.2231551

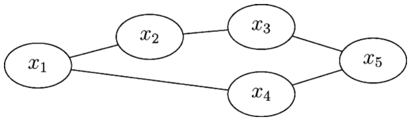


Fig. 1. Example of an interaction graph. For instance, $x_2 \perp x_4 \mid x_1, x_3, x_5$.

that satisfies the sparsity constraint (I.3). Here $\{e(t)\}_{t\in\mathbb{Z}}$ is a white noise process and $A_0, A_1, \ldots, A_n$ are $m \times m$ matrices such that the determinant of the matrix polynomial

$$A(z) = A_0 z^n + A_1 z^{n-1} + \cdots + A_n \tag{I.5}$$

has all its zeros in the unit disc $\mathbb{D}$ of the complex plane and $\det A_0 > 0$.

However, there are examples where AR models are insufficient, for example when the process has zeros close to the unit circle. Moreover, an AR model of exceedingly high order can be approximated by a low order ARMA model. Therefore, in this paper we take this problem one step further, namely to fit an autoregressive moving-average (ARMA) model, while respecting the sparsity constraint (I.3), to the same data. In fact, the problem at hand is equivalent to a *covariance extension problem*, namely a problem of finding an infinite extension $R_{n+1}, R_{n+2}, R_{n+3}, \ldots$ of the given sequence $R_0, R_1, \ldots, R_n$, preserving positivity of the corresponding block Toeplitz matrices; see, e.g., [8], [9]. In this context, the AR solution corresponds to very special covariance extension, namely the maximum entropy one. By allowing for ARMA models, we may choose from a continuum of infinitely many solutions, one of which might satisfy the required graph topology better.

The ARMA models that we shall consider here take the form

$$\sum_{j=0}^{n} A_j x(t-j) = \sum_{j=0}^{n} B_j e(t-j). \tag{I.6}$$

For technical reasons we shall here assume that the matrix coefficients of the moving-average part has the form

$$B_j = b_j I, \quad j = 0, 1, \ldots, n, \quad b_0 = 1 \tag{I.7}$$

where the scalar polynomial

$$b(z) = z^n + b_1 z^{n-1} + \cdots + b_n \tag{I.8}$$

has no zeros in the closed unit disc. Of course one or several of the coefficients $b_1, b_2, \ldots, b_n$ may be zero. Then the spectral density of the stationary vector process x becomes $\Phi(z) = W(z)W(z^{-1})^{\mathsf{T}}$, where $W(z) := b(z)A(z)^{-1}$.

Note that any ARMA model (I.6) can be reduced to a form where (I.7) holds. Naturally, in general this comes with an increase in the model order and ultimately of the number of parameters to be estimated. In particular, any ARMA model of order p can be represented by a ARMA model satisfying (I.7) of order $n = pm$. To see this note that an ARMA model of order p has a state space representation

$$\begin{cases} s(t+1) = Fs(t) + Ge(t) \\ x(t) = Hs(t) + De(t) \end{cases}$$

of dimension mp with transfer function $W(z) = H(sI - F)^{-1}G + D$. The inverse system

$$\begin{cases} s(t+1) = (F - GD^{-1}H)s(t) + GD^{-1}x(t) \\ e(t) = -D^{-1}Hs(t) + D^{-1}x(t) \end{cases}$$

has transfer function $W(z)^{-1} = b(z)^{-1}A(z)$ with $b(z) = \det(F - GD^{-1}H)$. The number of parameters of the representation will in general grow from $2m^2p$ to $mp(m^2+1)$. On the other hand, the representation satisfying (I.7) will be sparse whenever conditional independence relations hold, as we shall see in Section III. For highly sparse models the two ARMA representations will have a similar number of parameters.

Consequently, our basic problem is to determine a spectral density of the form

$$\Phi(z) = \psi(z)Q(z)^{-1} \tag{I.9}$$

satisfying the sparsity constraint (I.3) and the moment conditions

$$\int_{-\pi}^{\pi} e^{ik\theta}\Phi(e^{i\theta})\frac{d\theta}{2\pi} = R_k, \quad k = 0, 1, \ldots, n \tag{I.10}$$

where ψ is a scalar pseudo-polynomal[1] of degree at most n and Q is a real, para-hermitian $m \times m$ matrix-valued pseudo-polynomial; i.e., $Q(z)^* := Q(z^{-1})^{\mathsf{T}} = Q(z)$. Then the coefficients in the corresponding ARMA model (I.6) can be obtained by determining the minimum-phase spectral factors $A(z)$ and $b(z)$ from

$$A(z)A(z^{-1})^{\mathsf{T}} = Q(z) \tag{I.11a}$$
$$b(z)b(z^{-1}) = \psi(z). \tag{I.11b}$$

To deal with this problem we shall use the convex optimization approach to moment problems developed in various forms in [10]–[12], [16], [17], which we shall review in Section III together with the basic ideas on graphical models. At the same time we obtain an alternative motivation for the optimization procedure in [4], [6]. In Section IV we incorporate the sparsity constraint (I.3) in this optimization approach to yield one of our main theoretical results.

However, in applied problems, we are just given a string of measured data

$$\mathtt{x}_0, \mathtt{x}_1, \ldots, \mathtt{x}_N \in \mathbb{R}^m \tag{I.12}$$

from the ARMA model (I.6), and we want to estimate the parameters $A_0, A_2, \ldots, A_n, b_1, \ldots, b_n$ without prior knowledge of the topology of the graph. Hence we also need to estimate a suitable graphical structure E from the data. In fact, in many applications, determining the topology of the graph is the main task. This is the topic of Section V. Finally, in Section VI, we present some simulations and in Section VII our conclusions.

This is a extension of our previous conference paper [24].

II. Motivating Examples

We begin by presenting a selection of practical problems where the application of graphical models shows promise.

[1] A polynomial in positive and negative powers of z.

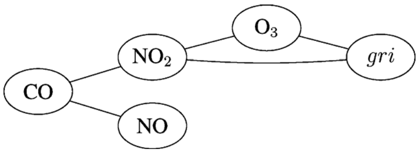


Fig. 2. Estimated graphical model for air pollution data.

For each of these applications, the interaction graphs, taken from the literature, have been determined by nonparametric methods. Using an ARMA approach as in this paper would also produce a dynamical model which can be used for prediction. It would also produce a more economical description for coding purposes.

A. Environmental Chemistry

In the environment the dynamics of chemical species concentrations are determined by a large number of highly interdependent reactions. Specifically, each compound may be involved in several reactions in the role of either a reactant or a product, hence resulting in a concentration decrease or increase respectively.

Graphical models constitute a very powerful tool in analyzing such complex systems. In fact, if the concentration dynamics of two chemical species are conditionally independent, it may be assumed that no significant direct reaction takes place between them. Hence, by estimating the interaction graph of a set of compounds one can gain insight about the reactions that govern the system dynamics.

Dahlhaus in [4] applied these concepts to a set of air pollutants. The data collected consisted of concentrations of CO and NO, which are produced by human activities (transportation, heating and industry), NO$_2$ and O$_3$, which are byproducts of atmospheric reactions, and the global radiation intesity (gri), which directly affects these reactions. The underlying graphical model was estimated by statistical testing on the inverse smoothed periodogram, and the resulting graph is depicted in Fig. 2.

B. Financial Markets Interdependence

In the design of investment portfolios it is crucial to have information about the levels of correlation between different asset prices in order to accurately estimate risk. On the other hand, during times of financial crisis, the correlation between different markets tends to increase. In order to be able to accurately assess risk in such situations, it is useful to consider conditional correlations instead. In fact, by enforcing certain conditional independence relations, one can eliminate the influence of the overall market dynamics.

In [26] Adelwahab *et al.* analyzed various stock markets utilizing graphical models. The data considered was the time series of day-to-day stock markets returns at closing time of the world's biggest financial markets. Here, the return for each market is computed as

$$x(t) = 100\,[\ln p(t) - \ln p(t-1)] \qquad \text{(II.1)}$$

where $p(t)$ is the corresponding closing price on day t. The countries considered were United States (S&P 500 index), United Kingdom (FTSE 100 index), Japan (Nikkei index), Germany (German Akien index), Canada (TSE 300 index), Hong Kong (Hang Seng index), France (CAC 40 index), Switzerland (SI index), Australia (AOAI) and Italy (MIBTel index). The period from January 2000 till December 2005 was considered and the interaction graph depicted in Fig. 3 was estimated by applying the nonparametric identification procedure proposed in [4].

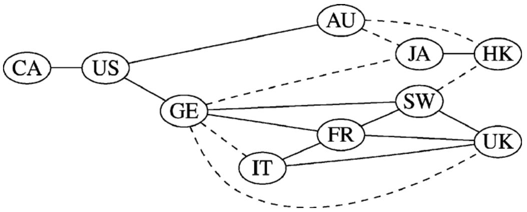


Fig. 3. Interaction graph for the international stock returns. The solid and dashed edges indicates strong and weak partial correlation respectively.

