Learning the Conditional Independence Structure of Stationary Time Series: A Multitask Learning Approach

Alexander Jung

***Abstract*—We propose a method for inferring the conditional independence graph (CIG) of a high-dimensional Gaussian vector time series (discrete-time process) from a finite-length observation. By contrast to existing approaches, we do not rely on a parametric process model (such as, e.g., an autoregressive model) for the observed random process. Instead, we only require certain smoothness properties (in the Fourier domain) of the process. The proposed inference scheme works even for sample sizes much smaller than the number of scalar process components if the true underlying CIG is sufficiently sparse. A theoretical performance analysis provides sufficient conditions on the sample size such that the new method is consistent asymptotically. Some numerical experiments validate our theoretical performance analysis and demonstrate superior performance of our scheme compared to an existing (parametric) approach in case of model mismatch.**

***Index Terms*—Graphical model selection, high-dimensional statistics, multitask LASSO, multitask learning, nonparametric time series, sparsity.**

I. Introduction

We consider a stationary discrete-time vector process or time series which could model, e.g., the time evolution of air pollutant concentrations [1], [2] or medical diagnostic data obtained in electrocorticography [3].

One specific way of representing the dependence structure of a vector process is via a graphical model [4], where the nodes of the graph represent the individual scalar process components, and the edges represent statistical relations between the individual process components. More precisely, the (undirected) edges of a *conditional independence graph (CIG)* associated with a process represent conditional independence statements about the process components [4], [1]. In particular, two nodes in the CIG are connected by an edge if and only if the two corresponding process components are conditionally dependent, given the remaining process components. Note that the so defined CIG for time series extends the basic notion of a CIG for random vectors by considering dependencies between entire scalar time series instead of dependencies between scalar random variables [5], [6]. However, for the special case of i.i.d. time series, these two concepts are fully equivalent.

In this work, we investigate the problem of graphical model selection (GMS), i.e., that of inferring the CIG of a time series, given a finite-length observation. Accurate GMS is an important preprocessing step for various inference tasks such as network anomaly detection or the prediction of future values of the time series [5].

Our work applies to the *high-dimensional* regime, where the model dimension, given by the number of process components, is allowed to be (much) larger than the amount of observed data, given by the sample size [7]–[9], [3], [10]–[12]. It is then intuitively clear that some additional problem structure is required in order to allow for the existence of consistent estimation schemes. Here, this structure is given by sparsity constraints placed on the CIG. More precisely, we assume that the underlying CIG has a small maximum node degree, i.e., each node has a relatively small number of neighbors.

A. Existing Work

GMS for high-dimensional processes with observations modeled as i.i.d. is now well developed [11], [9]. For continuous valued Gaussian Markov random fields, binary Ising models as well as mixed graphical models (containing both continuous and discrete random variables), efficient approaches (based on convex optimization) for inferring the underlying graphical model have been proposed [11], [9], [13], [14]. The authors of [11], [9], [14] present sufficient conditions such that their proposed recovery method is consistent in the high-dimensional regime. These sufficient conditions are complemented by the fundamental performance limits derived in [15], showing that in certain regimes the (computationally efficient) selection scheme put forward in [14] performs essentially optimal.

B. Contribution

In this paper, we develop and analyze a *nonparametric compressive GMS scheme* for general stationary time series. Thus, by contrast to most existing approaches [3], [10], [16], we do not rely on a specific finite dimensional model for the observed process. Instead, we require the observed process to be sufficiently smooth in the spectral domain. This smoothness constraint will be quantified by certain moments of the process autocovariance function (ACF) and requires the ACF to be effec-

Manuscript received January 15, 2015; revised June 19, 2015; accepted July 20, 2015. Date of publication July 23, 2015; date of current version September 23, 2015. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Alba Pages-Zamora. Most of the work on this paper was carried out while the author was with the Institute of Telecommunications, Vienna Institute of Technology.

The author is with the Department of Computer Science, Aalto University, 02150 Espoo, Finland (e-mail: alexander.jung@aalto.fi).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSP.2015.2460219

tively supported on a small interval, whose size is known beforehand, e.g., due to specific domain knowledge.

Inspired by a recent neighborhood regression approach to GMS for Gaussian Markov random fields based on i.i.d. samples [11], we propose a GMS method for time series by generalizing the neighborhood regression approach to the Fourier domain. Our approach exploits a specific problem structure, inherent to the GMS problem, corresponding to a special case of a *block-sparse recovery problem* [17]–[19], i.e., a *multitask learning problem* [20], [21].

Our main conceptual contribution is the formulation of GMS for time series as a multitask learning problem which is defined over a continuum of tasks, which are indexed by a continuous frequency variable $\theta \in [0,1)$. Based on this formulation, we develop a GMS scheme by combining a nonparametric Blackman-Tukey (BT) spectrum estimator with the *multitask LASSO (mLASSO)* [20], [22]. A theoretical performance analysis yields an upper bound on the probability of the proposed GMS method to deliver a wrong CIG. Moreover, we assess the empirical performance of the proposed scheme by means of illustrative numerical experiments.

C. Outline of the Paper

We formalize the problem of GMS for stationary time series in Section II. Our novel compressive GMS scheme for stationary processes is presented in Section III, which is organized in two parts: First, we discuss the spectrum estimator employed in our selection scheme. Then, we show how to apply the mLASSO for inferring the CIG, by formulating GMS for time series as a multitask learning problem. In Section IV, we present a theoretical performance guarantee in the form of an upper bound on the probability of our algorithm to fail in correctly recovering the true underlying CIG. Finally, the results of some illustrative numerical experiments are presented in Section V.

Notation and Basic Definitions

Boldface lowercase letters denote column vectors, whereas boldface uppercase letters denote matrices. The kth entry of a vector $\mathbf{a}$ is denoted by $(\mathbf{a})_k$, and the entry of a matrix $\mathbf{A}$ in the m-th row and n-th column by $(\mathbf{A})_{m,n}$. The submatrix of $\mathbf{A}$ comprised of the elements in rows $a,\ldots,b$ and columns $c,\ldots,d$ is denoted $\mathbf{A}_{a:b,c:d}$. The superscripts ^{T}, *, and ^{H} denote the transpose, (entry-wise) conjugate, and Hermitian transpose, respectively. The kth column of the identity matrix will be denoted by $\mathbf{e}_k$. For some, $p \in \mathbb{N}$, we write $[p] := \{1,\ldots,p\}$.

We denote by $\ell_q([0,1))$ the set of all vector-valued functions $\mathbf{c}(\cdot) : [0,1) \to \mathbb{C}^q$ such that each component $c_r(\theta)$ is square integrable, i.e., $c_r(\cdot) \in L^2([0,1))$ (we also use the shorthand L^2) with norm $\|c_r(\cdot)\|_{L^2} := \sqrt{\int_{\theta=0}^{1} |c_r(\theta)|^2 d\theta}$. We then define the generalized support of $\mathbf{c}(\cdot) \in \ell_q([0,1))$ as $\mathrm{gsupp}(\mathbf{c}(\cdot)) := \{r \in [p] | \|c_r(\cdot)\|_{L^2} > 0\}$. For $\mathbf{c}(\cdot) \in \ell_q([0,1))$ and a subset $\mathcal{I} \subseteq [q]$, we denote by $\mathbf{c}_{\mathcal{I}}(\cdot)$ the vector-valued function which is obtained by retaining only those components $c_r(\cdot)$ with $r \in \mathcal{I}$. Given $\mathbf{c}(\cdot) \in \ell_q([0,1))$, we define the norms $\|\mathbf{c}(\cdot)\|_2 := \sqrt{\sum_{r\in[q]} \|c_r(\cdot)\|_{L^2}^2}$ and $\|\mathbf{c}(\cdot)\|_1 := \sum_{r\in[q]} \|c_r(\cdot)\|_{L^2}$, respectively.

Given a matrix $\mathbf{H} \in \mathbb{C}^{p\times p}$, we denote its spectral norm as $\|\mathbf{H}\|_2 := \max_{\mathbf{x}\neq\mathbf{0}} \frac{\|\mathbf{H}\mathbf{x}\|_2}{\|\mathbf{x}\|_2}$. The norm $\|\mathbf{H}\|_\infty$ is defined as the largest magnitude of its entries, i.e., $\|\mathbf{H}\|_\infty := \max_{m,n} |(\mathbf{H})_{m,n}|$.

II. Problem Formulation

Consider a p-dimensional stationary Gaussian random process $\mathbf{x}[n]$ with (matrix-valued) ACF $\mathbf{R}_x[m] := \mathrm{E}\{\mathbf{x}[m]\mathbf{x}^T[0]\}$, which is assumed to be summable, i.e., $\sum_{m=-\infty}^{\infty} \|\mathbf{R}_x[m]\| < \infty$.[1]

The *spectral density matrix* (SDM) of the process $\mathbf{x}[n]$ is defined as

$$\mathbf{S}_x(\theta) := \sum_{m=-\infty}^{\infty} \mathbf{R}_x[m] \exp(-j2\pi\theta m). \tag{1}$$

The SDM may be interpreted as the multivariate generalization of the power spectral density of a scalar stationary random process [5], [24].

For our analysis, we require a mild technical condition for the eigenvalues $\lambda(\mathbf{S}_x(\theta))$ of the process SDM $\mathbf{S}_x(\theta)$.

Assumption 1: For known positive constants $U \geq L > 0$, we have

$$L \leq \lambda(\mathbf{S}_x(\theta)) \leq U \text{ for every } \theta \in [0,1). \tag{2}$$

We remark that the restriction induced by Assumption 1 is rather weak. E.g., the upper bound in (2) is already implied by the summability of the process ACF. The lower bound in (2) ensures that the CIG satisfies the global Markov property [4], [25]. An important and large class of processes satisfying (2) is given by the set of stable VAR processes [26]. In what follows, we will assume without loss of generality that $L = 1$, implying that $U \geq 1$.

The CIG of the p-dimensional vector process $\mathbf{x}[n]$ is the undirected graph $\mathcal{G}_x := (\mathcal{V}, \mathcal{E})$ with node set $\mathcal{V} = [p]$, corresponding to the scalar process components $\{x_r[n]\}_{r\in[p]}$, and edge set $\mathcal{E} \subseteq \mathcal{V}\times\mathcal{V}$. An edge between nodes r and r' is absent, i.e., $(r,r') \notin \mathcal{E}$, if the component processes $x_r[n]$ and $x_{r'}[n]$ are conditionally independent given the remaining components $\{x_t[n]\}_{t\in[p]\setminus\{r,r'\}}$ [1]. For a Gaussian stationary process $\mathbf{x}[n]$ whose SDM $\mathbf{S}_x(\theta)$ is invertible for every $\theta \in [0,1)$, which is implied by Assumption 1, the CIG of a process can be characterized conveniently via its SDM [1], [6]:

Lemma II.1: Consider a Gaussian stationary vector process $\mathbf{x}[n]$ with associated CIG $\mathcal{G}_x$ and SDM $\mathbf{S}_x(\theta)$ satisfying (2)*. The edge set $\mathcal{E}$ of the CIG $\mathcal{G}_x$ is then characterized by*

$$(r,r') \notin \mathcal{E} \text{ if and only if } \left[\mathbf{S}_x^{-1}(\theta)\right]_{r,r'} = 0 \quad \forall \theta \in [0,1). \tag{3}$$

Thus, in the Gaussian case, the edge set $\mathcal{E}$ corresponds to the zero entries of the inverse SDM $\mathbf{S}_x^{-1}(\theta)$, and the GMS problem is equivalent to detecting the zero entries of $\mathbf{S}_x^{-1}(\cdot)$.

We highlight that, by contrast to graphical models for random vectors, here we consider conditional independence relations between entire scalar time series and not between scalar random variables. In particular, the CIG $\mathcal{G}_x$ of a time series does not de-

[1]The precise choice of norm is irrelevant for the definition of summability, since in finite-dimensional vector spaces all norms are equivalent [23].

pend on time n but applies to the entire time series, as illustrated by the following example:

Consider the vector autoregressive (VAR) process [26]

$$\mathbf{x}[n] = \mathbf{A}\mathbf{x}[n-1]+\mathbf{w}[n] \text{ with } \mathbf{A} = (1/2)\begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}. \quad (4)$$

The noise process $\mathbf{w}[n]$ in (4) consists of i.i.d. Gaussian random vectors with zero mean and covariance matrix $\sigma^2\mathbf{I}$. A little calculation reveals that this stationary AR process has zero mean and its ACF is given by $\mathbf{R}_x[m] = \sigma^2 \sum_{l=0}^{\infty} \mathbf{A}^{m+l}\left(\mathbf{A}^T\right)^l$ [26]. Since the VAR parameter matrix $\mathbf{A}$ in (4) satisfies $\mathbf{A}^T\mathbf{A} = (1/2)\mathbf{I}$, we have $\mathbf{R}_x[0] = 2\sigma^2\mathbf{I}$. For an arbitrary but fixed time index $n = n_0$, the Gaussian random vector $\mathbf{x}[n_0]$ is zero mean with covariance matrix $\mathbf{C} = \mathbf{R}_x[0] = 2\sigma^2\mathbf{I}$. Thus, the scalar time samples $x_1[n]$ and $x_2[n]$ are marginally, i.e., for a fixed time index $n = n_0$, independent. However, since the inverse SDM of the process in (4) is given explicitly by [1]

$$\mathbf{S}_x^{-1}(\theta) = \frac{1}{\sigma^2}\left[\begin{pmatrix} 1.5 & 0 \\ 0 & 1.5 \end{pmatrix} - \begin{pmatrix} \cos\theta & j\sin\theta \\ -j\sin\theta & \cos\theta \end{pmatrix}\right]. \quad (5)$$

we have, upon comparing (5) with the relation (3), that the entire scalar process components $\{x_1[n]\}_{n\in\mathbb{Z}}$ and $\{x_2[n]\}_{n\in\mathbb{Z}}$ are dependent. In general, the marginal conditional independence structure at a fixed time $n = n_0$ is different from the conditional independence structure of the entire time series.

The problem of GMS considered in this paper can be stated as that of inferring the CIG $\mathcal{G}_x = (\mathcal{V}, \mathcal{E})$, or more precisely its edge set $\mathcal{E}$, based on an observed finite length data block $\big(\mathbf{x}[1], \ldots, \mathbf{x}[N]\big)$. Similar to [11], our approach to GMS is via separately estimating the neighborhood $\mathcal{N}(r) := \{r' \in [p] \,|(r, r') \in \mathcal{E}\}$ of each node $r \in \mathcal{V}$. For the specific neighborhood $\mathcal{N}(1)$, the edge set characterization (3) yields the following convenient characterization

$$\mathcal{N}(1) = \text{gsupp}\left(\left(\mathbf{S}_x^{-1}(\cdot)\right)_{1,2:p}\right) - 1. \quad (6)$$

The neighborhood characterization (6) can be generalized straightforwardly to the neighborhood $\mathcal{N}(r)$ of an arbitrary node $r \in [p]$ (cf. Section III-A).[2] For the derivation and analysis of the proposed GMS method, we will, besides Assumption 1, rely on three further assumptions on the CIG $\mathcal{G}_x$, inverse SDM $\mathbf{S}_x^{-1}(\theta)$ and ACF $\mathbf{R}_x[m]$ of the underlying process $\mathbf{x}[n]$.

