Learning Sparse High-Dimensional Matrix-Valued Graphical Models From Dependent Data

Jitendra K. Tugnait

***Abstract*—We consider the problem of inferring the conditional independence graph (CIG) of a sparse, high-dimensional, stationary matrix-variate Gaussian time series. All past work on high-dimensional matrix graphical models assumes that independent and identically distributed (i.i.d.) observations of the matrix-variate are available. Here we allow dependent observations. We consider a sparse-group lasso-based frequency-domain formulation of the problem with a Kronecker-decomposable power spectral density (PSD), and solve it via an alternating direction method of multipliers (ADMM) approach. The problem is bi-convex which is solved via flip-flop optimization. We provide sufficient conditions for local convergence in the Frobenius norm of the inverse PSD estimators to the true value. This result also yields a rate of convergence. We illustrate our approach using numerical examples utilizing both synthetic and real data.**

***Index Terms*—Sparse graph learning; matrix graph estimation; matrix time series; undirected graph; inverse spectral density estimation.**

I. Introduction

In graphical models, graphs display the conditional independence structure of the variables, and learning the graph structure is equivalent to learning a factorization of the joint probability distribution of these random variables [1]. In a vector graphical model, the conditional statistical dependency structure among p random variables $x_1, x_2, \cdots, x_p$, is represented using an undirected graph $\mathcal{G} = (V, \mathcal{E})$ with a set of p vertices (nodes) $V = \{1, 2, \cdots, p\} = [p]$, and a corresponding set of (undirected) edges $\mathcal{E} \subseteq [p] \times [p]$. There is no edge between nodes i and j iff x_i and x_j are conditionally independent given the remaining p-2 variables. Suppose $\boldsymbol{x} \sim \mathcal{N}_r(\mathbf{m}, \mathbf{\Sigma})$, with $\mathbf{m} \in \mathbb{R}^p$, $\mathbf{\Sigma} \in \mathbb{R}^{p\times p}$, positive definite $\mathbf{\Sigma} = \mathbf{\Omega}^{-1}$, where $\mathcal{N}_r(\mathbf{m}, \mathbf{\Sigma})$ denotes a real-valued Gaussian vector with mean $\boldsymbol{m}$ and covariance $\mathbf{\Sigma}$. Then Ω_{ij}, the (i, j)-th element of $\mathbf{\Omega}$, is zero iff x_i and x_j are conditionally independent [1]. Of much interest is the high-dimensional case where p is greater than or of the order of the data sample size n [2]. In particular, in a high-dimensional setting, as $n \uparrow \infty$, $p/n \to c > 0$, instead of $p/n \to 0$ as in classical low-dimensional statistical analysis framework [2, Chapter 1]. Such models for $\boldsymbol{x}$ have been extensively studied [2]–[5]. In this paper we address the problem of high-dimensional matrix graph estimation. If $p/n \ll 1$, we use the term low-dimensional for such cases in this paper.

Consider a stationary $p-$dimensional multivariate Gaussian time series $\boldsymbol{x}(t), t = 0, \pm1, \pm2, \cdots$, with ith component $x_i(t)$. In the corresponding time series graph $\mathcal{G} = (V, \mathcal{E})$, there is no edge between nodes i and j iff $\{x_i(t)\}$ and $\{x_j(t)\}$ are conditionally independent given the remaining p-2 scalar series $\{x_\ell(t), \ell \in [p], \ \ell \neq i, \ell \neq j\}$ [6]. Denote the power spectral density (PSD) matrix of zero-mean $\{\boldsymbol{x}(t)\}$ by $\boldsymbol{S}_x(f)$, where $\boldsymbol{S}_x(f) = \sum_{\tau=-\infty}^{\infty} \boldsymbol{R}_{xx}(\tau) e^{-\iota 2\pi f\tau}$, $\boldsymbol{R}_{xx}(\tau) = E\{\boldsymbol{x}(t+\tau)\boldsymbol{x}^\top(t)\}$ and $\iota = \sqrt{-1}$. In [6] it was shown that conditional independence of two time series components given all other components of the zero-mean time series, is encoded by zeros in the inverse PSD, that is, $\{i, j\} \notin \mathcal{E}$ iff the (i, j)-th element of $\boldsymbol{S}_x^{-1}(f)$, $[\boldsymbol{S}_x^{-1}(f)]_{ij} = 0$ for every f. In [6] the low-dimensional case is addressed whereas nonparametric frequency-domain approaches for graph estimation in high-dimensional settings have been considered in [7]–[9]. Refs. [7], [9] provide performance analysis and guarantees. Parametric modeling based approaches in low-dimensional settings for conditional independence graph (CIG) estimation for time series are discussed in [10]–[15]. These papers are focused on algorithm development and they do not provide performance guarantees (such as [9, Theorem 1]). Estimation of sparse high-dimensional parametric time series models is discussed in [16] where performance analysis in high-dimensions is carried out, but the graphical modeling aspect is not addressed.

