On the Identification of ARMA Graphical Models

Mattia Zorzi, *Senior Member, IEEE*

***Abstract*—This article considers the problem to estimate a graphical model corresponding to an autoregressive moving-average (ARMA) Gaussian stochastic process. We propose a new maximum entropy covariance and cepstral extension problem and we show that the problem admits an approximate solution, which represents an ARMA graphical model whose topology is determined by the selected entries of the covariance lags considered in the extension problem. Then, we show how the corresponding dual problem is connected with the maximum likelihood principle. Such connection allows us to design a Bayesian model and characterize an approximate maximum a posteriori estimator of the ARMA graphical model in the case the graph topology is unknown. We test the performance of the proposed method through some numerical experiments.**

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

I. Introduction

In VARIOUS scenarios, we gather data of high dimensionality that characterizes intricate systems with a multitude of variables. Consequently, there is the necessity to comprehend the interrelationships among these variables. Essentially, we are compelled to extract from the data the underlying structure of the graphical model that depicts these associations. Clearly, a graphical model holds significance when it exhibits a certain degree of sparsity, meaning it is not densely connected in terms of edges. Moreover, the network structure depends on how we define these interactions. The latter can describe conditional dependence relations between the stochastic processes characterizing the variables of the system and whose resulting network is undirected [1], [2], [3], [4]. Alternatively, the interactions can describe whether a variable of the system is necessary to reconstruct another variable and in this case the resulting network is directed [5], [6], [7], [8], [9], [10]. It is noteworthy that a related topic regards the identification of network modules solely from neighboring node data, as discussed in works like [11], [12].

In this article, we focus our attention on the identification of graphical models whose edges encode conditional dependence relations. The typical approach is to set up a multivariate regularized spectral estimation problem, which originates from a maximum entropy covariance extension problem. Many generalizations of the covariance extension paradigm have been proposed in the classic scenario, i.e., the case in which the spectrum is not associated to a graphical model [13], [14], [15], [16], [17]. Interestingly, in the case, one want to infer the moving-average (MA) part of the model, then it is sufficient to consider a maximum entropy covariance extension problem to which we add some moments constraints regarding the cepstral coefficients of the spectrum, see, for instance, the works in [18], [19], [20], and [21].

Most of the identification paradigms for graphical models restrict the attention to autoregressive (AR) processes, see, for instance, [22], [23], [24]. The reason is that the parametrization of the spectrum induced by this model class allows us to impose easily sparsity constraints in terms of edges of the graphical model.

Learning autoregressive moving-average (ARMA) graphical models has been a less-explored problem in the literature. The first paper [25], fixed the MA part using a scalar trigonometric polynomial, which is estimated through a preliminary step, then a maximum entropy covariance extension problem is considered. A generalization has been proposed in [26] where the MA part is fixed through a matrix trigonometric polynomial obeying a certain structure. Recently, an iterative procedure to estimate the AR and MA parts has been proposed in [27]. More precisely, the estimation of the AR and MA parts is preformed separately according two different paradigms. In addition, the sparsity of the graph is induced through a group Lasso regularization term which can introduce an adverse bias in general (see [28]).

In this article, we propose a new maximum entropy covariance and cepstral extension problem, i.e., the optimal spectrum maximizes the entropy rate while matching some cepstral coefficients and some entries of some covariance lags. We show that the problem admits an approximate solution, which represents an ARMA graphical model whose topology is determined by the selected entries of the covariance lags considered in the extension problem. The fact that the existence result regards only the approximate solution is in line with the results found in univariate case (see, for instance, [20]). It is worth noting that in this paradigm the AR and MA parts are jointly estimated. Then, we show how the corresponding dual problem is connected with the maximum likelihood principle. The latter allows us to design a Bayesian model and characterize an approximate maximum a posteriori (MAP) estimator of the ARMA graphical model in the case the graph topology is unknown. In particular, the sparsity of the graph is induced through a reweighted iterative scheme, earlier used for the AR case in [29], in which the weights are iteratively adapted in such a way to reduce the presence of the adverse bias.

Manuscript received 28 August 2023; revised 10 May 2024; accepted 30 June 2024. Date of publication 8 July 2024; date of current version 31 December 2024. Recommended by Associate Editor A. Mahajan.

The author is with the Department of Information Engineering, University of Padova, 35131 Padova, Italy (e-mail: zorzimat@dei.unipd.it).

Digital Object Identifier 10.1109/TAC.2024.3424657

The rest of this article is organized as follows. In Section II, we provide the necessary background about dynamic graphical models. In Section III, we introduce the maximum entropy covariance and cepstral extension problem, while in Section IV, we show its regularized dual problem admits solution. In Section V, we show how the dual problem is connected with the maximum likelihood estimator. In Section VI, we propose a Bayesian model to estimate an ARMA graphical model, including its topology, from the data. Section VII contains some numerical examples showing the performance of the proposed method. Finally, Section VIII concludes this article.

Notation: $\mathbb{N}$ and $\mathbb{Z}$ are the set of natural and integer numbers, respectively; $\mathbb{R}$ and $\mathbb{R}_+$ are the set of real and positive real numbers, respectively. $\mathcal{S}^m$ denotes the vector space of symmetric matrices of dimension m, $\mathcal{S}^m_+$ is the cone of positive definite symmetric matrices of dimension m. $\mathbb{L}^m_\infty(\mathbb{T})$ denotes the space of functions defined on $\mathbb{T} := [0, 2\pi)$, taking values in the space of Hermitian matrices of dimension m and bounded a.e. Let $\mathbb{S}^m_+ \subset \mathbb{L}^m_\infty(\mathbb{T})$ denote the family of spectral densities on $\mathbb{T}$ of dimension m, which are bounded and coercive a.e.. Given $\Phi(e^{i\theta}) \in \mathbb{L}^m_\infty(\mathbb{T})$, $\int_\mathbb{T} \Phi$ denotes the integral of Φ over $\mathbb{T}$ with respect to the normalized Lebesgue measure $\mathrm{d}\theta/2\pi$. Given $X \in \mathbb{R}^{m\times m}$, $[X]_{jh}$ denotes the entry of X in position (j, h); this notation applies also for $\Phi \in \mathbb{L}^m_\infty(\mathbb{T})$. We denote as $\mathrm{supp}(\Phi)$ the support of Φ, i.e., it is a binary matrix whose entries are equal to one if the corresponding entries in Φ are different from the null function.

II. Background on Graphical Models

A dynamic undirected graphical model is a graph $\mathcal{G}(V, E)$ associated to an m-dimensional Gaussian, stationary, purely nondeterministic stochastic process $y := \{y(t),\ t \in \mathbb{Z}\}$ having vertexes $V := \{1 \dots m\}$ corresponding to the components of y and edges $E \subseteq V \times V$ describing the conditional dependence relations between the components. More precisely, for any index set $I \subseteq V$, define $\mathcal{X}_I := \overline{\mathrm{span}}\{y_j(t) \text{ s.t. } j \in I,\ t \in \mathbb{Z}\}$ the closure of the set containing all the finite linear combinations of the random variables $y_j(t)$, $j \in I$, so that for any $j \neq h$, the notation

$$\mathcal{X}_j \perp \mathcal{X}_h \,|\, \mathcal{X}_{V\setminus\{j,h\}}$$

means that for all t, s, $y_j(t)$ and $y_h(s)$ are conditionally independent given the space linearly generated by $\{y_k(t),\ k \in V \setminus \{j, h\},\ t \in \mathbb{Z}\}$. Let $\Phi(e^{i\theta})$, $\theta \in \mathbb{T}$, denote the power spectral density of y. Then, it is possible to prove that [30]

$$\mathcal{X}_j \perp \mathcal{X}_h \,|\, \mathcal{X}_{V\setminus\{j,h\}} \iff [\Phi(e^{i\theta})^{-1}]_{jh} = 0$$

and thus

$$(j, h) \notin E \iff [\Phi(e^{i\theta})^{-1}]_{jh} \equiv 0.$$

In plain words, the topology of the graph corresponds to the support of the inverse of the spectral density. Notice that, $(j, j) \in E$ for any $j \in V$.

III. Covariance and Cepstral Extension for Graphical Models

Consider an m-dimensional Gaussian stationary stochastic process y with power spectral density Φ. Let $R_k := \mathbb{E}[y_{t+k}y_t^T]$ be the kth covariance lag of y. Moreover, we define as

$$c_k := \int_\mathbb{T} e^{i\theta k} \log\det\Phi \tag{1}$$

the kth cepstral coefficient of y [31]. Assume to know $[R_k]_{jh}, c_k \in \mathbb{R}$ with $k = 0, \dots, n$ and $(j, h) \in E$. Then, we consider the following maximum entropy covariance and cepstral extension problem:

$$\max_{\Phi\in\mathbb{S}^m_+} \int_\mathbb{T} \log\det\Phi \tag{2a}$$

$$\text{s.t.} \int_\mathbb{T} e^{i\theta k}[\Phi]_{jh} = [R_k]_{jh}, \quad k = 0, \dots, n,\ (j, h) \in E \tag{2b}$$

$$\int_\mathbb{T} e^{i\theta k} \log\det\Phi = c_k, \quad 1, \dots, n. \tag{2c}$$

It is worth noting that (2) is a generalization of the covariance selection problem proposed by Dempster [32]. The Lagrangian is

$$\begin{aligned} L(\Phi, \mathbf{p}, \mathbf{Q}) &= \int_\mathbb{T} \log\det\Phi \\ &\quad - \sum_{(j,h)\in E}\sum_{k=0}^{n}[Q_k]_{jh}\left(\int_\mathbb{T} e^{i\theta k}[\Phi]_{jh} - [R_k]_{jh}\right) \\ &\quad + \sum_{k=1}^{n} p_k\left(\int_\mathbb{T} e^{i\theta k}\log\det\Phi - c_k\right) - p_0c_0 \\ &= \int_\mathbb{T} \mathbf{p}\log\det\Phi - \mathrm{tr}(\mathbf{Q}\Phi) + \sum_{k=0}^{n}\mathrm{tr}(Q_k^T R_k) - \sum_{k=0}^{n} p_k c_k \end{aligned}$$

where $p_0 = 1$, c_0 is the zeroth cepstral coefficient defined as in (1) with $k = 0$ and the constant term p_0c_0 is only added for convenience in the following; Q_k, with $[Q_k]_{jh} = 0$ for $(j, h) \notin E$, and p_k, with $k \neq 0$, are the Lagrange multipliers and

$$\mathbf{p}(e^{i\theta}) := 1 + \frac{1}{2}\sum_{k=1}^{n} p_k(e^{-i\theta k} + e^{i\theta k})$$

$$\mathbf{Q}(e^{i\theta}) := Q_0 + \frac{1}{2}\sum_{k=1}^{n} Q_k e^{-i\theta k} + Q_k^T e^{i\theta k} \tag{3}$$

are the corresponding trigonometric polynomials. Let

$$\mathfrak{B}^m_+(E) := \Big\{ Q(e^{i\theta}) = Q_0 + \frac{1}{2}\sum_{k=1}^{n} Q_k e^{-i\theta k} + Q_k^T e^{i\theta k} \text{ s.t.}$$

$$[Q_k]_{jh} = 0 \text{ for } (j, h) \notin E, Q(e^{i\theta}) > 0 \text{ on } \mathbb{T} \Big\}$$

$$\mathfrak{B}^o_+ := \left\{ p(e^{i\theta}) = 1 + \frac{1}{2}\sum_{k=1}^{n} p_k(e^{-i\theta k} + e^{i\theta k}) \text{ s.t. } p(e^{i\theta}) > 0 \text{ on } \mathbb{T} \right\}.$$

Let $\overline{\mathfrak{B}^o_+}$ denote the closure of $\mathfrak{B}^o_+$; $\overline{\mathfrak{B}^m_+}(E)$ denotes the closure of $\mathfrak{B}^m_+(E)$ without the points $\mathbf{Q}$ for which $\det(\mathbf{Q}(e^{i\theta})) = 0$

for all $\theta \in \mathbb{T}$. The dual function $\sup_{\Phi \in \mathbb{S}_+^m} L(\Phi, \mathbf{p}, \mathbf{Q})$ is finite if and only if $\mathbf{p} \in \overline{\mathfrak{B}_+^o}$ and $\mathbf{Q} \in \overline{\mathfrak{B}_+^m}(E)$. Indeed, assume that there exists θ_0 for which $\mathbf{p}(e^{i\theta_0}) < 0$. Taking Φ which tends to be singular in a neighborhood of θ_0 we have that $L(\Phi, \mathbf{p}, \mathbf{Q}) \to \infty$. Hence, we must have $\mathbf{p} \in \overline{\mathfrak{B}_+^o}$. Moreover, assume that there exists θ_0 for which $\mathbf{Q}(e^{i\theta_0})$ has at least one negative eigenvalue. Let v be the corresponding eigenvector. Taking Φ which diverges along v in a neighborhood of θ_0 we have that $L(\Phi, \mathbf{p}, \mathbf{Q}) \to \infty$. Finally, if $\det(\mathbf{Q}(e^{i\theta})) = 0$ for all $\theta \in \mathbb{T}$ then we can exploit the fact that there exists θ_0 such that $\mathbf{p}(e^{i\theta}) > 0$ in a neighborhood of θ_0 because $\mathbf{p} \in \overline{\mathfrak{B}_+^o}$. Let $v(e^{i\theta})$ be a nonnull vector belonging to the null space of $\mathbf{Q}(e^{i\theta})$. Taking Φ which diverges along $v(e^{i\theta})$ in the neighborhood of θ_0 we have that $L(\Phi, \mathbf{p}, \mathbf{Q}) \to \infty$. Thus, we must have $\mathbf{Q} \in \overline{\mathfrak{B}_+^m}(E)$.