The first of these additional assumptions constrains the CIG of the observed process $\mathbf{x}[n]$ to be *sparse*, as made precise in

Assumption 2: The maximum node degree $\max_{r\in[p]} |\mathcal{N}(r)|$ *of the process CIG* $\mathcal{G}_x$ *is upper bounded by a known small constant* $s_{\max}$*, i.e.,*

$$\max_{r\in[p]} |\mathcal{N}(r)| \leq s_{\max} \ll p. \quad (7)$$

[2]Note that from the validity of (3) alone, we can only conclude that $\left\|\left[\mathbf{S}_x^{-1}(\cdot)\right]_{r,r'}\right\|_2 = 0$ for a pair $(r, r') \notin \mathcal{E}$. For $(r, r') \in \mathcal{E}$ it might be however that $\left[\mathbf{S}_x^{-1}(\theta)\right]_{r,r'} \neq 0$ only for θ belonging to a set of measure zero, which means $\left\|\left[\mathbf{S}_x^{-1}(\cdot)\right]_{r,r'}\right\|_2 = 0$ although $(r, r') \in \mathcal{E}$. However, this issue is resolved by the validity of Assumption 3. In particular, if for some positive $\rho_{\min} > 0$, (8) is in force, (3) becomes:

$$(r, r') \notin \mathcal{E} \text{ if and only if } \left\|\left[\mathbf{S}_x^{-1}(\cdot)\right]_{r,r'}\right\|_2 = 0.$$

The next assumption is necessary in order to allow for accurate selection schemes based on a finite length observation. In particular, we require the non-zero entries of $\mathbf{S}_x^{-1}(\theta)$ not being too small.

Assumption 3: For a known positive constant $\rho_{\min}$*,*

$$\min_{\substack{r\in[p] \\ r'\in\mathcal{N}(r)}} \left(\int_{\theta=0}^{1} \left|\left[\mathbf{S}_x^{-1}(\theta)\right]_{r,r'} / \left[\mathbf{S}_x^{-1}(\theta)\right]_{r,r}\right|^2 d\theta\right)^{1/2} \geq \rho_{\min}. \quad (8)$$

The integrand in (8) is well defined, since by (2) we have $\left[\mathbf{S}_x^{-1}(\theta)\right]_{r,r} \geq (1/U) > 0$ for all $\theta \in [0, 1)$ and any $r \in [p]$.

For the proposed selection method to be accurate, we require the process $\mathbf{x}[n]$ to be sufficiently smooth in the spectral domain. By *a smooth process* $\mathbf{x}[n]$, we mean a process $\mathbf{x}[n]$ such that the entries of its SDM $\mathbf{S}_x(\theta)$ are smooth functions of θ. These smoothness constraints will be expressed in terms of moments of the process ACF:

Assumption 4: For a small positive constant μ_0 *and a given non-negative weight function* $h[m]$*, that typically increases with* $|m|$*, we have the bound*

$$\mu_x^{(h)} := \sum_{m=-\infty}^{\infty} h[m]\|\mathbf{R}_x[m]\|_\infty \leq \mu_0. \quad (9)$$

For the particular weighting function $h[m] := |m|$, we will use the shorthand

$$\mu_x := \sum_{m=-\infty}^{\infty} |m|\|\mathbf{R}_x[m]\|_\infty. \quad (10)$$

We may interpret the moment μ_x as a measure for the effective ACF width of the process. Another particular choice for the weighting function will be given in Section IV. This choice is related to the window function of the BT estimator which is part of our GMS method (cf. Section III).

We note that Assumption 4 is similar in spirit to the underspread assumption for linear time varying systems and nonstationary processes [27] in that it allows to construct efficient decorrelation transformations. In particular, for a smooth process conforming to Assumption 4, one can verify that the discrete Fourier transform (DFT) of the observed block yields random vectors which are approximately uncorrelated for different frequencies. This decorrelation in the frequency domain is the key idea behind our Fourier based approach.

In what follows, we will formulate and analyze a GMS scheme for the class of p-dimensional Gaussian stationary processes $\mathbf{x}[n]$ conforming to Assumptions 1–4. This process class will be denoted as $\mathcal{M}$ for brevity.

III. The Selection Scheme

The GMS scheme developed in this section is inspired by the neighborhood regression approach in [11]. A main conceptual difference of our approach to [11] is that we perform neighborhood regression in the frequency domain. Moreover, while the approach in [11] is based on a standard sparse linear regression model, we formulate the neighborhood regression for time series as a multitask learning problem. This multitask learning problem is based on an estimator for the SDM, which will be discussed next.

A. SDM Estimation

Due to the direct relation (3) between the zero pattern of the inverse SDM and the edge set of the CIG, a naive approach to GMS would be to first estimate the SDM $\mathbf{S}_x(\theta)$, then invert this estimate and determine the location of the non-zero entries. With regards to the first step, it is natural to estimate $\mathbf{S}_x(\theta)$ by replacing the ACF in (1) with an empirical version $\widehat{\mathbf{R}}_x[m]$ which is based on sample averages. This yields the estimate

$$\widehat{\mathbf{S}}_x(\theta) := \sum_{m=-N+1}^{N-1} w[m]\widehat{\mathbf{R}}_x[m]e^{-j2\pi\theta m} \tag{11}$$

where, $\widehat{\mathbf{R}}_x^T[-m] = \widehat{\mathbf{R}}_x[m]$ and

$$\widehat{\mathbf{R}}_x[m] := \frac{1}{N}\sum_{n=1}^{N-m} \mathbf{x}[n+m]\mathbf{x}^T[n],\ m \in \{0,\ldots,N-1\}. \tag{12}$$

The real-valued window function $w[m]$ in (11), from now on assumed to satisfy

$$w[m] = 0 \text{ for } m \geq N \text{ and } w[0] = 1, \tag{13}$$

is chosen such that the estimate $\widehat{\mathbf{S}}_x(\theta)$ is guaranteed to be a psd matrix. A sufficient condition for this to be the case is non-negativity of the discrete-time Fourier transform (DTFT) $W(\theta)$ of the window function, i.e., $W(\theta) := \sum_{m=-\infty}^{\infty} w[m]\exp(-j2\pi\theta m) \geq 0$ [28, p. 40].

In what follows, we need a specific representation of the estimate $\widehat{\mathbf{S}}_x(\theta)$ in (11), which is stated in

Lemma III.1: Consider the estimate $\widehat{\mathbf{S}}_x(\theta)$ given by (11)*, for $\theta \in [0,1)$. Let us define the matrix*

$$\mathbf{A}(\theta) := \mathbf{W}(\theta)\mathbf{F}^T\mathbf{D}^T, \tag{14}$$

where $\mathbf{D} := (\mathbf{x}[1],\ldots,\mathbf{x}[N]) \in \mathbb{R}^{p\times N}$ is the data matrix, $\mathbf{F} \in \mathbb{C}^{N\times(2N-1)}$ denotes the first N rows of the size-$(2N-1)$ DFT matrix, i.e., $(\mathbf{F})_{k,l} = \exp(-j2\pi(k-1)(l-1)/(2N-1))$ and

$$\mathbf{W}(\theta) := (1/(2N-1))\,\mathrm{diag}\left\{\sqrt{W(\theta_f+\theta)}\right\}_{f\in[2N-1]}, \tag{15}$$

with $\theta_f := 2\pi(f-1)/(2N-1)$.

We then have the identity

$$\widehat{\mathbf{S}}_x(\theta) = (1/N)\mathbf{A}^H(\theta)\mathbf{A}(\theta). \tag{16}$$

Proof: Appendix A. □

As evident from the factorization (16), the rank of $\widehat{\mathbf{S}}_x(\theta)$ satisfies $\mathrm{rank}\{\widehat{\mathbf{S}}_x(\theta)\} \leq N$. Therefore, in the high-dimensional regime, where the number N of observations is much smaller than the number p of process components, the estimates $\widehat{\mathbf{S}}_x(\theta) \in \mathbb{C}^{p\times p}$ will be rank-deficient and cannot be inverted to obtain estimates of the edge set $\mathcal{E}$ via the relation (3).

In order to cope with the rank deficiency of the SDM estimate $\widehat{\mathbf{S}}_x(\theta)$, we next show that finding the support of the inverse SDM $\mathbf{S}_x^{-1}(\theta)$ based on the observation $\mathbf{x}[1],\ldots,\mathbf{x}[N]$ can be formulated as a *multitask learning problem*. For clarity, we detail this approach only for the problem of estimating the neighborhood $\mathcal{N}(1)$. The generalization to the neighborhood $\mathcal{N}(r)$ of an arbitrary node $r \in [p]$ is straightforward.

Indeed, consider the permuted process $\tilde{\mathbf{x}}[n] := \mathbf{P}_r\mathbf{x}[n]$, with the permutation matrix $\mathbf{P}_r := \left(\mathbf{e}_{\Pi_r(1)},\ldots,\mathbf{e}_{\Pi_r(p)}\right)$ where $\Pi_r(\cdot) : [p] \to [p]$ denotes the permutation exchanging entry 1 with entry r and leaving the remaining entries unchanged. As can be verified easily, the SDM $\mathbf{S}_{\tilde{\mathbf{x}}}(\theta)$ of the process $\tilde{\mathbf{x}}[n]$ is then given by $\mathbf{P}_r\mathbf{S}_x(\theta)\mathbf{P}_r$. Moreover, the CIG $\mathcal{G}_{\tilde{x}}$ of $\tilde{\mathbf{x}}[n]$ contains the edge (v,w) if and only if the CIG $\mathcal{G}_x$ of $\mathbf{x}[n]$ contains the edge $(\Pi_r(v),\Pi_r(w))$, i.e.,

$$(v,w) \in \mathcal{G}_{\tilde{x}} \text{ if and only if } (\Pi_r(v),\Pi_r(w)) \in \mathcal{G}_x. \tag{17}$$

Thus, the problem of determining the neighborhood $\mathcal{N}(r)$ in the CIG of the process $\mathbf{x}[n]$ is equivalent to the problem of determining the neighborhood $\mathcal{N}(1)$ in the CIG of the permuted process $\tilde{\mathbf{x}}[n] = \mathbf{P}_r\mathbf{x}[n]$.

Note that the SDM estimator (11) can be regarded as the natural adaptation, to the case of SDM estimation for vector process, of the BT estimator [28] for the power spectral density of a scalar process.

B. Multitask Learning Formulation

The basic intuition behind our approach is to perform a decorrelation of the time samples $\mathbf{x}[1],\ldots,\mathbf{x}[N]$ by applying a DFT. In particular, given the observation $\mathbf{D} = (\mathbf{x}[1],\ldots,\mathbf{x}[N]) \in \mathbb{R}^{p\times N}$, we compute the length-$(2N-1)$ DFT as

$$\hat{\mathbf{x}}[f] := \frac{1}{\sqrt{N}}\sum_{n\in[N]} \mathbf{x}[n]\exp\left(-j\frac{2\pi(n-1)(f-1)}{2N-1}\right), \tag{18}$$

for $f \in [2N-1]$. It can be shown that for a vector process $\mathbf{x}[n]$ conforming to Assumption 4 and a sufficiently large sample size N, the DFT vectors $\hat{\mathbf{x}}[1],\ldots,\hat{\mathbf{x}}[2N-1]$, which may be interpreted as random samples indexed by frequency f, are approximately independent. However, what hinders the straight application of the neighborhood regression method in [11], developed for the case of i.i.d. samples, is the fact that the samples $\hat{\mathbf{x}}[f]$ are not identically distributed. Indeed, the covariance matrix of the Gaussian random vector $\hat{\mathbf{x}}[f]$ is roughly equal to the SDM value $\mathbf{S}_x(\theta_f = 2\pi(f-1)/(2N-1))$, which in general varies with f. However, for processes with a smooth SDM, i.e., conforming to Assumption 4 with small μ_0, the SDM is approximately constant over small frequency intervals and therefore, in turn, the distribution of consecutive samples $\hat{\mathbf{x}}[f]$ is nearly identical. We exploit this by masking the DFT samples $\hat{\mathbf{x}}[f]$ such that, for a given center frequency $\theta \in [0,1)$, we only retain those samples $\hat{\mathbf{x}}[f]$ which fall into the pass band of the spectral window $W(\theta_f+\theta)$ in (15), which is the shifted (by the center frequency θ) DTFT of the window function $w[m]$ employed in the BT estimator (11). This spectral masking then yields the modified DFT samples

$$\tilde{\mathbf{x}}^{(\theta)}[f] := \sqrt{W(\theta_f+\theta)}\hat{\mathbf{x}}[f],\ \text{for } f \in [2N-1]. \tag{19}$$

By considering the significant DFT vectors $\tilde{\mathbf{x}}[f]$ approximately as i.i.d. samples of a Gaussian random vector with zero mean

and covariance matrix $\mathbf{S}_x(\theta)$, we can immediately apply the neighborhood regression approach in [11]. In particular, we formulate, for a specific center frequency θ, a sparse linear regression problem by regressing the first entry of the vector $\tilde{\mathbf{x}}^{(\theta)}[f]$ against its remaining entries. More precisely, based on the vector $\mathbf{y}(\theta) \in \mathbb{C}^{2N-1}$ and matrix $\mathbf{X}(\theta) \in \mathbb{C}^{(2N-1)\times(p-1)}$,

$$\mathbf{y}(\theta) := \begin{pmatrix} \tilde{x}_1^{(\theta)}[1] \\ \vdots \\ \tilde{x}_1^{(\theta)}[2N-1] \end{pmatrix}, \ \mathbf{X}(\theta) := \begin{pmatrix} \left(\tilde{\mathbf{x}}_{2:p}^{(\theta)}[1]\right)^T \\ \vdots \\ \left(\tilde{\mathbf{x}}_{2:p}^{(\theta)}[2N-1]\right)^T \end{pmatrix}, \tag{20}$$

we define, for each $\theta \in [0,1)$, the linear regression model

$$\mathbf{y}(\theta) := \mathbf{X}(\theta)\boldsymbol{\beta}(\theta) + \boldsymbol{\varepsilon}(\theta) \tag{21}$$

with the vector-valued parameter function $\boldsymbol{\beta}(\theta) \in \ell_q([0,1))$, with $q = p-1$, given by

$$\boldsymbol{\beta}(\theta) := \left[\left(\mathbf{S}_x(\theta)\right)_{2:p,2:p}\right]^{-1}(\mathbf{S}_x(\theta))_{2:p,1}. \tag{22}$$

Let us make the relation between the quantities $\mathbf{y}(\theta)$, $\mathbf{X}(\theta)$ and the observed data $\mathbf{D} = (\mathbf{x}[1], \ldots, \mathbf{x}[N])$ explicit by noting that, upon combining (18), (19) and (20), we have

$$\mathbf{y}(\theta) = \mathbf{W}(\theta)\mathbf{F}^T\left(\mathbf{D}_{1,1:N}\right)^T \tag{23}$$

and

$$\mathbf{X}(\theta) = \mathbf{W}(\theta)\mathbf{F}^T\left(\mathbf{D}_{2:p,1:N}\right)^T. \tag{24}$$

The product $\mathbf{F}^T\left(\mathbf{D}_{1,1:N}\right)^T$ in (23) just amounts to computing the DFT (of length $2N$–1) of the process component $x_1[n]$. Similarly, the rows of $\mathbf{F}^T\left(\mathbf{D}_{2:p,1:N}\right)^T$ in (24) are given by the DFTs (of length $2N-1$) of the process components $x_2[n], \ldots, x_p[n]$.