The need for matrix-valued graphical models arises in several applications [17]–[27] (see also related work of [28]). Here we observe matrix-valued time series $\{\boldsymbol{Z}(t)\}$ where $\boldsymbol{Z}(t) \in \mathbb{R}^{p\times q}$. If one vectorizes using vec($\boldsymbol{Z}$) where vec($\boldsymbol{Z}$) $\in \mathbb{R}^{pq}$ denotes column-wise vectorization of $\boldsymbol{Z}$, then use of vec($\boldsymbol{Z}$) will result in a pq-node graph with $(pq) \times (pq)$ precision matrix, which could be ultra-high-dimensional, and it ignores any structural information among rows and columns of $\boldsymbol{Z}(t)$ [17]. With $\otimes$ denoting the matrix Kronecker product, the basic idea in matrix-valued graphs is to model the covariance of vec($\boldsymbol{Z}$) as $\mathbf{\Psi} \otimes \mathbf{\Sigma}$ with $\mathbf{\Psi} \in \mathbb{R}^{q\times q}$ and $\mathbf{\Sigma} \in \mathbb{R}^{p\times p}$, reducing the number of unknowns from $\mathcal{O}(p^2q^2)$ to $\mathcal{O}(p^2 + q^2)$, while also preserving the structural information. Given data, one estimates two precision matrices $\mathbf{\Omega} = \mathbf{\Sigma}^{-1}$ and $\mathbf{\Upsilon} = \mathbf{\Psi}^{-1}$. In the matrix graph, conditional independence between $\boldsymbol{Z}_{ij}$ and $\boldsymbol{Z}_{k\ell}$ is determined by zeros in $\mathbf{\Omega}$ and $\mathbf{\Upsilon}$ [17]. This is the Kronecker graph model [29], [30]: If $\mathcal{G}_1$ and $\mathcal{G}_2$ are graphs with adjacency matrices $\mathcal{A}(\mathcal{G}_1)$ and $\mathcal{A}(\mathcal{G}_2)$, respectively, then the Kronecker product graph (KPG) $\mathcal{G}_1 \otimes \mathcal{G}_2$ is defined as the graph with adjacency matrix $\mathcal{A}(\mathcal{G}_1) \otimes \mathcal{A}(\mathcal{G}_2)$ [30, Def. 1]. In our context the nonzero entries of $\mathbf{\Upsilon}$ and $\mathbf{\Omega}$ determine the nonzero entries of the adjacency matrices of graphs $\mathcal{G}_1$ and $\mathcal{G}_2$, respectively, with KPG $\mathcal{G} = \mathcal{G}_1 \otimes \mathcal{G}_2$.

Our objective in this paper is to learn a conditional independence KPG associated with time-dependent matrix-valued zero-mean $p \times q$ Gaussian sequence $\boldsymbol{Z}(t)$, under high-dimensional settings, given observations of $\{\boldsymbol{Z}(t)\}_{t=0}^{n-1}$.

A. Related Work

Prior work on KPG estimation under high-dimensional settings [17]–[27] all assume that i.i.d. observations of $\boldsymbol{Z}$

J.K. Tugnait is with the Dept. of Elec. & Comp. Eng., 200 Broun Hall, Auburn University, Auburn, AL 36849, USA. Email: tugnajk@auburn.edu .

Supported by NSF Grants ECCS-2040536 and CCF-2308473.

are available for graphical modeling. Refs. [17], [18], [25] all solve the same bi-convex optimization problem, using an identical alternating minimization approach, but they differ in theoretical analysis. Ref. [21] uses a Kronecker sum model whereas we use a Kronecker-product separable covariance structure (see (4) later) for $\{\boldsymbol{Z}(t)\}$.

There is no prior reported work on high-dimensional matrix graph estimation with dependent data using a nonparametric approach. Parametric models using state-space models are estimated in [31], [32] for KPG estimation in low-dimensional settings using Bayesian approaches. Granger causality graphs (not the same as CIGs) for matrix time series are estimated in [33] using first-order AR models and in [34] using an infinite dimensional model class (which includes ARMAX models of any order), both in low-dimensional settings (i.e., $pq/n \ll 1$ and/or $\lim_{n\to\infty} pq/n = 0$, with pq representing number of nodes in KPG). In contrast, this paper considers conditional independence KPG's under high-dimensional settings. Ref. [35] investigates sum of Kronecker product AR models for matrix times series with no consideration of CIGs. Estimation of a KPG model corresponding to an AR Gaussian process is investigated in [36] in low-dimensional settings with no performance analysis or guarantees. A distinguishing aspect of [36] is that it imposes a Kronecker product decomposition on the support of the inverse PSD, not the inverse PSD of the time series. With regard to [34], [36], we note that in the synthetic data example using an ARMA model in [34, Sec. 6.1, Fig. 2], number of nodes is 16 ($= pq$) and sample size is $n = 3900$, leading to $pq/n = 0.004$, a low-dimensional setting. The real data example of [34, Sec. 6.2] does have pq =96 and $n = 500$, implying $pq/n = 0.19$. The distinction is that the ground truth is known in synthetic data examples permitting evaluation of the efficacy of the considered approach, whereas such is not the case in real data examples. Thus [34] does not address the high-dimensional scenario as relatively high $pq/n = 0.19$ in their real data example is not supported by any commensurate synthetic data example. In contrast, we provide such support, as seen in Table I, Sec. VI-A of this paper, where pq =225 and varying $n \in \{64, 128, 256, 512, 1024, 2148\}$, implying $pq/n = \{3.5, 1.76, 0.88, 0.44, 0.22, 0.11\}$. In our real data example (Sec. VI-B), we have $pq = 88$ and $n = 364$ with $pq/n = 0.24$. In [36, Sec. 6.1], the synthetic data example has $pq = 36$ and n =1000 or 2000 (pq/n=0.036 or 0.018), again a low-dimensional scenario. The real data example of [36, Sec. 6.2] has $pq = 36$ and $n = 389$ ($pq/n = 0.09$). The comments made pertaining to [34] regarding differences in pq/n ratios for real and synthetic data examples, apply to [36] as well. Finally, [37] considers a first-order matrix AR model for matrix time series where when vectorized, the vectorized time-series AR coefficient is expressed as a Kronecker product. Low-dimensional asymptotics are provided in [37] and the issue of the underlying CIG is not addressed.