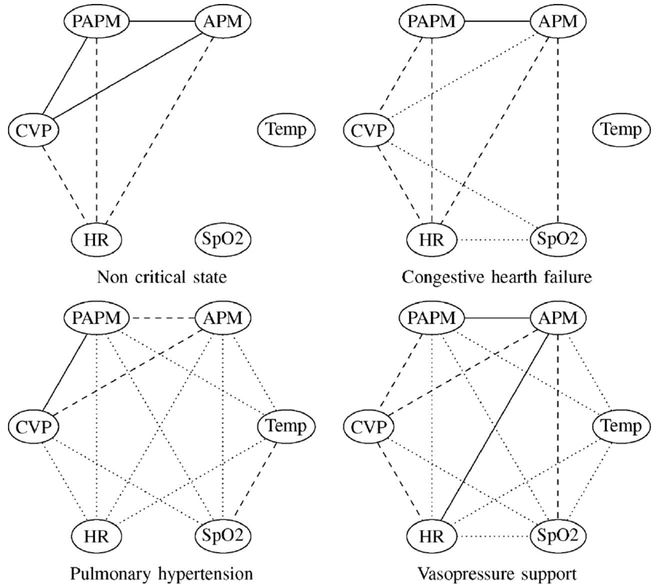
The authors of [26] observed that the markets are interacting by geographical proximity leading to three highly interacting subgraphs: one each for Europe, Asia and North America. Among these the European group is noteworthy by the strength of the partial correlation between its markets. This fact was attributed to the existence of the Economic and Monetary Union (EMU) since 1999. Another contributing factor is the small difference in time zones and the recent efforts in synchronizing the trading within Europe. Of particular interest is the role of Germany. In fact it appears from this model that Germany acts as a gateway of information between the European and American markets.

C. Physiological Monitoring

In intensive care, a high number – in the order of hundreds – of physiological parameters are recorded in order to monitor the condition of a patient. Such an high dimensional time series is then processed by an automated alarm system to ensure timely warnings whenever the patient enters into a critical condition.

In [25] Gather *et al.* explored the use of graphical models as a tool to build more robust intensive care monitoring systems. In fact, they have shown that different physiological conditions correspond to different interaction graphs of the monitored parameters. Therefore critical conditions can be detected by graphical model estimation. To validate this potential approach they considered a set of parameters concerning the haemodynamic system, namely: heart rate (HR), arterial mean pressure (APM), pulmonary arterial mean pressure (PAPM), central venous pressure (CVP), blood temperature (Temp) and pulsoximetry (SpO2). Such a choice of parameters was made because the interactions between them are well understood, and hence the estimation process can be validated against experts knowledge. The data about the above vital signs was collected every minute from a number of patients with pulmonary artery catheters. For each patient the data was divided in batches each corresponding to the different physiological condition occurring at that time. Then an estimated graphical models for each condition is obtained by averaging over the differerent patients as shown in Fig. 4. It was noted that such average graphs agreed with experts knowledge and thus could be used to potentially identify the patients conditions.

In order to apply this approach to online alarm systems we need to develop procedures for estimating the interaction graph that are robust even when the number of data points is low. In



Non critical state

Congestive hearth failure

Pulmonary hypertension

Vasopressure support

Fig. 4. Partial correlation graphs corresponding to different clinical states. Solid, dashed and dotted lines stands for different levels of partial correlation from high to low.

fact, the occurrence of a critical state must be detected immediately, and the longer data strings used for the estimation the longer is the delay in the eventual alarm.

III. Preliminaries

For any real, rational function F taking values in $\mathbb{C}^{m\times m}$

$$\Re\{F(z)\} := \frac{1}{2}\left[F(z) + F^*(z)\right]$$

with $F^*(z) = F(z^{-1})^\mathsf{T}$ is the Hermitian generalization of the real part in the scalar case. Moreover, for two $\mathbb{C}^{m\times m}$-valued functions F, G in $L_2(\mathbb{T})$, where $\mathbb{T}$ is the unit circle, define the inner product

$$\langle F, G\rangle = \int_{-\pi}^{\pi} \operatorname{tr}\{F(e^{i\theta})G^*(e^{i\theta})\}\frac{d\theta}{2\pi}$$

where tr denotes the trace.

Given the autocovariances $R_0, R_1, \ldots, R_n$ in $\mathbb{R}^{m\times m}$, define the matrix pseudo-polynomial

$$R(z) = R_0 + 2\Re\left\{\sum_{j=1}^{n} z^j R_j\right\}. \tag{III.1}$$

We also define the family $\mathcal{Q}(m,n)$ of $m \times m$ matrix pseudo-polynomials

$$\mathcal{Q}(m,n) = \left\{Q(z) = \Re\left\{\sum_{j=0}^{n} z^j Q_j\right\}\right.$$
$$\left. : Q(e^{i\theta}) > 0,\ \ \forall\theta \in [-\pi, \pi]\right\} \tag{III.2}$$

where $Q_0, Q_1, \ldots, Q_n \in \mathbb{R}^{m\times m}$. Then a straight-forward calculation shows that

$$\langle R, Q\rangle = \sum_{j=0}^{n} \operatorname{tr}(R_j Q_j) \tag{III.3}$$

for any $Q \in \mathcal{Q}(m,n)$, but, in view of (I.5) and (I.11), we also have

$$\langle R, Q\rangle = \operatorname{tr}\{\mathbf{A}^\mathsf{T} T(R)\mathbf{A}\} \tag{III.4}$$

where $\mathbf{A}^\mathsf{T} = [\,A_0^\mathsf{T} \quad A_1^\mathsf{T} \quad \ldots \quad A_n^\mathsf{T}\,]$ with A_k, $k = 0, 1, \ldots, n$, being the coefficients of the spectral factor $A(z)$ of $Q(z)$ as defined in (I.11), and

$$T(R) = \begin{bmatrix} R_0 & R_1 & \cdots & R_n \\ (R_1)^\mathsf{T} & R_0 & \cdots & R_{n-1} \\ \vdots & \vdots & \ddots & \vdots \\ (R_n)^\mathsf{T} & (R_{n-1})^\mathsf{T} & \cdots & R_0 \end{bmatrix} \tag{III.5}$$

is the block Toeplitz matrix corresponding to $R_0, R_1, \ldots, R_n$.

Proposition 1: Given the autocovariances $R_0, R_1, \ldots, R_n$, there exists a $\Phi \in \mathcal{S}_+^m$ satisfying the moment equations (I.10) if and only if $\langle R, Q\rangle > 0$ for all $Q \in \mathcal{Q}(m,n)$, or, equivalently, $T(R) > 0$; i.e., the block Toeplitz matrix $T(R)$ is positive definite.

Proof: Since (III.4) should hold for all $\mathbf{A}$ such that $Q \in \mathcal{Q}(m,n)$, $\langle R, Q\rangle > 0$ for all $Q \in \mathcal{Q}(m,n)$ if and only if $T(R) > 0$. Now, given (I.10)

$$\langle R, Q\rangle = \int_{-\pi}^{\pi} \operatorname{tr}\{\Phi(e^{i\theta})Q^*(e^{i\theta})\}\frac{d\theta}{2\pi} > 0$$

for any $\Phi \in \mathcal{S}_+^m$ and $Q \in \mathcal{Q}(m,n)$, which shows that the positivity condition is necessary. Sufficiency will follow from Theorem 2. ■

A. A Convex-Optimization Solution of the Moment Problem

We begin by reviewing the convex optimization approach to moment problems developed in [10]–[12], [16], [17]. The following result can be found in [17].

Theorem 2: Suppose that $T(R) > 0$, and let $\psi \in \mathcal{Q}(1,n)$. Then the optimization problem

$$(\mathbf{P})\left[\begin{array}{l} \max\limits_{\Phi\in\mathcal{S}_+^m}\ -\langle\psi I, \log(\psi\Phi^{-1})\rangle \\ \text{subject to } \int_{-\pi}^{\pi} e^{ij\theta}\Phi(e^{i\theta})\frac{d\theta}{2\pi} = R_j \\ \hfill j = 0, 1 \ldots, n \end{array}\right] \tag{III.6}$$

has a unique solution $\hat{\Phi}$, and it is rational of the form

$$\hat{\Phi}(z) = \psi(z)\hat{Q}(z)^{-1}. \tag{III.7}$$

Here $\hat{Q}$ is the unique solution of the dual optimization problem

$$(\mathbf{D})\ \min_{Q\in\mathcal{Q}(m,n)} \mathbb{J}_\psi(Q) \tag{III.8}$$

where the dual functional

$$\mathbb{J}_\psi(Q) := \langle R, Q\rangle - \langle\psi I, I + \log Q\rangle \tag{III.9}$$

is strictly convex.

For the benefit of the reader we provide a brief derivation of the dual functional (III.9). To this end, we note that the Lagrangian is given by

$$L(\Phi,Q) = -\langle \psi I, \log(\psi\Phi^{-1})\rangle + \operatorname{tr}\left\{\Re\left[\sum_{j=0}^{n} Q_j\left(R_j - \int_{-\pi}^{\pi} e^{ij\theta}\Phi(e^{i\theta})\frac{d\theta}{2\pi}\right)\right]\right\}$$

where $Q_0, Q_1, \dots, Q_n$ are $m \times m$ matrices of Lagrange multipliers. This in turn can be written in the more compact form

$$L(\Phi,Q) = \langle \psi I, \log(\psi^{-1}\Phi)\rangle + \langle R, Q\rangle - \langle \Phi, Q\rangle. \tag{III.10}$$

In the dual problem $\min_Q \sup_{\Phi\in\mathcal{S}_+^m} L(\Phi,Q)$, we need consider only $Q \in \overline{\mathcal{Q}(m,n)}$, as the supremum is infinite for $Q \notin \overline{\mathcal{Q}(m,n)}$. Since the Gateux differential of $\Phi \mapsto L(\Phi,Q)$ is $\delta L(\Phi,Q,\delta\Phi) = \langle \psi\Phi^{-1} - Q, \delta\Phi\rangle$, $\Phi = \psi Q^{-1}$ is a stationary point, which inserted into (III.10) yield the dual functional (III.9). For the rest of the proof we refer the reader to [17].