The error term $\boldsymbol{\varepsilon}(\theta)$ in (21) is defined implicitly via the definitions (22), (23), and (24). We show in Section IV that, if the SDM estimator (11) is accurate, i.e., $\widehat{\mathbf{S}}_x(\theta)$ is close to $\mathbf{S}_x(\theta)$ for all $\theta \in [0,1)$, the error term $\boldsymbol{\varepsilon}(\theta)$ will be small.

As can be verified easily, by comparing expressions (23) and (24) with (14), the vector $\mathbf{y}(\theta)$ and the matrix $\mathbf{X}(\theta)$ are given by the columns of the matrix $\mathbf{A}(\theta)$ defined in Lemma III.1. Therefore, according to (16), we have the identity

$$\left(\begin{array}{cc} \mathbf{y}(\theta) & \mathbf{X}(\theta) \end{array}\right)^H \left(\begin{array}{cc} \mathbf{y}(\theta) & \mathbf{X}(\theta) \end{array}\right) = \widehat{\mathbf{S}}_x(\theta), \text{ for } \theta \in [0,1), \tag{25}$$

where $\widehat{\mathbf{S}}_x(\theta)$ denotes the BT estimator in (11).

The link between the multitask learning problem (21) and the problem of determining $\mathcal{N}(1)$ is stated in

Lemma III.2: Consider the parameter vector $\boldsymbol{\beta}(\theta)$ defined for each $\theta \in [0,1)$ via (22). The generalized support of $\boldsymbol{\beta}(\cdot)$ is related to $\mathcal{N}(1)$ via

$$\operatorname{gsupp}(\boldsymbol{\beta}(\cdot)) = \mathcal{N}(1) - 1. \tag{26}$$

Proof: Partition the SDM $\mathbf{S}_x(\theta)$ and its inverse $\mathbf{S}_x^{-1}(\theta)$ as

$$\begin{pmatrix} \gamma(\theta) & \mathbf{c}^H(\theta) \\ \mathbf{c}(\theta) & \mathbf{G}(\theta) \end{pmatrix} := \mathbf{S}_x(\theta), \ \begin{pmatrix} \tilde{\gamma}(\theta) & \tilde{\mathbf{c}}^H(\theta) \\ \tilde{\mathbf{c}}(\theta) & \widetilde{\mathbf{G}}(\theta) \end{pmatrix} := \mathbf{S}_x^{-1}(\theta). \tag{27}$$

According to (3), we have

$$\operatorname{gsupp}(\tilde{\mathbf{c}}(\cdot)) = \mathcal{N}(1) - 1, \tag{28}$$

where $\tilde{\mathbf{c}}(\theta)$ is the lower left block of $\mathbf{S}_x^{-1}(\theta)$ (cf. (27)). Applying [29, Fact 2.17.3] to (27),

$$\tilde{\mathbf{c}}(\theta) = -\tilde{\gamma}(\theta)\mathbf{G}^{-1}(\theta)\mathbf{c}(\theta) \overset{(22)}{=} -\boldsymbol{\beta}(\theta)\tilde{\gamma}(\theta). \tag{29}$$

Note that $\tilde{\gamma}(\theta) \overset{(27)}{=} \left[\mathbf{S}_x^{-1}(\theta)\right]_{1,1} > 0$, since we assume $\mathbf{S}_x(\theta)$ to be strictly positive definite (cf. (2)), implying in turn that $\mathbf{S}_x^{-1}(\theta)$ is also strictly positive definite. Therefore,

$$\operatorname{gsupp}(\boldsymbol{\beta}(\cdot)) \overset{(29)}{=} \operatorname{gsupp}(\tilde{\mathbf{c}}(\cdot)) \overset{(28)}{=} \mathcal{N}(1) - 1.$$

□

Thus, the problem of determining the neighborhood $\mathcal{N}(1)$ of node $r = 1$ has been reduced to that of finding the joint support of the parameter vectors $\{\boldsymbol{\beta}(\theta)\}_{\beta\in[0,1)}$, of the linear model (21), using the observation of the vectors $\{\mathbf{y}(\theta)\}_{\theta\in[0,1)}$ given by (23).

Recovering the vector-valued parameter function $\boldsymbol{\beta}(\cdot) \in \ell_q([0,1))$ based on the model (21), is an instance of a *multitask learning problem* [20], [21], [30], [31], being, in turn, a special case of a block-sparse recovery problem [17]. Compared to existing work on multitask learning, the distinctive feature of the multitask learning problem given by (21) is that we have a continuum of individual tasks indexed by $\theta \in [0,1)$.

C. Multitask LASSO Based GMS

A popular approach for estimating a set of vectors with a small joint support, based on linear measurements, is the *group LASSO* [32]. Specializing the group LASSO to the multitask model (21) yields the *multitask LASSO* (mLASSO) [20], [22]. However, while [20], [22] consider a finite number of tasks, we consider a continuum of tasks indexed by $\theta \in [0,1)$. A natural generalization of the mLASSO to our setting is

$$\hat{\boldsymbol{\beta}}[\mathbf{y}(\cdot), \mathbf{X}(\cdot)] := \operatorname*{argmin}_{\boldsymbol{\beta}\in\ell_q([0,1))} \left\|\mathbf{y}(\cdot) - \mathbf{X}(\cdot)\boldsymbol{\beta}(\cdot)\right\|_2^2 + \lambda\|\boldsymbol{\beta}(\cdot)\|_1. \tag{30}$$

Note that the optimization in (30) has to be carried out over the Hilbert space $\ell_q([0,1))$ with inner product $\langle \mathbf{f}(\cdot), \mathbf{g}(\cdot)\rangle_{\ell_q} := \int_{\theta=0}^1 \mathbf{g}^H(\theta)\mathbf{f}(\theta)d\theta$, and induced norm $\|\mathbf{g}(\cdot)\|_2 = \sqrt{\sum_r \|g_r(\cdot)\|_{L^2}}$. Since the cost function in (30) is convex, continuous and coercive, i.e., $\lim_{\|\boldsymbol{\beta}(\cdot)\|\to\infty} f[\boldsymbol{\beta}(\cdot)] \to \infty$, it follows by convex analysis that a minimizer for (30) exists [33]. In the case of multiple solutions, we mean by $\hat{\boldsymbol{\beta}}(\cdot) = \operatorname{argmin}_{\boldsymbol{\beta}(\cdot)\in\ell_q([0,1))} f[\boldsymbol{\beta}(\cdot)]$ any of these solutions.[3]

If the design parameter $\lambda > 0$ in (30) is chosen suitably (cf. Section IV), the generalized support of $\hat{\boldsymbol{\beta}}(\cdot)$ coincides with that of the true parameter vector $\boldsymbol{\beta}(\cdot)$ in (21), i.e.,

$$\operatorname{gsupp}(\hat{\boldsymbol{\beta}}(\cdot)) = \operatorname{gsupp}(\boldsymbol{\beta}(\cdot)) \overset{(26)}{=} \mathcal{N}(1) - 1. \tag{31}$$

Thus, we can determine the neighborhood $\mathcal{N}(1)$ via computing the mLASSO based on the observation vector $\mathbf{y}(\theta)$ and

[3]Note that a sufficient condition for uniqueness of the solution to (30) would be strict convexity of the objective function. However, in the high-dimensional regime, where $N \ll p$, the system matrix $\mathbf{X}(\theta) \in \mathbb{C}^{(2N-1)\times(p-1)}$ defined by (33) is singular and therefore the objective function in (30) is not strictly convex. Thus, in this regime, uniqueness of the solution to (30) requires additional assumptions such as, e.g., incoherence conditions [34]. We emphasize, however, that our analysis does not require uniqueness of the solution to (30).

system matrix $\mathbf{X}(\theta)$ constructed via (23) and (24) from the observed data $\mathbf{D} = \big(\mathbf{x}[1], \ldots, \mathbf{x}[N]\big)$. The generalization to the determination of the neighborhood $\mathcal{N}(r)$ for an arbitrary node $r \in [p]$ is accomplished via (17) by using the permuted observation $\widetilde{\mathbf{D}} := \mathbf{P}_r\big(\mathbf{x}[1], \ldots, \mathbf{x}[N]\big)$ in (23) and (24) instead of $\mathbf{D}$. We arrive at the following algorithm for estimating the CIG of the observed process.

Algorithm 1:

1) *Given a specific node $r \in [p]$, form the permuted data matrix $\widetilde{\mathbf{D}} = \mathbf{P}_r\big(\mathbf{x}[1], \ldots, \mathbf{x}[N]\big)$, and compute the observation vector $\mathbf{y}(\theta)$ and system matrix $\mathbf{X}(\theta)$ according to*

$$\mathbf{y}(\theta) = \mathbf{W}(\theta)\mathbf{F}^T\big(\widetilde{\mathbf{D}}_{1,1:N}\big)^T \tag{32}$$

and

$$\mathbf{X}(\theta) = \mathbf{W}(\theta)\mathbf{F}^T\big(\widetilde{\mathbf{D}}_{2:p,1:N}\big)^T. \tag{33}$$

2) *Based on the observation vector $\mathbf{y}(\theta)$ and system matrix $\mathbf{X}(\theta)$ given by* (32) *and* (33)*, compute the mLASSO estimate $\hat{\boldsymbol{\beta}}(\theta)$ according to* (30) *and estimate the neighborhood $\mathcal{N}(r)$ by the index set*

$$\widehat{\mathcal{N}}(r) = \{\Pi_r(r'+1) \,|\, r' \in [p],\ \|\hat{\beta}_{r'}(\cdot)\|_{L^2} > \eta\}, \tag{34}$$

for some suitably chosen threshold η.

3) *Repeat step* 1) *and step* 2) *for all nodes $r \in [p]$ and combine the individual neighborhood estimates $\widehat{\mathcal{N}}(r)$ to obtain the final CIG estimate $\widehat{\mathcal{G}} = ([p], \widehat{\mathcal{E}})$.*

The proper choice for the mLASSO parameter λ in (30) and the threshold η in (34) will be discussed in Section IV.

For the last step of Algorithm 1, different ways of combining the individual neighborhood estimates $\widehat{\mathcal{N}}(r)$ to obtain the edge set of the CIG estimate $\widehat{\mathcal{G}}$ are possible. Two intuitive choices are the "AND" rule and the "OR" rule. For the AND (OR) rule, an edge (r, r') is present in $\widehat{\mathcal{G}}$, i.e., $(r, r') \in \widehat{\mathcal{E}}$, if and only if $r \in \widehat{\mathcal{N}}(r')$ and (or) $r' \in \widehat{\mathcal{N}}(r)$.

It is instructive to rewrite the mLASSO (30) computed in step 2 of Algorithm 1 in the form

$$\operatorname*{argmin}_{\boldsymbol{\beta} \in \ell_q([0,1))} \int_{\theta=0}^{1} [\boldsymbol{\beta}^H(\theta)\widehat{\mathbf{G}}(\theta)\boldsymbol{\beta}(\theta) - 2\Re\{\hat{\mathbf{c}}^H(\theta)\boldsymbol{\beta}(\theta)\}]d\theta + \lambda\|\boldsymbol{\beta}(\cdot)\|_1, \tag{35}$$

where the matrix $\widehat{\mathbf{G}}(\theta)$ and vector $\hat{\mathbf{c}}(\theta)$ are sub-blocks of the SDM estimate $\widehat{\mathbf{S}}_x(\theta)$ in (11), i.e.,

$$\begin{pmatrix} \hat{\gamma}(\theta) & \hat{\mathbf{c}}^H(\theta) \\ \hat{\mathbf{c}}(\theta) & \widehat{\mathbf{G}}(\theta) \end{pmatrix} := \widehat{\mathbf{S}}_x(\theta). \tag{36}$$

The equivalence of (35) and (30) can be verified easily by (25). The formulation (35) naturally suggests the generalization of our GMS method to a continuous-time process $x(t)$ (with effective bandwidth B_{e}). Indeed, we just need to replace the integration variable $\theta \in [0, 1)$ in (35) by the integration variable $f \in [-B_{\mathrm{e}}, B_{\mathrm{e}}]$ and the matrix $\widehat{\mathbf{G}}(\theta)$ and vector $\hat{\mathbf{c}}(\theta)$ with corresponding sub-blocks of a continuous-time SDM estimate $\widehat{\mathbf{S}}_x(f)$.

Let us highlight that Algorithm 1 is a nonparametric method as it relies on the nonparametric SDM estimator (11). In particular, it does not require to fit a finite dimensional parametric model to the observed process (which is done, e.g., in the VAR-based method put forward in [3]).

We finally mention that, in principle, Algorithm 1 can also be applied to non-Gaussian processes. However, the resulting graph estimate $\widehat{\mathcal{G}}$ is then not related to a CIG anymore but to a *partial correlation graph* of the process [1].