A frequency-domain formulation is used in this paper, following the approach of [9] for dependent vector time series. The resulting optimization problem is bi-convex, as in [17], [18], [25], but with complex variables, and is solved via an alternating minimization approach using Wirtinger calculus [38] for optimization of real functions of complex variables.

A preliminary version of parts of this paper appear in a workshop paper [39]. Theorems 1-3 and their proofs, and the real data example do not appear in [39].

B. Our Contributions, Outline and Notation

The underlying system model including a generative model (5) for time-dependent matrix Gaussian sequence, is presented in Sec. II. A frequency-domain based penalized log-likelihood objective function is derived in Sec. III for estimation of the matrix graph, resulting in a Kronecker-decomposable power spectral density representation (15). A flip-flop algorithm based on two ADMM algorithms is presented in Sec. IV to optimize the bi-convex objective function. In Sec. V the performance of the proposed optimization algorithm is analyzed under a high-dimensional large sample setting in Theorems 1-3, patterned after [22] and exploiting some results from [9], [40], [41]. Numerical results are presented in Sec. VI and proofs of Theorems 1, 2 and 3 are given in three appendices.

Notation. The superscripts $*$, $\top$ and H denote the complex conjugate, transpose and conjugate transpose operations, respectively, $\mathbb{R}$ and $\mathbb{C}$ denote the sets of real and complex numbers, respectively, and $\mathrm{Re}(\boldsymbol{x})$ is the real part of $\boldsymbol{x} \in \mathbb{C}^p$. We use $\iota := \sqrt{-1}$. A $p \times p$ identity matrix is denoted by $\boldsymbol{I}_p$. Given $\boldsymbol{A} \in \mathbb{C}^{p\times p}$, $\phi_{\min}(\boldsymbol{A})$, $\phi_{\max}(\boldsymbol{A})$, $|\boldsymbol{A}|$, $\mathrm{tr}(\boldsymbol{A})$ and $\mathrm{etr}(\boldsymbol{A})$ denote the minimum eigenvalue, maximum eigenvalue, determinant, trace, and exponential of trace of $\boldsymbol{A}$, respectively. We use $\boldsymbol{A} \succeq 0$ and $\boldsymbol{A} \succ 0$ to denote that Hermitian $\boldsymbol{A}$ is positive semi-definite and positive definite, respectively. For $\boldsymbol{B} \in \mathbb{C}^{p\times q}$, we define the operator norm, the Frobenius norm and the vectorized ℓ_1 norm, respectively, as $\|\boldsymbol{B}\| = \sqrt{\phi_{\max}(\boldsymbol{B}^H\boldsymbol{B})}$, $\|\boldsymbol{B}\|_F = \sqrt{\mathrm{tr}(\boldsymbol{B}^H\boldsymbol{B})}$ and $\|\boldsymbol{B}\|_1 = \sum_{i,j}|B_{ij}|$, where B_{ij} is the (i,j)-th element of $\boldsymbol{B}$, also denoted by $[\boldsymbol{B}]_{ij}$. For vector $\boldsymbol{\theta} \in \mathbb{C}^p$, we define $\|\boldsymbol{\theta}\|_1 = \sum_{i=1}^p |\theta_i|$ and $\|\boldsymbol{\theta}\|_2 = \sqrt{\sum_{i=1}^p |\theta_i|^2}$, and we also use $\|\boldsymbol{\theta}\|$ for $\|\boldsymbol{\theta}\|_2$. Given $\boldsymbol{A} \in \mathbb{C}^{p\times p}$, $\boldsymbol{A}^+ = \mathrm{diag}(\boldsymbol{A})$ is a diagonal matrix with the same diagonal as $\boldsymbol{A}$, and $\boldsymbol{A}^- = \boldsymbol{A} - \boldsymbol{A}^+$ is $\boldsymbol{A}$ with all its diagonal elements set to zero. We use $\boldsymbol{A}^{-*}$ for $(\boldsymbol{A}^*)^{-1}$, the inverse of complex conjugate of $\boldsymbol{A}$, and $\boldsymbol{A}^{-\top}$ for $(\boldsymbol{A}^\top)^{-1}$. Given $\boldsymbol{A} \in \mathbb{C}^{n\times p}$, column vector $\mathrm{vec}(\boldsymbol{A}) \in \mathbb{C}^{np}$ denotes the vectorization of $\boldsymbol{A}$ which stacks the columns of the matrix $\boldsymbol{A}$. The notation $\boldsymbol{y}_n = \mathcal{O}_P(\boldsymbol{x}_n)$ for random $\boldsymbol{y}_n, \boldsymbol{x}_n \in \mathbb{C}^p$ means that for any $\varepsilon > 0$, there exists $0 < T < \infty$ such that $P(\|\boldsymbol{y}_n\| \le T\|\boldsymbol{x}_n\|) \ge 1-\varepsilon\ \forall n \ge 1$. The notation $\boldsymbol{x} \sim \mathcal{N}_c(\mathbf{m}, \boldsymbol{\Sigma})$ denotes a complex random vector $\boldsymbol{x}$ that is circularly symmetric (proper), complex Gaussian with mean $\boldsymbol{m}$ and covariance $\boldsymbol{\Sigma}$, and $\boldsymbol{x} \sim \mathcal{N}_r(\mathbf{m}, \boldsymbol{\Sigma})$ denotes real-valued Gaussian $\boldsymbol{x}$ with mean $\boldsymbol{m}$ and covariance $\boldsymbol{\Sigma}$.