Theorem 2 provides a complete parameterization in terms of $\psi \in \mathcal{Q}(1,n)$ of all models (I.6) of the form (I.7) such that (I.9) satisfies the moment conditions (I.10). In particular, choosing $\psi \equiv 1$ we obtain the *maximum entropy solution* which corresponds to the AR model (I.4) and the solution of which is linear problem where $\mathbf{A}$ can be obtained from the normal equations.

B. Graphical Models of Stochastic Processes

The cross-spectrum

$$\Phi_{xy}(e^{i\theta}) = \sum_{k=-\infty}^{\infty} E\{x(k)y(0)^{\mathsf{T}}\}e^{ik\theta}$$

of two zero-mean, stationary Gaussian stochastic vector processes $\{x(t)\}_{t\in\mathbb{Z}}$ and $\{y(t)\}_{t\in\mathbb{Z}}$ plays an important role in the theory of graphical models. In particular, if x and y are scalar, the *coherence*

$$r_{xy}(e^{i\theta}) = \frac{\Phi_{xy}(e^{i\theta})}{\sqrt{\Phi_{xx}(e^{i\theta})\Phi_{yy}(e^{i\theta})}}$$

of x with y is useful in studying possible linear dynamic relations between x with y, as it measures the extent to which $y(t)$ may be predicted from $x(t)$ by an optimal linear least squares estimate.

Now, consider an m-dimensional, zero-mean, Gaussian, stationary stochastic vector process $\{x(t)\}_{t\in\mathbb{Z}}$. In order to build a graphical model, conditional independence needs to be characterized in way suitable for analysis. For a given $(k,\ell) \in V \times V$, $k \neq \ell$, let P be a permutation matrix such that

$$\tilde{x}(t) = Px(t) = \begin{bmatrix} y(t) \\ s(t) \end{bmatrix} \tag{III.11}$$

where

$$y(t) = \begin{bmatrix} x_k(t) \\ x_\ell(t) \end{bmatrix}$$

and s is formed by the remaining components ordered by their indices. The spectral density of $\tilde{x}$ can be evaluated from the one of x and partitioned in the following way:

$$\tilde{\Phi}(z) = P\Phi(z)P^{\mathsf{T}} = \begin{bmatrix} \Phi_{yy}(z) & \Phi_{ys}(z) \\ \Phi_{sy}(z) & \Phi_{ss}(z) \end{bmatrix}.$$

We are now interested in determining the part of y that is orthogonal to $X_{V\setminus\{k,\ell\}}$ by solving the following minimization problem:

$$\min_{\varepsilon} \; E[\varepsilon(t)^{\mathsf{T}}\varepsilon(t)]$$
$$\text{subject to} \quad y(t) - \varepsilon(t) \in X_{V\setminus\{k,\ell\}} \quad \forall t \in \mathbb{Z}$$

the optimal solution of which can be obtained as the output of an acasual filter with the input $\tilde{x}$ and the transfer function

$$W(z) = [\, I \quad -\Phi_{ys}(z)\Phi_{ss}^{-1}(z) \,]$$

[3], thus leading to the spectral density

$$\Phi_{\varepsilon\varepsilon}(z) = \Phi_{yy}(z) - \Phi_{ys}(z)\Phi_{ss}^{-1}(z)\Phi_{sy}(z)$$

the entries of which are the spectra and cross-spectra of the chosen components after removing the effects of all the other. In particular, we have

$$\Phi_{x_kx_\ell|s}(z) = \Phi_{x_kx_\ell}(z) - \Phi_{x_ks}(z)\Phi_{ss}^{-1}(z)\Phi_{sx_\ell}(z).$$

Clearly, if x is Gaussian, then so is ε, and hence x_k and x_ℓ are conditional independent if and only if $\Phi_{x_kx_\ell|s}(e^{i\theta}) = 0$ for all θ. The conditional coherence of x_k with x_ℓ can be defined as

$$r_{x_kx_\ell|s}(z) = \frac{\Phi_{x_kx_\ell|s}(z)}{\sqrt{\Phi_{x_kx_\ell|s}(z)\Phi_{x_kx_\ell|s}(z)}}$$

and, as proved by Dahlhaus in [4], satisfies

$$r_{x_kx_\ell|s}(e^{i\theta}) = \frac{[\Phi^{-1}(e^{i\theta})]_{k,\ell}}{\sqrt{[\Phi^{-1}(e^{i\theta})]_{k,k}[\Phi^{-1}(e^{i\theta})]_{\ell,\ell}}} \tag{III.12}$$

whenever $\Phi(e^{i\theta})$ is full rank for all θ. From this it follows that:

$$[\Phi^{-1}(e^{i\theta})]_{k,\ell} = 0 \quad \forall\theta \in [-\pi,\pi] \tag{III.13}$$

is a necessary and sufficient condition for x_k and x_ℓ to be conditionally independent. This is a dynamic version of a very important result first established by Dempster [14] and recently elaborated upon in [13], [15]

Using this characterization of conditional independence we can define subsets of $\mathcal{S}_+^m$ with a common graphical structure. To this end, let $\mathcal{S}_+^m(E) \subset \mathcal{S}_+^m$ be the set of all spectral densities such that (III.13) holds for all $(k,\ell) \notin E$.

IV. Covariance Extension for Graphical Models

We now turn to the basic problem of this paper, namely to find a model (I.6) that satisfies (I.2) and the sparsity condition (III.13). Now, by Theorem 2, all such solutions must have a spectral density of the form (I.9), and therefore the sparsity condition (III.13) can be reformulated as

$$Q_{k\ell} \equiv 0 \quad \text{for all } (k,\ell) \notin E. \tag{IV.1}$$

Indeed, the main reason for choosing the numerator in $W(z) := b(z)A(z)^{-1}$ to be scalar is to insure that the sparsity pattern of Φ

will be inheritied by Q. Now, unlike $\mathcal{S}_+^m$, the set $\mathcal{S}_+^m(E)$ is unfortunately not convex, so modifying the primal problem (III.6) by maximizing over $\mathcal{S}_+^m(E)$ is not a good idea. Instead, we modify the dual problem (III.8) by adding the constraint (IV.1). This gives us the convex optimization problem

$$(\mathbf{D_E})\begin{bmatrix} \min\limits_{Q\in\mathcal{Q}(m,n)} \langle R,Q\rangle - \langle \psi I, I+\log Q\rangle \\ \text{subject to } Q_{k\ell}\equiv 0 \quad (k,\ell)\notin E \end{bmatrix}. \tag{IV.2}$$

This optimization problem was used in the special maximum-entropy case $\psi\equiv 1$ in [6] to derive an AR model, but no theoretical justification was provided. In our setting, the dual problem (III.8) is just a device to solve the the primal problem (III.6), and *a priori* it is not clear how the added constraint (IV.1) affects the original problem. We need to formulate a problem for which (IV.2) is the dual. To this end, let $\mathcal{R}_+^m(E)$ denote the set of all R, expressed in the pseudo-polynomial form (III.1), with the property that there is an $\bar{R}$ such that $T(\bar{R})>0$ and $[\bar{R}_j]_{k\ell}=[R_j]_{k\ell}$ for all $(k,\ell)\in E$ and for $j=0,1,\ldots,n$.

Proposition 3: Suppose that $R\in\mathcal{R}_+^m(E)$. Then, for each $\psi\in\mathcal{Q}(1,n)$, (IV.2) is the dual of the optimization problem

$$(\mathbf{P_E})\begin{bmatrix} \max\limits_{\Phi\in\mathcal{S}_+^m} -\langle \psi I, \log(\psi\Phi^{-1})\rangle \\ \text{subject to} \\ \quad \int_{-\pi}^{\pi} e^{ij\theta}\Phi_{k\ell}(e^{i\theta})\frac{d\theta}{2\pi}=[R_j]_{k\ell} \\ \quad\quad \forall(k,\ell)\in E, \quad j=1,2\ldots n \end{bmatrix}. \tag{IV.3}$$

Moreover, strong duality holds for (IV.3) and (IV.2).

Proof: The Lagrangian of (IV.3) is given by

$$\begin{aligned} L(\Phi,Q) =& -\langle \psi I, \log(\psi\Phi^{-1})\rangle \\ &+\mathrm{Re}\left\{\sum_{(k,\ell)\in E}\sum_{j=0}^{n}[Q_j]_{k\ell}([R_j]_{k\ell}-\langle\Phi_{k\ell},z^j\rangle)\right\} \\ =& \langle\psi I,\log(\psi^{-1}\Phi)\rangle+\langle R,Q\rangle-\langle\Phi,Q\rangle \end{aligned}$$

where, for $j=1,2\ldots n$, $[Q_j]_{k\ell}$ are Lagrange multipliers for $(k,\ell)\in E$ and $[Q_j]_{k\ell}=0$ for $(k,\ell)\notin E$. Then the dual problem becomes

$$\min_{Q}\sup_{\Phi\in\mathcal{S}_+^m} L(\Phi,Q).$$

However, whenever Q fails to be positive semi-definite on the unit circle, the supremum takes the value $+\infty$. Moreover, as we shall see in the proof of Theorem 4, the dual functional will not have a minimum on the boundary of $\mathcal{Q}(m,n)$. Hence we need only minimize over $Q\in\mathcal{Q}(m,n)$. The Gateux differential of the Lagrangian with respect to Φ is

$$\delta L(\Phi,Q,\delta\Phi)=\langle\psi\Phi^{-1}-Q,\delta\Phi\rangle$$

and therefore $\Phi=\psi Q^{-1}$ is a stationary point of $\Phi\mapsto L(\Phi,Q)$. Then by substituting the stationary point into the Lagrangian we obtain the objective function of (IV.2).