D. Numerical Implementation

In order to numerically solve the optimization problem (30) we will use a simple discretization approach. More precisely, we require the optimization variable $\boldsymbol{\beta}(\cdot) \in \ell_q([0, 1))$ to be piecewise constant over the frequency intervals $[(f-1)/F, f/F)$, for $f \in [F]$, where the number F of intervals is chosen sufficiently large. As a rule of thumb, which is also justified by the results of some numerical experiments, we will use $F \approx 2\mu_x$, since the SDM $\mathbf{S}_x(\theta)$ is approximately constant over frequency intervals smaller than $1/\mu_x$. This may be verified by the Fourier relationship (1) between the process SDM and ACF. Thus, if we denote by $I_f(\theta)$ the indicator function of the frequency interval $[(f-1)/F, f/F)$, we represent the optimization variable $\boldsymbol{\beta}(\cdot) \in \ell_q([0, 1))$ as

$$\boldsymbol{\beta}(\theta) = \sum_{f \in [F]} \boldsymbol{\beta}_f I_f(\theta), \tag{37}$$

with the vector-valued expansion coefficients $\boldsymbol{\beta}_f \in \mathbb{C}^q$. Inserting (37) into (30) yields the finite-dimensional mLASSO

$$\hat{\boldsymbol{\beta}} = \operatorname*{argmin}_{\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_F)^T} \sum_{f \in [F]} \boldsymbol{\beta}_f^H \mathbf{G}_f \boldsymbol{\beta}_f - 2\Re\{\mathbf{c}_f^H \boldsymbol{\beta}_f\} + \lambda\|\boldsymbol{\beta}\|_1 \tag{38}$$

with $\mathbf{G}_f := \int_{\theta=(f-1)/F}^{f/F} \mathbf{X}^H(\theta)\mathbf{X}(\theta)d\theta$ and $\mathbf{c}_f := \int_{\theta=(f-1)/F}^{f/F} \times \mathbf{X}^H(\theta)\mathbf{y}(\theta)d\theta$. Here, we used $\|\boldsymbol{\beta}\|_1 := \sum_{r \in [q]} \|\boldsymbol{\beta}^{(r)}\|_2$ with the vectors $\boldsymbol{\beta}^{(r)} \in \mathbb{C}^F$ given elementwise as $\big(\boldsymbol{\beta}^{(r)}\big)_f := \big(\boldsymbol{\beta}_f\big)_r$. Based on the solution $\hat{\boldsymbol{\beta}} = \big(\hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_F\big)$ of (38), we replace the neighborhood estimate $\widehat{\mathcal{N}}(r)$ given by (34) in Algorithm 1 with

$$\widehat{\mathcal{N}}(r) = \{\Pi_r(r'+1) \,|\, r' \in [p],\ (1/\sqrt{F})\|\hat{\boldsymbol{\beta}}^{(r')}\|_2 > \eta\}, \tag{39}$$

where $\hat{\boldsymbol{\beta}}^{(r)} := \big(\big(\hat{\boldsymbol{\beta}}_1\big)_r, \ldots, \big(\hat{\boldsymbol{\beta}}_F\big)_r\big)$.

We note that Algorithm 1, based on the discretized version (38) of the mLASSO (30), scales well with the problem dimensions, i.e., it can be implemented efficiently for large sample size N and large number p of process components. Indeed, the expressions (32) and (33) can be evaluated efficiently using FFT algorithms. For a fast implementation of the mLASSO (38) we refer to [35].

IV. Selection Consistency of the Proposed Scheme

We will now analyze the probability of Algorithm 1 to deliver a wrong CIG. Our approach is to, separately for each node $r \in [p]$, bound the probability that the neighborhood $\mathcal{N}(r)$ is estimated incorrectly by Algorithm 1. Since the correct determination of all neighborhoods implies the delivery of the correct CIG, we can invoke a union bound over all p neighborhoods to finally obtain an upper bound on the error probability of the GMS method. For clarity, we detail the analysis only for the specific neighborhood $\mathcal{N}(1)$, the generalization to an arbitrary neighborhood $\mathcal{N}(r)$ being trivially obtained by considering the

permuted process $\tilde{\mathbf{x}}[n] = \mathbf{P}_r\mathbf{x}[n]$ (see our discussion around (17)).

The high-level idea is to divide the analysis into a deterministic part and a stochastic part. The deterministic part consists of a set of sufficient conditions on the multitask learning problem (21) such that the generalized support of the mLASSO $\hat{\boldsymbol{\beta}}[\mathbf{y}(\cdot), \mathbf{X}(\cdot)]$ (cf. (30)), coincides with the generalized support of the parameter vector $\boldsymbol{\beta}(\theta)$ in (22), which, in turn, is equal to $\mathcal{N}(1)-1$ (cf. (26)). These conditions are stated in Theorem IV.1 below. The stochastic part of the analysis amounts to controlling the probability that the sufficient conditions of Theorem IV.1 are satisfied. This will be accomplished by a large deviation analysis of the BT estimator in (11). By combining these two parts, we straightforwardly obtain our main result, i.e., Theorem IV.5 which presents a condition (lower bound) on the sample size N such that the error probability of our GMS method is upper bounded by a prescribed value.

1) Deterministic Part: The deterministic part of our analysis is based on the concept of the multitask compatibility condition [20]. Given an index set $\mathcal{S} \subseteq [q]$ of size $s_{\max}$, a system matrix $\mathbf{X}(\theta) \in \mathbb{C}^{(2N-1)\times(p-1)}$, defined for $\theta \in [0,1)$, is said to satisfy the multitask compatibility condition with constant $\phi(\mathcal{S})$ if

$$s_{\max}\frac{\|\mathbf{X}(\cdot)\boldsymbol{\beta}(\cdot)\|_2^2}{\|\boldsymbol{\beta}_{\mathcal{S}}(\cdot)\|_1^2} \geq \phi^2(\mathcal{S}) > 0 \tag{40}$$

for all vectors $\boldsymbol{\beta}(\cdot) \in \mathbb{A}(\mathcal{S}) \setminus \{\mathbf{0}\}$, where

$$\mathbb{A}(\mathcal{S}) := \big\{\boldsymbol{\beta}(\cdot) \in \ell_q([0,1)) \big| \|\boldsymbol{\beta}_{\mathcal{S}^c}(\cdot)\|_1 \leq 3\|\boldsymbol{\beta}_{\mathcal{S}}(\cdot)\|_1\big\}. \tag{41}$$

A quantity which is particularly relevant for the variable selection performance of the mLASSO is the minimum norm $\min_{r\in \mathrm{gsupp}(\boldsymbol{\beta}(\cdot))} \|\beta_r(\cdot)\|_{L^2}$ of the non-zero blocks of the parameter vector $\boldsymbol{\beta}(\cdot) \in \ell_q([0,1))$ in (22), which we require to be lower bounded by a known positive number $\beta_{\min}$, i.e.,

$$\min_{r\in \mathrm{gsupp}(\boldsymbol{\beta}(\cdot))} \|\beta_r(\cdot)\|_{L^2} \geq \beta_{\min}. \tag{42}$$

Based on $\phi(\mathcal{S})$ and $\beta_{\min}$, the following result characterizes the ability of the mLASSO $\hat{\boldsymbol{\beta}}[\mathbf{y}(\cdot), \mathbf{X}(\cdot)]$ (cf. (30)) to correctly identify $\mathrm{gsupp}(\boldsymbol{\beta}(\cdot)) = \mathcal{N}(1)-1$ (cf. (26)).

Theorem IV.1: Consider the multitask learning model (21) *with parameter vector* $\boldsymbol{\beta}(\cdot) \in \ell_q([0,1))$ *and system matrix* $\mathbf{X}(\theta)$. *The parameter vector* $\boldsymbol{\beta}(\cdot)$ *is assumed to satisfy (42) and having no more than* $s_{\max}$ *non-zero components, i.e.,*

$$\mathrm{gsupp}(\boldsymbol{\beta}(\cdot)) \subseteq \mathcal{S}, \text{ with } |\mathcal{S}| = s_{\max}. \tag{43}$$

Assume further that the system matrix possesses a positive multitask compatibility constant $\phi(\mathcal{S}) > 0$ *(cf.* (40)*), and the error term* $\boldsymbol{\varepsilon}(\theta)$ *in* (21) *satisfies*

$$\sup_{\theta\in[0,1)} \|\boldsymbol{\varepsilon}^H(\theta)\mathbf{X}(\theta)\|_\infty \leq \frac{\phi^2(\mathcal{S})\beta_{\min}}{32 s_{\max}}. \tag{44}$$

Denote by $\hat{\boldsymbol{\beta}}[\mathbf{y}(\cdot), \mathbf{X}(\cdot)]$ *the mLASSO estimate obtained from* (30) *with* $\lambda = \phi^2(\mathcal{S})\beta_{\min}/(8s_{\max})$. *Then, the index set*

$$\hat{\mathcal{S}} := \big\{r \in [q] \mid \|\hat{\beta}_r(\cdot)\|_{L^2} > \beta_{\min}/2\big\}, \tag{45}$$

coincides with the true generalized support of $\boldsymbol{\beta}(\cdot)$, *i.e.,* $\hat{\mathcal{S}} = \mathrm{gsupp}(\boldsymbol{\beta}(\cdot))$.

Proof: Appendix B. □

2) Stochastic Part: We now show that, for sufficiently large sample size N, the multitask learning problem (21) satisfies the condition (44) of Theorem IV.1 with high probability. To this end, we first verify that (44) is satisfied if the maximum SDM estimation error

$$E := \sup_{\theta\in[0,1)} \|\mathbf{E}(\theta)\|_\infty, \text{ with } \mathbf{E}(\theta) := \widehat{\mathbf{S}}_x(\theta) - \mathbf{S}_x(\theta), \tag{46}$$

is small enough. We then characterize the large deviation behavior of E to obtain an upper bound on the probability of Algorithm 1 to deliver a wrong neighborhood, i.e., we bound the probability $\mathrm{P}\{\widehat{\mathcal{N}}(r) \neq \mathcal{N}(r)\}$, for an arbitrary but fixed node $r \in [p]$.

In order to use Theorem IV.1, we need to ensure $\beta_{\min} = \min_{r\in\mathrm{gsupp}(\boldsymbol{\beta}(\cdot))} \|\beta_r(\cdot)\|_{L^2}$ (with $\boldsymbol{\beta}(\cdot)$ given by (22)) to be sufficiently large. This is accomplished by assuming (8), which is valid for any process $\mathbf{x}[n] \in \mathcal{M}$, and implying via (29) the lower bound

$$\beta_{\min} \geq \rho_{\min}. \tag{47}$$

In order to ensure validity of (44), we need the following relation between the maximum correlation $\sup_{\theta\in[0,1)} \|\boldsymbol{\varepsilon}^H(\theta)\mathbf{X}(\theta)\|_\infty$ and the estimation error E in (46).

Lemma IV.2: Consider the multitask learning problem (21), *with observation vector* $\mathbf{y}(\theta)$ *and system matrix* $\mathbf{X}(\theta)$ *given by* (32) *and* (33), *based on the permuted observation* $\mathbf{D} = \mathbf{P}_r\big(\mathbf{x}[1], \ldots, \mathbf{x}[N]\big)$ *of the process* $\mathbf{x}[n] \in \mathcal{M}$. *We have*

$$\sup_{\theta\in[0,1)} \|\boldsymbol{\varepsilon}^H(\theta)\mathbf{X}(\theta)\|_\infty \leq 2E\sqrt{s_{\max}}U. \tag{48}$$

Proof: Appendix C. □

Note that due to (48) and (47), a sufficient condition for (44) to be satisfied is

$$E \leq \phi^2(\mathcal{S})\rho_{\min}/(64U s_{\max}^{3/2}). \tag{49}$$

The following result characterizes the multitask compatibility condition $\phi(\mathcal{S})$ of the system matrix $\mathbf{X}(\theta)$ given by (24), for a process $\mathbf{x}[n] \in \mathcal{M}$, i.e., in particular satisfying (2).

Lemma IV.3: Consider the multitask learning problem (21) *which is constructed according to* (23), (24), *based on the observed process* $\mathbf{x}[n] \in \mathcal{M}$. *If the estimation error* E *in* (46) *satisfies*

$$E \leq 1/(32 s_{\max}), \tag{50}$$

then, for any subset $\mathcal{S} \subseteq [p]$ *with* $|\mathcal{S}| \leq s_{\max}$, *the system matrix* $\mathbf{X}(\theta)$, *given for any* $\theta \in [0,1)$ *by* (24), *satisfies the multitask compatibility condition* (40) *with a constant*

$$\phi(\mathcal{S}) \geq 1/\sqrt{2}. \tag{51}$$

Proof: Appendix D. □

Combining Lemma IV.3 with the sufficient condition (49), we have that the multitask learning problem (21) satisfies the requirement (44) of Theorem IV.1 if

$$E < \rho_{\min}/(128U s_{\max}^{3/2}). \tag{52}$$

Indeed, the validity of (52) implies (50) since $\rho_{\min} \leq U$ which can be verified from the assumption (8) and the relations $\left|\left(\mathbf{S}_x^{-1}(\theta)\right)_{r,r}\right| \geq \lambda_{\min}\left(\mathbf{S}_x^{-1}(\theta)\right) \overset{(2)}{\geq} 1/U$, and $\left|\left(\mathbf{S}_x^{-1}(\theta)\right)_{r,r'}\right| \leq \lambda_{\max}\left(\mathbf{S}_x^{-1}(\theta)\right) \overset{(2)}{\leq} 1$

In what follows, we derive an upper bound on the probability that (52) is not satisfied for a process $\mathbf{x}[n] \in \mathcal{M}$. This will be done with the aid of

Lemma IV.4: Let $\widehat{\mathbf{S}}_x(\theta)$ *be the estimate of* $\mathbf{S}_x(\theta)$*, obtained according to* (11) *with sample size* N *and window function* $w[\cdot] \in \ell_1(\mathbb{Z})$*. For* $\nu \in [0, 1/2)$*,*

$$\mathrm{P}\{E \geq \nu + \mu_x^{(h_1)}\} \leq 2e^{-\frac{N\nu^2}{8\|w[\cdot]\|_1^2 U^2} + 2\log p + \log 2N}. \quad (53)$$

where $\mu_x^{(h_1)}$ *denotes the ACF moment* (9) *obtained for the weighting function* $h_1[m] := |1 - w[m](1 - |m|/N)|$ *for* $|m| < N$ *and* $h_1[m] := 1$*, else.*

Proof: Appendix E. □

3) Main Result: Using Lemma IV.4, we can characterize the probability of the condition (52) to hold. Since validity of (52) allows to invoke Theorem IV.1, we arrive at

Theorem IV.5: Consider a process $\mathbf{x}[n] \in \mathcal{M}$ *and the corresponding SDM estimate* (11)*. Then, if*

$$\frac{N(\rho_{\min}/256)^2}{8s_{\max}^3 \|w[\cdot]\|_1^2 U^4} - \log(2N) \geq \log(2p^2/\delta), \text{ and} \quad (54)$$

$$\mu_x^{(h_1)} < \frac{\rho_{\min}}{256 U s_{\max}^{3/2}}, \quad (55)$$

the probability of Algorithm 1, using $\lambda = \rho_{\min}/(16 s_{\max})$ *in* (30) *and* $\eta = \rho_{\min}/2$ *in* (34)*, selecting the neighborhood of node* $r \in [p]$ *not correctly, i.e.,* $\widehat{\mathcal{N}}(r) \neq \mathcal{N}(r)$*, is upper bounded as* $\mathrm{P}\{\widehat{\mathcal{N}}(r) \neq \mathcal{N}(r)\} \leq \delta$.

