# Graphical Models of Autoregressive Moving-Average Processes

Enrico Avventi†, Anders Lindquist†, Bo Wahlberg‡

***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 which also spreads further light on previous approaches and results.**

## I. Introduction

Consider an $m$-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. 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

$$\Phi(e^{i\theta}) = \sum_{j=-\infty}^{\infty} R_j e^{ij\theta}, \tag{I.1}$$

where

$$R_k := E\{x(k)x(0)'\}, \tag{I.2}$$

and where we assume that $\Phi(e^{i\theta}) > 0$ for all $\theta \in [-\pi, \pi]$. Let $\mathbb{S}_+^m$ denote the class of such spectral densities that are integrable on $[-\pi, \pi]$. In fact, 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 $(k, \ell)$ such that $x_k$ and $x_j$ are conditionally independent give the rest of the components of the process $x$ [3], [4]. Such graphical models find applications in speech, bioinformatics, image processing, econometrics and many other fields; see [6] and references therein.

The problem to fit an autoregressive (AR) model to such a process has been considered [4], [6]. More precisely, given the autocovariances $R_0, R_1, \dots, R_n$, the problem in these paper were to find a multivariate autoregressive model

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

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

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

have no zeros in the closed unit disc.

In this paper we consider the problem to fit an autoregressive moving-average (ARMA) model, respecting the sparsity constraint (I.3), to the same data. Indeed, AR models of potentially exceedingly high order can often be approximated by a low order ARMA models. 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

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

where the scalar polynomial

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

has no zeros in the in the closed unit disc. Of course one or several of the coefficients $b_1, b_2, \dots, b_n$ may be zero.

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

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

satisfying (I.3) and the moment conditions

$$\int_{-\pi}^{\pi} e^{ij\theta} \Phi(e^{i\theta}) \frac{d\theta}{2\pi} = R_j, \quad j = 0, 1, \dots, n, \tag{I.10}$$

where $\psi$ is a scalar pseudo-polynomal of degree at most $n$ and $Q$ is a symmetric $m \times m$ matrix-valued pseudo-polynomial[1] of degree $n$. 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

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

respectively.

This research was supported by the Swedish Research Foundation through ACCESS and other VR grants and by the Swedish Foundation for Strategic Research through CIAM

†Division of Optimization and Systems Theory, Department of Mathematics, Center for Industrial and Applied Mathematics (CIAM), and ACCESS, Royal Institute of Technology, 100 44 Stockholm, Sweden

‡Automatic Control Lab and ACCESS, Royal Institute of Technology, 100 44 Stockholm, Sweden

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

ISBN 978-963-311-370-7

To deal with this problem we shall use the convex optimization approach to moment problems developed in various forms in [8], [9], [10], [11], [12], and we shall begin with reviewing some of this material in Section II, where we will also review the basic ideas on graphical models. In Section III we present our main results, and in Section IV we apply these results to ARMA modeling from statistical data.

## II. Preliminaries

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

$$\Re\{F(z)\} := \frac{1}{2}\left[F(z) + F^*(z)\right], \quad \text{where } F^*(z) = \overline{F(\bar{z}^{-1})}^{\mathsf{T}},$$

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 $L_2(\mathbb{T})$, 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 $\operatorname{tr}$ denotes the trace.

Given the autocovariances $R_0, R_1, \ldots, R_n$, define the matrix pseudo-polynomial

$$R(z) = \Re\left\{\sum_{j=0}^{n} z^j R_j\right\}. \tag{II.1}$$

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

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

where $Q_0, Q_1, \ldots, Q_n \in \mathbb{R}^{m\times m}$. Then a straight-forward calculation shows that, for any $Q \in \mathcal{Q}(m,n)$,

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

but, in view of (I.5) and (I.11), we also have

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

where $\mathbf{A} = ({A_0}^{\mathsf{T}}, \ldots, {A_n}^{\mathsf{T}})^{\mathsf{T}}$ and

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

*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 (II.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 convex-optimization solution of the moment problem*

We review the convex optimization approach to moment problems developed in [8], [9], [10], [11], [12]. The following result can be found in [12].