To prove the last statement note that that (IV.3) is a relaxation of (III.6) with $\bar{R}$ used in place of R the moment conditions. Since $T(\bar{R})>0$ (III.6) is feasible and hence so is also (IV.3). As the feasibility region of (IV.3) is the intersection between an open convex set and an affine set, any feasible point belongs to its relative interior so that Slater's condition holds. ■

It remains to show that the dual optimization problems $(\mathbf{P_E})$ and $(\mathbf{D_E})$ have unique solutions.

Theorem 4: Suppose that $R\in\mathcal{R}_+^m(E)$, and let $\psi\in\mathcal{Q}(1,n)$. Then the optimization problem (IV.3) has a unique solution $\hat{\Phi}$, which satisfies the sparsity condition (I.3), and this solution is rational of the form

$$\hat{\Phi}(z)=\psi(z)\hat{Q}(z)^{-1}. \tag{IV.4}$$

Here $\hat{Q}$ is the unique solution of the convex optimization problem (IV.2); i.e.

$$\min_{Q\in\mathcal{P}(m,n)} \mathbb{J}_\psi(Q) \tag{IV.5}$$

where the strictly convex functional $\mathbb{J}_\psi$ is given by (III.9) and $\mathcal{P}(m,n)$ is the subset of all $Q\in\mathcal{Q}(m,n)$ such that $Q_{k\ell}\equiv 0$ for $(k,\ell)\notin E$.

The proof of Theorem 4 is deferred to the Appendix. In the special case when $\psi\equiv 1$ the primal problem (IV.3) reduces to maximizing the entropy gain

$$\int_{-\pi}^{\pi}\log\det\Phi(e^{i\theta})\frac{d\theta}{2\pi} \tag{IV.6}$$

subject to all covariance conditions corresponding to edges in the graph; i.e., subject to

$$\begin{aligned}\int_{-\pi}^{\pi} e^{ij\theta}\Phi_{k\ell}(e^{i\theta})\frac{d\theta}{2\pi}&=[R_j]_{k\ell}, \\ \forall(k,\ell)\in E, \quad j&=1,2\ldots n.\end{aligned} \tag{IV.7}$$

This is the *maximum entropy solution* corresponding to the graph E. The corresponding dual problem amounts to minimizing

$$\mathbb{J}_1(Q)=\langle R,Q\rangle-\int_{-\pi}^{\pi}\log\det Q(e^{i\theta})\frac{d\theta}{2\pi} \tag{IV.8}$$

subject to $Q_{k\ell}\equiv 0$ for all $(k,\ell)\notin E$. Here we have used the fact that $\mathrm{tr}(\log M)=\log(\det M)$ for any positive definite matrix. Moreover, we have removed a constant term in dual cost criterion.

It turns out that (IV.8) is precisely equivalent to the problem considered in [6], as the following corollary states.

Corollary 5: Given the covariance sequence $R_0,R_1,\ldots,R_n$, let $T(R)$ be its block Toeplitz matrix (III.5). Then the maximum entropy solution corresponding to the graph E is

$$\hat{\Phi}(z)=\hat{Q}(z)^{-1} \tag{IV.9}$$

where

$$\hat{Q}(z)=\Re\left\{\sum_{j=0}^{n}z^j\hat{Q}_j\right\} \tag{IV.10}$$

is the unique optimal solution of the problem to minimize

$$J(X)=\mathrm{tr}\{T(R)X\}-\log\det X_{00} \tag{IV.11}$$

over all symmetric, positive semi-definite $(n+1)\times(n+1)$ matrices such that, for $j = 0, 1, \dots, n$

$$Q_j = \sum_{\nu=0}^{n-j} X_{\nu,\nu+j}, \tag{IV.12a}$$

$$[Q_j]_{k\ell} = 0, \quad \forall (k,\ell) \notin E. \tag{IV.12b}$$

Proof: We want to rewrite (IV.8) in the form (IV.11). To this end, we first observe that (III.4) yields

$$\langle R, Q\rangle = \operatorname{tr}\{T(R)X\}, \qquad \text{where} \qquad X = \mathbf{A}\mathbf{A}^\mathsf{T}$$

from which (IV.12a) readily follows. Next, from (I.11) we have

$$Q(z) = A_*(z^{-1})A_*(z)^\mathsf{T}, \ \text{ where } \ A_*(z) := z^n A(z^{-1})$$

and hence

$$\int_{-\pi}^{\pi} \log\det Q(e^{i\theta})\frac{d\theta}{2\pi} = \int_{-\pi}^{\pi} \log|\det A_*(e^{i\theta})|^2 \frac{d\theta}{2\pi}.$$

However, the real, rational function $\det A_*(z)$ is outer in the unit disc $\mathbb{D}$. Therefore, by Jensen's formula [27, p. 184]

$$\int_{-\pi}^{\pi} \log|\det A_*(e^{i\theta})|^2 \frac{d\theta}{2\pi} = 2\log|\det A_*(0)|.$$

Now, since $A_*(0) = A_0 > 0$, we have

$$2\log|\det A_*(0)| = \log\det(A_0A_0^\mathsf{T}) = \log(\det X_{00})$$

and hence (IV.8) is equivalent to (IV.11), as claimed. Finally, (IV.12b) is equivalent to $Q_{k\ell} \equiv 0$. ■

In a more recent paper [28], Songsiri and Vandenberghe add an ℓ_1 regularization term in X to the cost function (IV.11), thus replacing the sparsity constraint (IV.12). Since the ℓ_1 norm favors sparsity, this has the advantage of being a device for simultaneously estimating the graph topology, rather than estimating E separately. Following this lead, we might consider replacing (IV.2) in our ARMA setting by the problem to minimize

$$\mathbb{J}_\psi(Q) + \lambda\|Q\|_1 \tag{IV.13}$$

for some suitable ℓ_1 norm. However, in this paper we have chosen a different route, which we shall describe in the next section.

V. ARMA Identification of Graphical Models From Statistical Data

Given a string of measured data

$$\mathtt{x}_0, \mathtt{x}_1, \dots, \mathtt{x}_\mathtt{N} \in \mathbb{R}^m \tag{V.1}$$

from the ARMA model (I.6), we want to estimate the parameters $A_0, A_2, \dots, A_n, b_1, \dots, b_n$ and a suitable graphical structure E. To this end, we form the standard (biased) sample autocovariances

$$\hat{R}_k = \frac{1}{N-1}\sum_{j=1}^{N-k} \mathtt{x}_{\mathtt{k}+\mathtt{j}}\mathtt{x}_\mathtt{j}{}^\mathsf{T}, \quad k = 0, 1, \dots, n. \tag{V.2}$$

Such estimates are guaranteed to satisfy the condition $T(\hat{R}) > 0$. Moreover, we will consider a non-parametric Hermitian estimate $\hat{\Phi}_{NP}$ of the spectrum Φ, such as the (damped or smoothed) periodogram.

Our identification approach now proceeds in the following steps.

(a) Compile a list of the most likely sparsity patterns (graphical structures) E as described in detail below.

(b) For each E, estimate the numerator pseudo-polynomial ψ.

(c) Determine Q by solving the convex optimization problem (IV.2) with R and ψ given by (V.2) and (b) respecively. In this way we can estimate a spectral density $\hat{\Phi} = \psi Q^{-1} \in \mathcal{S}_+^m(E)$ for each E in the list compiled under point (a).

(d) Compare the estimates $\hat{\Phi}$ thus obtained by some fitness function.

(e) Determine the parameters $A_0, A_2, \dots, A_n$ and $b_1, \dots, b_n$ from (I.11) by spectral factorization.

It remains to provide procedures for the steps in points (a), (b) and (d), a task to which we turn next.

A. Compiling a List of Candidate Graphical Structures

We base our approach on a method to test the null hypothesis

$$X_{\{k\}} \perp X_{\{\ell\}} | X_{V/\{k,\ell\}}. \tag{H$_0$}$$

To this end, we form a nonparametric estimate of the conditional coherence (III.12) as

$$\hat{r}_{x_kx_\ell|s}(e^{i\theta}) = \frac{[\hat{\Phi}_{NP}^{-1}(e^{i\theta})]_{k,\ell}}{\sqrt{[\hat{\Phi}_{NP}^{-1}(e^{i\theta})]_{k,k}[\hat{\Phi}_{NP}^{-1}(e^{i\theta})]_{\ell,\ell}}} \tag{V.3}$$

where s is defined as in (III.11) and $\hat{\Phi}_{NP}$ is the (smoothed) nonparametric spectral estimate introduced above. It can be shown [5] that the real and imaginary parts of $\hat{r}_{x_kx_\ell|s}(e^{i\theta}) - r_{x_kx_\ell|s}(e^{i\theta})$ are asymptotically normally distributed with mean zero as $N \to \infty$ and that the limit variance σ depends only on the smoothing procedure used to determine $\hat{\Phi}_{NP}$. Moreover, as also shown in [5], we can select M frequencies $\theta_1, \theta_2, \dots, \theta_M \in [-\pi, \pi]$ so that $\hat{r}_{x_kx_\ell|s}(e^{i\theta_p}) \perp \hat{r}_{x_kx_\ell|s}(e^{i\theta_q})$ for all $p, q = 1, 2, \dots, M$ such that $p \neq q$.