Note that Theorem IV.5 applies to the infinite dimensional mLASSO optimization problem in (30), thereby ignoring any discretization or numerical implementation issue. Nevertheless, if the discretization is fine enough, i.e., the number F of frequency intervals used for the discretized mLASSO (38) is sufficiently large, we expect that Theorem IV.5 accurately predicts the performance of the GMS method obtained by using Algorithm 1 with the discretized mLASSO (38) instead of the infinite dimensional mLASSO (30).

Furthermore, Theorem IV.5 considers the probability of (the first two steps of) Algorithm 1 to fail in selecting the correct neighborhood $\mathcal{N}(r)$ of a specific node r. Since any reasonable combination strategy in step 3 of Algorithm 1 (such as the "AND" and the "OR" rule discussed below Algorithm 1) will yield the correct CIG if all neighborhoods are estimated correctly, we obtain, via a union bound over all nodes $r \in [p]$, the following bound on the probability of Algorithm 1 yielding a wrong CIG.

Corollary IV.6: Consider a process $\mathbf{x}[n] \in \mathcal{M}$ *and the corresponding SDM estimate* (11)*. Then, if*

$$\frac{N(\rho_{\min}/256)^2}{8s_{\max}^3 \|w[\cdot]\|_1^2 U^4} - \log(2N) \geq \log(2p^3/\delta), \text{ and} \quad (56)$$

$$\mu_x^{(h_1)} \leq \frac{\rho_{\min}}{256 U s_{\max}^{3/2}}, \quad (57)$$

the probability of Algorithm 1 using $\lambda = \rho_{\min}/(16 s_{\max})$ *in* (30) *and* $\eta = \rho_{\min}/2$ *in* (34)*, yielding a wrong CIG, i.e.,* $\widehat{\mathcal{G}} \neq \mathcal{G}$*, is upper bounded as* $\mathrm{P}\{\widehat{\mathcal{G}}_x \neq \mathcal{G}_x\} \leq \delta$.

According to (56), neglecting the term $\log(2N)$ and assuming $\rho_{\min}$ fixed, the sample size N has to grow polynomially with the maximum node degree and logarithmically with the number of process components p. This polynomial and logarithmic scaling of the sample size N on the maximum node degree $s_{\max}$ and number of process components p, respectively, is a typical requirement for accurate GMS in the high-dimensional regime [9], [11], [14].

Note also that, according to (56), the sample size N has to grow with the squared ℓ_1 norm $\|w[\cdot]\|_1^2$ of the window function $w[\cdot]$ employed in the BT estimator (11). For the inequality (57) to hold, one typically has to use a window function $w[\cdot]$ whose effective support matches those of the process ACF $\mathbf{R}_x[m]$. Therefore, Theorem IV.5 suggests that the sample size has to grow with the square of the effective process correlation width (effective size of the ACF support), which is quantified by μ_x. However, some first results on the fundamental limits of GMS for time series in indicate that the required sample size should be effectively independent of the correlation width μ_x [36].

One explanation of the discrepancy between the sufficient condition (56) and the lower bounds [36] on the required sample size is that the derivation of Theorem IV.5 is based on requiring the SDM estimator $\widehat{\mathbf{S}}_x(\theta)$, given by (11), to be accurate *simultaneously* for all $\theta \in [0, 1)$. According to [37], the achievable uniform estimation accuracy, measured by the minimax risk, depends inversely on the correlation width μ_x. However, the analysis in [36] suggests that it is not necessary to accurately estimate the SDM $\mathbf{S}_x(\theta)$ for all θ simultaneously. Indeed, for a process $\mathbf{x}[n]$ with underlying CIG $\mathcal{G}_x$, the SDM values $\mathbf{S}_x(\theta)$ are coupled over frequency $\theta \in [0, 1)$ via the relation (3). Due to this coupling, the SDM needs to be estimated accurately only on average (over frequency θ). A more detailed performance analysis of the selection scheme in Algorithm 1, taking the coupling effect due to (3) into account, is an interesting direction for future work.

V. Numerical Experiments

The performance of the GMS method given by Algorithm 1 is assessed by two complementary numerical experiments[4]. In the first experiment we measure the ability of our method to correctly identify the edge set of the CIG of a synthetically generated process. In a second experiment, we apply our GMS method to electroencephalography (EEG) measurements, demonstrating that the resulting CIG estimate may be used for detecting the eye state (open/closed) of a person.

A. Synthetic Process

We generated a Gaussian process $\mathbf{x}[n]$ of dimension $p = 64$ by applying a finite impulse response filter $g[m]$ of length 2 to a zero-mean stationary white Gaussian noise process $\mathbf{e}[n] \sim \mathcal{N}(\mathbf{0}, \mathbf{C}_0)$. The covariance matrix $\mathbf{C}_0$ was chosen such that the resulting CIG $\mathcal{G}_x = ([p], \mathcal{E})$ satisfies (7) with $s_{\max} = 3$. The

[4]Matlab code to reproduce the results in this section is available upon request from the author.

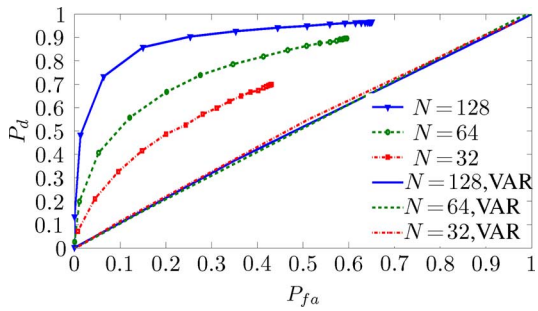

Fig. 1. ROC curves for the compressive selection scheme given by Algorithm 1 and for a VAR-model based GMS scheme presented in [3].

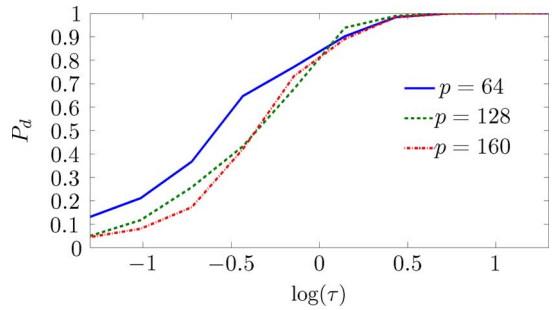

Fig. 2. Empirical detection probability P_d vs. rescaled sample size $\tau = N/(\log(p)s^3_{\max})$.

non-zero filter coefficients $g[0]$ and $g[1]$ are chosen such that the magnitude of the associated transfer function is uniformly bounded from above and below by positive constants, thereby ensuring condition (2).

We then computed the CIG estimate $\widehat{\mathcal{G}}_x$ using Algorithm 1 based on the discretized version (38) of the mLASSO (with $F = 4$) and the window function $w[m] = \exp(-m^2/44)$. In particular, we applied the *alternating direction method of multipliers (ADMM)* to the optimization problem (38) (cf. [38, Sec. 6.4]).[5] We set $\lambda = c_1\phi^2_{\min}\rho_{\min}/(18s_{\max}F)$ and $\eta = \rho_{\min}/2$, where c_1 was varied in the range $c_1 \in [10^{-3}, 10^3]$. Within the third step of Algorithm 1, we used the "OR" rule for combining the neighborhood estimates $\widehat{\mathcal{N}}(r)$.

In Fig. 1, we show receiver operating characteristic (ROC) curves with the empirical false alarm rate $P_{fa} := \frac{1}{M}\sum_{l\in[M]} \frac{|\widehat{\mathcal{E}}_l \backslash \mathcal{E}|}{p(p-1)/2-|\mathcal{E}|}$ and the empirical detection probability[6] $P_d := \frac{1}{M}\sum_{l\in[M]} \frac{|\widehat{\mathcal{E}}_l \cap \mathcal{E}|}{|\mathcal{E}|}$ for varying mLASSO parameter λ. Here, $\widehat{\mathcal{E}}_l$ denotes the edge set estimate obtained from Algorithm 1 during the l-th simulation run. We averaged over $M = 10$ i.i.d. simulation runs. As can be seen from Fig. 1, our selection scheme yields reasonable performance even if $N = 32$ only for a 64-dimensional process. We also adapted an existing VAR-based network learning method [3] in order to estimate the underlying CIG.[7]

The resulting ROC curves are also shown in Fig. 1. Note that the performance obtained for the VAR-based method is similar to a pure guess. The inferior performance of the VAR-based method is due to a model mismatch since the simulated process (being a moving average process) is not well approximated by a VAR process of order one.

We also evaluated the empirical detection probability P_d for fixed mLASSO parameter $\lambda = \rho_{\min}/10$ and varying rescaled sample size $\tau := N/(\log(p)s^3_{\max})$. According to Fig. 2, and as suggested by the bound (56) of Theorem IV.5, for a fixed squared norm $\|w[\cdot]\|_1^2$ (the window function $w[m]$ employed in (11) is fixed throughout the simulation), the rescaled sample size $\tau = N/(\log(p)s^3_{\max})$ seems to be an accurate performance indicator. In particular, the selection scheme in Algorithm 1 works well only if $\tau \gg 1$.

B. Eye State Detection

In this experiment, we evaluate the applicability of our GMS method for the problem of eye state detection based on EEG measurement data. This problem is relevant, e.g., for medical care or for driving drowsiness detection [39]. We used the EEG dataset donated by Oliver Roesler from Baden-Wuerttemberg Cooperative State University (DHBW), Stuttgart, Germany, and available at the UCI machine learning repository [40]. The dataset consists of 14980 time samples, each sample being a vector made up of 14 feature values. The true eye state was detected via a camera during the EEG recording.

As a first processing step, given the raw data, we removed parts of the time series which contain outliers. In a second step we performed a detrending operation by applying a boxcar filter of length 5. Based on the true eye state signal, which is equal to one if the eye was open and equal to zero if it was closed, we extracted two data blocks $\mathbf{D}_0$, $\mathbf{D}_1$, one corresponding to each state. We then applied Algorithm 1 with the discretized mLASSO (38) (with $F = 5$) instead of (30) and using again the OR-rule in the third step, i.e., $\widehat{\mathcal{G}}$ contains the edge (r, r') if either $r \in \widehat{\mathcal{N}}(r')$ or $r' \in \mathcal{N}(r)$. For the window function in the BT estimator (11) we used the choice $w[m] = \exp(-(m/59)^2)$. In Fig. 3, we show the two CIG estimates obtained for each of the two data blocks $\mathbf{D}_0, \mathbf{D}_1 \in \mathbb{R}^{14\times 1024}$ each corresponding to a sample size of $N = 1024$. As evident from Fig. 3, the resulting graph estimates for the two eye states differ significantly. In particular, the graph obtained for the "eye closed" state contains much more edges which are moreover localized at few nodes having relatively high degree. Thus, the CIG estimate delivered

[5]We used the all-zero initialization for the ADMM variables in our experiments. In general, the convergence of the ADMM implementation for LASSO type problems of the form (38) is not sensitive to the precise initialization of the optimization variables [38].

[6]The quantities P_{fa} and P_d may be interpreted as empirical proxies of a false alarm- and a true detection probability, respectively.

[7]In a nutshell, the authors of [3] set up a VAR model $\mathbf{x}[n] = \sum_{l\in[q]} \mathbf{A}^{(l)}\mathbf{x}[n-l] + \mathbf{w}[n]$, with known order q. The driving noise $\mathbf{w}[n]$ is modeled as i.i.d. $\mathbf{w}[n] \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ with known covariance matrix $\mathbf{\Sigma}$. It can be shown that the inverse SDM of such a VAR process is given by $\mathbf{S}_x^{-1}(\theta) = \mathbf{\Phi}^H(e^{-j\theta})\mathbf{\Sigma}^{-1}\mathbf{\Phi}(e^{-j\theta})$ with the characteristic polynomial $\mathbf{\Phi}(z) := \mathbf{I} - \sum_{l\in[q]} \mathbf{A}^{(l)}z^l$ [1]. Thus, for white driving noise $\mathbf{w}[n]$ with $\mathbf{\Sigma} = \sigma^2\mathbf{I}$, a Gaussian VAR-process with block-sparse VAR parameter matrices $\mathbf{A}^{(l)}$, implying a sparse matrix $\mathbf{\Phi}(e^{-j\theta})$, has a sparse CIG. The block-sparse parameter matrices $\mathbf{A}^{(l)}$ are estimated by solving a group LASSO problem (being similar to the mLASSO (30)). For the details of this VAR-based group LASSO approach, we refer to [3, Sec. II-A].

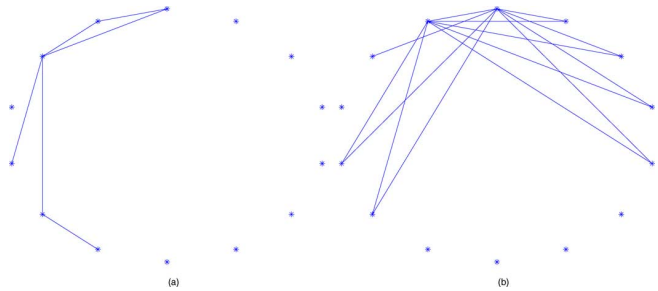

Fig. 3. Resulting CIG estimate for the EEG time series under different eye states. (a) "eye open". (b) "eye closed".

by Algorithm 1 could serve as an indicator for the eye state of a person based on EEG measurements.

VI. Conclusion

We proposed a nonparametric compressive selection scheme for inferring the CIG of a stationary discrete-time Gaussian vector process. This selection scheme is based on combining a BT estimator for the SDM with the mLASSO. The key idea behind this novel selection scheme is the formulation of the GMS problem for a stationary vector process as a multitask learning problem. This formulation lends itself to applying mLASSO to GMS for stationary vector processes. Drawing on an established performance characterization [20] of the mLASSO, we derived sufficient conditions on the observed sample size such that the probability of selecting a wrong CIG does not exceed a given (small) threshold. Some numerical experiments validate our theoretical performance analysis and show superior performance compared to an existing (VAR-based) method in case of model mismatch.

Our work may serve as the basis for some interesting avenues of further research, e.g., extending the concept of a CIG to processes with a singular SDM or introducing the notion of a frequency dependent CIG.