II. System Model

A random matrix $\boldsymbol{Z} \in \mathbb{R}^{p\times q}$ is said to have a matrix normal (Gaussian) distribution if its pdf $f(\boldsymbol{Z}|\boldsymbol{M}, \boldsymbol{\Sigma}, \boldsymbol{\Psi})$, characterized by $\boldsymbol{M} \in \mathbb{R}^{p\times q}$, $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$, $\boldsymbol{\Psi} \in \mathbb{R}^{q\times q}$, is [42, Chap. 2]

$$f(\boldsymbol{Z}|\boldsymbol{M}, \boldsymbol{\Sigma}, \boldsymbol{\Psi}) = \frac{\mathrm{etr}\Big(-\frac{1}{2}(\boldsymbol{Z}-\boldsymbol{M})\boldsymbol{\Psi}^{-1}(\boldsymbol{Z}-\boldsymbol{M})^\top\boldsymbol{\Sigma}^{-1}\Big)}{(2\pi)^{pq/2}\,|\boldsymbol{\Sigma}|^{q/2}\,|\boldsymbol{\Psi}|^{p/2}}. \tag{1}$$

We will use the notation $\boldsymbol{Z} \sim \mathcal{MVN}(\boldsymbol{M}, \boldsymbol{\Sigma}, \boldsymbol{\Psi})$ for the matrix normal distribution specified by (1). Equivalently,

$$\mathrm{vec}(\boldsymbol{Z}) \sim \mathcal{N}_r\big(\mathrm{vec}(\boldsymbol{M}), \boldsymbol{\Psi} \otimes \boldsymbol{\Sigma}\big). \tag{2}$$

Here $\boldsymbol{\Psi}$ is the row covariance matrix and $\boldsymbol{\Sigma}$ is the column covariance matrix [42] since the kth column $\boldsymbol{Z}_{\cdot k} \sim \mathcal{N}_r(\boldsymbol{0}, [\boldsymbol{\Psi}]_{kk}\boldsymbol{\Sigma})$ and the ith row $\boldsymbol{Z}_{i\cdot}^\top \sim \mathcal{N}_r(\boldsymbol{0}, [\boldsymbol{\Sigma}]_{ii}\boldsymbol{\Psi})$.

With $\boldsymbol{Z} \in \mathbb{R}^{p\times q}$ modeled as a zero-mean matrix normal vector and $\boldsymbol{z} = \text{vec}(\boldsymbol{Z})$, [17] assumes

$$E\{\boldsymbol{z}\boldsymbol{z}^\top\} = \boldsymbol{\Psi} \otimes \boldsymbol{\Sigma}\,, \tag{3}$$

implying a separable covariance structure [28]. Let $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ and $\boldsymbol{\Upsilon} = \boldsymbol{\Psi}^{-1}$ denote the respective precision matrices. Then $\boldsymbol{Z}_{ij}$ and $\boldsymbol{Z}_{k\ell}$ are conditionally independent given remaining entries in $\boldsymbol{Z}$ iff (i) at least one of $\boldsymbol{\Omega}_{ik}$ and $\boldsymbol{\Upsilon}_{j\ell}$ is zero when $i \neq k$, $j \neq \ell$, (ii) $\boldsymbol{\Omega}_{ik} = 0$ when $i \neq k$, $j = \ell$, and (iii) $\boldsymbol{\Upsilon}_{j\ell} = 0$ when $i = k$, $j \neq \ell$ [17].

In this paper we will model our time-dependent zero-mean matrix-valued, stationary, $p\times q$ Gaussian sequence $\boldsymbol{Z}(t)$, $\boldsymbol{z}(t) = \text{vec}(\boldsymbol{Z}(t))$, as having the separable covariance structure given by

$$E\{\boldsymbol{z}(t+\tau)\boldsymbol{z}^\top(t)\} = \boldsymbol{\Psi}(\tau) \otimes \boldsymbol{\Sigma} \tag{4}$$

where $\boldsymbol{\Psi}(\tau)$, $\tau = 0, \pm 1, \cdots$ models time-dependence while $\boldsymbol{\Sigma} \succ \boldsymbol{0}$ is fixed. Under (4), the row covariance sequence is $E\{\boldsymbol{Z}_{i\cdot}^\top(t+\tau)\boldsymbol{Z}_{i\cdot}(t)\} = [\boldsymbol{\Sigma}]_{ii}\,\boldsymbol{\Psi}(\tau)$ and the column covariance sequence is $E\{\boldsymbol{Z}_{\cdot k}(t+\tau)\boldsymbol{Z}_{\cdot k}^\top(t)\} = \boldsymbol{\Sigma}\,[\boldsymbol{\Psi}(\tau)]_{kk}$. Thus we allow possible temporal dependence in matrix observations via $\boldsymbol{\Psi}(\tau)$. With $\{\boldsymbol{e}(t)\}$ i.i.d., $\boldsymbol{e}(t) \sim \mathcal{N}_r(\boldsymbol{0}, \boldsymbol{I}_{pq})$, a generative model for $\boldsymbol{z}(t)$ is given by