Let $\mathbf{p} \in \overline{\mathfrak{B}_+^o}$ and $\mathbf{Q} \in \overline{\mathfrak{B}_+^m}(E)$, then $L(\cdot, \mathbf{p}, \mathbf{Q})$ is concave on $\mathbb{S}_+^m$ and the unconstrained maximum is also a stationary point, i.e., it is obtained by setting equal to zero the first variation of L along any direction $\delta\Phi \in \mathbb{L}_\infty^m(\mathbb{T})$

$$\delta L(\Phi, \mathbf{p}, \mathbf{Q}; \delta\Phi) = \operatorname{tr} \int_{\mathbb{T}} \mathbf{p}\Phi^{-1}\delta\Phi - \mathbf{Q}\delta\Phi.$$

Accordingly, the stationary point satisfies the condition

$$\Phi^\circ(e^{i\theta}) = \mathbf{p}(e^{i\theta})\mathbf{Q}(e^{i\theta})^{-1} \text{ a.e.} \tag{4}$$

The dual function takes the form

$$\begin{aligned} J(\mathbf{p}, \mathbf{Q}) &:= L(\Phi^\circ, \mathbf{p}, \mathbf{Q}) \\ &= \int_{\mathbb{T}} \mathbf{p} \log\det(\mathbf{p}\mathbf{Q}^{-1}) - \mathbf{p}m + \sum_{k=0}^{n} \operatorname{tr}(Q_k^T R_k) - \sum_{k=0}^{n} p_k c_k. \end{aligned}$$

Therefore, the dual problem is

$$\begin{aligned} &\min_{\mathbf{p},\mathbf{Q}} J(\mathbf{p}, \mathbf{Q}) \\ &\text{s.t. } \mathbf{p} \in \overline{\mathfrak{B}_+^o}, \ \mathbf{Q} \in \overline{\mathfrak{B}_+^m}(E). \end{aligned} \tag{5}$$

If the dual problem admits an interior point solution, then Problem (2) admits a unique solution of the form in (4). The latter represents the power spectral density of an ARMA process whose support of the inverse of its spectral density corresponds to E. Accordingly, model (4) describes an ARMA graphical model whose conditional dependence relations are specified by E. At this point, it is worth comparing the extension problem (2) with the one proposed in [25]; more precisely, the following maximum entropy covariance extension problem has been proposed:

$$\begin{aligned} &\max_{\Phi \in \mathbb{S}_+^m} \int_{\mathbb{T}} \mathbf{p} \log\det\Phi \\ &\text{s.t. } \int_{\mathbb{T}} e^{i\theta k}[\Phi]_{jh} = [R_k]_{jh}, \quad k = 0, \dots, n, \ (j, h) \in E \end{aligned} \tag{6}$$

where $\mathbf{p}(e^{i\theta}) = \sum_{k=-n}^{n} p_k e^{-i\theta k} > 0$ is given, e.g., it is usually estimated from the data though a preliminary step. It turns out that the solution to (6) has the same form of (4), and thus, it corresponds to an ARMA graphical model, but the main difference is that $\mathbf{p}$ is fixed by the user. On the contrary, our formulation is more general in the sense that $\mathbf{p}$ and $\mathbf{Q}$ are jointly optimized.

However, to prove the existence of an interior point solution to (5) is not trivial, see the scalar case (i.e., $m = 1$) in [19] and [33]. Drawing inspiration from the scalar case [21], in the following section, we will address this issue considering a regularized version of the dual problem and whose regularizer forces the optimal solution $(\mathbf{p}, \mathbf{Q}) \in \mathfrak{B}_+^o \times \mathfrak{B}_+^m(E)$.

In the case one want to consider the general setup in which the trigonometric polynomials $\mathbf{p}$ and $\mathbf{Q}$ have different degree, say n_p and n_q, then it is sufficient to replace n with n_p and n_q in (2c) and (2b), respectively. For ease of exposition, we consider the case $n = n_p = n_q$, but all the following results trivially hold also for the case $n_p \neq n_q$.

IV. Existence of the Approximate Solution

We consider the following regularized dual problem:

$$\begin{aligned} &\min_{\mathbf{p},\mathbf{Q}} J(\mathbf{p}, \mathbf{Q}) + g_\lambda(\mathbf{p}) \\ &\text{s.t. } \mathbf{p} \in \overline{\mathfrak{B}_+^o}, \ \mathbf{Q} \in \overline{\mathfrak{B}_+^m}(E) \end{aligned} \tag{7}$$

where

$$g_\lambda(\mathbf{p}) := \lambda \int_{\mathbb{T}} \frac{1}{\mathbf{p}} \tag{8}$$

is the regularization term and $\lambda > 0$ is the regularization parameter, which is taken sufficiently small. The aim of this section is to prove the following result showing that Problem (7) admits a unique solution belonging to the interior of the feasible set.

Assumption 1: There exists a power spectral density $\bar{\Phi}$, which is positive definite on some open interval in $\mathbb{T}$ and such that

$$\int_{\mathbb{T}} [\bar{\Phi}]_{jh} e^{i\theta k} = [R_k]_{jh}, \quad (j, h) \in E, \ k = 0, \dots, n.$$

Theorem 1: Under Assumption 1, Problem (7) admits a unique solution $(\hat{\mathbf{p}}, \hat{\mathbf{Q}}) \in \mathfrak{B}_+^o \times \mathfrak{B}_+^m(E)$ and the rational spectral density $\hat{\Phi} = \hat{\mathbf{p}}\hat{\mathbf{Q}}^{-1} \in \mathbb{S}_+^m$ is such that

$$\int_{\mathbb{T}} e^{i\theta k}[\hat{\Phi}]_{jh} = [R_k]_{jh}, \quad k = 0, \dots, n, \ (j, h) \in E \tag{9a}$$

$$\int_{\mathbb{T}} e^{i\theta k} \log\det\hat{\Phi} = c_k + \varepsilon_k, \quad 1, \dots, n \tag{9b}$$

where $\varepsilon_k = \lambda \int_{\mathbb{T}} e^{i\theta k}\hat{\mathbf{p}}^{-2}$, i.e., the cepstral coefficients are matched only approximately and the approximation error is ε_k which depends on λ.

It is worth noting that $\hat{\Phi}$ represents an approximate solution of Problem (2) and it corresponds to an ARMA graphical model whose graph is $\mathcal{G}(V, E)$. Let

$$\tilde{J}_\lambda(\mathbf{p}, \mathbf{Q}) := J(\mathbf{p}, \mathbf{Q}) + g_\lambda(\mathbf{p}) \tag{10}$$

be the regularized dual function in Problem (7).

Proposition 1: $\tilde{J}_\lambda$ is strictly convex on $\mathfrak{B}_+^o \times \mathfrak{B}_+^m(E)$.

Proof: Let $\mathcal{S}^m(E) := \{\, Y \in \mathcal{S}^m \text{ s.t. } [Y]_{jh} = 0 \text{ for } (j, h) \notin E\}$ and $\mathcal{S}_+^m(E) := \mathcal{S}^m(E) \cap \mathcal{S}_+^m$. Consider the function

$$h_\lambda(X, Y) := x \log\det(xY^{-1}) + \lambda x^{-1}$$

with $x \in \mathbb{R}_+$ and $Y \in \mathcal{S}_+^m(E)$. Then, its first and second variations along the direction $(\delta x, \delta Y) \in \mathbb{R} \times \mathcal{S}^m(E)$ are

$$\begin{aligned}&\delta h_\lambda(x, Y; \delta x, \delta Y)\\ &= (m + m\log x - \log\det(Y) - \lambda x^{-2})\delta x - x\mathrm{tr}(Y^{-1}\delta Y)\\ &\delta^2 h_\lambda(x, Y; \delta x, \delta Y) = (mx^{-1} + 2\lambda x^{-3})\delta x^2\\ &\quad - 2\mathrm{tr}(Y^{-1}\delta Y)\delta x + x\mathrm{tr}(Y^{-1}\delta Y Y^{-1}\delta Y)\\ &= \begin{bmatrix}\delta x & \mathrm{vec}(\delta Y)^T\end{bmatrix} M \begin{bmatrix}\delta x\\ \mathrm{vec}(\delta Y)\end{bmatrix}\end{aligned}$$

where

$$M := \begin{bmatrix} mx^{-1} + 2\lambda x^{-3} & \mathrm{vec}(Y^{-1})^T\\ \mathrm{vec}(Y^{-1}) & x(Y^{-1}\otimes Y^{-1})\end{bmatrix}.$$

Matrix M is positive definite because $mx^{-1} + 2\lambda x^{-3} > 0$ and the Schur complement $M/mx^{-1} + 2\lambda x^{-3}$ is positive

$$\begin{aligned}&mx^{-1} + 2\lambda x^{-3} - x^{-1}\mathrm{vec}(Y^{-1})^T(Y^{-1}\otimes Y^{-1})^{-1}\mathrm{vec}(Y^{-1})\\ &\quad = mx^{-1} + 2\lambda x^{-3} - x^{-1}\mathrm{vec}(Y^{-1})^T(Y\otimes Y)\mathrm{vec}(Y^{-1})\\ &\quad = mx^{-1} + 2\lambda x^{-3} - x^{-1}\mathrm{vec}(Y^{-1})^T\mathrm{vec}(Y)\\ &\quad = mx^{-1} + 2\lambda x^{-3} - x^{-1}\mathrm{tr}(Y^{-1}Y) = 2\lambda x^{-3} > 0.\end{aligned}$$

We conclude that $h_\lambda(x, Y)$ is strictly convex in the domain $\mathbb{R}_+ \times \mathcal{S}_+^m(E)$ because $\delta^2 h_\lambda(x, Y; \delta x, \delta Y) > 0$ for any nonnull direction $(\delta x, \delta Y) \in \mathbb{R} \times \mathcal{S}^m(E)$. Finally, notice that

$$\tilde{J}_\lambda(\mathbf{p}, \mathbf{Q}) := \int_{\mathbb{T}} h_\lambda(\mathbf{p}, \mathbf{Q}) - \mathbf{p}m + \sum_{k=0}^n \mathrm{tr}(Q_k^T R_k) - \sum_{k=0}^n p_k c_k$$

where the first term is strictly convex on $\mathfrak{B}_+^o \times \mathfrak{B}_+^m(E)$ because it is the integral of a strictly convex function while the remaining terms are linear in $(\mathbf{p}, \mathbf{Q})$. Accordingly, the claim holds. ∎

In what follows, we denote by $\partial\mathfrak{B}_+^m(E)$ the boundary of $\mathfrak{B}_+^m(E)$, which is defined without the points $\mathbf{Q}$ for which $\det(\mathbf{Q}(e^{i\theta})) = 0$ for all $\theta \in \mathbb{T}$. Moreover, we denote by $\partial\mathfrak{B}_+^o$ the boundary of $\mathfrak{B}_+^o$.

Proposition 2: $\tilde{J}_\lambda$ is lower semicontinuous on $\overline{\mathfrak{B}_+^o} \times \overline{\mathfrak{B}_+^m}(E)$ with values on the extended reals.

