ARMA Identification of Graphical Models

Joint work with Anders Lindquist and 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 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.

C.1   Introduction

Graphical models represent families of probability distributions in the form of graphs which expose relative conditional independences. In this paper we 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. 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 [13]. As was shown in [5],

this is manifested by a sparsity pattern in the inverse of the $m \times m$ matrix-valued spectral density

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

where

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

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 C.3, it can be shown that

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

for pairs (k, ℓ) such that x_k and x_j are conditionally independent give the rest of the components of the process x; also see [5, 13]. Such graphical models find applications in speech, bioinformatics, image processing, econometrics and many other fields; see [20] and references therein.

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

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

for an arbitrary index set $I \subset V$, (C.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

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

A typical such graph is depicted in Figure 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 C.2 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 [13, 20] 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
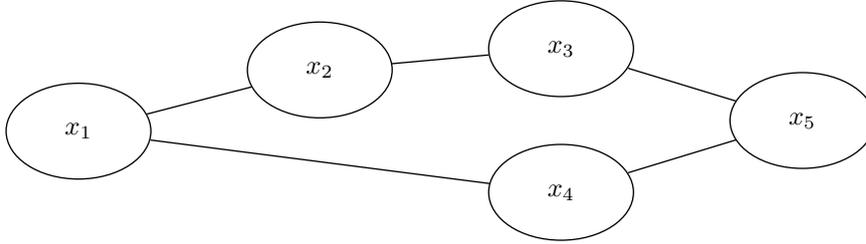
Figure C.1:   Example of an interaction graph.

precisely, given the (estimates of) autocovariances $R_0, R_1, \ldots, R_n$, the problem in these papers is to find a multivariate autoregressive model

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

that satisfies the sparsity constraint (C.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

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

have no zeros in the closed unit disc and $\det A_0 > 0$.

However, there are examples where AR models are insufficient. 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 (C.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., [17, 12]. 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

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

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

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

where the scalar polynomial

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

has no zeros in the 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 $\Phi(z) = W(z)W(z^{-1})^\top$, where $W(z) := b(z)A(z)^{-1}$.

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

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

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

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

where ψ is a scalar pseudo-polynomial 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 (C.6) can be obtained by determining the minimum-phase spectral factors $A(z)$ and $b(z)$ from

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

respectively. To deal with this problem we shall use the convex optimization approach to moment problems developed in various forms in [10, 11, 9, 18, 4], which we shall review in Section C.3 together with the basic ideas on graphical models. At the same time we obtain an alternative motivation for the optimization procedure in [13, 20]. In Section C.4 we incorporate the sparsity constraint (C.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

$$x_0, x_1, \ldots, x_N \in \mathbb{R}^n \tag{C.12}$$

from the ARMA model (C.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 C.5. Finally, in Section C.6, we present some simulations and in Section C.7 our conclusions.

This is a considerable extension of our previous conference paper [3].

C.2   Motivating examples

We begin by presenting a selection of practical problems where the application of graphical models shows promise. For each of these applications, the interaction

---

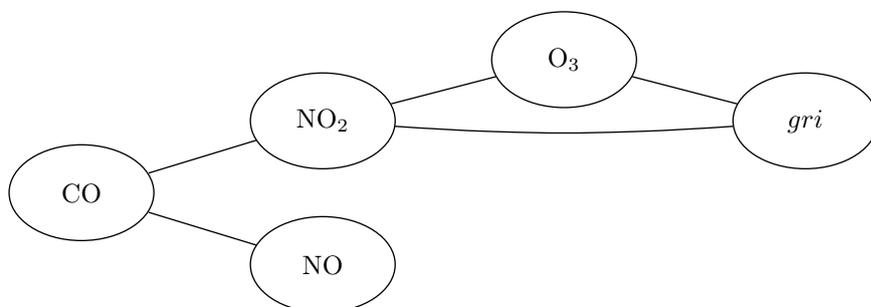[1]A polynomial in positive and negative powers of z.

Figure C.2:   The estimated graphical model for air pollution data.

graphs, taken from the literature, have been determined by nonparametric meth-
ods. 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.

C.2.1   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 condi-
tionally independent, it may be assumed that no significant direct reaction takes
place between them. Hence, by estimating the interaction graph of a set of com-
pounds one can gain insight about the reactions that govern the system dynamics.

Dahlhaus in [13] 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 byprod-
ucts 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 figure C.2. It has been noted that part of the estimated graph can be explained
by, and thus validates, a photodynamical theory governing those reactions; cf. [24].