$$\boldsymbol{z}(t) = \sum_{i=0}^{L} (\boldsymbol{B}_i \otimes \boldsymbol{F})\boldsymbol{e}(t-i)\,,\ \boldsymbol{B}_i \in \mathbb{R}^{q\times q}\,,\ \boldsymbol{F} \in \mathbb{R}^{p\times p} \tag{5}$$

$$\Rightarrow\ E\{\boldsymbol{z}(t+\tau)\boldsymbol{z}^\top(t)\} = \underbrace{\big(\sum_{i=0}^{L} \boldsymbol{B}_i\boldsymbol{B}_{i-\tau}^\top\big)}_{=\boldsymbol{\Psi}(\tau)} \otimes \underbrace{(\boldsymbol{F}\boldsymbol{F}^\top)}_{=\boldsymbol{\Sigma}}\,. \tag{6}$$

In (5), we can have $L \uparrow \infty$ so long as assumption (A2) stated in Sec. III holds. In sequel, we exploit (4) in our approach without considering (6), the latter is used only for synthetic data generation.

The PSD of $\{\boldsymbol{z}(t)\}$ is $\boldsymbol{S}_z(f) = \bar{\boldsymbol{S}}(f) \otimes \boldsymbol{\Sigma}$ where $\bar{\boldsymbol{S}}(f) = \sum_\tau \boldsymbol{\Psi}(\tau)e^{-\iota 2\pi f\tau}$. Then $\boldsymbol{S}_z^{-1}(f) = \bar{\boldsymbol{S}}^{-1}(f) \otimes \boldsymbol{\Sigma}^{-1}$, and by [6], in the $pq-$node graph $\mathcal{G} = (V, \mathcal{E})$, $|V| = pq$, associated with $\{\boldsymbol{z}(t)\}$, edge $\{i,j\} \notin \mathcal{E}$ iff $[\boldsymbol{S}_z^{-1}(f)]_{ij} = 0$ for every f. This does not account for the separable structure of our model. Noting that $\bar{\boldsymbol{S}}^{-1}(f)$, $f \in [0, 0.5]$, plays the role of $\boldsymbol{\Upsilon} = \boldsymbol{\Psi}^{-1}$, using [6], [17] (also [30, Observation 1]), we deduce that $\{\boldsymbol{Z}_{ij}(t)\}$ and $\{\boldsymbol{Z}_{k\ell}(t)\}$ are conditionally independent given remaining entries in $\{\boldsymbol{Z}(t)\}$ iff (i) at least one of $\boldsymbol{\Omega}_{ik}$ and $[\bar{\boldsymbol{S}}^{-1}(f)]_{j\ell}$, for every $f \in [0, 0.5]$, is zero when $i \neq k$, $j \neq \ell$, (ii) $\boldsymbol{\Omega}_{ik} = 0$ when $i \neq k$, $j = \ell$, and (iii) $[\bar{\boldsymbol{S}}^{-1}(f)]_{j\ell} = 0$, for every $f \in [0, 0.5]$ when $i = k$, $j \neq \ell$. That is, we have a KPG $\mathcal{G} = \mathcal{G}_1 \otimes \mathcal{G}_2$ where the adjacency matrix of $\mathcal{G}_1$ is specified by the nonzero entries of $\bar{\boldsymbol{S}}^{-1}(f)$ $\forall f \in [0, 0.5]$, and that of $\mathcal{G}_2$ follows from the nonzero entries of $\boldsymbol{\Omega}$.

Our objective is to learn the graph associated with $\{\boldsymbol{Z}(t)\}$ under some sparsity constraints on $\boldsymbol{\Omega}$ and $\bar{\boldsymbol{S}}^{-1}(f)$, $f \in [0, 0.5]$. Since $\alpha\bar{\boldsymbol{S}}^{-1}(f) \otimes (\alpha^{-1}\boldsymbol{\Omega}) = \bar{\boldsymbol{S}}^{-1}(f) \otimes \boldsymbol{\Omega}$, to resolve scaling ambiguity, we could normalize $\|\boldsymbol{\Omega}\|_F = 1$ or $\|\bar{\boldsymbol{S}}^{-1}(f_1) \ \cdots \ \bar{\boldsymbol{S}}^{-1}(f_M)]\|_F = 1$ for suitably placed M frequencies in $(0, 0.5)$; we will follow the latter as stated later in step 2 of Sec. IV-A.