Proof: From Proposition 1, it follows that $\tilde{J}_\lambda$ is differentiable and, thus, continuous in the interior of $\overline{\mathfrak{B}_+^o} \times \overline{\mathfrak{B}_+^m}(E)$. Accordingly, we only need to show the lower semicontinuity on the boundary of $\overline{\mathfrak{B}_+^o} \times \overline{\mathfrak{B}_+^m}(E)$. More precisely, we only need to show that the three terms

$$\tilde{J}_1 = -\int_{\mathbb{T}} \mathbf{p}\log\det(\mathbf{Q}),\ \tilde{J}_2 = \int_{\mathbb{T}} m\mathbf{p}\log\mathbf{p},\ \tilde{J}_3 = \int_{\mathbb{T}} \lambda\mathbf{p}^{-1}$$

are lower semicontinuous because the other terms in $\tilde{J}_\lambda$ are obviously continuous.

First term: Take a sequence $\{(\mathbf{p}_j, \mathbf{Q}_j) \in \overline{\mathfrak{B}_+^o} \times \overline{\mathfrak{B}_+^m}(E),\ j \in \mathbb{N}\}$, which converges to $(\mathbf{p}, \mathbf{Q})$ and the latter belongs to the boundary of the feasible set. Assume that $\mathbf{Q} \in \partial\mathfrak{B}_+^m(E)$ and $\mathbf{p} \in \overline{\mathfrak{B}_+^o}$. Since $\mathbf{Q}$ is bounded and the convergence $\mathbf{Q}_j \to \mathbf{Q}$ is uniform, then we have that $\mu := \sup_j \max_\theta \sigma_{\max}(\mathbf{Q}_j) < \infty$ where $\sigma_{\max}(Y)$ denotes the maximum eigenvalue of $Y \in \mathcal{S}_+^m$. Accordingly, $0 \le \mu^{-1}\mathbf{Q}_j \le I$ for any $j \in \mathbb{N}$. Let $\mathcal{Z}(\mathbf{Q}) := \{\,\theta \in \mathbb{T} \text{ s.t. } \det(\mathbf{Q}(e^{i\theta})) = 0\,\}$ and $\tilde{\mathcal{Z}} := (\cup_{j\in\mathbb{N}}\mathcal{Z}(\mathbf{Q}_j)) \cup \mathcal{Z}(\mathbf{Q})$. Since $\tilde{\mathcal{Z}}$ has Lebesgue measure equal to zero, we have

$$\tilde{J}_1 = \int_{\mathbb{T}} \mathbf{p}_j \log\det(\mathbf{Q}_j) = \int_{\mathbb{T}\setminus\tilde{\mathcal{Z}}} \mathbf{p}_j \log\det(\mathbf{Q}_j)$$

thus we can consider the integrands over $\mathbb{T} \setminus \tilde{\mathcal{Z}}$ in which they are well defined. Then

$$\mathbf{p}_j \log\det(\mu^{-1}\mathbf{Q}_j) \to \mathbf{p}\log\det(\mu^{-1}\mathbf{Q})$$

pointwise in $\mathbb{T} \setminus \tilde{\mathcal{Z}}$ as $j \to \infty$. Since $-\mathbf{p}_j \log\det(\mu^{-1}\mathbf{Q}_j)$ is a nonnegative function, by Fatou's lemma [34], page 23, we have that

$$\int_{\mathbb{T}\setminus\tilde{\mathcal{Z}}} \mathbf{p}\log\det(\mu^{-1}\mathbf{Q}) \le \liminf_{j\to\infty} \int_{\mathbb{T}\setminus\tilde{\mathcal{Z}}} \mathbf{p}_j \log\det(\mu^{-1}\mathbf{Q}_j)$$

which proves that $\tilde{J}_1$ is lower semicontinuous in these points. In the remaining case $\mathbf{Q} \in \mathfrak{B}_+^m(E)$, $\mathbf{p} \in \partial\mathfrak{B}_+^o$ we have that $\tilde{J}_1$ is continuous in these points.

Other terms: The lower semicontinuity of $\tilde{J}_2$ has been proved in [20], proof of Lemma 5.1. The lower semicontinuity of $\tilde{J}_3$ follows from [33], see proof of Lemma 5.5 (first case). ∎

The aim of the next propositions is to analyze the behavior of $\tilde{J}_\lambda$ when the optimization variables diverge or are on the boundary of the feasible set. We define the norm of the Lagrange multipliers

$$\|\mathbf{p}\| := \sqrt{\sum_{k=1}^n p_k^2},\quad \|\mathbf{Q}\| := \sqrt{\sum_{k=0}^n \mathrm{tr}(Q_k Q_k^T)}$$

and $\|(\mathbf{p}, \mathbf{Q})\| = \|\mathbf{p}\| + \|\mathbf{Q}\|$.

Proposition 3: Let Assumption 1 hold. Consider a sequence $\{(\mathbf{p}_j, \mathbf{Q}_j) \in \overline{\mathfrak{B}_+^o} \times \overline{\mathfrak{B}_+^m}(E),\ j \in \mathbb{N}\}$ such that $\|(\mathbf{p}_j, \mathbf{Q}_j)\| \to \infty$ as $j \to \infty$. Then, $\tilde{J}_\lambda(\mathbf{p}_j, \mathbf{Q}_j) \to \infty$.

Proof: Since $\mathbf{p}_j$ is a nonnegative function and its integral on $\mathbb{T}$ is equal to one, then it follows that $\|\mathbf{p}_j\|$ cannot diverge as $j \to \infty$. Accordingly, it must happen that $\|\mathbf{Q}_j\| \to \infty$. By Assumption 1, we have that

$$\begin{aligned}\tilde{J}_\lambda(\mathbf{p}, \mathbf{Q}) &\ge \int_{\mathbb{T}} -\mathbf{p}\log\det(\mathbf{Q}) + \sum_{k=0}^n \mathrm{tr}(Q_k R_k^T) + s(\mathbf{p})\\ &= \int_{\mathbb{T}} -\mathbf{p}\log\det(\mathbf{Q}) + \mathrm{tr}(\bar{\Phi}\mathbf{Q}) + s(\mathbf{p}) \qquad (11)\end{aligned}$$

where

$$s(\mathbf{p}) := \int_{\mathbb{T}} m(\mathbf{p}\log\mathbf{p} - \mathbf{p}) - \sum_{k=0}^n p_k c_k.$$

Accordingly, $s(\mathbf{p}_j)$ cannot diverge as $j \to \infty$. Let $\mathbf{Q}_j^0 := \mathbf{Q}_j^0/\|\mathbf{Q}_j^0\| \in \overline{\mathfrak{B}_+^m}(E)$. Let

$$\eta := \liminf_{j\to\infty} \int_{\mathbb{T}} \mathrm{tr}(\bar{\Phi}\mathbf{Q}_j^0) \ge 0.$$

Then, the sequence $\{\mathbf{Q}_j^0,\ j \in \mathbb{N}\}$ has a subsequence $\{\mathbf{Q}_{j_m}^0,\ j_m \in \mathbb{N}\}$ such that $\int_{\mathbb{T}} \mathrm{tr}(\bar{\Phi}\mathbf{Q}_{j_m}^0) \to \eta$ as $j_m \to \infty$. Moreover, since such subsequence belongs to the surface of the unit ball, which is a compact set, it has a subsubsequence $\{\mathbf{Q}^0_{j_t},\ j_t\in\mathbb{N}\}$, which is convergent. Let

$$\mathbf{Q}^0_\infty=\lim_{j_t\to\infty}\mathbf{Q}^0_{j_t}\in\overline{\mathfrak{B}^m_+}(E).$$

Then, $\eta=\int_{\mathbb{T}}\operatorname{tr}(\bar\Phi\mathbf{Q}^0_\infty)$ and $\eta>0$. The latter property can be proved by contradiction. Assume that $\eta=0$. Since $\operatorname{tr}(\bar\Phi\mathbf{Q}^0_\infty)\geq 0$ a.e. on $\mathbb{T}$, it follows that $\operatorname{tr}(\bar\Phi\mathbf{Q}^0_\infty)=0$ a.e. on $\mathbb{T}$. By Assumption 1, then $\mathbf{Q}^0_\infty$ must vanish on some open interval in $\mathbb{T}$. By [35], Lemma 1, it follows that $\mathbf{Q}^0_\infty\equiv 0$, which is a contradiction because $\|\mathbf{Q}^0_\infty\|=1$. Taking into account (11), we have

$$\begin{aligned}\liminf_{j\to\infty}\tilde J_\lambda(\mathbf{p}_j,\mathbf{Q}_j)&\geq\liminf_{j\to\infty}\Big[\,\|\mathbf{Q}_j\|\int_{\mathbb{T}}\operatorname{tr}(\bar\Phi\mathbf{Q}^0_j)\\&\quad-m\log\|\mathbf{Q}_j\|-\int_{\mathbb{T}}\mathbf{p}_j\log\det(\mathbf{Q}^0_j)+s(\mathbf{p}_j)\Big]\\&=\liminf_{j\to\infty}[\,\eta\|\mathbf{Q}_j\|-m\log\|\mathbf{Q}_j\|]\\&\quad+\liminf_{j\to\infty}\Big[s(\mathbf{p}_j)-\int_{\mathbb{T}}\mathbf{p}_j\log\det(\mathbf{Q}^0_j)\Big]=\infty\end{aligned}$$

where we exploited the fact that the second term does not diverge as $j\to\infty$. ■

Proposition 4: Consider a sequence $\{(\mathbf{p}_j,\mathbf{Q}_j)\in\mathfrak{B}^o_+\times\mathfrak{B}^m_+(E),\ j\in\mathbb{N}\}$ converging to $(\mathbf{p},\mathbf{Q})$, which belongs to the boundary of $\overline{\mathfrak{B}^o_+}\times\overline{\mathfrak{B}^m_+}(E)$. Then, it cannot be an infimizing sequence for $\tilde J_\lambda$.

Proof: We have two cases.

First case: $\mathbf{p}\in\partial\mathfrak{B}^o_+$ and $\mathbf{Q}\in\overline{\mathfrak{B}^m_+}(E)$. Let $Q_{k,j}\in\mathbb{R}^{m\times m}$ and $p_{j,k}\in\mathbb{R}$, with $k=0\ldots n$, be the coefficients of the trigonometric polynomials $\mathbf{Q}_j$ and $\mathbf{p}_j$, respectively. Notice that

$$\begin{aligned}\tilde J_\lambda(\mathbf{p}_j,\mathbf{Q}_j)&=\int_{\mathbb{T}}\mathbf{p}_jD_{IS}(\mathbf{p}_j^{-1}\mathbf{Q}_j,I)-\operatorname{tr}(\mathbf{Q}_j)\\&\quad+\sum_{k=0}^n\operatorname{tr}(Q_{j,k}^TR_k)-\sum_{k=0}^n p_{j,k}c_k+g_\lambda(\mathbf{p}_j)\end{aligned}$$

where

$$D_{IS}(X,Y)=\log\det(X^{-1}Y)+\operatorname{tr}(XY^{-1})-m \tag{12}$$

denotes the Itakura–Saito distance between $X,Y\in\mathcal{S}^m_+$. Since $\mathbf{p}_j(e^{i\theta})>0$ and $D_{IS}(\mathbf{p}_j^{-1}\mathbf{Q}_j)\geq 0$ on $\mathbb{T}$, it follows that

$$\begin{aligned}&\tilde J_\lambda(\mathbf{p}_j,\mathbf{Q}_j)\geq\\&-\int_{\mathbb{T}}\operatorname{tr}(\mathbf{Q}_j)+\sum_{k=0}^n\operatorname{tr}(Q_{j,k}^TR_k)-\sum_{k=0}^n p_{j,k}c_k+g_\lambda(\mathbf{p}_j)\\&\to\infty\end{aligned}$$

where we exploited the fact that $g_\lambda(\mathbf{p}_j)\to\infty$ as $j\to\infty$, while the other terms converge to a finite value.