C.2.2   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 – see e.g [22]. In order to be able to accurately
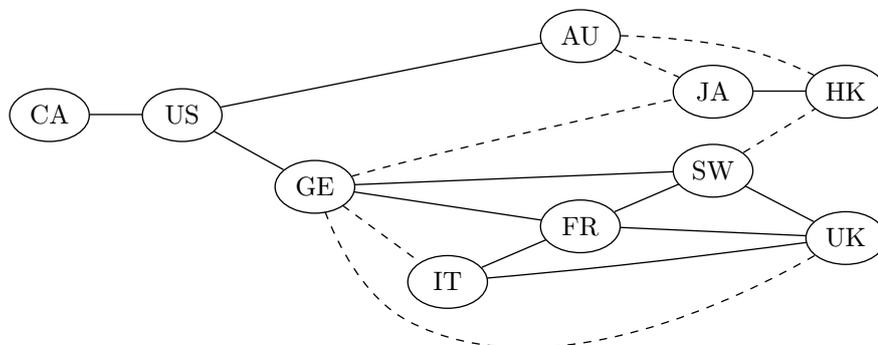
Figure C.3: The interaction graph for the international stock returns. The solid and dashed edges indicates strong and weak partial correlation respectively.

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 [1] 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

$$x(t) = 100 \left[\ln p(t) - \ln p(t-1) \right] \tag{C.13}$$

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 Figure C.2.2 was estimated by applying the nonparametric identification procedure proposed in [13].

The authors of [1] 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.2.3 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 [16] 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 different patients as shown in figure C.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 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.

C.3 Preliminaries

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

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

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

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

where tr denotes the trace.
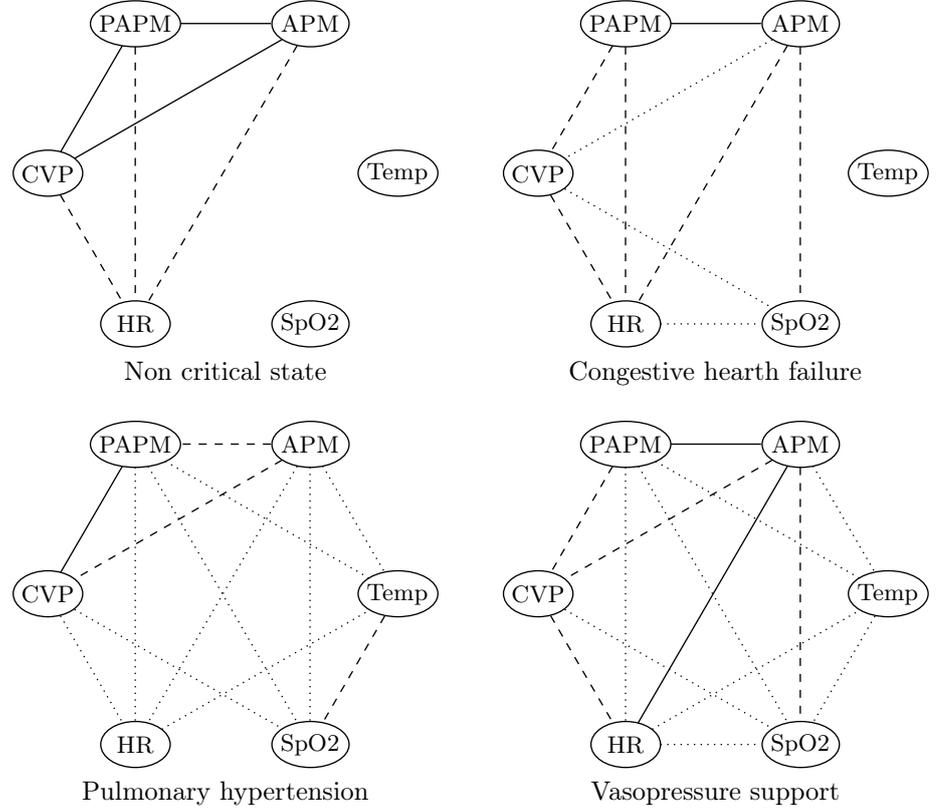
Figure C.4: Partial correlation graphs corresponding to different clinical states. Solid, dashed and dotted lines stands for different levels of partial correlation from high to low.

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

$$R(z) = \Re \left\{ \sum_{j=0}^{n} z^j R_j \right\}. \tag{C.14}$$

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

$$\mathcal{Q}(m, n) = \left\{ Q(z) = \Re \left\{ \sum_{j=0}^{n} z^j Q_j \right\} \ : \ Q_j \in \mathbb{R}^{m \times m}, \right.$$
$$\left. Q(e^{i\theta}) > 0, \ \forall \theta \in [-\pi, \pi] \right\}, \tag{C.15}$$