Under the the null hypothesis (H$_0$) the real and imaginary parts of $\hat{r}_{x_kx_\ell|s}(e^{i\theta_j})$, $j = 1, 2, \dots, M$, are asymptotically independent and normally distributed with mean zero and variance σ. Hence the probability that the absolute values of these random variables are all less than or equal to γ is

$$p(\gamma) := \prod_{j=1}^{M}\Big\{ P\left\{|\mathrm{Re}[\hat{r}_{x_kx_\ell|s}(e^{i\theta_j})]| \le \gamma\right\} \times P\left\{|\mathrm{Im}[\hat{r}_{x_kx_\ell|s}(e^{i\theta_j})]| \le \gamma\right\}\Big\}$$

which asymptotically equals

$$p_{\lim}(\gamma) = [G(\gamma) - G(-\gamma)]^{2M}$$

where G is the cumulative distribution function of a Gaussian variable with mean zero and variance σ. Now, following Dahlhaus in [4], let $\gamma(\alpha)$ be such that $p_{\lim}(\gamma(\alpha)) = 1 - \alpha$. Then we reject the null hypothesis (H$_0$) at the significance level α if any of the random variables $\mathrm{Re}\{\hat{r}_{x_kx_\ell|s}(e^{i\theta_j})\}$,

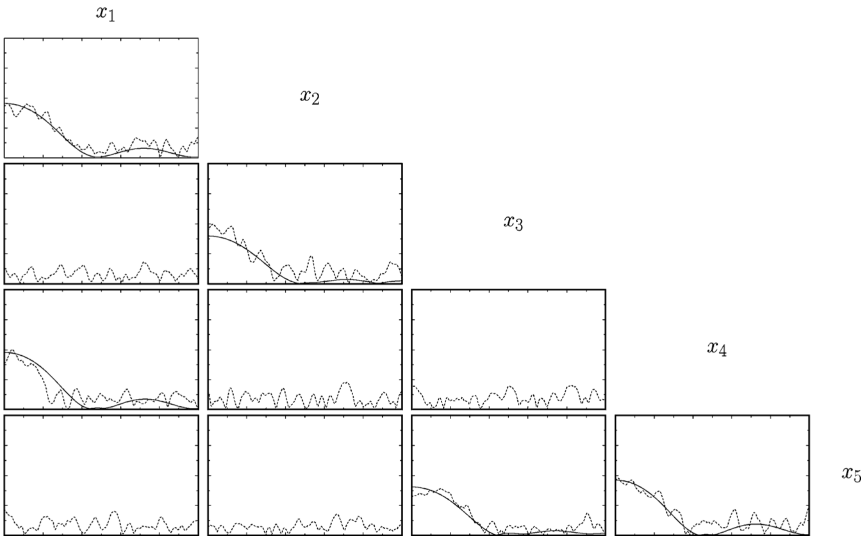


Fig. 5. Partial coherences of a stochastic process with an interaction graph as in Fig. 1. The estimated counterparts computed via a smoothed periodogram from $N = 2^{10}$ data points are plotted in dashed lines.

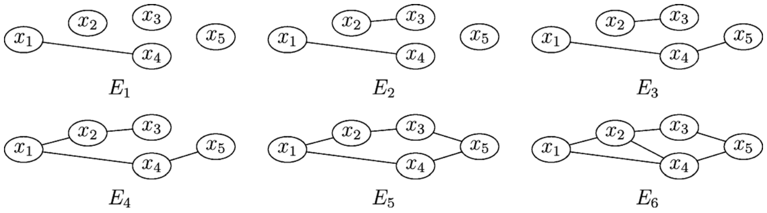


Fig. 6. Some of the graph candidates generated by the estimated partial coherence of Fig. 5.

$\mathrm{Im}\{\hat{r}_{x_k x_\ell|s}(e^{i\theta_j})\}, j = 1, 2, \ldots, M$, has absolute value greater than $\gamma(\alpha)$.

In this paper, instead of considering a fixed α, we now vary α from 0 to 1, and let $E(\alpha)$ be the set of $(k, \ell) \in V \times V$ such that the null hypothesis is rejected by this test at the significance level α. If (H_0) is rejected at the significance level $\bar{\alpha}$, it will also be rejected for all $\alpha > \bar{\alpha}$; i.e., $E(\alpha) \supset E(\bar{\alpha})$. Therefore the family of graphical structure $\{E(\alpha) : \alpha \in [0, 1]\}$ will consist of a finite number of distinct graphical structures

$$E_0 \subset E_1 \subset \cdots \subset E_{m(m-1)/2}$$

ordered by significance levels. In particular, $E_0 = E(0)$ requires Φ^{-1}, and hence Q, to be diagonal, whereas $E_{m(m-1)/2} = E(1)$ does not enforce any conditional independence. Note that the number of different graphical structures considered above is polynomial in m. This is very advantageous compared to an exahaustive list, as, for example, considered in [6], which grows exponentially in m.

To illustrate the proposed method, let us consider an example. Fig. 5 depicts the theoretically partial coherences of a spectrum of the form (I.9) corresponding to a graph as in Fig. 1. The corresponding estimates derived as in (V.3) are reproduced with dashed lines. Now, for each subplot, we need to consider the infinity norm of the estimated partial coherence and order edges accordingly. From Fig. 5 we obtain

$$\begin{aligned}&(x_2, x_5), (x_1, x_3), (x_3, x_4), (x_1, x_5), (x_2, x_4),\\ &\qquad (x_3, x_5), (x_1, x_2), (x_4, x_5), (x_2, x_3), (x_1, x_4)\end{aligned} \quad \text{(V.4)}$$

ordered from low to high. Finally we build the list of candidate graphs in the following way. Start with $E_0 = \{\}$ corresponding to a diagonal spectrum, and then form $E_1, E_2, \ldots, E_{10}$ by successively adding edges from the end of the list (V.4) one at the time. In particular we will have $E_1 = \{(x_1, x_4)\}$, $E_2 = \{(x_1, x_4),\ (x_2, x_3)\}$, etc. as depicted in Fig. 6. Note that, with the given estimates, all edges of the graph in Fig. 1 are ranked higher than the missing edges and hence the original graph is present in the compiled list of candidates labeled as E_5.

B. Estimating the Pseudo-Polynomial ψ

Given a graphical structure E, consider a matrix version of the procedure in [20, page 689], which amounts to solving

$$\begin{bmatrix} \min\limits_{\substack{Q \in \mathcal{Q}(m,n)\\ \psi \in \mathcal{Q}(1,n)}} \ \max_j \left\| Q(\zeta_j) - \psi(\zeta_j)\hat{\Phi}_{NP}(\zeta_j)^{-1} \right\| \\ \text{subject to } Q_{k\ell} \equiv 0 \quad (k, \ell) \notin E \end{bmatrix} \quad \text{(V.5)}$$

with $\zeta_j := e^{i\theta_j}$, $j = 1, 2, \ldots, M$, where the angles $\theta_1, \theta_2, \ldots, \theta_M \in [-\pi, \pi]$ are suitable frequencies, possibly, but not necessarily, the same as the ones above, and $\| \cdot \|$ is the induced L_2 norm. It is not hard to see that (V.5) is equivalent to the following semi-definite programing problem:

$$\begin{bmatrix} \min\limits_{(\psi, Q, \epsilon) \in \mathcal{Q}(1,n) \times \mathcal{Q}(m,n) \times [0,\infty)} \ \epsilon \\ \text{subject to} \\ -\epsilon I \leq Q(\zeta_j) - \psi(\zeta_j)\hat{\Phi}_{NP}(\zeta_j)^{-1} \leq \epsilon I \\ j = 1, 2, \ldots, M \\ Q_{k\ell} \equiv 0 \quad (k, \ell) \notin E \end{bmatrix}. \quad \text{(V.6)}$$

In order to insure that ψ and Q are positive definite, one may need to add the constraints $\psi(\zeta_j) \geq \delta$ and $Q(\zeta_j) \geq \delta I$, $j = 1, 2, \ldots, M$, for some $\delta > 0$. Note, however, that the positivity of these quantities on frequencies between the sampling points is not automatically guaranteed. Hence a more advanced approach would be to use the Kalman-Yakubovich-Popov equations to express the positivity constraints in terms of the coefficients of Q and ψ. In the solution we are of course only interested in ψ, as a more accurate Q will be determined in step (c). However, the Q obtained here may be used as a starting point in an iterative algorithm solving (IV.2).