Second case: $\mathbf{p}\in\mathfrak{B}^o_+$ and $\mathbf{Q}\in\partial\mathfrak{B}^m_+(E)$. In this case, the limit of $\tilde J_\lambda(\mathbf{p}_j,\mathbf{Q}_j)$ takes a finite value, however, we will show that $(\mathbf{p},\mathbf{Q})$ cannot be a point of minimum for $\tilde J_\lambda$. By continuity we have

$$\mathbf{p}(e^{i\theta})\geq\mu:=\min_{\theta\in\mathbb{T}}\mathbf{p}(e^{i\theta})>0.$$

Now, consider the first variation of $\tilde J_\lambda(\mathbf{p},\cdot)$ at $\mathbf{Q}+tI\in\mathfrak{B}^m_+(E)$, with $t>0$, along the direction $\delta\mathbf{Q}=I$

$$\begin{aligned}\delta\tilde J_\lambda(\mathbf{p},\mathbf{Q}+tI;I)&=-\int_{\mathbb{T}}\mathbf{p}\operatorname{tr}(\mathbf{Q}+tI)^{-1}+\operatorname{tr}(R_0)\\&\leq-\mu\int_{\mathbb{T}}\operatorname{tr}(\mathbf{Q}+tI)^{-1}+\operatorname{tr}(R_0).\end{aligned}$$

By the Lebesgue's monotone convergence theorem, we have

$$\lim_{t\to0^+}\delta\tilde J_\lambda(\mathbf{p},\mathbf{Q}+tI;I)\leq-\mu\int_{\mathbb{T}}\operatorname{tr}(\mathbf{Q}^{-1})+\operatorname{tr}(R_0)=-\infty$$

because $\mu>0$ and $\operatorname{tr}(\mathbf{Q}^{-1})$ is a positive rational function having some poles on $\mathbb{T}$. We conclude that $(\mathbf{p},\mathbf{Q})$ cannot be a point of minimum for $\tilde J_\lambda$. ■

We are now ready to prove the existence result.

Proof of Theorem 1: Consider the nonempty sublevel set

$$\tilde J_\lambda^{-1}(-\infty,r]:=\{(\mathbf{p},\mathbf{Q})\in\overline{\mathfrak{B}^o_+}\times\overline{\mathfrak{B}^m_+}(E)\text{ s.t. }\tilde J_\lambda(\mathbf{p},\mathbf{Q})\leq r\}$$

where the real number r is taken sufficiently large. Such sublevel set is closed and bounded by Proposition 4 and Proposition 3, respectively, and thus, compact because $(\mathbf{p},\mathbf{Q})$ belongs to a finite-dimensional vector space. Since $\tilde J_\lambda$ is lower semicontinuous, see Proposition 2, it follows that the regularized dual problem (7) admits solution by the Weierstrass' theorem. Moreover, the minimum belongs to the interior of the feasible set $\tilde J_\lambda^{-1}(-\infty,r]\subset\mathfrak{B}^o_+\times\mathfrak{B}^m_+(E)$ by Proposition 4 and 3. Finally, this point of minimum is unique because $\tilde J_\lambda$ is strictly convex over $\mathfrak{B}^o_+\times\mathfrak{B}^m_+(E)$ by Proposition 1. Moreover, this point of minimum satisfies the stationarity conditions (i.e., the first variation of $\tilde J_\lambda$ must vanish along any direction), which implies the moment conditions in (9). □

Remark 1: Theorem 1 also suggests a simple procedure to choose λ in a proper way. First, we choose a set of L candidate values, say λ_l with $l=1,\ldots,L$. For each λ_l, we solve the regularized dual problem in (7) and, in view of Theorem 1, the corresponding approximation errors in matching the cepstral coefficients. Let $\varepsilon_{k,l}$ be the approximation error corresponding to the kth cepstral coefficient using λ_l. Then, we can choose the best λ_l as the one maximizing a score function, which measures how well we match the cepstral coefficients, e.g.,

$$f(\lambda_l)=-\sum_{k=1}^n\varepsilon^2_{k,l}.$$

Remark 2: The choice of g_λ in (8) is not unique. The class of alternative regularizers takes the form

$$g_\lambda(\mathbf{p})=\int_{\mathbb{T}}\frac{1}{\mathbf{p}^\nu}$$

where $\nu\in\mathbb{N}$ (see [36]). Using reasonings similar to the ones used before, the results of this section still hold with these alternative regularizers.

V. Connection With the ML Estimator

Consider a graphical model $\mathcal{G}(V,E)$ associated to the ARMA process y whose spectral density $\Phi\in\mathbb{S}^m_+$ is of the form in (4). Assume that the data $y^N:=\{y(1),\ldots,y(N)\}$ generated by this model is available. Then, an estimate of this graphical model can be obtained through Problem (2): it is sufficient to replace R_ks and c_ks with their sample estimators using y^N

$$\hat{R}_k = \hat{R}_{-k}^T = \frac{1}{N}\sum_{t=1}^{N-k} y(t+k)y(t)^T$$

$$\hat{c}_k = \hat{c}_{-k} = \int_{\mathbb{T}} e^{i\theta k}\log\det(\hat{\Phi}_P)$$

where

$$\hat{\Phi}_P(e^{i\theta}) = \sum_{k=1-h(N)}^{h(N)-1} \hat{R}_k e^{-i\theta k}$$

denotes the f-periodogram computed from the data y^N with $h(N)$ such that $h(N)\to\infty$ and $h(N)^2/N \to 0$ as $N\to\infty$. It is worth noting that $\hat{\Phi}_P$ is a mean-square consistent estimator under mild assumptions on Φ [37]. The dual problem coincides with (5) where R_ks and c_ks are replaced with $\hat{R}_k$s and $\hat{c}_k$s. Let

$$\hat{\Phi}_n(e^{i\theta}) := \sum_{k=-n}^{n} \hat{R}_k e^{-i\theta k}$$

$$\hat{\psi}(e^{i\theta}) := \sum_{k=-h(N)+1}^{h(N)-1} \hat{c}_k e^{-i\theta k} \approx \log\det(\hat{\Phi}_P).$$

It is not difficult to see that the dual function in (5) can be written as

$$J(\mathbf{p},\mathbf{Q}) = \int_{\mathbb{T}} -\mathbf{p}\log\det(\mathbf{p}^{-1}\mathbf{Q}) + \mathrm{tr}(\mathbf{Q}\hat{\Phi}_P) - \mathbf{p}\hat{\psi} - \mathbf{p}m$$

where we exploited the identities

$$\mathrm{tr}\sum_{k=0}^{n} Q_k^T\hat{R}_k = \mathrm{tr}\int_{\mathbb{T}}\mathbf{Q}\hat{\Phi}_n = \mathrm{tr}\int_{\mathbb{T}}\mathbf{Q}\hat{\Phi}_P$$

$$\sum_{k=0}^{n} p_k\hat{c}_k = \int_{\mathbb{T}}\mathbf{p}\sum_{k=-n}^{n}\hat{c}_k e^{-i\theta k} = \int_{\mathbb{T}}\mathbf{p}\hat{\psi}$$

because $\mathbf{p}\in\overline{\mathfrak{B}_+^o}$ and $\mathbf{Q}\in\overline{\mathfrak{B}_+^m}(E)$. Thus

$$J(\mathbf{p},\mathbf{Q}) = \int_{\mathbb{T}}\mathbf{p}\log\det(\mathbf{p}\mathbf{Q}^{-1}) + \mathrm{tr}(\mathbf{Q}\hat{\Phi}_P) - \mathbf{p}\hat{\psi} - \mathbf{p}m$$

$$\approx \int_{\mathbb{T}}\mathbf{p}[\log\det(\mathbf{p}\mathbf{Q}^{-1}) + \mathrm{tr}(\mathbf{p}^{-1}\mathbf{Q}\hat{\Phi}_P) - \log\det(\hat{\Phi}_P) - m]$$

$$= \int_{\mathbb{T}}\mathbf{p}D_{IS}(\hat{\Phi}_P,\mathbf{p}\mathbf{Q}^{-1})$$

where we recall that D_{IS} is the Itakura–Saito distance defined in (12). Let $(\mathbf{p},\mathbf{Q})\in\mathfrak{B}_+^o\times\mathfrak{B}_+^m(E)$. Then, $\mathbf{p}(e^{i\theta})>0$ and $D_{\mathrm{IS}}(\hat{\Phi}_P,\mathbf{p}\mathbf{Q}^{-1})\geq 0$ on $\mathbb{T}$. By the weighted mean value theorem it follows that there exists $\bar{\theta}$ such that

$$J(\mathbf{p},\mathbf{Q}) \approx \int_{\mathbb{T}}\mathbf{p}D_{IS}(\hat{\Phi}_P,\mathbf{p}\mathbf{Q}^{-1})$$

$$= \mathbf{p}(e^{i\bar{\theta}})\int_{\mathbb{T}}D_{IS}(\hat{\Phi}_P,\mathbf{p}\mathbf{Q}^{-1})$$

$$= \mathbf{p}(e^{i\bar{\theta}})[\int_{\mathbb{T}}\log\det(\mathbf{p}\mathbf{Q}^{-1}) + \mathrm{tr}(\mathbf{p}^{-1}\mathbf{Q}\hat{\Phi}_P) - \log\det(\hat{\Phi}_P) - m]. \tag{13}$$

Proposition 5: Consider an ARMA process with spectral density $\Phi = \mathbf{p}\mathbf{Q}^{-1}$ with $\mathbf{p}\in\mathfrak{B}_+^o$ and $\mathbf{Q}\in\mathfrak{B}_+^m(E)$. For N sufficiently large, the negative log-likelihood of the data y^N under this ARMA model is (up to constant terms not depending on $\mathbf{p}$ and $\mathbf{Q}$)

$$\ell(y^N;\mathbf{p},\mathbf{Q}) \approx \frac{N}{2}\int_{\mathbb{T}}\log\det(\mathbf{p}\mathbf{Q}^{-1}) + \mathrm{tr}(\mathbf{p}^{-1}\mathbf{Q}\hat{\Phi}_P). \tag{14}$$

Proof: Let $\mathbf{W}$ be a causal rational right spectral factor of $\mathbf{p}^{-1}\mathbf{Q}$, which is analytic on $\mathbb{T}$. Let s be the practical length of $\mathbf{W}$, that is a natural number for which

$$\mathbf{W}(z) \approx \sum_{k=0}^{s} W_k z^{-k}. \tag{15}$$

Since $\mathbf{p}(z)\mathbf{Q}(z)^{-1} = \mathbf{W}^{-1}(z)\mathbf{W}(z^{-1})^{-T}$, then a model for the ARMA process is

$$\mathbf{W}(z)y(t) = e(t) \tag{16}$$

where e is normalized white Gaussian noise. By (15), we can approximate (16) with the high order AR process

$$\sum_{k=0}^{s} W_k y(t-k) = e(t).$$

The negative log-likelihood of this model is [38]

$$\ell(y^N;\mathbf{p},\mathbf{Q})$$

$$= \frac{N-s}{2}\int_{\mathbb{T}} -\log\det(\mathbf{G}^*\mathbf{G}) + \frac{N}{N-s}\mathrm{tr}(\mathbf{G}^*\mathbf{G}\hat{\Phi}_P)$$

$$\approx \frac{N}{2}\int_{\mathbb{T}} -\log\det(\mathbf{G}^*\mathbf{G}) + \mathrm{tr}(\mathbf{G}^*\mathbf{G}\hat{\Phi}_P)$$

where $\mathbf{G}(z) = \sum_{k=0}^{s} W_k z^{-k}$ and we exploited the fact that $N \gg s$. Finally, since $\mathbf{G}\approx\mathbf{W}$ and, thus, $\mathbf{G}^*\mathbf{G}\approx\mathbf{W}^*\mathbf{W} = \mathbf{p}^{-1}\mathbf{Q}$ we obtain (14). ■

Taking into account (13), we conclude that for N sufficiently large

$$\ell(y^N;\mathbf{p},\mathbf{Q}) \approx \frac{N}{2}\left[\frac{J(\mathbf{p},\mathbf{Q})}{\mathbf{p}(e^{i\bar{\theta}})} + \int\log\det(\hat{\Phi}_P) + m\right]. \tag{17}$$

Although Problem (2) provides a tool to estimate an ARMA graphical model from the data, it is required to know the topology, i.e., the set E. In the following, we will show how to address the problem in the case the topology is not known.