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

$$\langle R, Q \rangle = \sum_{j=0}^{n} \operatorname{tr}(R_j Q_j), \tag{C.16}$$

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

$$\langle R, Q \rangle = \operatorname{tr}\{\mathbf{A}^\top T(R)\mathbf{A}\}, \tag{C.17}$$

where $\mathbf{A} = \begin{bmatrix} A_0 \\ \vdots \\ A_n \end{bmatrix}$ and

$$T(R) = \begin{bmatrix} R_0 & R_1 & \cdots & R_n \\ (R_1)^\top & R_0 & \cdots & R_{n-1} \\ \vdots & \vdots & \ddots & \vdots \\ (R_n)^\top & (R_{n-1})^\top & \cdots & R_0 \end{bmatrix} \tag{C.18}$$

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

Proposition C.3.1 *Given the autocovariances $R_0, R_1, \ldots, R_n$, there exists a $\Phi \in \mathcal{S}_+^m$ satisfying the moment equations (C.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 (C.17) 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 (C.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 C.3.1. $\qquad\square$

C.3.1   A convex-optimization solution of the moment problem

We begin by reviewing the convex optimization approach to moment problems developed in [10, 11, 9, 18, 4]. The following result can be found in [4].

Theorem C.3.1 *Suppose that $T(R) > 0$, and let $\psi \in \mathcal{Q}(1, n)$. Then the optimization problem*

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

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

$$\hat{\Phi}(z) = \psi(z)\hat{Q}(z)^{-1}. \tag{C.19}$$

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

$$\mathbb{J}_\psi(Q) := \langle R, Q \rangle - \langle \psi I, I + \log Q \rangle \tag{C.20}$$

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,

$$\begin{aligned}
\Delta &= \langle R, \hat{Q} \rangle - \langle \psi I, I + \log \hat{Q} \rangle + \langle \psi I, \log(\psi \hat{\Phi}^{-1}) \rangle \\
&= \langle R, \hat{Q} \rangle - \langle \psi I, I + \log \hat{Q} \rangle + \langle \psi I, \log \hat{Q} \rangle \\
&= \langle \hat{\Phi}, \hat{Q} \rangle - \langle \psi I, I \rangle = 0.
\end{aligned} \tag{C.21}$$

Theorem C.3.1 provides a complete parameterization in terms of $\psi \in \mathcal{Q}(1,n)$ of all models (C.6) of the form (C.7) such that (C.9) satisfies the moment conditions (C.10). In particular, choosing $\psi \equiv 1$ we obtain the *maximum entropy solution* which corresponds to the AR model (C.4) and the solution of which is linear problem where $\mathbf{A}$ can be obtained from the normal equations.

C.3.2   Graphical models of stochastic processes

The cross-spectrum

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

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

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

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

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

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

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

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

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

[5], 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

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

Clearly, if x is Gaussian, then so is ε, and hence x_k and x_ℓ are conditional independent if and only if $\Phi_{x_k x_\ell | s}(e^{i\theta}) = 0$ for all θ. The conditional coherence of x_k with x_ℓ 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 [13], satisfies

$$r_{x_k x_\ell | s}(e^{i\theta}) = \frac{[\Phi^{-1}(e^{i\theta})]_{k,\ell}}{\sqrt{[\Phi^{-1}(e^{i\theta})]_{k,k}[\Phi^{-1}(e^{i\theta})]_{\ell,\ell}}} \tag{C.22}$$

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

$$[\Phi^{-1}(e^{i\theta})]_{k,\ell} = 0 \quad \forall \theta \in [-\pi, \pi] \tag{C.23}$$

is a necessary and sufficient condition for x_k and x_ℓ 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 (C.23) holds for all $(k, \ell) \notin E$.

C.4 Covariance extension for graphical models

We now turn to the basic problem of this paper, namely to find a model (C.6) that satisfies (C.2) and the sparsity condition (C.23). Now, by Theorem C.3.1, all such solutions must have a spectral density of the form (C.9), and therefore the sparsity condition (C.23) can be reformulated as

$$Q_{k\ell} \equiv 0 \quad \text{for all } (k, \ell) \notin E. \tag{C.24}$$

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 (C.24). This gives us the convex optimization problem

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

This optimization problem was used in the special maximum-entropy case $\psi \equiv 1$ in [20] to derive an AR model, but no theoretical justification was provided. In our setting, 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 (C.24) affects the original problem. We need to formulate a problem for which (D$_E$) is the dual.

Proposition C.4.1 *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*

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

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

Proof. The Lagrangian of (P_E) is given by

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

$$+ \sum_{(k,\ell) \in E} \sum_{j=0}^{n} [Q_j]_{k\ell}([R_j]_{k\ell} - \langle \Phi_{k\ell}, z^j \rangle) =$$

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

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 C.4.1, 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 (D_E).

To prove the last statement note that that (P_E) is a relaxation of (P) with $\bar{R}$ 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 (C.14), 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 Π_E is a projection, and, since the diagonal elements of X are unaffected by Π_E, we have $\text{tr}(\Pi_E X) = \text{tr}(X)$.