A more natural, but more complicated, method for determining ψ could be to solve the quasi-convex optimization problem

$$\begin{bmatrix} \displaystyle\min_{\substack{Q\in\mathcal{Q}(m,n)\\ \psi\in\mathcal{Q}(1,n)}} \max_j \left\| \psi(\zeta_j) Q(\zeta_j)^{-1} - \hat{\Phi}_{NP}(\zeta_j) \right\| \\ \text{subject to } Q_{k\ell} \equiv 0 \quad (k,\ell) \notin E \end{bmatrix}.$$

Proceeding in the same way as the step from (V.5) to (V.6) we obtain the constraints

$$-\epsilon I \leq \psi(\zeta_j) Q(\zeta_j)^{-1} - \hat{\Phi}_{NP}(\zeta_j) \leq \epsilon I$$

$j = 1, 2, \ldots, M$, which are not linear, or, equivalently

$$-\epsilon\, Q(\zeta_j) \leq \psi(\zeta_j) I - \hat{\Phi}_{NP}(\zeta_j) Q(\zeta_j) \leq \epsilon\, Q(\zeta_j) \qquad \text{(V.7)}$$

which is linear only if we disallow ϵ from being a variable. Therefore, this problem needs to be solved in steps. First fix ϵ and solve the feasibility problem to find $\psi \in \mathcal{Q}(1, n)$ and $Q \in \mathcal{P}(m, n)$ satisfying (V.7), after which ϵ is decreased in steps (e.g., by the bisection algorithm) until we obtain the smallest ϵ for which feasibility problem is solvable. Recall that $\mathcal{P}(m, n)$ is the subset of all $Q \in \mathcal{Q}(m, n)$ such that $Q_{k\ell} \equiv 0$ for $(k, \ell) \notin E$. It remains to determine bounds for this method. This could be done along the lines proposed in [21]–[23].

C. Graphical Model Selection

It remains to grade each of the models obtained with some fitness function that weights both the sparsity of the model and the adherence to the observed data. For instance in [6] information theoretic criteria that originate from order selection were considered. Here we propose a fitness function that naturally stems from the theory presented so far.

Let ψ_E and Q_E be calculated as in (b) and (c) above with respect to the edge list E. Since, in solving (IV.2), only a subset of the estimated covariance lags are matched, one can utilize the remaining ones as data for validation. Specifically, we consider

$$\Delta(E) = \max_{j=0,1,\ldots,n;\ (k,\ell)\notin E} \left| [\hat{R}_j]_{k,\ell} - \int_{-\pi}^{\pi} e^{ij\theta} \psi_E(e^{i\theta}) \left[Q_E(e^{i\theta})^{-1} \right]_{k\ell} \frac{d\theta}{2\pi} \right| \qquad \text{(V.8)}$$

which is the maximum gap between the estimated covariance lags and the ones corresponding to $\psi_E Q_E^{-1}$, as a criterion to judge how close the model matches the data.

However, (V.8) alone is not effective as a fitness function. In fact, by choosing models that minimize $\Delta(E)$ we end up favoring dense models over sparse ones. Hence we need also to weight in the sparsity pattern, which is the cardinality of E, here denoted as $C(E)$. In particular, we considered the following fitness function:

$$F(E) = [(1+\epsilon)\Delta(E_0) - \Delta(E)] \times \left[\frac{m(m-1)}{2} - C(E) \right] \qquad \text{(V.9)}$$

to be maximized, where ϵ is some small positive constant. By using a product instead of a sum in (V.9) we obtain robustness to scaling. In fact, any scaling performed on one of the two terms becomes just a scaling factor of the fitness function itself and, thus does not change the optimal solution. Note that both factors are nonnegative. The ϵ has been introduced not to exclude diagonal spectra, as the fitness function $F(E)$ would otherwise always be zero for them.

Remark 6: By replacing the dual problem (IV.2) by its ℓ_1-regularized counterpart (IV.13), we could potentially combine steps (a), (c) and (d). We have not investigated whether such a procedure would be more efficient at identifying the graphical structure than the one presented above – this will be a matter of future study.

In any case, it appears that a fixed value of λ in (IV.13) is not likely to be optimal for all data records. Hence, as proposed in [28] in the context of AR models, the regularized problem would need to be solved for different values of λ and the resulting models compared. The computational complexity of the two approaches would then largely depend on the number of candidate models to be evaluated.

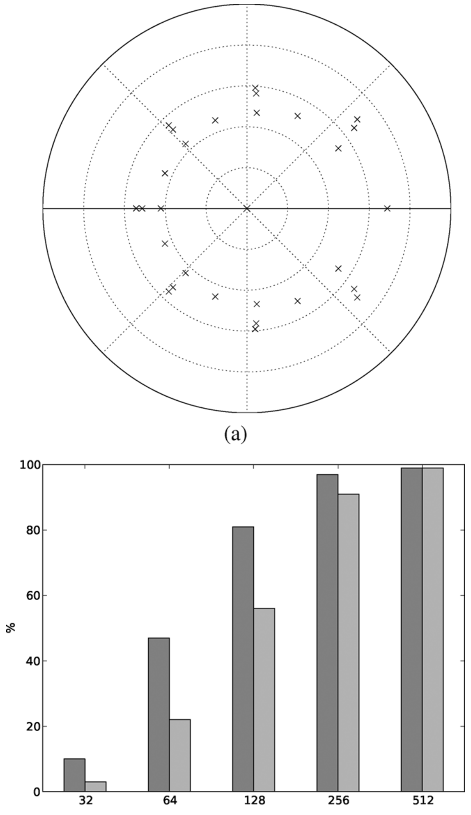
VI. Simulations

In order to validate our approach we present some results from simulations. In particular, we focus on the ability of the proposed method to estimate the underlying graphical structure correctly. To this end, we generated data sequences of various length by feeding white noise through two different systems (I.6) with $m = 4$. Here, the two models considered have the same simple graphical structure, a connected chain. For each data length, the procedure described in Section V was executed with data from 100 different realizations. In this paper we have used the Broyden–Fletcher–Goldfarb–Shanno (BFGS) method, which approximates Newton's method, to solve (IV.2).

The first system considered is an AR filter with $n = 8$ whose pole positions are depicted in Fig. 7(a). The estimation procedure was performed using ARMA models with $n = 6$. To determine the effect of the zero estimation, we executed the procedure also with $\psi = 1$ fixed, thus limiting the model class considered to AR models. In this case, the true value $n = 8$ of the order was used instead.

In Fig. 7(b) the percentage of runs that completed successfully estimating the correct graphical structure are reported, for each data length. Note that for $N = 256$ or higher we obtained a similar performance using our ARMA procedure as we did when restricting the model class to AR models. Still, with short data records, the extra estimation of zeros clearly has an effect on the success rate.

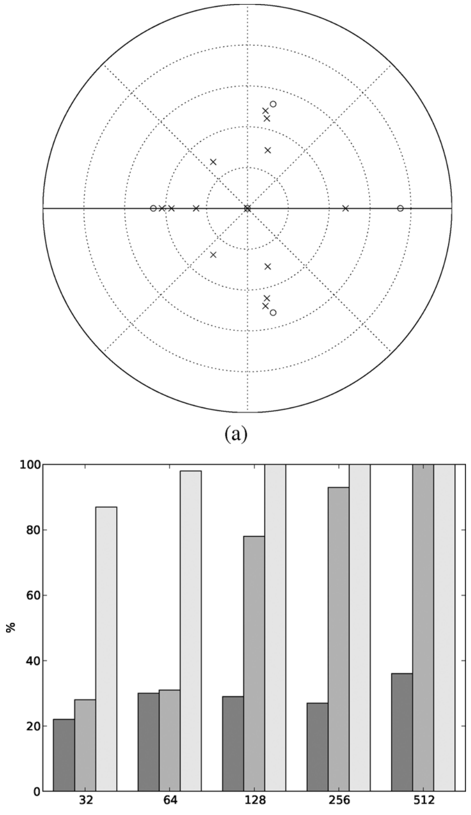
The second set of simulations was run with data generated by an ARMA system with $n = 4$. In Fig. 8(a), the positions of the zeros and poles are depicted. Our ARMA procedure was executed with the same order $n = 4$ and compared to the AR



(a)

(b)

Fig. 7. Simulation results obtained with data generated by an AR filter. (a) Poles of the AR model. (b) Success rate of the proposed procedure using $\psi = 1$ (dark grey) and an estimated ψ (grey), respectively.



(a)

(b)

Fig. 8. Simulation results obtained with data generated by an ARMA system. (a) Poles and zeros of the ARMA model. (b) Success rate of the proposed procedure using $\psi = 1$ (dark grey), an estimated ψ (grey), and the true ψ (light grey), respectively.

procedure using the higher order $n = 5$ to allow for a comparable number of parameters. To gain further insight, a third execution of the procedure was performed for each data set with ψ set equal to that of the original system. The results are reported in Fig. 8(b).

In this case, the AR estimation performs poorly without showing a clear improvement as the data length increases. This is most probably due to the difficulty of a low order AR model to describe accurately pairs of a pole and a zero close to each other. Clearly, a higher order AR model class can be considered instead, although this will lead to models of much higher dimensions. On the other hand, it is interesting to compare the difference in performance between using an estimated and the true ψ. In the same way as for the first simulation, an estimated ψ yields a worse success rate with shorter data sequences while becoming comparable as the data length increases.

These results suggest that the proposed identification procedure is promising. Nonetheless, for short data records, the estimation of the MA part has a negative effect on the performance of the graphical structure identification. This problem may be overcome with alternative zeros estimation techniques.

VII. Conclusion

In this paper we have extended the results in [4], [6] to graphical models of ARMA process. This has been done by posing the problem in the moment-problem framework of [11], [12], [16]–[18]. In particular we have shown that, given the MA part, a minimum-phase ARMA model with graphical structure is uniquely determined, up to a scalar factor, by a particular subset of covariance values and that the corresponding set of interpolation conditions is the largest such set that guarantees the desired graphical structure. Except for allowing for tuning the solution within the same ARMA model class, our approach extends and spreads further light on previous approaches and results. Finally, we apply this parameterization to the problem of system identification with sparsity constraints. We provide a step-by-step procedure to estimate the graphical structure and the corresponding ARMA model respecting the sparsity pattern. Some of these results are preliminary in nature, and further work is needed to test numerical algorithms and statistical procedures. For short data records, as in Application C in Section II, one may consider using a THREE-type approach [12], for which there now are efficient algorithms in the multivariable case [29]–[31].