III. Penalized Negative Log-Likelihood

Given $\boldsymbol{z}(t)$ for $t = 0, 1, 2, \cdots, n-1$. Define the (normalized) DFT's $\boldsymbol{d}_z(f_m)$ and $\boldsymbol{D}_z(f_m)$ of $\boldsymbol{z}(t)$ and $\boldsymbol{Z}(t)$, respectively, as (recall $\iota = \sqrt{-1}$),

$$\boldsymbol{d}_z(f_m) = \frac{1}{\sqrt{n}} \sum_{t=0}^{n-1} \boldsymbol{z}(t) \exp\left(-\iota 2\pi f_m t\right)\,, \tag{7}$$

$$\boldsymbol{D}_z(f_m) = \frac{1}{\sqrt{n}} \sum_{t=0}^{n-1} \boldsymbol{Z}(t) \exp\left(-\iota 2\pi f_m t\right)\,, \tag{8}$$

$$f_m = m/n,\ m = 0, 1, \cdots, n-1. \tag{9}$$

Then $\boldsymbol{d}_z(f_m) = \text{vec}(\boldsymbol{D}_z(f_m))$. It is established in [43] (see also [9]) that, for even n, the set of random vectors $\{\boldsymbol{d}_z(f_m)\}_{m=0}^{n/2}$ is a sufficient statistic for any inference problem based on dataset $\{\boldsymbol{z}(t)\}_{t=0}^{n-1}$. Suppose $\boldsymbol{S}_z(f_k)$ is locally smooth, so that $\boldsymbol{S}_z(f_k)$ is (approximately) constant over $K = 2m_t + 1$ consecutive frequency points f_m's where m_t is the half-window size; in our case, this assumption applies to $\bar{\boldsymbol{S}}(f_k)$. Pick $M = \big\lfloor (\frac{n}{2} - m_t - 1)/K \big\rfloor$ and

$$\tilde{f}_k = \big((k-1)K + m_t + 1\big)/n, \qquad k \in [M], \tag{10}$$

yielding M equally spaced frequencies $\tilde{f}_k$ in the interval $(0, 0.5)$. We state the local smoothness assumption as assumption (A1).

(A1) Assume that for $\ell = -m_t, -m_t+1, \cdots, m_t$,

$$\boldsymbol{S}_z(\tilde{f}_{k,\ell}) = \boldsymbol{S}_z(\tilde{f}_k)\,, \tag{11}$$

$$\text{where}\ \ \tilde{f}_{k,\ell} = \big((k-1)K + m_t + 1 + \ell\big)/n\,. \tag{12}$$

We will invoke [44, Theorem 4.4.1] for distribution of $\boldsymbol{d}_z(f_m)$. To this end we need assumption (A2).

(A2) The matrix time series $\{\boldsymbol{Z}(t)\}_{t=-\infty}^{\infty}$ is zero-mean stationary, Gaussian, satisfying $\sum_{\tau=-\infty}^{\infty} |[\boldsymbol{\Psi}(\tau)]_{k\ell}| < \infty$ for every $k, \ell \in [q]$.

By [44, Theorem 4.4.1], under assumption (A1), asymptotically (as $n \to \infty$), $\boldsymbol{d}_z(f_m)$, $m = 1, 2, \cdots, (n/2)-1$, ($n$ even), are independent proper, complex Gaussian $\mathcal{N}_c(\boldsymbol{0}, \boldsymbol{S}_z(f_m))$ random vectors, respectively. Denote the joint probability density function of $\boldsymbol{d}_z(f_m)$, $m = 1, 2, \cdots, (n/2) - 1$, as $f_{\boldsymbol{\mathcal{D}}}(\boldsymbol{\mathcal{D}})$ where $\boldsymbol{\mathcal{D}} = \{\boldsymbol{D}_z(f_m)\}_{m=1}^{(n/2)-1}$. Then we have [9], [43]

$$f_{\boldsymbol{\mathcal{D}}}(\boldsymbol{\mathcal{D}}) = \prod_{k=1}^{M} \left[\prod_{\ell=-m_t}^{m_t} \frac{\exp\left(-g_{kl} - g_{kl}^*\right)}{\pi^{pq}\,|\boldsymbol{B}_k|^{1/2}|\boldsymbol{B}_k^*|^{1/2}} \right]\,, \tag{13}$$

$$g_{kl} = \frac{1}{2}\boldsymbol{d}_z^H(\tilde{f}_{k,\ell})\big(\bar{\boldsymbol{S}}^{-1}(\tilde{f}_k) \otimes \boldsymbol{\Sigma}^{-1}\big)\boldsymbol{d}_z(\tilde{f}_{k,\ell})\,, \tag{14}$$

$$\boldsymbol{B}_k = \bar{\boldsymbol{S}}(\tilde{f}_k) \otimes \boldsymbol{\Sigma}\,. \tag{15}$$

Using $\text{tr}\big(\boldsymbol{A}^\top\boldsymbol{B}\boldsymbol{C}\boldsymbol{G}^\top\big) = (\text{vec}(\boldsymbol{A}))^\top(\boldsymbol{G} \otimes \boldsymbol{B})\text{vec}(\boldsymbol{C})$ and parametrizing in terms of $\boldsymbol{\Phi}_k := \bar{\boldsymbol{S}}^{-1}(\tilde{f}_k)$ and $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$, we have

$$\begin{aligned} g_{kl} &= \frac{1}{2}\boldsymbol{D}_z^H(\tilde{f}_{k,\ell})\boldsymbol{\Sigma}^{-1}\boldsymbol{D}_z(\tilde{f}_{k,\ell})(\bar{\boldsymbol{S}}^{-1}(\tilde{f}_k))^\top \\ &= \frac{1}{2}\boldsymbol{D}_z^H(\tilde{f}_{k,\ell})\boldsymbol{\Omega}\boldsymbol{D}_z(\tilde{f}_{k,\ell})\boldsymbol{\Phi}_k^*\,. \end{aligned} \tag{16}$$