VI. Bayesian Viewpoint

The aim of this section is to show how to regularize the dual problem (5) in order to estimate an ARMA graphical model in the case E is not known. More precisely, we take the Bayesian perspective proposed in [29], and thus, we model $\mathbf{p}$ and $\mathbf{Q}$ as stochastic trigonometric polynomials (i.e., p_ks and the entries of Q_ks are random variables) taking values

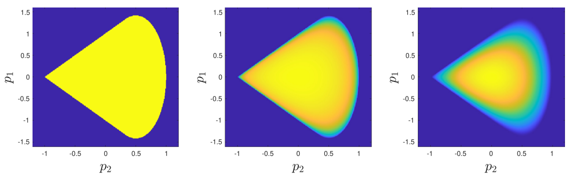

Fig. 1. *Left.* Yellow region denotes the values of p_1 and p_2 such that the corresponding $p \in \mathfrak{B}_+^o$; *Center.* Prior with $\lambda = 0.2$, hot colors indicate high probability while cold colors low probability; *Right.* Prior with $\lambda = 1$.

in $\overline{\mathfrak{B}_+^o}$ and $\overline{\mathfrak{B}_+^m} := \overline{\mathfrak{B}_+^m}(V \times V)$, respectively. We also assume they are independent. Let $f(\mathbf{Q}|\boldsymbol{\gamma})$ be the prior of $\mathbf{Q}$ where $\boldsymbol{\gamma}$ denotes the vector containing $\gamma_{jh} > 0$, $j \geq h$, which are referred to as hyperparameters. Recall that $[\mathbf{Q}]_{jh}$ is the entry of $\mathbf{Q}$ in position (j, h), that is $[\mathbf{Q}(e^{i\theta})]_{jh} = [Q_0]_{jh} + 0.5\sum_{k=1}^n [Q_k]_{jh}e^{-i\theta k} + [Q_k]_{hj}e^{i\theta k}$. We assume that $[\mathbf{Q}]_{jh}$s are independent each other. Thus, $f(\mathbf{Q}|\boldsymbol{\gamma}) = \prod_{j\geq h} f([\mathbf{Q}]_{jh}|\boldsymbol{\gamma})$ and $f([\mathbf{Q}]_{jh}|\boldsymbol{\gamma}) = e^{-\gamma_{jh}q_{jh}(\mathbf{Q})}/c_{jh}$ where

$$q_{jh}(\mathbf{Q}) := \max\{|[Q_0]_{jh}|, \max_{k=1\ldots n} |[Q_k]_{jh}|, \max_{k=1\ldots n} |[Q_k]_{hj}|\}$$

and c_{jh} is the normalizing constant. In plain words, $[\mathbf{Q}]_{jh}$ given $\boldsymbol{\gamma}$ follows an exponential distribution, which is the most principled way to incorporate regularization into Bayesian models. The hyperparameter γ_{jh} determines the likelihood of the trigonometric polynomial $[\mathbf{Q}]_{jh}$ being identically equal to zero. For further details, including the motivation of this prior in the Fisherian viewpoint, see [29]. Therefore, we obtain

$$f(\mathbf{Q}|\boldsymbol{\gamma}) = e^{-h_W(\mathbf{Q})}/\prod_{j\geq h} c_{jh}$$

where

$$h_W(\mathbf{Q}) = \sum_{j\geq h} \gamma_{jh} q_{jh}(\mathbf{Q}).$$

Moreover, we model γ_{jh}s as independent random variables with exponential distribution:

$$f(\boldsymbol{\gamma}) = \prod_{j\geq h} \varepsilon e^{-\varepsilon\gamma_{jh}}$$

where $\varepsilon > 0$ is a fixed small constant. Let $f(\mathbf{p})$ be the prior of $\mathbf{p}$. We assume that

$$f(\mathbf{p}) = c^{-1}e^{-g_\lambda(\mathbf{p})} \tag{18}$$

where c is the normalizing constant and $\lambda > 0$ is the corresponding hyperparameter. It is worth noting that λ plays the same role of the regularization parameter in (7), and thus, we assume it is fixed.

Example 1: Consider the trigonometric polynomial $\mathbf{p}(e^{i\theta}) = 1 + p_1\cos(\theta) + p_2\cos(2\theta)$ with $n = 2$. Fig. 1 shows the prior (18) for $\lambda = 0.2$ and $\lambda = 1$. We notice that the larger λ is, the more the neighborhood of high-probability around $\mathbf{p}(e^{i\theta}) = 1$ is shrunk.

The negative log-likelihood of y^N, $\mathbf{p}$, $\mathbf{Q}$, and $\boldsymbol{\gamma}$ takes the form

$$\ell(y^N, \mathbf{p}, \mathbf{Q}, \boldsymbol{\gamma}) = -\log f(y^N, \mathbf{p}, \mathbf{Q}, \boldsymbol{\gamma})$$

$$= -\log f(y^N|\mathbf{p}, \mathbf{Q}) - \log f(\mathbf{p}) - \log f(\mathbf{Q}|\boldsymbol{\gamma}) - \log f(\boldsymbol{\gamma}).$$

Since, $-\log f(y^N|\mathbf{p}, \mathbf{Q})$ coincides with $\ell(y^N; \mathbf{p}, \mathbf{Q})$ (i.e., the likelihood function considered in Section V), we have

$$\ell(y^N, \mathbf{p}, \mathbf{Q}; \boldsymbol{\gamma}) = \ell(y^N; \mathbf{p}, \mathbf{Q}) + g_\lambda(\mathbf{p}) + h_W(\mathbf{Q}) + \sum_{j\geq h} \log c_{jh} + \varepsilon\gamma_{jh} \tag{19}$$

where we have neglected the constant terms not depending on $\mathbf{p}$, $\mathbf{Q}$, and $\boldsymbol{\gamma}$. The MAP estimator of $(\mathbf{p}, \mathbf{Q})$ is given by minimizing $\ell(y^N, \mathbf{p}, \mathbf{Q}, \boldsymbol{\gamma})$ with respect to $(\mathbf{p}, \mathbf{Q}) \in \overline{\mathfrak{B}_+^o} \times \overline{\mathfrak{B}_+^m}$. The hyperparameter vector $\boldsymbol{\gamma}$, however, is not known. According to the empirical Bayes method [39], it is possible to find an estimate $\hat{\boldsymbol{\gamma}}$ by minimizing the so called negative log-marginal likelihood with respect to $\boldsymbol{\gamma}$

$$-\log \int_{\overline{\mathfrak{B}_+^o}\times\overline{\mathfrak{B}_+^m}} f(y^N, \mathbf{p}, \mathbf{Q}, \boldsymbol{\gamma})\mathrm{d}\mathbf{p}\mathrm{d}\mathbf{Q}. \tag{20}$$

Then, the MAP estimator of $(\mathbf{p}, \mathbf{Q})$ is given by minimizing $\ell(y^N, \mathbf{p}, \mathbf{Q}, \hat{\boldsymbol{\gamma}})$. However, it is difficult to find an analytic expression for (20) making challenging the optimization of $\boldsymbol{\gamma}$. One way to overcome this issue is to consider an upper bound for (19). By (17) and Proposition 3 in [29], the following inequality approximately holds for N sufficiently large

$$\ell(y^N, \mathbf{p}, \mathbf{Q}, \boldsymbol{\gamma}) \leq \tilde{\ell}(y^N, \mathbf{p}, \mathbf{Q}, \alpha, \boldsymbol{\gamma}) \tag{21}$$

where

$$\begin{aligned}\tilde{\ell}(y^N, \mathbf{p}, \mathbf{Q}, \alpha, \boldsymbol{\gamma}) :=\ & \frac{N}{2}[\alpha^{-1}J(\mathbf{p}, \mathbf{Q}) + \int \log\det(\hat{\Phi}_P) + m] \\ & + g_\lambda(\mathbf{p}) + h_W(\mathbf{Q}) - \sum_{j>h}(2n+1)\log\gamma_{jh} \\ & - \sum_{j=1}^n (n+1)\log\gamma_{jj} + \sum_{j\geq h}\varepsilon\gamma_{jh}\end{aligned} \tag{22}$$

$\alpha = \mathbf{p}(e^{i\bar{\theta}})$ and we neglected the constant terms in inequality (21). Then, a simplified approach to estimate $\boldsymbol{\gamma}$ is the generalized maximum likelihood (GML) method [40]: this hyperparameter is jointly optimized with $\mathbf{p}$ and $\mathbf{Q}$ minimizing (22). However, the optimization with respect to $\mathbf{p}$ is difficult because of the presence of α. Accordingly, we can find a preliminary estimate $(\mathbf{p}^{(0)}, \mathbf{Q}^{(0)})$ and then, in view of (13), an estimate of α is given by

$$\hat{\alpha} = \frac{J(\mathbf{p}^{(0)}, \mathbf{Q}^{(0)})}{\int_{\mathbb{T}} \log\det(\mathbf{p}^{(0)}(\mathbf{Q}^{(0)}\hat{\Phi}_P)^{-1}) + \mathrm{tr}((\mathbf{p}^{(0)})^{-1}\mathbf{Q}^{(0)}\hat{\Phi}_P) - m}. \tag{23}$$

Theorem 2: Let $\alpha > 0$ and assume that there exists a power spectral density $\bar{\Phi}$, which is positive definite on some open interval in $\mathbb{T}$ and such that

$$\int_{\mathbb{T}} \bar{\Phi}e^{i\theta k} = R_k, \quad k = 0, \ldots, n. \tag{24}$$

Then, the optimization problem

$$\min_{(\mathbf{p}, \mathbf{Q}, \boldsymbol{\gamma})\in\mathcal{C}} \tilde{\ell}(y^N, \mathbf{p}, \mathbf{Q}, \alpha, \boldsymbol{\gamma})$$

with $\mathcal{C} := \{(\mathbf{p}, \mathbf{Q}, \boldsymbol{\gamma}) \text{ s.t. } \mathbf{p} \in \overline{\mathfrak{B}_+^o},\ \mathbf{Q} \in \overline{\mathfrak{B}_+^m},\ \gamma_{jh} > 0\}$, admits solution. Moreover, the latter is such that $(\mathbf{p}, \mathbf{Q}) \in \mathfrak{B}_+^o \times \mathfrak{B}_+^m$.

The proof is deferred after Theorem 3. Since the joint optimization of $\mathbf{p}, \mathbf{Q}, \boldsymbol{\gamma}$ is still difficult, we perform the optimization of $(\mathbf{p}, \mathbf{Q})$ and $\boldsymbol{\gamma}$ iteratively through a two-step algorithm where at the l-iteration the variables are estimated as follows:

$$(\hat{\mathbf{p}}^{(l)}, \hat{\mathbf{Q}}^{(l)}) = \underset{(\mathbf{p},\mathbf{Q}) \in \overline{\mathfrak{B}_+^o} \times \overline{\mathfrak{B}_+^m}}{\operatorname{argmin}} \tilde{\ell}(y^N, \mathbf{p}, \mathbf{Q}, \hat{\alpha}, \hat{\boldsymbol{\gamma}}^{(l)}) \tag{25}$$

$$\hat{\boldsymbol{\gamma}}^{(l+1)} = \underset{\gamma_{jh} > 0}{\operatorname{argmin}}\, \tilde{\ell}(y^N, \hat{\mathbf{p}}^{(l)}, \hat{\mathbf{Q}}^{(l)}, \hat{\alpha}, \boldsymbol{\gamma}). \tag{26}$$

It is not difficult to see that Problem (25) is equivalent to the following problem:

$$\begin{aligned} &\min_{\mathbf{p},\mathbf{Q}} J(\mathbf{p}, \mathbf{Q}) + \frac{2\alpha}{N}[g_\lambda(\mathbf{p}) + h_W(\mathbf{Q})] \\ &\text{s.t. } \mathbf{p} \in \overline{\mathfrak{B}_+^o},\ \mathbf{Q} \in \overline{\mathfrak{B}_+^m} \end{aligned} \tag{27}$$

with $\alpha = \hat{\alpha}$ and $\boldsymbol{\gamma} = \hat{\boldsymbol{\gamma}}^{(l)}$ which is a regularized version of the dual problem in (5). In particular, the regularization term $g_\lambda(\mathbf{p})$ avoids that the optimal solution $\mathbf{p}$ is on the boundary, while $h_W(\mathbf{Q})$ induces sparsity in $\mathbf{Q}$ in a soft way; finally, the factor $2\alpha/N$ tunes the best tradeoff between the "fit function" and the regularization terms.

Theorem 3: Let $\alpha > 0$ and $\boldsymbol{\gamma}$ be such that $\gamma_{jh} \geq 0$. Moreover, assume that there exists a power spectral density $\bar{\Phi}$, which is positive definite on some open interval in $\mathbb{T}$ and such that (24) holds. Then, Problem (27) admits a unique solution $(\hat{\mathbf{p}}, \hat{\mathbf{Q}}) \in \mathfrak{B}_+^o \times \mathfrak{B}_+^m$.