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

$$\begin{bmatrix} \displaystyle\max_{\Phi\in\mathcal{S}_+^m} -\langle \psi I, \log(\psi\Phi^{-1})\rangle \\ \displaystyle\int_{-\pi}^{\pi} e^{ij\theta}\Phi(e^{i\theta})\frac{d\theta}{2\pi} = R_j, \quad j = 1, 2 \ldots n \end{bmatrix} \tag{P}$$

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

$$\hat{\Phi}(z) = \psi(z)\hat{Q}(z)^{-1}. \tag{II.6}$$

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

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

where the dual functional

$$\mathbb{J}_\psi(Q) := \langle R, Q\rangle - \langle \psi I, I + \log Q\rangle \tag{II.7}$$

is strictly convex.

It can be shown that strong duality holds; i.e., the maximum value in (P) equals the minimum value in (D). In fact,

$$\begin{aligned} \Delta &= \langle R, \hat{Q}\rangle - \langle \psi I, I + \log\hat{Q}\rangle + \langle \psi I, \log(\psi\hat{\Phi}^{-1})\rangle \\ &= \langle R, \hat{Q}\rangle - \langle \psi I, I + \log\hat{Q}\rangle + \langle \psi I, \log\hat{Q}\rangle \\ &= \langle \hat{\Phi}, \hat{Q}\rangle - \langle \psi I, I\rangle = 0. \end{aligned} \tag{II.8}$$

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.

### *Graphical models of stochastic processes*

The cross-spectrum

$$\Phi_{xy}(e^{i\theta}) = \sum_{j=-\infty}^{\infty} E[x(j)y(0)^T]e^{ij\theta}$$

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 optimum linear least squares function.

Now, consider an $m$-dimensional, zero-mean, Gaussian, stationary stochastic vector process $\{x(t)\}_{t\in\mathbb{Z}}$. Then two distinct components $x_k$ and $x_\ell$ are said to be independent conditionally on all the other components, or in short *conditionally independent*, if

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

where $V = \{1, 2, ...m\}$ and

$$X_I = \operatorname{span}\{x_k(t) : k \in I, t \in \mathbb{Z}\}.$$

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

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

Such a graph is referred to in the literature as a partial/conditional independence graph or, more simply, as interaction graph. A model of the process $x$ which takes conditional independence relations into consideration is commonly referred as a *graphical model*. 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

$$\tilde{x}(t) = Px(t) = \begin{bmatrix} y(t) \\ s(t) \end{bmatrix} \quad \text{where} \quad y(t) = \begin{bmatrix} x_k(t) \\ x_\ell(t) \end{bmatrix}$$

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

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

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

$$\begin{aligned} &\min_{\varepsilon} E[\varepsilon(t)^{\mathsf{T}}\varepsilon(t)] \\ &\text{s.t} \quad y(t) - \varepsilon(t) \in X_{V\setminus\{k,\ell\}} \quad \forall t \in \mathbb{Z} \end{aligned}$$

When $x$ is a Gaussian process the problem can be solved [3], and the optimal time series $\varepsilon$ can be obtained from $\tilde{x}$ by an acasual filter with the transfer function

$$W(z) = \begin{bmatrix} I & -\Phi_{ys}(z)\Phi_{ss}^{-1}(z) \end{bmatrix}.$$

We can evaluate its spectral density as

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

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

$$\Phi_{x_k x_\ell|s}(z) = \Phi_{x_k x_\ell}(s) - \Phi_{x_k s}(z)\Phi_{ss}^{-1}(z)\Phi_{s x_\ell}(z).$$

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

$$r_{x_k x_\ell|s}(z) = \frac{\Phi_{x_k x_\ell|s}(z)}{\sqrt{\Phi_{x_k x_\ell|s}(z)\Phi_{x_k x_\ell|s}(z)}}.$$