Appendix

Proof of Theorem 4: For any symmetric $m \times m$ matrix M, let $\Pi_E M$ be the matrix formed from M by replacing all elements corresponding to $(k, \ell) \notin E$ by zero. Then Π_E is a projection, and, since the diagonal elements of M are unaffected by Π_E, we have $\mathrm{tr}(\Pi_E M) = \mathrm{tr}(M)$. Moreover, since M symmetric

$$\Pi_E M = M - \sum_{(k,\ell)\notin E} H_{k\ell} M H_{k\ell} \qquad \text{(A1)}$$

where $H_{k\ell}$ is an $m\times m$ matrix in which element (k,ℓ) is one and all other elements are zero. Therefore, since $\mathrm{tr}\{H_{k\ell}RH_{k\ell}Q\} = \mathrm{tr}\{RH_{k\ell}QH_{k\ell}\}$

$$\langle \Pi_E R, Q\rangle = \langle R, \Pi_E Q\rangle. \tag{A2}$$

For the proof of Theorem 4 we need four lemmas.

Lemma 7: Let $\mathcal{P}(m,n)$ be the subset of all $Q \in \mathcal{Q}(m,n)$ such that $Q_{k\ell} \equiv 0$ for $(k,\ell) \notin E$. Then $\Pi_E\mathcal{R}_+^m(E)$ and $\mathcal{P}(m,n)$ are convex sets of the same dimension.

Proof: Clearly the space of all $\bar{R}$ such that $T(\bar{R}) > 0$ has the same dimension as $\mathcal{Q}(m,n)$, and consequently $\Pi_E\mathcal{R}_+^m(E)$ and $\mathcal{P}(m,n)$ have the same dimension. Convexity is immediate. ■

Next, define the map $F_\psi : \mathcal{P}(m,n) \to \Pi_E\mathcal{R}_+^m(E)$ defined as in (III.1) by

$$F_\psi(Q) = R_0 + 2\Re\left\{\sum_{j=1}^{n} z^j R_j\right\} \tag{A3a}$$

where, for $k = 0, 1, \dots, n$

$$R_k = \int_{-\pi}^{\pi} e^{ik\theta}\psi(e^{i\theta})\Pi_E Q(e^{i\theta})^{-1}\frac{d\theta}{2\pi}. \tag{A3b}$$

Lemma 8: The map $F_\psi : \mathcal{P}(m,n) \to \Pi_E\mathcal{R}_+^m(E)$ is injective.

Proof: Since the dual functional $\mathbb{J}_\psi$, defined by (III.9), is strictly convex (Theorem 2), then so is its restriction to $\mathcal{P}(m,n)$. Hence it has at most one stationary point, which would then be the solution of $F_\psi(Q) = \Pi_E R$ for some $R \in \mathcal{R}_+^m(E)$. In fact, in view of (III.3), $\mathbb{J}_\psi$ can be written

$$\mathbb{J}_\psi(Q) = \sum_{k=0}^{n} \mathrm{tr}(R_k Q_k) - \int_{-\pi}^{\pi} \psi\,\mathrm{tr}(\log Q)\frac{d\theta}{2\pi}$$

which has the Gateaux derivative

$$\delta\mathbb{J}_\psi(Q;\delta Q) = \mathrm{tr}\left\{\sum_{k=0}^{n}\left(R_k - \int_{-\pi}^{\pi} e^{ik\theta}\psi Q^{-1}\frac{d\theta}{2\pi}\right)\delta Q_k\right\}.$$

Hence, any stationary point has to satisfy (I.10), which after projection yields $F_\psi(Q) = \Pi_E R$. ■

Lemma 9: The inverse image $F_\psi^{-1}(K)$ is bounded for any compact $K \subset \Pi_E\mathcal{R}_+^m(E)$.

Proof: First note that, in view of (A2), (A3) and the fact that $\Pi_E Q = Q$

$$\begin{aligned}\langle Q, F_\psi(Q)\rangle &= \sum_{k=0}^{n} \mathrm{tr}\left(Q_k \int_{-\pi}^{\pi} e^{ik\theta}\psi Q^{-1}\frac{d\theta}{2\pi}\right)\\ &= m\int_{-\pi}^{\pi}\psi\frac{d\theta}{2\pi} =: \kappa \end{aligned} \tag{A4}$$

for each $Q \in \mathcal{P}(m,n)$, where κ is a constant. Next, for each $R \in K$, we choose an $\bar{R}$ such that $T(\bar{R}) > 0$ and $[\bar{R}_j]_{k\ell} = [R_j]_{k\ell}$ for all $(k,\ell) \in E$ and for $j = 0, 1, \dots, n$. Since the assignment of $\bar{R}_{k\ell}(z)$, $(k,\ell) \notin E$, is arbitrary, we select $\bar{R}$ so that the smallest eigenvalue of $T(\bar{R})$ is maximized. Then, since K is compact, these eigenvalues are bounded away from zero, and, in particular, there is an $\varepsilon > 0$ such that $T(\bar{R}) \geq \varepsilon I$ for all $R \in K$. Moreover, since $Q = \Pi_E Q$ and $\Pi_E \bar{R} = R$, (A2) yields $\langle Q, R\rangle = \langle Q, \bar{R}\rangle$. Therefore, by (III.4)

$$\langle Q, R\rangle = \mathrm{tr}\{\mathbf{A}^{\mathsf{T}} T(\bar{R})\mathbf{A}\} \geq \varepsilon\|\mathbf{A}\|^2. \tag{A5}$$

If $F_\psi^{-1}(R)$ is nonempty, by Lemma 8, it is a singleton Q, and $\langle Q, R\rangle = \langle Q, F_\psi(Q)\rangle$. Hence it follows from (A4) and (A5) that $\|\mathbf{A}\|^2 \leq \kappa/\varepsilon$. Thus $\mathbf{A}$, and hence Q, is bounded for all $Q \in F_\psi^{-1}(K)$. ■

Lemma 10: The moment map $F_\psi : \mathcal{P}(m,n) \to \Pi_E\mathcal{R}_+^m(E)$ is proper; i.e., the inverse image $F_\psi^{-1}(K)$ is compact for any compact $K \subset \Pi_E\mathcal{R}_+^m(E)$.

Proof: We first note that, in view of (III.3), the fact that $\mathrm{tr}(\Pi_E Q^{-1}) = \mathrm{tr}(Q^{-1})$, and Cramer's rule

$$\langle \psi I, F_\psi(Q)\rangle = \langle \psi^2 I, Q^{-1}\rangle = \int_{-\pi}^{\pi} \frac{\psi^2}{\det Q}\mathrm{tr}(\mathrm{Adj}\ Q)\frac{d\theta}{2\pi} \tag{A6}$$

where $\mathrm{Adj}\ A$ denotes the adjugate matrix of A. Next, let $R^{(k)}$ be a sequence in K converging to $\hat{R} \in K$. If the inverse image of this sequence is empty or finite, it is trivially compact, so we assume it is infinite. Since $F_\psi^{-1}(K)$ is bounded (Lemma 9), there is a convergent subsequence $(Q^{(k)})$ in the inverse image of the sequence $(R^{(k)})$ converging to some limit $\hat{Q}$. (To simplify notations we use k as an index also for these subsequences.) We want to show that $\hat{Q} \in F_\psi^{-1}(K)$. The only way this can fail is that $\hat{Q}$ belongs to the the boundary of $\mathcal{P}(m,n)$, that is, $\det Q$ has a zero on the unit circle. We need to rule this out. From (A6) we have

$$\lim_{k\to\infty}\langle \psi I, R^{(k)}\rangle = \lim_{k\to\infty}\int_{-\pi}^{\pi}\frac{\psi^2}{\det Q^{(k)}}\mathrm{tr}(\mathrm{Adj}\ Q^{(k)})\frac{d\theta}{2\pi}$$

which is the same as

$$\langle \psi I, \hat{R}\rangle = \int_{-\pi}^{\pi}\frac{\psi^2}{\det \hat{Q}}\mathrm{tr}(\mathrm{Adj}\ \hat{Q})\frac{d\theta}{2\pi}. \tag{A7}$$

Suppose $\hat{Q}$ belongs to the the boundary of $\mathcal{P}(m,n)$; i.e., there is a a θ_0 such that $\det\hat{Q}(e^{i\theta_0}) = 0$. Then, if $\lambda_1, \dots, \lambda_m$ are the eigenvalues of $\hat{Q}(e^{i\theta_0})$, we have $\mathrm{tr}\{\mathrm{Adj}\ \hat{Q}(e^{i\theta_0})\} = \sum_{k=1}^{m}\prod_{j\neq k}\lambda_j$. Therefore, if θ_0 is a simple zero, $\mathrm{tr}\{\mathrm{Adj}\ \hat{Q}(e^{i\theta_0})\} > 0$, and there is an $\varepsilon > 0$ such that the Lipschitz condition $\det\hat{Q} \leq |\theta - \theta_0|$ holds for $|\theta - \theta_0| < \varepsilon$. Therefore, in view of (A7)

$$\langle \psi I, \hat{R}\rangle \geq \int_{\theta_0-\varepsilon}^{\theta_0+\varepsilon}\frac{\psi^2}{|\theta-\theta_0|}\mathrm{tr}(\mathrm{Adj}\ \hat{Q})\frac{d\theta}{2\pi} = +\infty$$

which is a contradiction. If θ_0 is a multiple zero of order p, then $\det\hat{Q} \leq |\theta - \theta_0|^p$ holds for $|\theta - \theta_0| < \varepsilon$. Then $p-1$ zeros can be canceled, reducing the problem to the one already treated. Hence $\hat{Q} \in \mathcal{P}(m,n)$, establishing that F_ψ is proper. ■

We are now in a position to prove Theorem 4. By Lemma 7, $\Pi_E\mathcal{R}_+^m(E)$ and $\mathcal{P}(m,n)$ are Euclidean spaces of the same dimension; i.e., they are diffeomorphic to $\mathbb{R}^N$ for the appropriate N. Moreover, the map $F_\psi : \mathcal{P}(m,n) \to \Pi_E\mathcal{R}_+^m(E)$ is injective (Lemma 8) and proper (Lemma 10). Consequently, by Theorem 2.1 (or, in a simpler form, Corollary 2.3) in [19], F_ψ is a homeomorphism. In particular, the dual optimization problem

(IV.5) has a unique solution. The rest follows from strong duality (Proposition 3).