Proof: First, notice that the results in Section IV also holds in the particular case $E = V \times V$. Moreover, the objective function in Problem (27) can be written as

$$\check{J}_{\lambda,\alpha,\boldsymbol{\gamma}}(\mathbf{p}, \mathbf{Q}) = \tilde{J}_{\frac{2\alpha\lambda}{N}}(\mathbf{p}, \mathbf{Q}) + \frac{2\alpha}{N} h_W(\mathbf{Q})$$

where $\tilde{J}$ is the regularized dual function defined in (10) and recall that $\lambda > 0$. In plain words, the difference between $\tilde{J}$ and $\check{J}$ is that in the latter appears the term $s(\mathbf{Q}) := 2\alpha N^{-1} h_W(\mathbf{Q})$. Then, it follows that $\check{J}$ is strictly convex in $\mathfrak{B}_+^o \times \mathfrak{B}_+^m$ because $\tilde{J}$ is strictly convex, by Proposition 1, and $s(\mathbf{Q})$ is convex. Moreover, $\check{J}$ is lower semicontinuous on $\overline{\mathfrak{B}_+^o} \times \overline{\mathfrak{B}_+^m}$ because $\tilde{J}$ is lower semicontinuous, by Proposition 2, and $s(\mathbf{Q})$ is continuous.

Next, take a sequence $\{(\mathbf{p}_j, \mathbf{Q}_j) \in \overline{\mathfrak{B}_+^o} \times \overline{\mathfrak{B}_+^m},\ j \in \mathbb{N}\}$ such that $\|(\mathbf{p}_j, \mathbf{Q}_j)\| \to \infty$ as $j \to \infty$. Then, in view of Proposition 3, we have

$$\liminf_{j\to\infty} \check{J}_{\lambda,\alpha,\boldsymbol{\gamma}}(\mathbf{p}_j, \mathbf{Q}_j) \geq \liminf_{j\to\infty} \tilde{J}_{\frac{2\alpha\lambda}{N}}(\mathbf{p}_j, \mathbf{Q}_j) \to \infty.$$

Thus, this sequence cannot be an infimizing one.

Finally, consider a sequence $\{(\mathbf{p}_j, \mathbf{Q}_j) \in \mathfrak{B}_+^o \times \mathfrak{B}_+^m,\ j \in \mathbb{N}\}$ converging to $(\mathbf{p}, \mathbf{Q})$, which belongs to the boundary of $\overline{\mathfrak{B}_+^o} \times \overline{\mathfrak{B}_+^m}$. The latter cannot be an infimizing sequence. This statement can be proved using the same arguments exploited in the proof of Proposition 4: the unique substantial difference regards the case $\mathbf{p} \in \mathfrak{B}_+^o$ and $\mathbf{Q} \in \partial\mathfrak{B}_+^m$. The function $\check{J}_{\lambda,\alpha,\boldsymbol{\gamma}}(\mathbf{p}, \cdot)$ is differentiable at $\mathbf{Q} + tI$, with $t > 0$, along the direction $\delta\mathbf{Q} = I$. More precisely, the first variation is

$$\begin{aligned} &\delta\check{J}_{\lambda,\alpha,\boldsymbol{\gamma}}(\mathbf{p}, \mathbf{Q} + tI; I) \\ &\quad = -\int_{\mathbb{T}} \mathbf{p}\operatorname{tr}(\mathbf{Q} + tI)^{-1} + \operatorname{tr}(R_0) + \sum_j \frac{2\alpha\gamma_{jj}}{N} \\ &\quad \leq -\mu \int_{\mathbb{T}} \operatorname{tr}(\mathbf{Q} + tI)^{-1} + \operatorname{tr}(R_0) + \sum_j \frac{2\alpha\gamma_{jj}}{N} \end{aligned}$$

and by the Lebesgue's monotone convergence theorem, we have

$$\begin{aligned} &\lim_{t\to 0^+} \delta\check{J}_{\lambda,\alpha,\boldsymbol{\gamma}}(\mathbf{p}, \mathbf{Q} + tI; I) \\ &\quad \leq -\mu \int_{\mathbb{T}} \operatorname{tr}(\mathbf{Q}^{-1}) + \operatorname{tr}(R_0) + \sum_j \frac{2\alpha\gamma_{jj}}{N} = -\infty. \end{aligned}$$

In the light of the abovementioned results, taken $r \in \mathbb{R}$ sufficiently large, we can restrict the search of the minimum over the sublevel set $\{\, (\mathbf{p}, \mathbf{Q}) \in \mathfrak{B}_+^o \times \mathfrak{B}_+^m \text{ s.t. } \check{J}_{\lambda,\alpha,\boldsymbol{\gamma}}(\mathbf{p}, \mathbf{Q}) \leq r \,\}$, which is compact. Since $\check{J}$ is strictly convex over this sublevel set and by the Weierstrass' theorem, we conclude that Problem (27) admits solution which is also unique. ■

Remark 3: It is not difficult to see that, keeping γ_{jj}s fixed, for any $j > h$ it is possible to make the solution to (27) such $[\hat{\mathbf{Q}}]_{jh}$ is arbitrarily close to zero taking γ_{jh} sufficiently large.

We are ready to prove Theorem 2.

Proof of Theorem 2: Since $\mathcal{C}$ is open and unbounded, we need to show that we can restrict the search of the minimum of $\tilde{\ell}$ over a compact sect $\mathcal{C}^\star \subset \mathcal{C}$ for which $\tilde{\ell}$ is continuous, then the claim follows by the Weierstrass' theorem. Moreover, we must show that $(\mathbf{p}, \mathbf{Q}) \in \mathfrak{B}_+^o \times \mathfrak{B}_+^m$ for any $(\mathbf{p}, \mathbf{Q}, \boldsymbol{\gamma}) \in \mathcal{C}^\star$. Notice that

$$\begin{aligned} &\tilde{\ell}(y^N, \mathbf{p}, \mathbf{Q}, \alpha, \boldsymbol{\gamma}) = \frac{N}{2\alpha} \check{J}_{\lambda,\alpha,\boldsymbol{\gamma}}(\mathbf{p}, \mathbf{Q}) + \kappa \\ &\quad - \sum_{j>h} (2n+1)\log\gamma_{jh} - \sum_{j=1}^n (n+1)\log\gamma_{jj} + \sum_{j\geq h} \varepsilon\gamma_{jh} \\ &\quad \geq \frac{N}{2\alpha} \check{J}_{\lambda,\alpha,\mathbf{0}}(\mathbf{p}, \mathbf{Q}) + \kappa - \sum_{j>h} (2n+1)\log\gamma_{jh} \\ &\quad - \sum_{j=1}^n (n+1)\log\gamma_{jj} + \sum_{j\geq h} \varepsilon\gamma_{jh} \end{aligned} \tag{28}$$

where κ contains the terms not depending on $\mathbf{p}$, $\mathbf{Q}$, and $\boldsymbol{\gamma}$. Now, consider a sequence $\{(\mathbf{p}_j, \mathbf{Q}_j, \boldsymbol{\gamma}_j) \in \mathcal{C},\ j \in \mathbb{N}\}$ such that there exists at least one $\gamma_{jh} \to \infty$. Then, in view of (28)

$$\lim_{j\to\infty} \tilde{\ell}(y^N, \mathbf{p}_j, \mathbf{Q}_j, \alpha, \boldsymbol{\gamma}_j) = \infty$$

because, by Theorem 3, $\check{J}_{\lambda,\alpha,\mathbf{0}}$ is bounded from below on $\overline{\mathfrak{B}_+^o} \times \overline{\mathfrak{B}_+^m}$. Next, take a sequence $\{(\mathbf{p}_j, \mathbf{Q}_j, \boldsymbol{\gamma}_j) \in \mathcal{C},\ j \in \mathbb{N}\}$ such that there exists at least one $\gamma_{jh} \to 0$. Then, by (28), we have that $\tilde{\ell}(y^N, \mathbf{p}_j, \mathbf{Q}_j, \alpha, \boldsymbol{\gamma}_j) \to \infty$ as $j \to \infty$ because $\check{J}_{\lambda,\alpha,\mathbf{0}}$ is bounded from below. Therefore, we can restrict the search of the minimum over the set $\tilde{\mathcal{C}} := \{(\mathbf{p}, \mathbf{Q}, \boldsymbol{\gamma}) \in \mathcal{C} \text{ s.t. } \delta \leq \gamma_{jh} \leq \mu\}$ where $\mu > \delta > 0$. Finally, by Theorem 3, we can restrict the search of the minimum over

$$\mathcal{C}^\star := \{(\mathbf{p}, \mathbf{Q}, \boldsymbol{\gamma}) \in \mathcal{C} \text{ s.t. } (\mathbf{p}, \mathbf{Q}) \in \mathcal{A},\ \delta \leq \gamma_{jh} \leq \mu\}$$

Algorithm 1: GML estimator.

Compute $(\mathbf{p}^{(0)}, \mathbf{Q}^{(0)})$ as solution of (7) with $E = V \times V$
Compute $\hat{\alpha}$ as in (23)
Set $\hat{\gamma}^{(1)} = 0$ and $l = 0$
Repeat
  $l = l + 1$
  Compute $(\hat{\mathbf{p}}^{(l)}, \hat{\mathbf{Q}}^{(l)})$ as solution of (27)
  Compute $\hat{\gamma}^{(l+1)}$ through (29)
until $\|\hat{\mathbf{Q}}^{(l)} - \hat{\mathbf{Q}}^{(l-1)}\| \leq \epsilon$

and $\mathcal{A} \subset \mathfrak{B}_+^o \times \mathfrak{B}_+^m$ is a compact set. Accordingly, $\mathcal{C}^\star$ is a compact set. ∎

It is not difficult to see that (26) is the same optimization problem in Section IV of [29] and it admits the following closed form solution:

$$\hat{\gamma}_{jh}^{(l+1)} = \begin{cases} \frac{n+1}{q_{jh}(\hat{\mathbf{Q}}^{(l)})+\varepsilon}, & \text{if } j = h \\ \frac{2n+1}{q_{jh}(\hat{\mathbf{Q}}^{(l)})+\varepsilon}, & \text{if } j > h. \end{cases} \tag{29}$$

The resulting estimator, hereafter called GML estimator, is described in Algorithm 1 where $\epsilon > 0$ is a fixed threshold.

Corollary 1: Let $(\hat{\mathbf{p}}^{(l)}, \hat{\mathbf{Q}}^{(l)}, \hat{\gamma}^{(l+1)})$, with $l \in \mathbb{N}$, be the sequence generated by Algorithm 1 where $\hat{\alpha}$ is fixed. If there is a limit point $(\hat{\mathbf{p}}, \hat{\mathbf{Q}}, \hat{\gamma})$ of such a sequence, then it is a coordinatewise minimum point of $\tilde{\ell}$, i.e.,

$$\tilde{\ell}(y^N, \hat{\mathbf{p}}, \hat{\mathbf{Q}}, \hat{\alpha}, \hat{\gamma}) \leq \tilde{\ell}(y^N, \mathbf{p}, \mathbf{Q}, \hat{\alpha}, \hat{\gamma})$$
$$\tilde{\ell}(y^N, \hat{\mathbf{p}}, \hat{\mathbf{Q}}, \hat{\alpha}, \hat{\gamma}) \leq \tilde{\ell}(y^N, \hat{\mathbf{p}}, \hat{\mathbf{Q}}, \hat{\alpha}, \gamma)$$

for any $\mathbf{p} \in \overline{\mathfrak{B}_+^o}$, $\mathbf{Q} \in \overline{\mathfrak{B}_+^m}$ and $\gamma_{jh} > 0$.