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

$$r_{x_k x_\ell|s}(e^{i\theta}) = \frac{[\Phi^{-1}(e^{i\theta})]_{k,\ell}}{\sqrt{[\Phi^{-1}(e^{i\theta})]_{k,k}[\Phi^{-1}(e^{i\theta})]_{\ell,\ell}}} \tag{II.9}$$

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

$$[\Phi^{-1}(e^{i\theta})]_{k,\ell} = 0 \quad \forall\theta \in [-\pi, \pi] \tag{II.10}$$

is a necessary and sufficient condition for $x_k$ and $x_\ell$ to be conditionally independent.

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 (II.10) holds for all $(k, \ell) \notin E$.

## III. Main results

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 (II.10). Now, by Theorem 2, all such solutions must have a spectral density of the form (I.9), and therefore the sparsity condition (II.10) can be reformulated as

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

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

$$\begin{bmatrix} \displaystyle\min_{Q\in\mathcal{Q}(m,n)} \langle R, Q\rangle - \langle \psi I, I + \log Q\rangle \\ \text{subject to} \ \ Q_{k\ell} \equiv 0 \quad (k, \ell) \notin E \end{bmatrix} \tag{D$_E$}$$

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 fact, the dual problem (D) is just a device to solve the the primal problem (P), and *a priori* it is not clear how the added constraint (III.1) affects the original problem. We need to formulate a problem for which (D$_E$) is the dual.

*Proposition 3:* Suppose that there exist $\bar{R}_0, \bar{R}_1, \ldots, \bar{R}_n \in \mathbb{R}^{m\times m}$ such that $T(\bar{R}) > 0$ and $[\bar{R}_j]_{k\ell} = [R_j]_{k\ell}$ for all $(k, \ell) \in E$ and for $j = 0, 1, \ldots, n$. Then, for each $\psi \in \mathcal{Q}(1, n)$, (D$_E$) is the dual of the optimization problem

$$\begin{bmatrix} \displaystyle\max_{\Phi\in\mathcal{S}_+^m} -\langle \psi I, \log(\psi\Phi^{-1})\rangle \\ \text{s.t} \begin{cases} \displaystyle\int_{-\pi}^{\pi} e^{ij\theta}\Phi_{k\ell}(e^{i\theta})\frac{d\theta}{2\pi} = [R_j]_{k\ell} \\ \forall(k, \ell) \in E, \quad j = 1, 2 \ldots n \end{cases} \end{bmatrix}. \tag{P$_E$}$$

Moreover, strong duality holds for (P$_E$) and (D$_E$).

*Proof:* The Lagrangian of (P$_E$) is given by

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

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 7, 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 $\Phi$ 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 (D$_E$).

To prove the last statement note that that (P$_E$) is a relaxation of (P) with $\bar{R}$ is used in place of $R$ the moment conditions. Since $T(\bar{R}) > 0$ (P) is feasible and hence so is also (P$_E$). As the feasibility region of (P$_E$) 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. ■

Now, let $\mathcal{R}_+^m(E)$ denote the set of all $R$, expressed in the pseudo-polynomial form (II.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 $j = 0, 1, \ldots, n$, and let $\mathcal{P}(m,n)$ be the subset of all $Q \in \mathcal{Q}(m,n)$ such that $Q_{k\ell} \equiv 0$ for $(k,\ell) \notin E$. Moreover, for any $m \times m$ matrix $X$, let $\Pi_E X$ be the matrix formed from $X$ by replacing all elements corresponding to $(k,\ell) \notin E$ by zero. Then $\Pi_E$ is a projection, and, since the diagonal elements of $X$ are unaffected by $\Pi_E$, we have $\operatorname{tr}(\Pi_E X) = \operatorname{tr}(X)$.