Appendix A
Proof of Lemma III.1

Let $\tilde{x}_r[n]$ and $\tilde{x}_t[n]$ denote $(2N-1)$-periodic discrete-time signals, with one period given by

$$\tilde{x}_{\{r,t\}}[n-1] := \begin{cases} (\mathbf{x}[n])_{\{r,t\}} & \text{for } n \in [N], \\ 0 & \text{for } n \in [2N-1] \setminus [N] \end{cases} \quad (58)$$

and corresponding DFTs

$$\tilde{X}_{\{r,t\}}[k] := \sum_{n=0}^{2N-2} \tilde{x}_{\{r,t\}}[n] \exp(-j2\pi kn/(2N-1)) \overset{(58)}{=} \sum_{n\in[N]} \big(\mathbf{x}[n]\big)_{\{r,t\}} \exp(-j2\pi k(n-1)/(2N-1)),$$

for $k = 0, \ldots, 2N-2$. Note that (cf. (14))

$$\tilde{X}_{\{r,t\}}[k] = \big(\mathbf{DF}\big)_{\{r,t\},k+1}. \quad (59)$$

Let us verify the equivalence of (16) and (11) entry-wise. To this end, for arbitrary but fixed $r, t \in [p]$, consider the entry $\hat{s} := \big(\widehat{\mathbf{S}}_x(\theta)\big)_{r,t}$ of the SDM estimate given by (11). By inspecting (11),

$$\hat{s} = (1/N) \sum_{m=-N+1}^{N-1} w[m] \exp(-2\pi m\theta) \cdot (\tilde{x}_r \otimes \tilde{x}_t)[m], \quad (60)$$

where $(\tilde{x}_r \otimes \tilde{x}_t)[m] = \sum_{n=0}^{2N-2} \tilde{x}_r[n+m]\tilde{x}_t[n]$ denotes the periodic autocorrelation function of $\tilde{x}_r[n]$ and $\tilde{x}_t[n]$. The DFTs $W[k]$ and $V[k]$ of the $(2N-1)$-periodic signals $w[m]\exp(-2\pi m\theta)$ and $(\tilde{x}_r \otimes \tilde{x}_t)[m]$ are given by [41, Ch. 8], using $\theta_k := 2\pi(k-1)/(2N-1)$,

$$W[k] = W(\theta + \theta_{k+1}) \text{ and } V[k] = \tilde{X}_r[k]\tilde{X}_t^*[k], \quad (61)$$

respectively. Here, $W(\theta) = \sum_{m=-\infty}^{\infty} w[m]\exp(-j2\pi\theta m)$ denotes the DTFT of the window function $w[m]$ employed in the BT estimator (cf. (11)). Using again [41, Ch. 8], we obtain from (60) that

$$\begin{aligned} \hat{s} &= \frac{1}{N(2N-1)} \sum_{k=0}^{2N-2} W[k]V^*[k] \\ &\overset{(61)}{=} \frac{1}{N(2N-1)} \sum_{k\in[2N-1]} W(\theta+\theta_k)\tilde{X}_r^*[k]\tilde{X}_t[k] \\ &\overset{(59),(15)}{=} \frac{1}{N} \sum_{k\in[2N-1]} (\mathbf{DF})_{t,k}\big(\mathbf{W}^2(\theta)\big)_{k,k}\big((\mathbf{DF})^H\big)_{k,r}. \end{aligned} \quad (62)$$

Note that the last expression is nothing but the (r,t)-th entry of the RHS in (16).

Appendix B
Proof of Theorem IV.1

We will need the following lemma, which is a straightforward generalization of [20, Thm. 6.1].

Lemma B.1: Consider the multitask learning problem (21) *with parameter vector* $\boldsymbol{\beta}(\cdot) \in \ell_q([0,1))$, *observation vector* $\mathbf{y}(\theta)$ *and system matrix* $\mathbf{X}(\theta)$ *defined by* (22), (32) *and* (33), *respectively. Suppose,*

$$\sup_{\theta\in[0,1)} \|\boldsymbol{\varepsilon}^H(\theta)\mathbf{X}(\theta)\|_\infty < \frac{\lambda}{4}, \text{ and } \operatorname{gsupp}(\boldsymbol{\beta}(\cdot)) \subseteq \mathcal{S}, \quad (63)$$

with an index set $\mathcal{S} \subseteq [q]$ *of size* $s_{\max} = |\mathcal{S}|$. *If the system matrix* $\mathbf{X}(\theta)$ *possesses a positive multitask compatibility constant* $\phi(\mathcal{S}) > 0$, *the mLASSO estimate* $\hat{\boldsymbol{\beta}}[\mathbf{y}(\cdot), \mathbf{X}(\cdot)]$ *given by* (30) *satisfies*

$$\|\boldsymbol{\beta}(\cdot) - \hat{\boldsymbol{\beta}}(\cdot)\|_1 < \frac{4\lambda s_{\max}}{\phi^2(\mathcal{S})}. \quad (64)$$

Evaluating Lemma B.1 for the specific choice $\lambda = \frac{\phi^2(\mathcal{S})\beta_{\min}}{8s_{\max}}$, we have that, under condition (44) (which ensures (63)), the mLASSO estimate $\hat{\boldsymbol{\beta}}[\mathbf{y}(\cdot), \mathbf{X}(\cdot)]$ satisfies

$$\|\boldsymbol{\beta}(\cdot) - \hat{\boldsymbol{\beta}}(\cdot)\|_1 < \beta_{\min}/2. \quad (65)$$

This implies, in turn, for any $r \in \operatorname{gsupp}(\boldsymbol{\beta}(\cdot))$,

$$\|\hat{\beta}_r(\cdot)\|_{L^2} \geq \|\beta_r(\cdot)\|_{L^2} - |\|\beta_r(\cdot)\|_{L^2} - \|\hat{\beta}_r(\cdot)\|_{L^2}| \overset{(42),(65)}{>} \beta_{\min}/2$$

and similarly for any $r \in [p] \setminus \operatorname{gsupp}(\boldsymbol{\beta}(\cdot))$,

$$\|\hat{\beta}_r(\cdot)\|_{L^2} \leq \|\beta_r(\cdot)\|_{L^2} + |\|\beta_r(\cdot)\|_{L^2} - \|\hat{\beta}_r(\cdot)\|_{L^2}| \overset{(65)}{<} \beta_{\min}/2.$$

Thus, the set $\{r : \|\hat{\beta}_r(\cdot)\|_{L^2} \geq \beta_{\min}/2\}$ coincides with the true generalized support $\operatorname{gsupp}(\boldsymbol{\beta}(\cdot))$.

Appendix C
Proof of Lemma IV.2

Let us recall the partitioning (27) and (36) of the SDM $\mathbf{S}_x(\theta)$ and SDM estimate $\widehat{\mathbf{S}}_x(\theta)$, respectively. For the sake of light notation, we consider throughout the remainder of this proof an arbitrary but fixed frequency θ and drop the argument of the frequency dependent variables, e.g., $\mathbf{S}_x(\theta)$, $\mathbf{G}(\theta)$, $\mathbf{c}(\theta)$, $\widehat{\mathbf{S}}_x(\theta)$, $\widehat{\mathbf{G}}(\theta)$, $\hat{\mathbf{c}}(\theta)$ and so on. Moreover, we denote the rth columns of $\mathbf{X}$, $\mathbf{G}$ and $\widehat{\mathbf{G}}$ by $\mathbf{x}_r$, $\mathbf{g}_r$ and $\hat{\mathbf{g}}_r$, respectively.

First observe if we define the matrix $\mathbf{J} \in \mathbb{R}^{(p-1)\times p}$ by setting $J_{k,l} = 1$ if $l = k+1$ and $J_{k,l} = 0$ else, we have

$$\mathbf{c} = \mathbf{J}\mathbf{S}_x\mathbf{e}_1. \tag{66}$$

Consider the system matrix $\mathbf{X}$ given by (33) and note that, by comparing (25) with (36), we have

$$\mathbf{X}^H\mathbf{X} = \widehat{\mathbf{G}}. \tag{67}$$

We also require a helpful identity for certain sub-matrices of the SDM:

$$\big(\mathbf{S}_x\big)_{r+1,1} = \mathbf{g}_r^H\mathbf{G}^{-1}\mathbf{c}. \tag{68}$$

This can be verified by

$$\mathbf{g}_r^H\mathbf{G}^{-1}\mathbf{c} \overset{(66)}{=} \mathbf{e}_r^H\mathbf{G}\mathbf{G}^{-1}\mathbf{J}\mathbf{S}_x\mathbf{e}_1 = \mathbf{e}_r^H\mathbf{J}\mathbf{S}_x\mathbf{e}_1 = \big(\mathbf{S}_x\big)_{r+1,1}.$$

Note that

$$\begin{aligned}|\mathbf{x}_r^H\boldsymbol{\varepsilon}| &\overset{(21)}{=} |\mathbf{x}_r^H(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})| \\ &\overset{(25),(22)}{=} \big|\big(\widehat{\mathbf{S}}_x\big)_{r+1,1} - \big(\hat{\mathbf{g}}_r - \mathbf{g}_r\big)^H\mathbf{G}^{-1}\mathbf{c} - \mathbf{g}_r^H\mathbf{G}^{-1}\mathbf{c}\big|. \end{aligned}\tag{69}$$

Combining (69) with (68),

$$\begin{aligned}|\mathbf{x}_r^H\boldsymbol{\varepsilon}| &= \big|\big(\widehat{\mathbf{S}}_x - \mathbf{S}_x\big)_{r+1,1} - \big(\hat{\mathbf{g}}_r - \mathbf{g}_r\big)^H\mathbf{G}^{-1}\mathbf{c}\big| \\ &\overset{(22)}{\leq} \big|\big(\widehat{\mathbf{S}}_x - \mathbf{S}_x\big)_{r+1,1}\big| + \big|\big(\hat{\mathbf{g}}_r^H - \mathbf{g}_r^H\big)\boldsymbol{\beta}\big|. \end{aligned}\tag{70}$$

Applying the Cauchy-Schwarz inequality to the second term in (70) and using

$$|\operatorname{gsupp}(\boldsymbol{\beta}(\cdot))| \overset{(26)}{=} |\mathcal{N}(1)| \overset{(7)}{\leq} s_{\max}, \tag{71}$$

we obtain

$$|\mathbf{x}_r^H\boldsymbol{\varepsilon}| \leq \big\|\mathbf{S}_x - \widehat{\mathbf{S}}_x\big\|_\infty(1 + \sqrt{s_{\max}}\|\boldsymbol{\beta}\|_2). \tag{72}$$

Inserting the bound

$$\|\boldsymbol{\beta}\|_2 \overset{(22)}{=} \big\|\mathbf{G}^{-1}\mathbf{c}\big\|_2 \overset{(2)}{\leq} U,$$

into (72), finally yields

$$|\mathbf{x}_r^H\boldsymbol{\varepsilon}| \leq \big\|\mathbf{S}_x - \widehat{\mathbf{S}}_x\big\|_\infty(1 + \sqrt{s_{\max}}U) \leq 2\big\|\mathbf{S}_x - \widehat{\mathbf{S}}_x\big\|_\infty\sqrt{s_{\max}}U.$$

Appendix D
Proof of Lemma IV.3

We first state an inequality which applies to any vector function $\boldsymbol{\beta}(\cdot) \in \ell_q([0,1))$ for some q. In particular,

$$\begin{aligned}\int_{\theta=0}^1 \|\boldsymbol{\beta}(\theta)\|_1^2 d\theta &= \int_{\theta=0}^1 \sum_{r\in[q]} |\beta_r(\theta)| \sum_{r'\in[q]} |\beta_{r'}(\theta)| d\theta \\ &\overset{(a)}{\leq} \sum_{r\in[q]}\sum_{r'\in[q]} \|\beta_r(\cdot)\|_{L^2}\|\beta_{r'}(\cdot)\|_{L^2} \\ &= \|\boldsymbol{\beta}(\cdot)\|_1^2, \end{aligned}\tag{73}$$

where step (a) is due to the Cauchy-Schwarz inequality. This, in turn, implies for any $\boldsymbol{\beta}'(\cdot) \in \mathbb{A}(\mathcal{S})$ (cf. (41)) that

$$\int_{\theta=0}^1 \|\boldsymbol{\beta}'(\theta)\|_1^2 d\theta \overset{(73)}{\leq} \|\boldsymbol{\beta}'(\cdot)\|_1^2 \overset{(41)}{\leq} 16\|\boldsymbol{\beta}'_{\mathcal{S}}(\cdot)\|_1^2. \tag{74}$$

Observe that

$$\begin{aligned}&\|\mathbf{X}(\cdot)\boldsymbol{\beta}(\cdot)\|_2^2 \\ &= \int_{\theta=0}^1 \boldsymbol{\beta}^H(\theta)\mathbf{X}^H(\theta)\mathbf{X}(\theta)\boldsymbol{\beta}(\theta)d\theta \\ &\overset{(67)}{=} \int_{\theta=0}^1 \boldsymbol{\beta}^H(\theta)\mathbf{G}(\theta)\boldsymbol{\beta}(\theta)d\theta + \boldsymbol{\beta}^H(\theta)\big[\widehat{\mathbf{G}}(\theta) - \mathbf{G}(\theta)\big]\boldsymbol{\beta}(\theta)d\theta. \end{aligned}\tag{75}$$

Since $\mathbf{a}^H\mathbf{M}\mathbf{a} \leq \|\mathbf{M}\|_\infty\|\mathbf{a}\|_1^2$ for any vector $\mathbf{a} \in \mathbb{C}^q$ and matrix $\mathbf{M} \in \mathbb{C}^{q\times q}$, we obtain further

$$\begin{aligned}&\|\mathbf{X}(\cdot)\boldsymbol{\beta}(\cdot)\|_2^2 \\ &\overset{(75)}{\geq} \int_{\theta=0}^1 \boldsymbol{\beta}^H(\theta)\mathbf{G}(\theta)\boldsymbol{\beta}(\theta) - \|\widehat{\mathbf{G}}(\theta) - \mathbf{G}(\theta)\|_\infty\|\boldsymbol{\beta}(\theta)\|_1^2 d\theta \\ &\geq \int_{\theta=0}^1 \boldsymbol{\beta}^H(\theta)\mathbf{G}(\theta)\boldsymbol{\beta}(\theta)d\theta \\ &\quad - \sup_{\theta\in[0,1)} \|\widehat{\mathbf{S}}_x(\theta) - \mathbf{S}_x(\theta)\|_\infty \int_{\theta=0}^1 \|\boldsymbol{\beta}(\theta)\|_1^2 d\theta \\ &\overset{(50)}{\geq} \int_{\theta=0}^1 \boldsymbol{\beta}^H(\theta)\mathbf{G}(\theta)\boldsymbol{\beta}(\theta)d\theta - \frac{1}{32s_{\max}}\|\boldsymbol{\beta}(\cdot)\|_1^2. \end{aligned}\tag{76}$$

Combining (76) with (74), we have for any $\boldsymbol{\beta}(\cdot) \in \mathbb{A}(\mathcal{S})$,

$$\begin{aligned}s_{\max}\frac{\|\mathbf{X}(\cdot)\boldsymbol{\beta}(\cdot)\|_2^2}{\|\boldsymbol{\beta}_{\mathcal{S}}(\cdot)\|_1^2} &\geq s_{\max}\frac{\int_{\theta=0}^1 \boldsymbol{\beta}^H(\theta)\mathbf{G}(\theta)\boldsymbol{\beta}(\theta)d\theta}{\|\boldsymbol{\beta}_{\mathcal{S}}(\cdot)\|_1^2} - 1/2 \\ &\overset{(2)}{\geq} 1 - 1/2 = 1/2. \end{aligned}\tag{77}$$

Appendix E
Proof of Lemma IV.4

We will establish Lemma IV.4 by bounding $\big|\big(\widehat{\mathbf{S}}_x(\theta) - \mathbf{S}_x(\theta)\big)_{k,l}\big|$ for a fixed pair $k, l \in [p]$ and then appealing to a union bound over all pairs $k, l \in [p]$.