REFERENCES

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

[2] S. L. Lauritzen, *Graphical Models*. Oxford, U.K.: Oxford Univ. Press, 1996.

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

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

[5] R. Dahlhaus, M. Eichler, and J. Sandkühler, "Identification of synaptic connections in neural ensembles by graphical models," *J. Neurosci. Methods*, vol. 77, pp. 93–107, 1997.

[6] J. Songsiri, J. Dahl, and L. Vandenberghe, "Graphical models of autoregressive processes," in *Convex Optimization in Signal Processing and Communications*, D. P. Palomar and Y. C. Eldar, Eds. Cambridge, U.K.: Cambridge Univ. Press, 2010.

[7] M. Eichler, "Fitting graphical interaction models to multivariate time serie," in *Proc. 22nd Conf. Uncertainty Artificial Intell.*, 2006, pp. 147–154.

[8] T. T. Georgiou, "Realization of power spectra from partial covariance sequences," *IEEE Trans. Acoust., Speech Signal Process.*, vol. 35, pp. 438–449, 1987.

[9] C. I. Byrnes, A. Lindquist, S. V. Gusev, and A. S. Matveev, "A complete parameterization of all positive rational extensions of a covariance sequence," *IEEE Trans. Autom. Control*, vol. 40, pp. 1841–1857, 1995.

[10] C. I. Byrnes, S. Gusev, and A. Lindquist, "A convex optimization approach to the rational covariance extension problem," *SIAM J. Control Optim.*, vol. 37, pp. 211–229, 1999.

[11] C. I. Byrnes, S. Gusev, and A. Lindquist, "From finite covariance windows to modeling filters: A convex optimization approach," *SIAM Rev.*, vol. 43, pp. 645–675, 2001.

[12] C. I. Byrnes, T. Georgiou, and A. Lindquist, "A new approach to spectral estimation: A tunable high-resolution spectral estimator," *IEEE Trans. Signal Process.*, vol. 48, pp. 3189–3205, 2000.

[13] F. P. Carli, A. Ferrante, M. Pavon, and G. Picci, "A maximum entropy solution of the covariance extension problem for reciprocal processes," *IEEE Trans. Autom. Control*, vol. 56, no. 9, pp. 1999–2012, Sep. 2011.

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

[15] A. Ferrante and M. Pavon, "Matrix completionà la Dempster by the principle of parsimony," *IEEE Trans. Inform. Theory*, vol. 57, no. 6, pp. 3925–3931, Jun. 2011.

[16] T. Georgiou and A. Lindquist, "Kullback-Leibler approximation of spectral density functions," *IEEE Trans. Inform. Theory*, vol. 49, pp. 2910–2917, 2003.

[17] A. Blomqvist, A. Lindquist, and R. Nagamune, "Matrix-valued Nevanlinna-Pick interpolation with complexity constraint: An optimization approach," *IEEE Trans. Autom. Control*, vol. 48, no. 12, pp. 2172–2190, Dec. 2003.

[18] C. I. Byrnes and A. Lindquist, "Important moments in systems and control," *SIAM J. Control Optim.*, vol. 47, no. 5, pp. 2458–2469, 2008.

[19] C. I. Byrnes and A. Lindquist, "Interior point solutions of variational problems and global inverse function theorems," *Int. J. Robust Nonlin. Control*, vol. 17, pp. 463–481, 2007.

[20] C. I. Byrnes, P. Enqvist, and A. Lindquist, "Cepstral coefficients, covariance lags and pole-zero models for finite data strings," *IEEE Trans. Signal Proces.*, vol. 50, no. 4, pp. 677–693, Apr. 2001.

[21] J. Karlsson and A. Lindquist, "Stability-preserving rational approximation subject to interpolation constraints," *IEEE Trans. Autom. Control*, vol. 53, no. 7, pp. 1724–1730, Jul. 2008.

[22] J. Karlsson, T. T. Georgiou, and A. Lindquist, "The inverse problem of analytic interpolation with degree constraint and weight selection for control synthesis," *IEEE Trans. Autom. Control*, vol. 55, pp. 405–418, 2010.

[23] G. Fanizza, "Modeling and Model Reduction by Analytic Interpolation and Optimization," Ph.D. dissertation, KTH, Stockholm, Sweden, 2008.

[24] E. Avventi, A. Lindquist, and B. Wahlberg, "Graphical models of autoregressive moving-average processes," in *Proc. 19th Int. Symp. Math. Theory Networks Syst. (MTNS'10)*, Budapest, Hungary, Jul. 5–9, 2010, pp. 915–921.

[25] U. Gather, R. Fried, M. Imhoff, and R. Fried, "Multivariate time series, partial spectral coherence, graphical interaction model, substantive research hypotheses, haemodynamic variables," *Statist. Medicine*, vol. 21, pp. 2685–2701, 2002.

[26] A. Abdelwahab, O. Amor, and T. Abdelwahed, "The analysis of the interdependence structure in Int. financial markets by graphical models," *Int. Res. J. Finance Econ.*, vol. 15, pp. 291–306, 2008.

[27] L. V. Ahlfors, *Complex Analysis*. New York: McGraw-Hill, 1953.

[28] J. Songsiri and L. Vandenberghe, "Topology selection in graphical models of autoregressive processes," *J. Mach. Learning Res.*, vol. 11, pp. 2671–2705, 2010.

[29] A. Ferrante, C. Masiero, and M. Pavon, "Time and spectral domain relative entropy: A new approach to multivariate spectral estimation," *IEEE Trans. Autom. Control*, to be published.

[30] A. Ferrante, M. Pavon, and F. Ramponi, "Hellinger vs. Kullback-Leibler multivariable spectrum approximation," *IEEE Trans. Autom. Control*, vol. 53, pp. 954–967, 2008.

[31] F. Ramponi, A. Ferrante, and M. Pavon, "A globally convergent matricial algorithm for multivariate spectral estimation," *IEEE Trans. Autom. Control*, vol. 54, no. 10, pp. 2376–2388, Oct. 2009.

Enrico Avventi received the M.Sc. degree in computer engineering from the University of Padova, Padova, Italy, in 2005 and the Ph.D. degree in optimization and system theory from the Royal Institute of Technology, Stockholm, Sweden, in 2011.

His main areas of research interest are analytic interpolation with degree constraints and its various applications.

Anders G. Lindquist (M'77–SM'86–F'89–LF'10) received the Ph.D. degree in from the Royal Institute of Technology (KTH), Stockholm, Sweden, in 1972.

From 1972 to 1974 he held visiting positions at the University of Florida and Brown University. In 1974 he became an Associate Professor and in 1980 a Professor at the University of Kentucky, where he remained until 1983. In 1982 he was appointed as Professor of the Chair of Optimization and Systems Theory at KTH, and he remained in this position until 2010. Between 2000 and 2009 he was also the Head of the Mathematics Department there. He is now Zhiyuan Chair Professor and Qian Ren Scholar at Shanghai Jiao Tong University, Shanghai, China, and the Director of the Strategic Research Center for Industrial and Applied Mathematics at KTH.

Dr. Lindquist received the 2009 W.T. and Idalia Reid Prize in Mathematics from SIAM, the George S. Axelby Outstanding Paper Award for the year 2003, and the SIGEST paper award from SIAM Review in 2001. He received an Honorary Doctorate (Doctor Scientiarum Honoris Causa) from Technion (Israel Institute of Technology), Haifa, in June 2010. He is a Member of the Royal Swedish Academy of Engineering Sciences, a Foreign Member of the Russian Academy of Natural Sciences, an Honorary Member the Hungarian Operations Research Society, a Fellow of SIAM, and a Fellow of IFAC.

Bo Wahlberg (S'85–M'87–SM'02–F'07) was born in Norrköping, Sweden, in 1959. He received the M.Sc. degree in electrical engineering and the Ph.D. degree in from Linköping University, Sweden, in 1983 and 1987, respectively.

In December 1991, he became Professor of the Chair of Automatic Control at KTH Royal Institute of Technology, Stockholm, Sweden. He was a Visiting Professor with the Department of Electrical Engineering, Stanford University, Stanford, CA, August 1997 to July 1998 and August 2009 to June 2010, and Vice President of KTH 1999 to 2001. His research interests include modeling, system identification, estimation and control of industrial processes.