*Lemma 4:* $\Pi_E\mathcal{R}_+^m(E)$ 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

$$F_\psi(Q) = \Re\left\{\sum_{k=0}^{n} e^{ik\theta} \int_{-\pi}^{\pi} e^{ik\omega}\psi(e^{i\omega})\Pi_E Q(e^{i\omega})^{-1}\frac{d\omega}{2\pi}\right\}. \tag{III.2}$$

*Lemma 5:* 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 (II.7), 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 (II.3), $\mathbb{J}_\psi$ can be written

$$\mathbb{J}_\psi(Q) = \sum_{k=0}^{n} \operatorname{tr}(R_k Q_k) - \int_{-\pi}^{\pi} \psi \operatorname{tr}(\log Q)\frac{d\theta}{2\pi}, \tag{III.3}$$

which has the Gateaux derivative

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

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

*Lemma 6:* The 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 (II.3), the fact that $\operatorname{tr}(\Pi_E Q^{-1}) = \operatorname{tr}(Q^{-1})$, and Cramer's rule,

$$\begin{aligned} \langle \psi I, F_\psi(Q)\rangle &= \langle \psi^2 I, Q^{-1}\rangle \\ &= \int_{-\pi}^{\pi} \frac{\psi^2}{\det Q}\operatorname{tr}(\text{Adj } Q)\frac{d\theta}{2\pi}, \end{aligned} \tag{III.4}$$

where Adj $A$ denotes the adjugate matrix of $A$. We now proceed as in the proof of Lemma 6.3 in [13] to show that $F_\psi^{-1}(K)$ is bounded for any compact $K$ in $\Pi_E\mathcal{R}_+^m(E)$. To this end, let $R^{(k)}$ be a sequence in $K$ converging to $\hat{R}$, and suppose that its preimage contains an infinite number of points $Q^{(k)}$. (For simplicity of notation, $k$ will also be used as the index of the corresponding subsequence $R^{(k)}$, which of course also tends to $\hat{R}$.) This is no restriction, since the cases that the preimage is empty or contains only a finite number of points cannot contradict boundedness. Now set $M_k := \|Q^{(k)}\|$, where $\|\cdot\|$ is any norm in the finite-dimensonal space $\mathcal{Q}(m,n)$ and $\tilde{Q}^{(k)} := Q^{(k)}/M_k$. Then

$$\begin{aligned} \lim_{k\to\infty} M_k\langle \psi^2 I, (\tilde{Q}^{(k)})^{-1}\rangle &= \lim_{k\to\infty}\langle \psi I, F_\psi(Q^{(k)})\rangle \\ &= \lim_{k\to\infty}\langle \psi I, R^{(k)}\rangle \\ &= \langle \psi I, \hat{R}\rangle. \end{aligned} \tag{III.5}$$

However, $\hat{R} \in K \subset \Pi_E\mathcal{R}_+^m(E) \subset \mathcal{R}_+^m(E)$, and hence there there is an $\bar{R} \in \mathcal{R}_+^m(E)$ with the property $T(\bar{R}) > 0$ differing from $\hat{R}$ only at off-diagonal elements. Consequently, by Proposition 1,

$$\begin{aligned} \langle \psi I, \hat{R}\rangle &= \int_{-\pi}^{\pi} \psi \operatorname{tr}(\hat{R})\frac{d\theta}{2\pi} \\ &= \int_{-\pi}^{\pi} \psi \operatorname{tr}(\bar{R})\frac{d\theta}{2\pi} = \langle \psi I, \bar{R}\rangle > 0. \end{aligned} \tag{III.6}$$