Lemma C.4.1 $\Pi_E \mathcal{R}_+^m(E)$ 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

$$F_\psi(Q) = \Re \left\{ \sum_{k=0}^{n} e^{ik\theta} \int_{-\pi}^{\pi} e^{ik\omega} \psi(e^{i\omega}) \Pi_E Q(e^{i\omega})^{-1} \frac{d\omega}{2\pi} \right\}. \tag{C.25}$$

Lemma C.4.2 *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 (C.20), is strictly convex (Theorem C.3.1), 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 (C.16), $\mathbb{J}_\psi$ can be written

$$\mathbb{J}_\psi(Q) = \sum_{k=0}^{n} \operatorname{tr}(R_k Q_k) - \int_{-\pi}^{\pi} \psi \operatorname{tr}(\log Q)\frac{d\theta}{2\pi}, \qquad (C.26)$$

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 (C.10), which after projection yields $F_\psi(Q) = \Pi_E R$. $\qquad\square$

Lemma C.4.3 *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 (C.16), the fact that $\operatorname{tr}(\Pi_E Q^{-1}) = \operatorname{tr}(Q^{-1})$, and Cramer's rule,

$$\begin{aligned}
\langle \psi I, F_\psi(Q)\rangle &= \langle \psi^2 I, Q^{-1}\rangle \\
&= \int_{-\pi}^{\pi} \frac{\psi^2}{\det Q} \operatorname{tr}(\operatorname{Adj} Q)\frac{d\theta}{2\pi},
\end{aligned} \qquad (C.27)$$

where $\operatorname{Adj} A$ denotes the adjugate matrix of A. We now proceed as in the proof of Lemma 6.3 in [6] 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

$$\begin{aligned}
\lim_{k\to\infty} M_k \langle \psi^2 I, (\tilde{Q}^{(k)})^{-1}\rangle &= \lim_{k\to\infty} \langle \psi I, F_\psi(Q^{(k)})\rangle \\
&= \lim_{k\to\infty} \langle \psi I, R^{(k)}\rangle \\
&= \langle \psi I, \hat{R}\rangle.
\end{aligned} \qquad (C.28)$$

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 C.3.1,

$$
\begin{aligned}
\langle \psi I, \hat{R} \rangle &= \int_{-\pi}^{\pi} \psi \operatorname{tr}(\hat{R}) \frac{d\theta}{2\pi} \\
&= \int_{-\pi}^{\pi} \psi \operatorname{tr}(\bar{R}) \frac{d\theta}{2\pi} = \langle \psi I, \bar{R} \rangle > 0.
\end{aligned}
\tag{C.29}
$$

Since the sequence $\langle \psi^2 I, (\tilde{Q}^{(k)})^{-1} \rangle$ in (C.28) 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 θ_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 $\operatorname{tr}\{ \operatorname{Adj} \hat{Q}(e^{i\theta_0}) \} = \sum_{k=1}^{n} \prod_{j \neq k} \lambda_j$. Therefore, if θ_0 is a simple zero, $\operatorname{tr}\{ \operatorname{Adj} \hat{Q}(e^{i\theta_0}) \} > 0$, 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 (C.27),

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

which is a contradiction. If θ_0 is a multiple zero of order p, then $\det \hat{Q} \leq |\theta - \theta_0|^p$ and $p-1$ zeros can be used to cancel at most $p-1$ zeros in $\operatorname{tr}\{ \operatorname{Adj} \hat{Q}(e^{i\theta_0}) \}$, reducing the problem to the one already treated. Hence $\hat{Q} \in \mathcal{P}(m,n)$, establishing that F_ψ is proper. $\qquad \square$

Theorem C.4.1 *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 (C.3), and this solution is rational of the form*

$$
\hat{\Phi}(z) = \psi(z) \hat{Q}(z)^{-1}.
\tag{C.30}
$$

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

$$
\min_{Q \in \mathcal{P}(m,n)} \mathbb{J}_\psi(Q),
\tag{C.31}
$$

where the strictly convex functional $\mathbb{J}_\psi$ is given by (C.20).