Define $q \times (qM)$ matrix $\mathbf{\Gamma}$ as

$$\mathbf{\Gamma} = [\mathbf{\Phi}_1 \ \mathbf{\Phi}_2 \ \cdots \ \mathbf{\Phi}_M] \,. \tag{17}$$

Using $|\boldsymbol{B}_k| = |\bar{\boldsymbol{S}}(\tilde{f}_k) \otimes \mathbf{\Sigma}| = |\bar{\boldsymbol{S}}(\tilde{f}_k)|^p \, |\mathbf{\Sigma}|^q$, up to some constants the negative log-likelihood follows from (13) as

$$\begin{aligned} -\frac{1}{KMpq} \ln f_{\boldsymbol{\mathcal{D}}}(\boldsymbol{\mathcal{D}}) &\propto G(\mathbf{\Omega}, \mathbf{\Gamma}, \mathbf{\Gamma}^*) \\ &:= -\frac{1}{p} \ln(|\mathbf{\Omega}|) - \frac{1}{2Mq} \sum_{k=1}^{M} \big(\ln(|\mathbf{\Phi}_k|) + \ln(|\mathbf{\Phi}_k^*|) \big) \\ &\quad + \frac{1}{2Mq} \sum_{k=1}^{M} \mathrm{tr}(\boldsymbol{A}_k + \boldsymbol{A}_k^*) \,, \end{aligned} \tag{18}$$

$$\boldsymbol{A}_k = \frac{1}{Kp} \sum_{\ell=-m_t}^{m_t} \boldsymbol{D}_z^H(\tilde{f}_{k,\ell}) \mathbf{\Omega} \boldsymbol{D}_z(\tilde{f}_{k,\ell}) \mathbf{\Phi}_k^* \,. \tag{19}$$

In the high-dimension case, to enforce sparsity and to make the problem well-conditioned, we propose to minimize a penalized version $\mathcal{L}(\mathbf{\Omega}, \mathbf{\Gamma})$ w.r.t. $\mathbf{\Omega}$ and $\mathbf{\Gamma}$,

$$\mathcal{L}(\mathbf{\Omega}, \mathbf{\Gamma}) = G(\mathbf{\Omega}, \mathbf{\Gamma}, \mathbf{\Gamma}^*) + P_p(\mathbf{\Omega}) + P_q(\{\mathbf{\Phi}\}), \tag{20}$$

$$P_p(\mathbf{\Omega}) = \lambda_p \sum_{i \neq j}^{p} |\mathbf{\Omega}_{ij}| = \lambda_p \|\mathbf{\Omega}^-\|_1 \,, \tag{21}$$

$$\begin{aligned} P_q(\{\mathbf{\Phi}\}) &= \alpha \lambda_q \sum_{k=1}^{M} \sum_{i \neq j}^{q} \big| [\mathbf{\Phi}_k]_{ij} \big| \\ &\quad + (1-\alpha) \sqrt{M} \lambda_q \sum_{i \neq j}^{q} \|\mathbf{\Phi}^{(ij)}\| \,, \end{aligned} \tag{22}$$

$$\mathbf{\Phi}^{(ij)} := [[\mathbf{\Phi}_1]_{ij} \ [\mathbf{\Phi}_2]_{ij} \ \cdots \ [\mathbf{\Phi}_M]_{ij}]^\top \in \mathbb{C}^M \,, \tag{23}$$

where $\{\mathbf{\Phi}\} := \{\mathbf{\Phi}_k\}_{k=1}^M$, $\alpha \in [0,1]$, $\lambda_p, \lambda_q > 0$ are tuning parameters, $P_p(\mathbf{\Omega})$ is the lasso constraint, $P_q(\{\mathbf{\Phi}\})$ is a sparse-group lasso sparsity constraint (cf. [45]–[47]) and $\sqrt{M}$ in $P_q(\{\mathbf{\Phi}\})$ reflects number of group variables [47].

IV. Optimization

The objective function $\mathcal{L}(\mathbf{\Omega}, \mathbf{\Gamma})$ in (20) is biconvex: (strictly) convex in $\mathbf{\Gamma}$, $\mathbf{\Phi}_k \succ \mathbf{0}$, for fixed $\mathbf{\Omega}$, and (strictly) convex in $\mathbf{\Omega}$, $\mathbf{\Omega} \succ \mathbf{0}$, for fixed $\mathbf{\Gamma}$. As is a general approach for biconvex function optimization [48], we will use an iterative and alternating minimization approach where we optimize w.r.t. $\mathbf{\Omega}$ with $\mathbf{\Gamma}$ fixed, and then optimize w.r.t. $\mathbf{\Gamma}$ with $\mathbf{\Omega}$ fixed at the last optimized value, and repeat the two optimizations (flip-flop). The algorithm is only guaranteed to converge to a local stationary point of $\mathcal{L}(\mathbf{\Omega}, \mathbf{\Gamma})$ [48, Sec. 4.2.1].