Since the sequence $\langle \psi^2 I, (\tilde{Q}^{(k)})^{-1}\rangle$ in (III.5) is bounded away from zero, the sequence $M_k$ must be bounded. Therefore the inverse image of a convergent sequence in $K$ has a cluster point $\hat{Q}$ in the closure of $\mathcal{P}(m,n)$. Now, the lemma would follow if we could show that $\hat{Q}$ lies in the open set $\mathcal{P}(m,n)$ and not on the boundary. To this end, suppose that $\hat{Q}$ lies on the boundary; i.e., there is a a $\theta_0$ such that $\det\hat{Q}(e^{i\theta_0}) = 0$. Now, if $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of $\hat{Q}(e^{i\theta_0})$, then $\mathrm{tr}\{\mathrm{Adj}\ \hat{Q}(e^{i\theta_0})\} = \sum_{k=1}^{n}\prod_{j\neq k}\lambda_j$. Therefore, if $\theta_0$ is a simple zero, $\mathrm{tr}\{\mathrm{Adj}\ \hat{Q}(e^{i\theta_0})\} > 0$, and there is a $\varepsilon > 0$ such that the Lipschitz condition $\det \hat{Q} \leq |\theta - \theta_0|$ holds for $|\theta - \theta_0| < \varepsilon$. Therefore, in view of (III.4),

$$\langle \psi I, \hat{R}\rangle \geq \frac{1}{M}\int_{\theta-\varepsilon}^{\theta+\varepsilon} \frac{\psi^2}{|\theta-\theta_0|}\,\mathrm{tr}(\mathrm{Adj}\ \hat{Q})\frac{d\theta}{2\pi} = +\infty,$$

which is a contradiction. If $\theta_0$ is a multiple zero of order $p$, then $\det \hat{Q} \leq |\theta - \theta_0|^p$ and $p-1$ zeros can be used to cancel at most $p-1$ zeros in $\mathrm{tr}\{\mathrm{Adj}\ \hat{Q}(e^{i\theta_0})\}$, reducing the problem to the one already treated. Hence $\hat{Q} \in \mathcal{P}(m,n)$, establishing that $F_\psi$ is proper. ■

*Theorem 7:* Suppose that $R \in \mathcal{R}_+^m(E)$, and let $\psi \in \mathcal{Q}(1,n)$. Then the optimization problem (P$_E$) has a unique solution $\hat{\Phi}$, which satisfies the sparsity condition (I.3), and this solution is rational of the form

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

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

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

where the strictly convex functional $\mathbb{J}_\psi$ is given by (II.7).

*Proof:* By Lemma 4, $\Pi_E\mathcal{R}_+^m(E)$ 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 5) and proper (Lemma 6). Consequently, by Theorem 2.1 (or, in a simpler form, Corollary 2.3) in [14], $F_\psi$ is a homeomorphism. In particular, the dual optimization problem (III.8) has a unique solution. The rest follows from strong duality (Proposition 3). ■

## IV. ARMA IDENTIFICATION OF GRAPHICAL MODELS FROM STATISTICAL DATA

Given a string of measured data

$$x_0, x_1, \ldots, x_N \in \mathbb{R}^n \tag{IV.1}$$

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

$$\hat{R}_k = \frac{1}{N-1}\sum_{j=1}^{N-k} x_{k+j}x_j^*, \quad k =, 1, \ldots, n. \tag{IV.2}$$

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 $\Phi$, such as the (damped or smoothed) periodogram.

Our identification approach now proceeds in the following steps.

(i) Compile a list of the most likely sparsity patterns (graphical structures) $E$.

(ii) For each $E$, estimate the numerator pseudo-polynomial $\psi$.

(iii) Determine $Q$ by solving the convex optimization problem (D$_E$) with $R$ and $\psi$ given by (IV.2) and (ii) 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 (i).

(iv) Compare the estimates $\hat{\Phi}$ thus obtained by some information theoretic criteria as in [6] (and references therein) to choose a suitable graphical structure. In this selection process one may also consider the errors

$$[\Delta_j]_{k\ell} = \hat{R}_j - \int_{-\pi}^{\pi} e^{ij\theta}\psi(e^{i\theta})\left(Q(e^{i\theta})^{-1}\right)_{k\ell}\frac{d\theta}{2\pi},$$

For $j = 1, 2\ldots n$ and $(k,\ell) \notin E$. If any such $[\Delta_j]_{k\ell}$ is too large in absolute value, we may reject the corresponding solution.