Proof: The statement is supported by the fact that at every stage of the sequential procedure, we are able to identify a specific variable for which the function $\tilde{\ell}$ attains its minimum value. ∎

It is worth noting one could consider the following upper bound for $\ell(y^N, \mathbf{p}, \mathbf{Q}, \gamma)$, which considers the likelihood of the data

$$\breve{\ell}(y^N, \mathbf{p}, \mathbf{Q}, \gamma) := \ell(y^N; \mathbf{p}, \mathbf{Q}) + g_\lambda(\mathbf{p}) + h_W(\mathbf{Q}) - \sum_{j>h}(2n+1)\log\gamma_{jh} - \sum_{j=1}^{n}(n+1)\log\gamma_{jj} + \sum_{j\geq h}\varepsilon\gamma_{jh} \tag{30}$$

with $\ell(y^N; \mathbf{p}, \mathbf{Q})$ given by the approximation in (14). This leads to a two-step algorithm: in the first step $\breve{\ell}$ is optimized with respect to $(\mathbf{p}, \mathbf{Q})$, while in the second step with respect to γ. However, $\breve{\ell}$ is not convex in $(\mathbf{p}, \mathbf{Q})$, and thus, it is plausible to obtain a local minimum, which is not a global one in the first step. As a consequence, the sequence generated by the corresponding sequential algorithm does not enjoy the property in Corollary 1. Accordingly, the upper bound $\tilde{\ell}$ aims to "convexify" the problem in $(\mathbf{p}, \mathbf{Q})$ with the introduction of the additional parameter α and guarantees the sequence satisfies the property in Corollary 1.

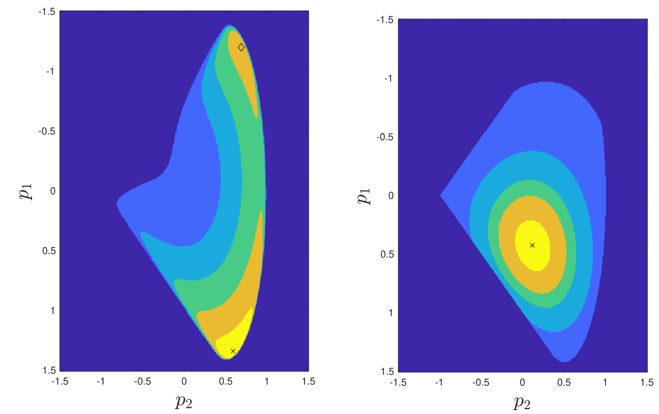

Fig. 2. Some sublevel sets of $(p_1, p_2) \mapsto \breve{\ell}(y^N, \mathbf{p}, \mathbf{Q}^\star, \gamma)$, left, and $(p_1, p_2) \mapsto \tilde{\ell}(y^N, \mathbf{p}, \mathbf{Q}^\star, \gamma)$, right. The sublevel sets depicted with hot colors correspond to a smaller constant than the one of the sublevel sets depicted with cold colors. The former are overlapped to the latter. The global minima are denoted by a cross, while the other local minima by a diamond.

VII. Numerical Experiments

A. Local Minima Using $\breve{\ell}$

As we noticed in Section VI, an upper bound for the negative log-likelihood $\ell(y^N, \mathbf{p}, \mathbf{Q}, \gamma)$ is given also by $\breve{\ell}$ defined in (30). However, the minimization of $\breve{\ell}$ with respect to $(\mathbf{p}, \mathbf{Q})$ could lead to a local minimum, which is not a global one. Next, we provide a numerical example in which this situation happens.

We consider an ARMA graphical model with $m = 6$ nodes, order $n = 2$, and with $E := \{(1,1), (1,4), (1,6), (2,2), (3,3), (3,4), (3,6), (4,4), (4,6), (5,5), (5,6), (6,6)\}$. From this model, we generate a data sequence y^N of length $N = 500$. Then, we fix the hyperparameter vector γ as follows:

$$\gamma_{jh} = \begin{cases} 10^3, & \text{if } (j,h) \in E \\ 5 \cdot 10^4, & \text{if } (j,h) \notin E \end{cases}$$

that is we consider a hyperparameter, which induces the correct sparsity pattern on $\mathbf{Q}$, but it induces a little more bias than required in the other entries. Finally, we set $\lambda = 1$. Then, we solve numerically the minimization problem

$$(\hat{\mathbf{p}}^\star, \hat{\mathbf{Q}}^\star) = \underset{(\mathbf{p},\mathbf{Q}) \in \overline{\mathfrak{B}_+^o} \times \overline{\mathfrak{B}_+^m}}{\operatorname{argmin}} \breve{\ell}(y^N, \mathbf{p}, \mathbf{Q}, \gamma). \tag{31}$$

Fig. 2(left) shows some sublevel sets of the function $(p_1, p_2) \mapsto \breve{\ell}(y^N, \mathbf{p}, \mathbf{Q}^\star, \gamma)$, recall that $\mathbf{p}$ is characterized by the variables p_1, p_2 see (3). We can see that there are two local minima and one of them is not a global minimum. Now, we do the same thing with $\tilde{\ell}$. More precisely, we solve numerically the minimization problem (31) where replace the objective function with $\tilde{\ell}(y^N, \mathbf{p}, \mathbf{Q}, \hat{\alpha}, \gamma)$ and $\hat{\alpha}$ is computed as in the preliminary step of Algorithm 1. Fig. 2(right) shows some sublevel sets of the function $(p_1, p_2) \mapsto \tilde{\ell}(y^N, \mathbf{p}, \mathbf{Q}^\star, \hat{\alpha}, \gamma)$. In this case, we only have the global minimum. Fig. 3 shows the pointwise Frobenius norm of the spectrum $\Phi^\star = \mathbf{p}(\mathbf{Q}^\star)^{-1}$ when: $(\mathbf{p}, \mathbf{Q}^\star)$ is the global minimum of $\breve{\ell}$; $(\mathbf{p}, \mathbf{Q}^\star)$ is the other local minimum of

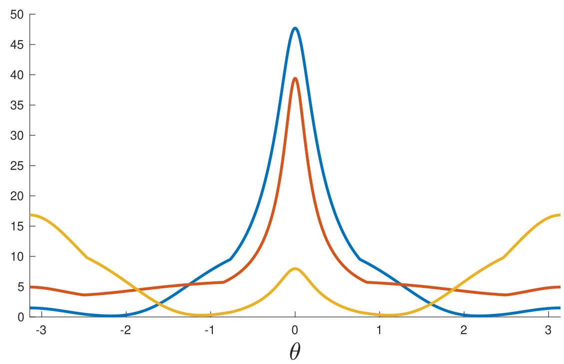

Fig. 3. Pointwise Frobenius norm of $\Phi^\star = \mathbf{p}(\mathbf{Q}^\star)^{-1}$ when: $(\mathbf{p}, \mathbf{Q}^\star)$ is the global minimum of $\check{\ell}(y^N, \cdot, \cdot, \gamma)$ (blue line); $(\mathbf{p}, \mathbf{Q}^\star)$ is the other local minimum of $\check{\ell}(y^N, \cdot, \cdot, \gamma)$ (yellow line); $(\mathbf{p}, \mathbf{Q}^\star)$ is the global minimum of $\tilde{\ell}(y^N, \cdot, \cdot, \hat{\alpha}, \gamma)$ (red line).

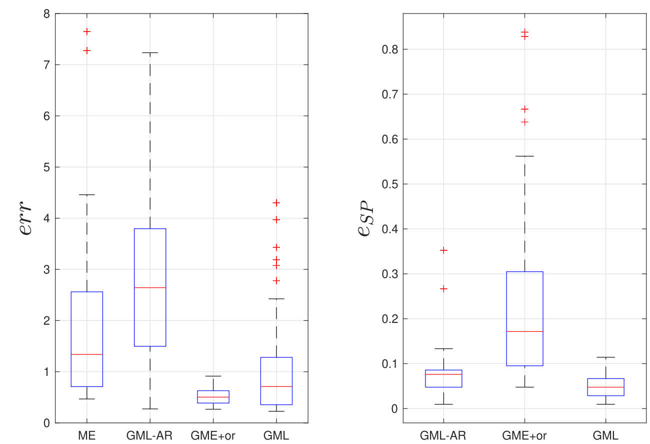

Fig. 4. First Monte Carlo experiment with $N = 500$. Boxplots of the relative error (left) and the fraction of misspecified edges (right).

$\check{\ell}$; $(\mathbf{p}, \mathbf{Q}^\star)$ is the global minimum of $\tilde{\ell}$. While the two spectra corresponding to the global minima of $\check{\ell}$ and $\tilde{\ell}$ are similar, the spectrum corresponding to the local minimum of $\check{\ell}$ is very different.

B. Performance Analysis

We aim to test the performance of the proposed estimator in respect to the methods available in the literature through some Monte Carlo experiments. Each experiment is composed by $M = 100$ trials and in each trial we randomly generate an ARMA graphical model with power spectral density $\Phi_T = \mathbf{p}_T \mathbf{Q}_T^{-1}$ with $m = 15$ nodes and order $n = 2$. More precisely, the matrix coefficients of $\mathbf{Q}_T$ are generated in such a way that the fraction of nonnull entries in $\mathbf{Q}_T$ is equal to 0.17. The coefficients of $\mathbf{p}_T$ are generated in such a way that $\mathbf{p}_T$ has a stable zero whose modulus is greater than 0.98. In this way, the resulting model will not be well approximated by a low-order AR model. Finally, we generate a data sequence y^N of length N from such model. Then, we estimate the model from y^N using the following estimators.

1) *ME* that denotes the maximum entropy estimator for ARMA models and whose parameters are given by solving (7) with $E = V \times V$; this method represents the multivariate extension to the maximum entropy method proposed in [20]; here $\lambda = 1$.
2) *GML-AR* that denotes the GML method for estimating AR graphical models [29]; here $\varepsilon = 10^{-4}$ and $\epsilon = 10^{-8}$.
3) *GME+or* that denotes the generalized maximum entropy method to estimate ARMA graphical models, proposed in [27], and equipped with an "oracle"; more precisely, the oracle knows the true trigonometric polynomial $\mathbf{p}_T$; here the grid for the values of the regularization parameter γ has been chosen as in [27], that is $\{0, 0.02, 0.04, 0.06, 0.08, 0.1, 0.3, 0.5, 1\}$; the best γ is selected according to the BIC criterium; finally, the threshold for the stopping condition is $\epsilon = 10^{-8}$.
4) *GML* that denotes the estimator proposed in Section VI with $\lambda = 1$, $\varepsilon = 10^{-4}$, and $\epsilon = 10^{-8}$.

To assess the performance of the estimators, we compute the fraction of misspecified edges with respect to the true model

$$e_{\text{SP}} = \frac{\|\mathbf{E} - \hat{\mathbf{E}}\|_0}{m^2}$$

where $\mathbf{E} = \operatorname{supp}(\Phi_T^{-1})$, $\hat{\mathbf{E}} = \operatorname{supp}(\hat{\Phi}^{-1})$, $\hat{\Phi}$ denotes the estimator of Φ_T, and $\|\cdot\|_0$ denotes the ℓ_0 matrix norm. Moreover, we compute the relative error of $\hat{\Phi}^{-1}$ with respect to the true one Φ_T^{-1}

$$\text{err} = \frac{\int_{\mathbb{T}} \|\hat{\Phi}^{-1} - \Phi_T^{-1}\|_F}{\int_{\mathbb{T}} \|\Phi_T^{-1}\|_F}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. In the first Monte Carlo experiment, we have considered $N = 500$. Fig. 4 (left) shows the boxplots of err for the different estimators. *GML-AR* exhibits the worst performance because it is not able to approximate the ARMA model through a low order AR model. *ME* performs worse than *GME+or* and *GML* because it estimates a full graphical model, i.e., this method is not able to control the model complexity in terms of number of edges. Finally, *GME+or* exhibits the best performance thanks to its oracle. Fig. 4(right) shows the boxplots of e_{SP} for *GML-AR*, *GME+or*, and *GML*. Notice that, we did not plot the boxplot for *ME* because it always estimates a full graph, i.e., the fraction of misspecified edges is close to 1. *GME+or* exhibits the worst performance because the regularization term used to induce sparsity is

$$\gamma \sum_{j \geq h} q_{jh}(\mathbf{Q})$$

which penalizes the elements of $\mathbf{Q}$ in the same way. Thus, it introduces an adverse bias in the case $q_{jh}(\mathbf{Q})$s take very different values. In addition, the fact that the grid is fixed (otherwise the user should change it for any trial) could degradate the performance of the estimator. Finally, *GML* performs slightly better than *GML-AR*. Their performance is similar because they use the same regularization term to induce sparsity. In the second Monte Carlo experiment, we have considered $N = 1000$. Fig. 5

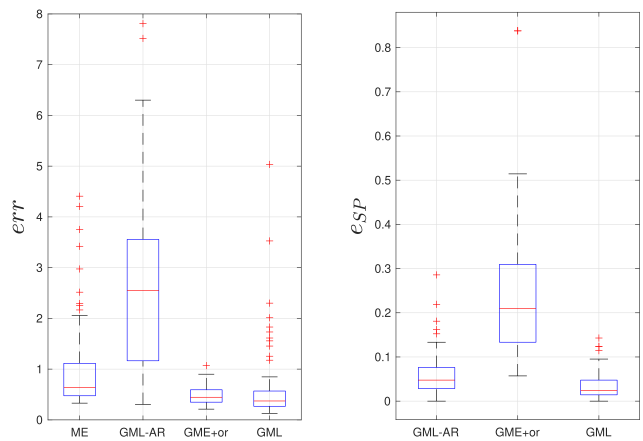

Fig. 5. Second Monte Carlo experiment with $N = 500$. Boxplots of the relative error (left) and the fraction of misspecified edges (right).