Proof. By Lemma C.4.1, $\Pi_E \mathcal{R}^m_+(E)$ 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 C.4.2) and proper (Lemma C.4.3). Consequently, by Theorem 2.1 (or, in a simpler form, Corollary 2.3) in [7], F_ψ is a homeomorphism. In particular, the dual optimization problem (C.31) has a unique solution. The rest follows from strong duality (Proposition C.4.1). $\square$

In the special case when $\psi \equiv 1$ the primal problem (P$_E$) reduces to maximizing the entropy gain

$$\int_{-\pi}^{\pi} \log \det \Phi(e^{i\theta}) \frac{d\theta}{2\pi} \tag{C.32}$$

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

$$\int_{-\pi}^{\pi} e^{ij\theta} \Phi_{k\ell}(e^{i\theta}) \frac{d\theta}{2\pi} = [R_j]_{k\ell}, \quad \forall (k,\ell) \in E, \quad j = 1, 2 \ldots n. \tag{C.33}$$

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

$$\mathbb{J}_1(Q) = \langle R, Q \rangle - \int_{-\pi}^{\pi} \log \det Q(e^{i\theta}) \frac{d\theta}{2\pi} \tag{C.34}$$

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

It turns out that (C.34) is precisely equivalent to the problem considered in [20], as the following corollary states.

Corollary 1 *Given the covariance sequence $R_0, R_1, \ldots, R_n$, let $T(R)$ be its block Toeplitz matrix (C.18). Then the maximum entropy solution corresponding to the graph E is*

$$\hat{\Phi}(z) = \hat{Q}(z)^{-1}, \tag{C.35}$$

where

$$\hat{Q}(z) = \Re \left\{ \sum_{j=0}^{n} z^j \hat{Q}_j \right\} \tag{C.36}$$

is the unique optimal solution of the problem to minimize

$$J(X) = \text{tr}\{T(R)X\} - \log \det X_{00} \tag{C.37}$$

over all symmetric, positive semi-definite $(n+1) \times (n+1)$ matrices such that

$$Q_j = \sum_{\nu=0}^{n-j} X_{\nu,\nu+j}, \quad j = 0, 1, \ldots, n \tag{C.38a}$$

$$[Q_j]_{k\ell} = 0, \quad \forall (k,\ell) \notin E, \quad j = 1, 2 \ldots n \tag{C.38b}$$

Proof. We want to rewrite (C.34) in the form (C.37). To this end, we first observe that (C.17) yields

$$\langle R, Q \rangle = \text{tr}\{T(R)X\},$$

where

$$X = AA^\top,$$

from which (C.38a) readily follows. Next, from (C.11) we have

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

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 $\mathbb{D}$. Therefore, by Jensen's formula [2, 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

$$2 \log |\det A_*(0)| = \log \det(A_0 A_0^\top) = \log(\det X_{00}),$$

and consequently (C.34) is equivalent to (C.37), as claimed. Finally, (C.38b) is equivalent to $Q_{k\ell} \equiv 0$. $\qquad\square$

In a more recent paper [25], Songsiri and Vandenberghe add an ℓ_1 regularization term in X to the cost function (C.37), thus replacing the sparsity constraint (C.38). 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 (D_E) in our ARMA setting by the problem to minimize

$$\mathbb{J}_\psi(Q) + \lambda \|Q\|_1 \tag{C.39}$$

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

C.5   ARMA identification of graphical models from statistical data

Given a string of measured data

$$x_0, x_1, \ldots, x_N \in \mathbb{R}^n \tag{C.40}$$

from the ARMA model (C.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

$$\hat{R}_k = \frac{1}{N-1} \sum_{j=1}^{N-k} x_{k+j} x_j^*, \quad k =, 1, \ldots, n. \tag{C.41}$$

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.

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

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

(iii) Determine Q by solving the convex optimization problem (D_E) with R and ψ given by (C.41) 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 fitness function.

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

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

C.5.1   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 (C.22) as

$$\hat{r}_{x_k x_\ell | s}(e^{i\theta}) = \frac{[\hat{\Phi}_{NP}^{-1}(e^{i\theta})]_{k,\ell}}{\sqrt{[\hat{\Phi}_{NP}^{-1}(e^{i\theta})]_{k,k}[\hat{\Phi}_{NP}^{-1}(e^{i\theta})]_{\ell,\ell}}}, \tag{C.42}$$

where $\hat{\Phi}_{NP}$ is the (smoothed) nonparametric spectral estimate introduced above. It can be shown [23] that the real and imaginary parts of $\hat{r}_{x_k x_\ell | s}(e^{i\theta}) - r_{x_k x_\ell | s}(e^{i\theta})$ 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 [23], we can select M frequencies $\theta_1, \theta_2, \ldots, \theta_M \in [-\pi, \pi]$ so that $\hat{r}_{x_k x_\ell | s}(e^{i\theta_p}) \perp \hat{r}_{x_k x_\ell | s}(e^{i\theta_q})$ for all $p, q = 1, 2, \ldots, M$ such that $p \neq q$.