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

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

### *Estimating the graphical structure E*

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 (II.9) as

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

where $\hat{\Phi}_{NP}$ 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 $\sigma$ depends only on the smoothing procedure used to determine $\hat{\Phi}_{NP}$. Moreover, as also shown in [5], we can select $M$ frequencies $\theta_1, \theta_2, \ldots, \theta_M \in [-\pi, \pi]$ 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 $p, q = 1, 2, \ldots, M$ such that $p \neq q$.

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

$$\begin{aligned} p(\gamma) := \prod_{j=1}^{M}\Big\{ & P\left\{|\mathrm{Re}[\hat{r}_{x_kx_\ell|s}(e^{i\theta_j})]| \leq \gamma\right\}\cdot \\ & \cdot P\left\{|\mathrm{Im}[\hat{r}_{x_kx_\ell|s}(e^{i\theta_j})]| \leq \gamma\right\}\Big\} = \\ = & [G(\gamma) - G(-\gamma)]^{2M} \end{aligned}$$

where $G$ is the c.d.f of a gaussian variable with mean zero and variance $\sigma$. Now let $\gamma(\alpha)$ be such that $p(\gamma(\alpha)) = 1-\alpha$. Then we reject the null hypotesis (H$_0$) at the significance level $\alpha$ if any of the random variables $\mathrm{Re}\{\hat{r}_{x_kx_\ell|s}(e^{i\theta_j})\}$,

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

Suppose now that we vary $\alpha$ 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 $\alpha$. 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 $\Phi^{-1}$, and hence $Q$, to be diagonal, whereas $E_{m(m-1)/2} = E(1)$ allows for no 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, e.g as considered in [6], that grows exponentially in $m$.

### *Estimating the pseudo-polynomial $\psi$*

Given a graphical structure $E$, consider a matrix version of the procedure in [15, page 689], which amounts to solving

$$\begin{bmatrix} \min\limits_{\substack{Q \in \mathcal{Q}(m,n) \\ \psi \in \mathcal{Q}(1,n)}} \max\limits_{j} \left\| Q(e^{i\theta_j}) - \psi(e^{i\theta_j}) \hat{\Phi}_{NP}(e^{i\theta_j})^{-1} \right\|_2 \\ \text{subject to} \quad Q_{k\ell} \equiv 0 \quad (k, \ell) \notin E \end{bmatrix} \tag{IV.4}$$

where $\theta_1, \theta_2, \dots, \theta_M \in [-\pi, \pi]$ are suitable frequencies, possibly, but not necessarily, the same as the ones above. It is not hard to see that (IV.4) is equvalent to the following semi-definite programing problem

$$\begin{bmatrix} \min\limits_{(\psi, Q, \epsilon) \in \mathcal{Q}(1,n) \times \mathcal{Q}(m,n) \times [0,\infty)} \quad \epsilon \\ \text{s. t.} \begin{cases} -\epsilon I \le Q(e^{i\theta_j}) - \psi(e^{i\theta_j}) \hat{\Phi}_{NP}(e^{i\theta_j})^{-1} \le \epsilon I, \\ \qquad\qquad\qquad j = 1, 2, \dots, M \\ Q_{k\ell} \equiv 0 \quad (k, \ell) \notin E \end{cases} \end{bmatrix}. \tag{IV.5}$$

In order to insure that $\psi$ and $Q$ are positive definite, one may need to add the constraints $\psi(e^{i\theta_j}) \ge \delta$ and $Q(e^{i\theta_j}) \ge \delta I$, $j = 1, 2, \dots, M$, for some $\delta > 0$. In the solution we are of course only interested in $\psi$, as a more accurate $Q$ will be determined in step (iii). However, the $Q$ obtained here may be used as a starting point in an algorithm solving (D$_E$) such as Newton's method.