Set $\hat{\sigma}(\theta) := [\widehat{\mathbf{S}}_x(\theta)]_{k,l}$, $\bar{\sigma}(\theta) := \mathrm{E}\{\hat{\sigma}(\theta)\}$, $\sigma(\theta) := [\mathbf{S}_x(\theta)]_{k,l}$ and the bias $b(\theta) := \sigma(\theta) - \bar{\sigma}(\theta)$. By the triangle inequality,

$$\begin{aligned}
&\mathrm{P}\{\sup_{\theta\in[0,1)} |\hat{\sigma}-\sigma| \geq \nu + \mu_x^{(h_1)}\} \\
&\leq \mathrm{P}\{\sup_{\theta\in[0,1)} |\hat{\sigma}(\theta)-\bar{\sigma}(\theta)| + \sup_{\theta\in[0,1)} |b(\theta)| \geq \nu + \mu_x^{(h_1)}\} \\
&\leq \mathrm{P}\{\sup_{\theta\in[0,1)} |\hat{\sigma}(\theta)-\bar{\sigma}(\theta)| \geq \nu\}, \qquad (78)
\end{aligned}$$

where the last inequality holds since for any $\theta \in [0,1)$, the bias satisfies $|b(\theta)| \leq \mu_x^{(h_1)}$, with $h_1[m]$ as defined in Lemma IV.4, which is verified next.

From

$$\begin{aligned}
&\mathrm{E}\{\widehat{\mathbf{S}}_x(\theta)\} \\
&\overset{(11)}{=} \mathrm{E}\left\{ \frac{1}{N} \sum_{m=0}^{N-1} w[m] \sum_{n\in[N-|m|]} \mathbf{x}[n+m]\mathbf{x}^T[n] e^{-j2\pi\theta m} \right. \\
&\quad \left. + \frac{1}{N} \sum_{m=-N+1}^{-1} w[m] \sum_{n\in[N-|m|]} \mathbf{x}[n]\mathbf{x}^T[n-m] e^{-j2\pi\theta m} \right\} \\
&\overset{(13)}{=} \sum_{m\in\mathbb{Z}} w[m](1-|m|/N)\mathbf{R}_x[m] e^{-j2\pi\theta m},
\end{aligned}$$

we obtain

$$|\sigma(\theta) - \bar{\sigma}(\theta)| = \Big| \sum_{m\in\mathbb{Z}} h_1[m] \big[\mathbf{R}_x[m]\big]_{k,l} e^{-j2\pi\theta m} \Big| \overset{(9)}{\leq} \mu_x^{(h_1)}. \qquad (79)$$

Similarly, with $\widetilde{\mathcal{N}} := \{-N+1, \ldots, N-1\}$,

$$\hat{\sigma}(\theta) - \bar{\sigma}(\theta) \overset{(11)}{=} (1/N) \sum_{m\in\widetilde{\mathcal{N}}} w[m] q_{k,l}[m] e^{-j2\pi\theta m}, \qquad (80)$$

where $q_{k,l}[m] := \mathbf{x}_k^T \mathbf{J}_m \mathbf{x}_l - \mathrm{E}\{\mathbf{x}_k^T \mathbf{J}_m \mathbf{x}_l\}$. Here, $\mathbf{x}_k := (x_k[1], \ldots, x_k[N])^T \in \mathbb{R}^N$, $\mathbf{x}_l := (x_l[1], \ldots, x_l[N])^T \in \mathbb{R}^N$ and the matrix $\mathbf{J}_m \in \{0,1\}^{N\times N}$ is defined element-wise as $\big(\mathbf{J}_m\big)_{v,w} = 1$ for $w - v = m$ and $\big(\mathbf{J}_m\big)_{v,w} = 0$ else. Note that $\mathbf{J}_m = \mathbf{J}_{-m}^T$ and $\|\mathbf{J}_m\|_2 \leq 1$. By (80), for any $\theta \in [0,1)$,

$$\begin{aligned}
|\hat{\sigma}(\theta) - \bar{\sigma}(\theta)\}| &\leq (1/N) \sum_{m\in\widetilde{\mathcal{N}}} w[m] |q_{k,l}[m]| \\
&\leq (1/N) \|w[\cdot]\|_1 \max_{m\in\widetilde{\mathcal{N}}} |q_{k,l}[m]| \qquad (81)
\end{aligned}$$

In order to upper bound the probability $\mathrm{P}\{\sup_{\theta\in[0,1)} |\hat{\sigma}(\theta) - \bar{\sigma}(\theta)| \geq \nu\}$, we now bound the probability of the event $\mathcal{A} := \{\max_{m\in\widetilde{\mathcal{N}}} (1/N)|q_{k,l}[m]| \geq \tilde{\nu}\}$, by first considering the large deviation behavior of $(1/N)|q_{k,l}[m]|$ for a specific m and then using a union bound over all $m \in \widetilde{\mathcal{N}}$.

Since we assume the process $\mathbf{x}[n]$ to be Gaussian and stationary, the random vectors $\mathbf{x}_k$ and $\mathbf{x}_l$, defining the random variable $q_{k,l}[m]$, are zero-mean normally distributed with Toeplitz covariance matrices $\mathbf{C}_k = \mathrm{E}\{\mathbf{x}_k \mathbf{x}_k{}^T\}$ and $\mathbf{C}_l = \mathrm{E}\{\mathbf{x}_l \mathbf{x}_l{}^T\}$, whose first row is given by $\big\{\big(\mathbf{R}_x[m]\big)_{k,k}\big\}_{m\in[N]}$ and $\big\{\big(\mathbf{R}_x[m]\big)_{l,l}\big\}_{m\in[N]}$, respectively. According to [42, Lemma 4.1], and due to the Fourier relationship (1), we can bound the spectral norm of $\mathbf{C}_k$ as

$$\|\mathbf{C}_k\|_2 \leq \max_{\theta\in[0,1]} \big|\big(\mathbf{S}_x(\theta)\big)_{k,k}\big| \overset{(a)}{\leq} U.$$

Here, step (a) follows from (2) together with the matrix norm inequality $\|\cdot\|_\infty \leq \|\cdot\|_2$ [43, p. 314]. Similarly, one can also verify that $\|\mathbf{C}_l\|_2 \leq U$.

Therefore, for any $\tilde{\nu} < 1/2$, we can invoke Lemma F.2 with the choices $\mathbf{x} = \mathbf{x}_k$, $\mathbf{y} = \mathbf{x}_l$, $\lambda_{\max} = U \geq 1$, $\mathbf{Q} = \mathbf{J}_m$ and $\lambda'_{\max} = \|\mathbf{J}_m\|_2 \leq 1$, yielding

$$\mathrm{P}\{(1/N)|q_{k,l}[m]| \geq \tilde{\nu}\} \overset{(95)}{\leq} 2\exp\left(-\frac{N\tilde{\nu}^2}{8U^2}\right). \qquad (82)$$

Then, by a union bound over all $m \in \widetilde{\mathcal{N}}$,

$$\mathrm{P}\{\mathcal{A}\} \leq 2\exp\left(-\frac{N\tilde{\nu}^2}{8U^2} + \log(2N)\right), \qquad (83)$$

and, in turn,

$$\begin{aligned}
&\mathrm{P}\{\sup_{\theta\in[0,1)} |\hat{\sigma}(\theta)-\bar{\sigma}(\theta)| \geq \nu\} \\
&\overset{(81)}{\leq} \mathrm{P}\{\max_{m\in\widetilde{\mathcal{N}}} \frac{1}{N}|q_{k,l}[m]| \geq \frac{\nu}{\|w[\cdot]\|_1}\} \\
&\overset{(83)}{\leq} 2\exp(-N\nu^2/(8\|w[\cdot]\|_1^2 U^2) + \log 2N)). \qquad (84)
\end{aligned}$$

Applying (84) to (78), we have for any $\nu < 1/2$ that

$$\mathrm{P}\{\sup_{\theta\in[0,1)} |\hat{\sigma}(\theta)-\sigma(\theta)| \geq \nu + \mu_x^{(h_1)}\} \leq 2e^{-\frac{N\nu^2}{8\|w[\cdot]\|_1^2 U^2} + \log(2N)}.$$

Another application of the union bound (over all p^2 pairs $(k,l) \in [p] \times [p]$) finally yields (53).

Appendix F
Large Deviations of a Gaussian Quadratic Form

Lemma F.1: Consider the quadratic form $q := \mathbf{w}^T\mathbf{Q}\mathbf{w}$ with real-valued standard normal vector $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and a real-valued symmetric matrix $\mathbf{Q} \in \mathbb{R}^{N\times N}$ with $\|\mathbf{Q}\|_2 \leq \lambda_{\max}$. For any $\nu < 1/2$,

$$\mathrm{P}\{q - \mathrm{E}\{q\} \geq N\nu\} \leq \exp\big(-N\nu^2/(8\max\{\lambda_{\max}^2, 1\})\big). \qquad (85)$$

Proof: Our argument closely follows the techniques used in [10]. Consider the eigenvalue decomposition of $\mathbf{Q}$, i.e.,

$$\mathbf{Q} = \sum_{l\in[N]} q_l \mathbf{v}_l \mathbf{v}_l^T, \qquad (86)$$

with eigenvalues $q_l \in \mathbb{R}$ and eigenvectors $\{\mathbf{v}_l\}_{l\in[N]}$ forming an orthonormal basis for $\mathbb{R}^N$. Note that, for any $l \in [N]$, we have $|q_l| \leq \|\mathbf{Q}\|_2 \leq \lambda_{\max}$. Based on (86), we can rewrite the quadratic form $q = \mathbf{w}^T\mathbf{Q}\mathbf{w}$ as

$$q = \sum_{l\in[N]} q_l z_l^2, \qquad (87)$$

with i.i.d. standard Gaussian random variables $z_l \sim \mathcal{N}(0,1)$, for $l \in [N]$. We then obtain

$$\begin{aligned}
&\mathrm{P}\{q - \mathrm{E}\{q\} \geq N\nu\} \\
&= \mathrm{P}\{\mathbf{w}^T\mathbf{Q}\mathbf{w} - \mathrm{E}\{\mathbf{w}^T\mathbf{Q}\mathbf{w}\} \geq N\nu\} \\
&\overset{(87)}{=} \mathrm{P}\{\sum_{l\in[N]} q_l(z_l^2 - 1) \geq N\nu\} \\
&\overset{\gamma\geq 0}{=} \mathrm{P}\big\{\gamma\big[\sum_{l\in[N]} q_l(z_l^2 - 1) - N\nu\big] \geq 0\big\} \\
&\leq \mathrm{E}\big\{\exp\big(\gamma\big[\sum_{l\in[N]} q_l(z_l^2 - 1) - N\nu\big]\big)\big\}, \qquad (88)
\end{aligned}$$

for any positive $\gamma > 0$. In what follows, we set

$$\gamma = \nu/(4\max\{\lambda_{\max}^2, 1\}), \tag{89}$$

which implies, since $|q_l| < \lambda_{\max}$ and $\nu < 1/2$ by assumption,

$$2|q_l|\gamma = 2|q_l|\nu/(4\max\{\lambda_{\max}^2, 1\}) < 1/2. \tag{90}$$

Due to (90), we also have $|\gamma q_l| < 1/2$ and can therefore use the identity

$$\mathrm{E}\{\exp(az_l^2)\} = 1/\sqrt{1-2a}, \tag{91}$$

valid for a standard Gaussian random variable $z_l \sim \mathcal{N}(0,1)$ and $|a| < 1/2$. Observe that

$$\begin{aligned}
&\mathrm{P}\{q - \mathrm{E}\{q\} \geq N\nu\} \\
&\overset{(88)}{\leq} \mathrm{E}\big\{\exp\big(\gamma\big[\sum_{l\in[N]} q_l(z_l^2-1) - N\nu\big]\big)\big\} \\
&= \exp\big(-\gamma\big[\sum_{l\in[N]} q_l + N\nu\big]\big)\mathrm{E}\big\{\exp\big(\gamma\sum_{l\in[N]} q_l z_l^2\big)\big\}.
\end{aligned} \tag{92}$$

Since the variables z_l are i.i.d.,

$$\mathrm{E}\big\{\exp\big(\gamma\sum_{l\in[N]} q_l z_l^2\big)\big\} \overset{(91)}{=} \exp\left(\sum_{l\in[N]} -(1/2)\log(1-2\gamma q_l)\right),$$

which, upon inserting into (92) yields

$$\begin{aligned}
&\mathrm{P}\{q - \mathrm{E}\{q\} \geq N\nu\} \\
&\qquad \leq \exp\big(-\sum_{l\in[N]} \big[\gamma q_l + \frac{1}{2}\log(1-2\gamma q_l)\big] - \gamma N\nu\big).
\end{aligned} \tag{93}$$

By (90), we can then apply the inequality $\log(1-a) > -a - a^2$ (valid for $|a| < 1/2$) to (93), yielding

$$\begin{aligned}
&\mathrm{P}\{q - \mathrm{E}\{q\} \geq N\nu\} \\
&\leq \exp\big(\sum_{l\in[N]} -\gamma q_l + \gamma q_l + 2\gamma^2 q_l^2 - \gamma N\nu\big) \\
&\overset{|q_l|\leq\lambda_{\max}}{\leq} \exp\big(-N(\gamma\nu - 2\gamma^2\lambda_{\max}^2)\big).
\end{aligned} \tag{94}$$

Putting together the pieces,

$$\begin{aligned}
&\mathrm{P}\{q - \mathrm{E}\{q\} \geq N\nu\} \\
&\overset{(94)}{\leq} \exp\big(-N(\gamma\nu - 2\gamma^2\lambda_{\max}^2)\big) \\
&\overset{(89)}{\leq} \exp\big(-N(\gamma\nu - (1/2)\gamma\nu\lambda_{\max}^2/\max\{\lambda_{\max}^2, 1\})\big) \\
&\leq \exp\big(-N\gamma\nu/2\big) \\
&\overset{(89)}{=} \exp\big(-N\nu^2/(8\max\{\lambda_{\max}^2, 1\})\big).
\end{aligned}$$

□

Lemma F.2: Consider two real-valued zero-mean random vectors $\mathbf{x} \in \mathbb{R}^N$ and $\mathbf{y} \in \mathbb{R}^N$, such that the stacked vector $\mathbf{z} := \big(\mathbf{x}^T\ \mathbf{y}^T\big)^T \in \mathbb{R}^{2N}$ is zero-mean multivariate normally distributed, i.e., $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{C}_z)$ with covariance matrix $\mathbf{C}_z := \mathrm{E}\{\mathbf{z}\mathbf{z}^T\}$. Let the individual covariance matrices satisfy $\|\mathbf{C}_x\|_2 \leq \lambda_{\max}$, $\|\mathbf{C}_y\|_2 \leq \lambda_{\max}$. We can then characterize the large deviations of the quadratic form $q := \mathbf{y}^T\mathbf{Q}\mathbf{x}$, with an arbitrary (possibly non-symmetric) real-valued matrix $\mathbf{Q} \in \mathbb{R}^{N\times N}$ satisfying $\|\mathbf{Q}\|_2 \leq \lambda'_{\max}$, as

$$\begin{aligned}
&\mathrm{P}\{|q - \mathrm{E}\{q\}| \geq N\nu\} \\
&\qquad \leq 2\exp\big(-N\nu^2/(8\max\{\lambda_{\max}'^2\lambda_{\max}^2, 1\})\big),
\end{aligned} \tag{95}$$

valid for any $\nu < 1/2$.