Under the the null hypothesis (H_0) the real and imaginary parts of $\hat{r}_{x_k x_\ell | s}(e^{i\theta_j})$, $j = 1, 2, \ldots, M$, are asymptotically independent and normally distributed with mean zero and variance σ. Hence, asymptotically, the probability that the absolute values of these random variables are all less or equal to γ is

$$p(\gamma) := \prod_{j=1}^{M} \left\{ P\left\{ |\mathrm{Re}[\hat{r}_{x_k x_\ell | s}(e^{i\theta_j})]| \leq \gamma \right\} \cdot \right.$$

$$\left. \cdot P\left\{ |\mathrm{Im}[\hat{r}_{x_k x_\ell | s}(e^{i\theta_j})]| \leq \gamma \right\} \right\} =$$

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

where G is the cumulative distribution function of a Gaussian variable with mean zero and variance σ. Now, following Dahlhaus in [13], let $\gamma(\alpha)$ be such that $p(\gamma(\alpha)) = 1 - \alpha$. Then we reject the null hypothesis (H$_0$) at the significance level α if any of the random variables $\mathrm{Re}\{\hat{r}_{x_k x_\ell | s}(e^{i\theta_j})\}$, $\mathrm{Im}\{\hat{r}_{x_k x_\ell | s}(e^{i\theta_j})\}$, $j = 1, 2, \ldots, M$, 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 $E_{m(m-1)/2} = E(1)$ 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, as, for example, considered in [20], which grows exponentially in m.

To illustrate the proposed method, let us consider an example. Figure C.5 depicts the partial coherences of a spectrum of the form (C.9) corresponding to a graph as in Figure C.1. The corresponding estimates derived as in (C.42) are reproduced with a dashed line. Now, for each subplot, we need to consider the infinity norm of the estimated partial coherence and order edges accordingly. From figure C.5 we obtain

$$(x_2, x_5), (x_1, x_3), (x_3, x_4), (x_1, x_5), (x_2, x_4), (x_3, x_5), (x_1, x_2), (x_4, x_5), (x_2, x_3), (x_1, x_4)$$
(C.43)

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 (C.43) 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 Figure C.6. Note that, with the given estimates, all edges of the graph in Figure C.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.

C.5.2 Estimating the pseudo-polynomial ψ

Given a graphical structure E, consider a matrix version of the procedure in [8, page 689], which amounts to solving

$$\left[\begin{array}{c} \min\limits_{\substack{Q \in \mathcal{Q}(m,n) \\ \psi \in \mathcal{Q}(1,n)}} \max\limits_{j} \left\| Q(e^{i\theta_j}) - \psi(e^{i\theta_j}) \hat{\Phi}_{NP}(e^{i\theta_j})^{-1} \right\|_2 \\ \text{subject to } \ Q_{k\ell} \equiv 0 \quad (k, \ell) \notin E \end{array} \right]$$
(C.44)

where $\theta_1, \theta_2, \ldots, \theta_M \in [-\pi, \pi]$ are suitable frequencies, possibly, but not necessarily, the same as the ones above. It is not hard to see that (C.44) is equivalent to the

Figure C.5:   Coherence (above the diagonal) and partial coherence (below the diagonal) of a stochastic process with an interaction graph as in Figure C.1. The estimated counterparts computed via a smoothed periodogram from $N = 2^{10}$ data points are plotted in dashed lines.

following semi-definite programing problem

$$
\left[
\begin{array}{l}
\min_{(\psi,Q,\epsilon)\in\mathcal{Q}(1,n)\times\mathcal{Q}(m,n)\times[0,\infty)} \epsilon \\[6pt]
\text{subject to} \begin{cases} -\epsilon I \le Q(e^{i\theta_j}) - \psi(e^{i\theta_j})\hat{\Phi}_{NP}(e^{i\theta_j})^{-1} \le \epsilon I, \\ \qquad\qquad\qquad\qquad\qquad\qquad j = 1,2,\ldots,M \\ Q_{k\ell} \equiv 0 \quad (k,\ell) \notin E \end{cases}
\end{array}
\right]. \qquad (\text{C.45})
$$

In order to insure that ψ and Q are positive definite, one may need to add the constraints $\psi(e^{i\theta_j}) \ge \delta$ and $Q(e^{i\theta_j}) \ge \delta I$, $j = 1,2,\ldots,M$, for some $\delta > 0$. In the solution we are of course only interested in ψ, 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 ψ could be to solve the quasi-convex optimization problem