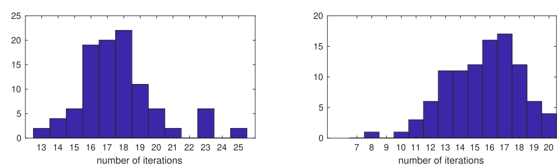

Fig. 6. Distribution of the number of performed iterations of the reweighting scheme for GML in the Monte Carlo experiments with $N = 500$ (left) and $N = 1000$ (right).

shows the boxplots of err and $e_{\rm SP}$ for the different estimators. In regard to the relative error, the performance of *ME* is improved in respect to the case $N = 500$ because increasing the data length the overfitting phenomenon is less evident. However, *ME* is still worse than *GML+or* and *GML*. Finally, *GML* is able to achieve a performance similar to the one of *GME+or*. In regard to the fraction of misspecified edges, the scenario is similar to the case $N = 500$, but the performance for all the estimators now is slightly improved. To conclude, these two Monte Carlo experiments show the superiority of the proposed estimator in the case the data is generated by an ARMA model.

Remark 4: The computational complexity of the GML estimator depends on the number of iterations performed by the reweighting scheme (which cannot be determined a priori). Fig. 6 shows how the number of performed iterations of the reweighting scheme is distributed in the two Monte Carlo experiments of before: we can see that the number of iterations decreases when the data length N increases. For any iteration, the updating of γ is very fast, indeed, as shown in (29), it has a closed-form expression. The bottleneck is the computation of the solution to Problem (27). Such solution is found numerically through a gradient descent method whose computational complexity of each iteration is $O((m^3 + m^2 n)M)$. Here, $M = 1000$ is the number of points used to grid the frequencies interval $[0, 2\pi]$ for computing numerically the integrals needed to evaluate the objective function and its gradient.

VIII. Conclusion

In this article, we have introduced a maximum entropy covariance and cepstral extension problem for graphical models. Under mild assumptions, we showed that the problem admits an approximate solution, which represents an ARMA graphical model whose topology is determined by the selected entries of the covariance lags considered in the extension problem. Then, we showed how its dual problem is linked with the maximum likelihood estimator. From such connection, we have proposed a Bayesian model and an approximate MAP estimator to learn ARMA graphical models in the case the topology is not known. Finally, we have conducted some numerical experiments to show the effectiveness of the proposed method. It is worth noting the concepts introduced in this article can be extended to other types of ARMA graphical models, e.g., the latent-variable graphical models [41], for which only the AR case has been addressed.

References

[1] S. Maanan, B. Dumitrescu, and C. Giurcăeneanu, "Maximum entropy expectation-maximization algorithm for fitting latent-variable graphical models to multivariate time series," *Entropy*, vol. 20, no. 1, 2018, Art. no. 76.

[2] J. You and C. Yu, "Sparse plus low-rank identification for dynamical latent-variable graphical AR models," *Automatica*, vol. 159, 2023, Art. no. 111405.

[3] M. Zorzi and R. Sepulchre, "AR identification of latent-variable graphical models," *IEEE Trans. Autom. Control*, vol. 61, no. 9, pp. 2327–2340, Sep. 2016.

[4] M. Zorzi, "Autoregressive identification of kronecker graphical models," *Automatica*, vol. 119, 2020, Art. no. 109053.

[5] D. Materassi and M. V. Salapaka, "Signal selection for estimation and identification in networks of dynamic systems: A graphical model approach," *IEEE Trans. Autom. Control*, vol. 65, no. 10, pp. 4138–4153, Oct. 2020.

[6] M. S. Veedu, H. Doddi, and M. V. Salapaka, "Topology learning of linear dynamical systems with latent nodes using matrix decomposition," *IEEE Trans. Autom. Control*, vol. 67, no. 11, pp. 5746–5761, Nov. 2022.

[7] M. S. Veedu and M. V. Salapaka, "Topology identification under spatially correlated noise," *Automatica*, vol. 156, 2023, Art. no. 111182.

[8] H. Doddi, D. Deka, S. Talukdar, and M. Salapaka, "Estimating linear dynamical networks of cyclostationary processes," 2020, *arXiv:2009.12667*.

[9] M. Zorzi and A. Chiuso, "Sparse plus low rank network identification: A nonparametric approach," *Automatica*, vol. 76, pp. 355–366, 2017.

[10] M. Zorzi, "Nonparametric identification of kronecker networks," *Automatica*, vol. 145, 2022, Art. no. 110518.

[11] K. R. Ramaswamy, G. Bottegal, and P. M. Van den Hof, "Learning linear modules in a dynamic network using regularized kernel-based methods," *Automatica*, vol. 129, 2021, Art. no. 109591.

[12] K. R. Ramaswamy and P. M. J. V. D. Hof, "A local direct method for module identification in dynamic networks with correlated noise," *IEEE Trans. Autom. Control*, vol. 66, no. 11, pp. 5237–5252, Nov. 2021.

[13] A. Ferrante, M. Pavon, and F. Ramponi, "Hellinger versus Kullback–Leibler multivariable spectrum approximation," *IEEE Trans. Autom. Control*, vol. 53, no. 4, pp. 954–967, May 2008.

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

[15] A. Ferrante, F. Ramponi, and F. Ticozzi, "On the convergence of an efficient algorithm for Kullback–Leibler approximation of spectral densities," *IEEE Trans. Autom. Control*, vol. 56, no. 3, pp. 506–515, Mar. 2011.

[16] B. Zhu, "On the well-posedness of a parametric spectral estimation problem and its numerical solution," *IEEE Trans. Autom. Control*, vol. 65, no. 3, pp. 1089–1099, Mar. 2020.

[17] B. Zhu, A. Ferrante, J. Karlsson, and M. Zorzi, "M^2-spectral estimation: A flexible approach ensuring rational solutions," *SIAM J. Control Optim.*, vol. 59, no. 4, pp. 2977–2996, 2021.

[18] A. Ringh, J. Karlsson, and A. Lindquist, "Multidimensional rational covariance extension with approximate covariance matching," *SIAM J. Control Optim.*, vol. 56, no. 2, pp. 913–944, 2018.

[19] J. Karlsson, A. Lindquist, and A. Ringh, "The multidimensional moment problem with complexity constraint," *Integral Equ. Operator Theory*, vol. 84, no. 3, pp. 395–418, 2016.

[20] A. Ringh, J. Karlsson, and A. Lindquist, "Multidimensional rational covariance extension with applications to spectral estimation and image compression," *SIAM J. Control Optim.*, vol. 54, no. 4, pp. 1950–1982, 2016.

[21] P. Enqvist, "A convex optimization approach to ARMA(n, m) model design from covariance and cepstral data," *SIAM J. Control Optim.*, vol. 43, no. 3, pp. 1011–1036, 2004.

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

[23] S. Maanan, B. Dumitrescu, and C. Giurcăeneanu, "Conditional independence graphs for multivariate autoregressive models by convex optimization: Efficient algorithms," *Signal Process.*, vol. 133, pp. 122–134, 2017.

[24] D. Alpago, M. Zorzi, and A. Ferrante, "A scalable strategy for the identification of latent-variable graphical models," *IEEE Trans. Autom. Control*, vol. 67, no. 7, pp. 3349–3362, Jul. 2022.

[25] E. Avventi, A. G. Lindquist, and B. Wahlberg, "ARMA identification of graphical models," *IEEE Trans. Autom. Control*, vol. 58, no. 5, pp. 1167–1178, May 2013.

[26] D. Alpago, M. Zorzi, and A. Ferrante, "Data-driven link prediction over graphical models," *IEEE Trans. Autom. Control*, vol. 68, no. 4, pp. 2215–2228, Apr. 2023.

[27] J. You, C. Yu, J. Sun, and J. Chen, "Generalized maximum entropy based identification of graphical ARMA models," *Automatica*, vol. 141, 2022, Art. no. 110319.

[28] D. Wipf and S. Nagarajan, "Iterative reweighted ℓ_1 and ℓ_2 methods for finding sparse solutions," *J. Sel. Topics Signal Process.*, vol. 4, no. 2, pp. 317–329, 2010.

[29] M. Zorzi, "Empirical Bayesian learning in AR graphical models," *Automatica*, vol. 109, 2019, Art. no. 108516.

[30] D. R. Brillinger, "Remarks concerning graphical models for time series and point processes," *Braz. Rev. Econometrics*, vol. 16, no. 1, pp. 1–23, 1996.

[31] A. Lindquist, C. Masiero, and G. Picci, "On the multivariate circulant rational covariance extension problem," in *Proc. 52nd IEEE Conf. Decis. Control*, 2013, pp. 7155–7161.

[32] A. Dempster, "Covariance selection," *Biometrics*, vol. 28, no. 1, pp. 157–175, 1972.

[33] B. Zhu and M. Zorzi, "A well-posed multidimensional rational covariance and generalized cepstral extension problem," *SIAM J. Control Optim.*, vol. 61, no. 3, pp. 1532–1556, 2023.

[34] W. Rudin, *Real and Complex Analysis*, 3 rd ed. New York, NY, USA: McGraw-Hill, 1987.

[35] A. Ringh, J. Karlsson, and A. Lindquist, "The multidimensional circulant rational covariance extension problem: Solutions and applications in image compression," in *Proc. IEEE 54th Annu. Conf. Decis. Control*, 2015, pp. 5320–5327.

[36] B. Zhu and M. Zorzi, "A weaker regularity condition for the multidimensional ν-moment problem," *IEEE Control Syst. Lett.*, vol. 7, pp. 1795–1800, 2023.

[37] L. Falconi, A. Ferrante, and M. Zorzi, "Mean-square consistency of the f-truncated M^2-periodogram," *Automatica*, vol. 147, 2023, Art. no. 110672.

[38] P. Whittle, "Estimation and information in stationary time series," *Arkiv För Matematik*, vol. 2, no. 5, pp. 423–434, 1953.

[39] C. Rasmussen and C. Williams, *Gaussian Processes for Machine Learning.* Cambridge, MA, USA: MIT Press, 2006.

[40] Z. Zhou, R. Leahy, and J. Qi, "Approximate maximum likelihood hyperparameter estimation for gibbs priors," *IEEE Trans. Image Process.*, vol. 6, no. 6, pp. 844–861, Jun. 1997.

[41] V. Ciccone, A. Ferrante, and M. Zorzi, "Learning latent variable dynamic graphical models by confidence sets selection," *IEEE Trans. Autom. Control*, vol. 65, no. 12, pp. 5130–5143, Dec. 2020.

Mattia Zorzi (Senior Member, IEEE) received the M.S. degree in automation engineering and the Ph.D. degree in information engineering from the University of Padova, Padova, Italy, in 2009 and 2013, respectively.

He held Postdoctoral appointments with the Department of Electrical Engineering and Computer Science, University of Liege, Liege, Belgium, and with the Human Inspired Technology Research Centre, University of Padova, Padova, Italy. He held visiting positions with the Department of Electrical and Computer Engineering, University of California, Davis, Davis, CA, USA, and with the Department of Engineering, University of Cambridge, Cambridge, U.K., in 2011 and 2013–2014, respectively. He is currently an Associate Professor with the Department of Information Engineering, University of Padova. His current research interests include machine learning, robust estimation, identification theory.

Dr. Zorzi has been an Associate Editor for *Automatica* since 2021 and IEEE CONTROL SYSTEMS LETTERS since 2019. He has also been an Associate Editor on the IEEE Control System Society Conference Editorial Board since 2017 and the EUCA Conference Editorial Board since 2020. He is a member of the IFAC Technical Committee on Modelling, Identification and Signal Processing.

Open Access provided by 'Università degli Studi di Padova' within the CRUI CARE Agreement