Proof: Introducing the shorthand $p(\nu) := \mathrm{P}\{|q - \mathrm{E}\{q\}| \geq N\nu\}$, an application of the union bound yields

$$p(\nu) \leq \underbrace{\mathrm{P}\{q - \mathrm{E}\{q\} \geq N\nu\}}_{:=p_+(\nu)} + \underbrace{\mathrm{P}\{q - \mathrm{E}\{q\} \leq -N\nu\}}_{:=p_-(\nu)}. \tag{96}$$

We will derive an upper bound on $p(\nu)$ by separately upper bounding $p_+(\nu)$ and $p_-(\nu)$. The derivations are completely analogous and we will only detail the derivation of the upper bound on $p_+(\nu)$.

Defining the matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{N\times 2N}$ via the matrix square root of the covariance matrix $\mathbf{C}_z$, i.e.,

$$\begin{pmatrix}\mathbf{A}\\ \mathbf{B}\end{pmatrix} := \mathbf{C}_z^{1/2}, \tag{97}$$

we have the following innovation representation for the random vectors $\mathbf{x}$ and $\mathbf{y}$:

$$\mathbf{x} = \mathbf{A}\mathbf{v}, \text{ and } \mathbf{y} = \mathbf{B}\mathbf{v}, \tag{98}$$

where $\mathbf{v} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is a standard normally distributed random vector of length $2N$. Note that $\mathbf{C}_x = \mathbf{A}\mathbf{A}^T$ and $\mathbf{C}_y = \mathbf{B}\mathbf{B}^T$, implying $\|\mathbf{A}\|_2 = \sqrt{\|\mathbf{C}_x\|_2}$ and $\|\mathbf{B}\|_2 = \sqrt{\|\mathbf{C}_y\|_2}$. Thus,

$$\|\mathbf{A}\|_2 \leq \sqrt{\lambda_{\max}}, \text{ and } \|\mathbf{B}\|_2 \leq \sqrt{\lambda_{\max}}. \tag{99}$$

Let us further develop

$$\begin{aligned}
p_+(\nu) &= \mathrm{P}\{\mathbf{y}^T\mathbf{Q}\mathbf{x} - \mathrm{E}\{\mathbf{y}^T\mathbf{Q}\mathbf{x}\} \geq N\nu\} \\
&\overset{(98)}{=} \mathrm{P}\{\mathbf{v}^T\mathbf{B}^T\mathbf{Q}\mathbf{A}\mathbf{v} - \mathrm{E}\{\mathbf{v}^T\mathbf{B}^T\mathbf{Q}\mathbf{A}\mathbf{v}\} \geq N\nu\} \\
&\overset{(a)}{=} \mathrm{P}\{\mathbf{v}^T\mathbf{M}\mathbf{v} - \mathrm{E}\{\mathbf{v}^T\mathbf{M}\mathbf{v}\} \geq N\nu\},
\end{aligned} \tag{100}$$

with the symmetric matrix

$$\mathbf{M} = (1/2)[\mathbf{B}^T\mathbf{Q}\mathbf{A} + \mathbf{A}^T\mathbf{Q}^T\mathbf{B}]. \tag{101}$$

In (100), step (a) follows from the identity $\mathbf{v}^T\mathbf{D}\mathbf{v} = (1/2)[\mathbf{v}^T\mathbf{D}\mathbf{v} + \mathbf{v}^T\mathbf{D}^T\mathbf{v}]$, which holds for an arbitrary matrix $\mathbf{D} \in \mathbb{R}^{2N\times 2N}$. Combining (101) with (99) yields

$$\begin{aligned}
\|\mathbf{M}\|_2 &\overset{(101)}{=} (1/2)\|\mathbf{B}^T\mathbf{Q}\mathbf{A} + \mathbf{A}^T\mathbf{Q}^T\mathbf{B}\|_2 \\
&\overset{(a)}{\leq} (1/2)(\|\mathbf{B}^T\|_2\|\mathbf{Q}\|_2\|\mathbf{A}\|_2 + \|\mathbf{A}^T\|_2\|\mathbf{Q}^T\|_2\|\mathbf{B}\|_2) \\
&= \|\mathbf{B}\|_2\|\mathbf{Q}\|_2\|\mathbf{A}\|_2 \overset{(99)}{\leq} \lambda_{\max}\lambda'_{\max},
\end{aligned} \tag{102}$$

where step (a) is due to the triangle inequality and submultiplicativity of the spectral norm. Using (102), the application of Lemma F.1 to (100) yields

$$p_+(\nu) \leq \exp\big(-N\nu^2/(8\max\{\lambda_{\max}'^2\lambda_{\max}^2, 1\})\big), \tag{103}$$

and, similarly,

$$p_-(\nu) \leq \exp\big(-N\nu^2/(8\max\{\lambda_{\max}'^2\lambda_{\max}^2, 1\})\big). \tag{104}$$

Inserting (103) and (104) into (96) finally yields

$$p(\nu) \leq 2\exp\big(-N\nu^2/(8\max\{\lambda'^2_{\max}\lambda^2_{\max},1\})\big).$$

□

ACKNOWLEDGMENT

The author is grateful to R. Heckel who performed a careful review of some early manuscripts, thereby pointing to some errors in the consistency analysis and the formulation of Lemma B.1. Moreover, some helpful comments from and discussions with H. Bölcskei and F. Hlawatsch, resulting in an improved presentation of the main ideas, are appreciated sincerely.

REFERENCES

[1] R. Dahlhaus, "Graphical interaction models for multivariate time series," *Metrika*, vol. 51, pp. 151–172, 2000.
[2] R. Dahlhaus and M. Eichler, "Causality and graphical models for time series," in *Highly Struct. Stochast. Syst.*, P. Green, N. Hjort, and S. Richardson, Eds. Oxford, U.K.: Oxford Univ. Press, 2003, pp. 115–137.
[3] A. Bolstad, B. D. van Veen, and R. Nowak, "Causal network inference via group sparse regularization," *IEEE Trans. Signal Process.*, vol. 59, no. 6, pp. 2628–2641, Jun. 2011.
[4] S. L. Lauritzen, *Graphical Models*. Oxford, U.K.: Clarendon, 1996.
[5] F. R. Bach and M. I. Jordan, "Learning graphical models for stationary time series," *IEEE Trans. Signal Process.*, vol. 52, no. 8, pp. 2189–2199, Aug. 2004.
[6] R. Brillinger, "Remarks concerning graphical models for time series and point processes," *Revista de Econometria*, vol. 16, pp. 1–23, 1996.
[7] N. E. Karoui, "Operator norm consistent estimation of large-dimensional sparse covariance matrices," *Ann. Statist.*, vol. 36, no. 6, pp. 2717–2756, 2008.
[8] N. P. Santhanam and M. J. Wainwright, "Information-theoretic limits of selecting binary graphical models in high dimensions," *IEEE Trans. Inf. Theory*, vol. 58, no. 7, pp. 4117–4134, Jul. 2012.
[9] P. Ravikumar, M. J. Wainwright, and J. Lafferty, "High-dimensional Ising model selection using ℓ_1-regularized logistic regression," *Ann. Statist.*, vol. 38, no. 3, pp. 1287–1319, 2010.
[10] J. Bento, M. Ibrahimi, and A. Montanari, "Learning networks of stochastic differential equations," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, Vancouver, Canada, 2010, vol. 23, pp. 172–180.
[11] N. Meinshausen and P. Bühlmann, "High-dimensional graphs and variable selection with the Lasso," *Ann. Statist.*, vol. 34, no. 3, pp. 1436–1462, 2006.
[12] J. H. Friedmann, T. Hastie, and R. Tibshirani, "Sparse inverse covariance estimation with the graphical LASSO," *Biostatist.*, vol. 9, no. 3, pp. 432–441, Jul. 2008.
[13] J. D. Lee and T. J. Hastie, "Learning mixed graphical models," May 2012, arXiv:1205.5012 [Online]. Available: http://arxiv.org/abs/1205.5012
[14] P. Ravikumar, M. J. Wainwright, B. Raskutti, and G. Yu, "High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence," *Electron. J. Statist.*, vol. 5, pp. 935–980, 2011.
[15] W. Wang, M. J. Wainwright, and K. Ramchandran, "Information-theoretic bounds on model selection for Gaussian Markov random fields," in *Proc. IEEE ISIT*, Austin, TX, USA, Jun. 2010, pp. 1373–1377.
[16] J. Songsiri and L. Vandenberghe, "Topology selection in graphical models of autoregressive processes," *J. Mach. Learn. Res.*, vol. 11, pp. 2671–2705, 2010.
[17] Y. C. Eldar, P. Kuppinger, and H. Bölcskei, "Block-sparse signals: Uncertainty relations and efficient recovery," *IEEE Trans. Signal Process.*, vol. 58, no. 6, pp. 3042–3054, Jun. 2010.
[18] M. Mishali and Y. C. Eldar, "Reduce and boost: Recovering arbitrary sets of jointly sparse vectors," *IEEE Trans. Signal Process.*, vol. 56, no. 10, pp. 4692–4702, Oct. 2008.
[19] Y. C. Eldar and H. Rauhut, "Average case analysis of multichannel sparse recovery using convex relaxation," *IEEE Trans. Inf. Theory*, vol. 56, no. 1, pp. 505–519, Jan. 2009.
[20] P. Bühlmann and S. van de Geer, *Statistics for High-Dimensional Data*. New York, NY, USA: Springer, 2011.
[21] K. Lounici, M. Pontil, A. B. Tsybakov, and S. van de Geer, "Taking advantage of sparsity in multi-task learning," in *Proc. 22nd Annu. Conf. Learn. Theory (COLT)*, 2009, pp. 73–82.
[22] S. Lee, J. Zhu, and E. P. Xing, "Adaptive multi-task Lasso: With application to eQTL detection," in *Proc. Adv. Neural Inf. Process. Syst.*, Vancouver, Cananda, 2010, vol. 23, pp. 1306–1314.
[23] C. D. Meyer, *Matrix Analysis and Applied Linear Algebra*. Philadelphia, PA, USA: SIAM, 2000.
[24] P. J. Brockwell and R. A. Davis, *Time Series: Theory and Methods*. New York, NY, USA: Springer, 1991.
[25] M. Eichler, "Graphical models in time series analysis," Ph.D. dissertation, Universität Heidelberg, Heidelberg, Germany, 1999.
[26] H. Lütkepohl, *New Introduction to Multiple Time Series Analysis*. New York, NY, USA: Springer, 2005.
[27] G. Matz and F. Hlawatsch, "Nonstationary spectral analysis based on time-frequency operator symbols and underspread approximations," *IEEE Trans. Inf. Theory*, vol. 52, no. 3, pp. 1067–1086, Mar. 2006.
[28] P. Stoica and R. Moses, *Introduction to Spectral Analysis*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1997.
[29] D. S. Bernstein, *Matrix Mathematics: Theory, Facts, and Formulas*, 2nd ed. Princeton, NJ, USA: Princeton Univ. Press, 2009.
[30] A. Argyriou, T. Evgeniou, and M. Pontil, "Convex multi-task feature learning," *Mach. Learn.*, vol. 73, no. 3, pp. 243–272, 2008.
[31] G. Obozinski, M. J. Wainwright, and M. I. Jordan, "Support union recovery in high-dimensional multivariate regression," *Ann. Statist.*, vol. 39, no. 1, pp. 1–47, 2011.
[32] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *J. Roy. Statist. Soc. Ser. B*, vol. 68, no. 1, pp. 49–67, 2006.
[33] A. Bossavit, *A Course in Convex Analysis*. Warsaw, Poland: Interdisciplinary Centre for Mathematical and Computational Modelling (ICM), 2003.
[34] F. R. Bach, "Consistency of the group LASSO and multiple kernel learning," *J. Mach. Learn. Res.*, vol. 9, pp. 1179–1225, 2008.
[35] Z. T. Qin and D. Goldfarb, "Structured sparsity via alternating direction methods," *J. Mach. Learn. Res.*, no. 13, pp. 1435–1468, May 2012.
[36] G. Hannak, A. Jung, and N. Görtz, "On the information-theoretic limits of graphical model selection for Gaussian time series," presented at the EUSIPCO, Lisbon, Portugal, 2014.
[37] T. T. Cai, Z. Ren, and H. H. Zhou, "Optimal rates of convergence for estimating Toeplitz covariance matrices," *Probab. Theory Relat. Fields*, vol. 156, no. 1, pp. 101–143, Jun. 2013.
[38] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, *Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers*. Hanover, MA, USA: Now Publishers, 2010, vol. 3, Foundations and Trends in Machine Learning.
[39] T. Wang, S.-U. Guan, K. L. Man, and T. Ting, "Time series classification for EEG eye state identification based on incremental attribute learning," in *Proc. Int. Symp. Comput., Consum., Control (IS3C)*, Jun. 2014, pp. 158–161.
[40] K. Bache and M. Lichman, UCI Machine Learning Repository, 2013 [Online]. Available: http://archive.ics.uci.edu/ml
[41] A. V. Oppenheim, R. W. Schafer, and J. R. Buck, *Discrete-Time Signal Processing*, 2nd ed. Englewood Cliffs, NJ, USA: Prentice-Hall, 1998.
[42] R. Gray, "Toeplitz and circulant matrices: A review," *Found. Trends Commun. Inf. Theory*, vol. 2, no. 3, pp. 155–239, 2006.
[43] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 1985.

Alexander Jung received the Diplom-Ingenieur and Dr. techn. degrees in electrical engineering/signal processing from Vienna University of Technology, Vienna, Austria, in 2008, and 2011, respectively. He is currently assistant professor for machine learning within the Department of Computer Science at Aalto University. His research interests are in statistical signal processing and machine learning for big data with emphasis on sparse models as well as trade-offs between accuracy and computational/ommunication complexity of learning algorithms. The quality of his research is documented by several national and international awards including a Best Student Paper award at the worlds largest signal processing conference, the IEEE ICASSP 2011.