$$
\left[
\begin{array}{l}
\min_{\substack{Q\in\mathcal{Q}(m,n) \\ \psi\in\mathcal{Q}(1,n)}} \max_j \left\| \psi(e^{i\theta_j})Q(e^{i\theta_j})^{-1} - \hat{\Phi}_{NP}(e^{i\theta_j}) \right\|_2 \\[10pt]
\qquad \text{subject to } Q_{k\ell} \equiv 0 \quad (k,\ell) \notin E
\end{array}
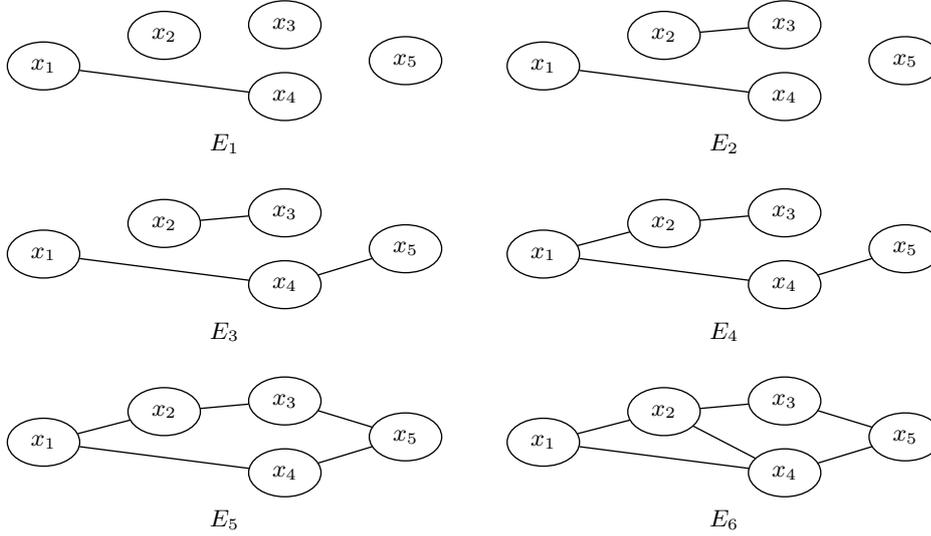\right] \qquad (\text{C.46})
$$

Figure C.6:   Some of the graph candidates generated by the estimated partial coherence of Figure C.5.

Proceeding in the same way as the step from (C.44) to (C.45) we obtain the constraints

$$- \epsilon I \leq \psi(e^{i\theta_j})Q(e^{i\theta_j})^{-1} - \hat{\Phi}_{NP}(e^{i\theta_j}) \leq \epsilon I, \qquad (C.47)$$

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

$$- \epsilon\, Q(e^{i\theta_j}) \leq \psi(e^{i\theta_j})I - \hat{\Phi}_{NP}(e^{i\theta_j})Q(e^{i\theta_j}) \leq \epsilon\, Q(e^{i\theta_j}), \qquad (C.48)$$

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 (C.48), 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. It remains to determine bounds for this method. This could be done along the lines proposed in [21, 19, 14].

C.5.3   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 [20] 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 (ii) and (iii) above with respect to the edge list E. Since, in solving (D_E), only a subset of the estimated covariance lags are

matched, one can utilize the remaining ones as data for validation. Specifically, we consider

$$\Delta(E) = \max_{j,k,\ell} \left| [\hat{R}_j]_{k,\ell} - \int_{-\pi}^{\pi} e^{ij\theta} \psi_E(e^{i\theta}) \left[Q_E(e^{i\theta})^{-1} \right]_{k\ell} \frac{d\theta}{2\pi} \right|, \qquad \text{(C.49)}$$

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

Unfortunately (C.49) 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

$$F(E) = [(1 + \epsilon)\Delta(E_0) - \Delta(E)] \left[\frac{m(m-1)}{2} - C(E) \right] \qquad \text{(C.50)}$$

to be maximized, where ϵ is some small positive constant. The advantages of such a choice is mainly its robustness to scaling.

Remark C.5.1 *By replacing the dual problem* (D_E) *by its* ℓ_1-*regularized counterpart* (C.39), *we could potentially combine steps (i), (iii) and (iv). 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 is not likely to be optimal for all data records. Hence, as proposed in [25] 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.

C.6   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 underling graphical structure correctly. To this end, we fed the filter corresponding to Figure C.5 with white noise and recorded various data strings of different length N, from $N = 256$ to $N = 4096$.

For each N, we repeated the procedure described in section C.5 twenty times with different data strings of the same length. Finally the resulting estimated graphs were compared to the one in Figure C.1. Here we use

$$d(E, \bar{E}) = \max\{C(E), C(\bar{E})\} - C(E \cap \bar{E}) \qquad \text{(C.51)}$$
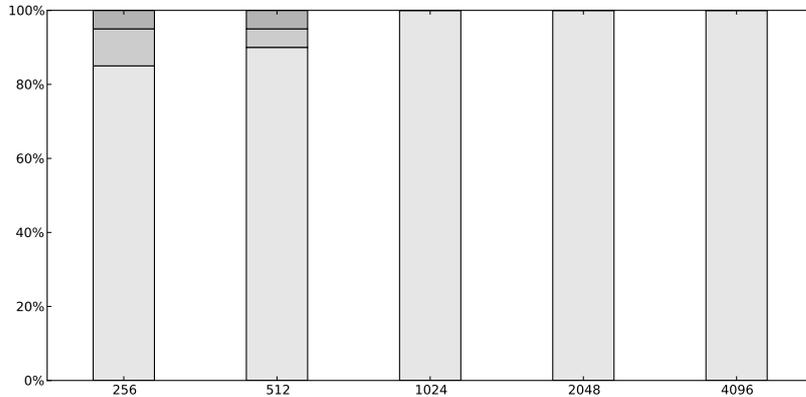
Figure C.7:   Simulation results for various for various data length.

as a measure of distance between interaction graphs corresponding to the edge sets E and $\bar{E}$ respectively.