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

$$\begin{bmatrix} \min\limits_{\substack{Q \in \mathcal{Q}(m,n) \\ \psi \in \mathcal{Q}(1,n)}} \max\limits_{j} \left\| \psi(e^{i\theta_j}) Q(e^{i\theta_j})^{-1} - \hat{\Phi}_{NP}(e^{i\theta_j}) \right\|_2 \\ \text{subject to} \quad Q_{k\ell} \equiv 0 \quad (k, \ell) \notin E \end{bmatrix} \tag{IV.6}$$

Proceeding in the same way as the step from (IV.4) to (IV.5) we obtain the constraints

$$-\epsilon I \le \psi(e^{i\theta_j}) Q(e^{i\theta_j})^{-1} - \hat{\Phi}_{NP}(e^{i\theta_j}) \le \epsilon I, \tag{IV.7}$$

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

$$-\epsilon\, Q(e^{i\theta_j}) \le \psi(e^{i\theta_j}) I - \hat{\Phi}_{NP}(e^{i\theta_j}) Q(e^{i\theta_j}) \le \epsilon\, Q(e^{i\theta_j}), \tag{IV.8}$$

which is linear only if we disallow $\epsilon$ from being a variable. Therefore, this problem needs to be solved in steps. First fix $\epsilon$ and solve the feasibility problem to find $\psi \in \mathcal{Q}(1, n)$ and $Q \in \mathcal{P}(m, n)$ satisfying (IV.8), after which $\epsilon$ is decreased in steps (e.g., by the bisection algorithm) until we obtain the smallest $\epsilon$ for which feasibility problem is solvable. It remains to determine bounds for this method. This could be done along the lines proposed in [16], [17], [18].

## V. Conclusions

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 [9], [10], [11], [12], [13]. 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 biggest such set that guarantees the desired graphical structure. 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.

## References

[1] J. Whittaker, *Graphical models in applied multivariate statistics*, Wiley New York, 1990.

[2] S. L. Lauritzen, *Graphical models*, Oxford University 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, pp157-172, 2000.

[5] R. Dahlhaus, M. Eichler and J. Sandkühler, Identification of synaptic connections in neural ensembles by graphical models, *J. Neuroscience 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 University Press, 2010.

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

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

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

[10] C. I. Byrnes, T. Georgiou, and A. Lindquist, A new approach to spectral estimation: A tunable high-resolution spectral estimator, *IEEE Trans. Sig. Proc.* vol. 49, pp. 3189-3205, 2000.

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

[12] A. Blomqvist, A. Lindquist and R. Nagamune, Matrix-valued Nevanlinna-Pick interpolation with complexity constraint: An optimization approach, *IEEE Transactions on Automatic Control* 48 (Dec. 2003), 2172–2190.

[13] C. I. Byrnes and A. Lindquist, Important moments in systems and control, *SIAM J. Control and Optimization* 47(5) (2008), 2458–2469.

[14] C. I. Byrnes and A. Lindquist, Interior point solutions of variational problems and global inverse function theorems, *International Journal of Robust and Nonlinear Control* 17 (2007), 463–481.

[15] C.I. Byrnes, P. Enqvist and A. Lindquist, Cepstral coefficients, covariance lags and pole-zero models for finite data strings, *IEEE Trans. Signal Processing* **SP-50** (April 2001), 677–693.

[16] J. Kalsson and A. Lindquist, Stability-preserving rational approximation subject to interpolation constraints, IEEE Trans. Automatic Control 53(7) (2008), 1724–1730.

[17] J. Kalsson, T.T. Georgiou and A. Lindquist, The inverse problem of analytic interpolation with degree constraint and weight selection for control synthesis, *IEEE Trans. Automatic Control* AC-55 (2010), 405–418.

[18] G. Fanizza, *Modeling and Model Reduction by Analytic Interpolation and Optimization*, Ph.D., Royal Institute of Technology, 2008.