The outcomes are reported in Figure C.7. Here the different shades of grey represent the portion of simulations yielding the same distance between the estimated and actual graph. The light gray corresponds to simulations that recovered the original graph perfectly, while the middle and dark grays indicate that the estimated graph differed from the original by one or two, respectively, by the measure (C.51). It appears that the proposed method is quite effective at recovering the graphical structure in figure C.1 from data, having done so correctly 85% of the times with as little as 256 data measurements and having been always correct with $N = 1024$ or more.

C.7   Conclusions

In this paper we have extended the results in [13, 20] to graphical models of ARMA process. This has been done by posing the problem in the moment-problem framework of [11, 9, 18, 4, 6]. 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. 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 C.2, one may consider using a THREE-type approach [9], for which there now are efficient algorithms in the multivariate case [15].

C.8   Bibliography

[1] A. Abdelwahab, O. Amor, and T. Abdelwahed. The analysis of the inter-dependence structure in international financial markets by graphical models. *International Research Journal of Finance and Economics*, 15:291–306, 2008.

[2] L.V. Ahlfors. *Complex Analysis*. McGraw-Hill, 1953.

[3] E. Avventi, A. Lindquist, and A. Wahlberg. Graphical models of autoregressive moving-average processes. In *Proc.19th International Symposium on the Mathematical Theory of Networks and Systems. (MTNS 2010) Budapest, Hungary, July 5–9, 2010*.

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

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

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

[7] Christopher I. Byrnes and Anders Lindquist. Interior point solutions of variational problems and global inverse function theorems. *International Journal of Robust and Nonlinear Control*, 17:463–481, 2007.

[8] C.I. Byrnes, P. Enqvist, and A. Lindquist. Cepstral coefficients, covariance lags and pole-zero models for finite data strings. *IEEE Trans. Signal Processing*, April 2001.

[9] C.I. Byrnes, T.T. Georgiou, and A. Lindquist. A new approach to spectral estimation: A tunable high-resolution spectral estimator. *IEEE Trans. on Signal Processing*, 48(11):3189–3205, Nov 2000.

[10] C.I. Byrnes, S.V. Gusev, and A. Lindquist. A convex optimization approach to the rational covariance extension problem. *SIAM Journal on Control and Optimization*, 37(1):211–229, 1999.

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

[12] 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. Automatic Control*, 40(11):1841–1857, 1995.

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

[14] G. Fanizza. *Modeling and Model Reduction by Analytic Interpolation and Optimization*. PhD thesis, Royal Institute of Technology, 2008.

[15] A. Ferrante, C. Masiero, and M. Pavon. Time and spectral domain relative entropy: A new approach to multivariate spectral estimation. *to appear*.

[16] Ursula Gather, Michael Imhoff, and Roland Fried. Graphical models for multivariate time series from intensive care monitoring. *Statistics in Medicine*, 21(18):2685–2701, 2002.

[17] T.T. Georgiou. Realization of power spectra from partial covariance sequences. *IEEE Trans. Acoustics, Speech and Signal Processing*, ASSP-35:438–449, 1987.

[18] T.T. Georgiou and A. Lindquist. Kullback-Leibler approximation of spectral density functions. *IEEE Transactions on Information Theory*, 49:2910–2917, Nov 2003.

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

[20] J. Dahl J. Songsiri and L. Vandenberghe. *Graphical models of autoregressive processes*. Cambridge University Press, 2010.

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

[22] R.W. Kolb. *Financial Contagion: The Viral Threat to the Wealth of Nations*. Robert W. Kolb Series in Finance. John Wiley & Sons, 2011.

[23] M. Eichler R. Dahlhaus and J. Sandkühler. Identification of synaptic connections in neural ensembles by graphical models. *J. Neuroscience Methods*, 77:93–107, 1997.

[24] J.H. Seinfeld. *Atmospheric chemistry and physics of air pollution*. A Wiley-Interscience publication. Wiley, 1986.

[25] Jitkomut Songsiri and Lieven Vandenberghe. Topology selection in graphical models of autoregressive processes. *J. Mach. Learn. Res.*, 11:2671–2705, December 2010.