With $\hat{\mathbf{\Gamma}} = [\hat{\mathbf{\Phi}}_1 \ \hat{\mathbf{\Phi}}_2 \ \cdots \ \hat{\mathbf{\Phi}}_M]$ denoting the estimate of $\mathbf{\Gamma}$, fix $\mathbf{\Gamma} = \hat{\mathbf{\Gamma}}$ and let $\mathcal{L}_1(\mathbf{\Omega})$ denote $\mathcal{L}(\mathbf{\Omega}, \hat{\mathbf{\Gamma}})$ up to some irrelevant constants. We minimize $\mathcal{L}_1(\mathbf{\Omega})$ w.r.t. $\mathbf{\Omega}$ to estimate $\hat{\mathbf{\Omega}}$, where

$$\mathcal{L}_1(\mathbf{\Omega}) = -\frac{1}{p} \ln(|\mathbf{\Omega}|) + \frac{1}{p} \mathrm{tr}\left(\mathbf{\Omega} \check{\mathbf{\Theta}}\right) + P_p(\mathbf{\Omega}) \,, \tag{24}$$

$$\check{\mathbf{\Theta}} = \frac{1}{MKq} \sum_{k=1}^{M} \sum_{\ell=-m_t}^{m_t} \mathrm{Re}\big\{ \boldsymbol{D}_z(\tilde{f}_{k,\ell}) \hat{\mathbf{\Phi}}_k^* \boldsymbol{D}_z^H(\tilde{f}_{k,\ell}) \big\} \,. \tag{25}$$

Fix $\mathbf{\Omega} = \hat{\mathbf{\Omega}}$ and and let $\mathcal{L}_2(\mathbf{\Gamma})$ denote $\mathcal{L}(\hat{\mathbf{\Omega}}, \mathbf{\Gamma})$ up to some irrelevant constants. We minimize $\mathcal{L}_2(\mathbf{\Gamma})$ w.r.t. $\mathbf{\Gamma}$ to obtain estimate $\hat{\mathbf{\Gamma}}$, where

$$\begin{aligned} \mathcal{L}_2(\mathbf{\Gamma}) &= -\frac{1}{2Mq} \sum_{k=1}^{M} (\ln(|\mathbf{\Phi}_k|) + \ln(|\mathbf{\Phi}_k^*|)) \\ &\quad + \frac{1}{2Mq} \sum_{k=1}^{M} \mathrm{tr}\left(\tilde{\mathbf{\Theta}}_k \mathbf{\Phi}_k + \tilde{\mathbf{\Theta}}_k^* \mathbf{\Phi}_k^* \right) + P_q(\{\mathbf{\Phi}\}) \,, \end{aligned} \tag{26}$$

$$\tilde{\mathbf{\Theta}}_k = \frac{1}{Kp} \sum_{\ell=-m_t}^{m_t} \boldsymbol{D}_z^\top(\tilde{f}_{k,\ell}) \hat{\mathbf{\Omega}} \boldsymbol{D}_z^*(\tilde{f}_{k,\ell}) \,. \tag{27}$$

Our optimization algorithm is as in Sec. IV-A.

A. Flip-Flop Optimization

1. Initialize $m = 1$, $\mathbf{\Omega}^{(0)} = \boldsymbol{I}_p$, $\mathbf{\Phi}_k^{(0)} = \boldsymbol{I}_q$, $k \in [M]$.
2. Set $\hat{\mathbf{\Omega}} = \mathbf{\Omega}^{(m-1)}$ in (27). Use the iterative ADMM algorithm [49], as outlined in [9, Sec. 4] and based on Wirtinger calculus [38], to minimize $\mathcal{L}_2(\mathbf{\Gamma})$ (given by (26)) w.r.t. $\mathbf{\Gamma}$ to obtain estimates $\mathbf{\Phi}_k^{(m)}$, $k \in [M]$, the M component matrices of the estimate $\mathbf{\Gamma}^{(m)}$. Details are in Sec. IV-B and step II of Sec. IV-D. Normalize $\mathbf{\Gamma}^{(m)} \leftarrow \mathbf{\Gamma}^{(m)} / \|\mathbf{\Gamma}^{(m)}\|_F$ to resolve the scaling ambiguity. Let $m \leftarrow m+1$.
3. Set $\hat{\mathbf{\Gamma}} = \mathbf{\Gamma}^{(m)}$ in (25). Use the ADMM algorithm of [40, Sec. III] (with $\alpha = 1$ therein, no group-lasso penalty) to minimize $\mathcal{L}_1(\mathbf{\Omega})$ w.r.t. $\mathbf{\Omega}$, to obtain estimate $\mathbf{\Omega}^{(m)}$. Details are in Sec. IV-C and step IV of Sec. IV-D.
4. Repeat steps 2 and 3 until convergence.

B. ADMM for Estimation of $\mathbf{\Gamma}$

After variable splitting, the scaled augmented Lagrangian for minimization of $\mathcal{L}_2(\mathbf{\Gamma})$ is [9]

$$\begin{aligned} \mathcal{L}_2^{aL}(\{\mathbf{\Phi}\}, \{\boldsymbol{W}\}, \{\boldsymbol{U}\}) &= \frac{1}{2Mq} \sum_{k=1}^{M} \mathrm{tr}\big(\tilde{\mathbf{\Theta}}_k \mathbf{\Phi}_k + \tilde{\mathbf{\Theta}}_k^* \mathbf{\Phi}_k^* \big) \\ &\quad - \frac{1}{2Mq} \sum_{k=1}^{M} (\ln(|\mathbf{\Phi}_k|) + \ln(|\mathbf{\Phi}_k^*|)) + P_q(\{\boldsymbol{W}\}) \\ &\quad + \frac{\rho}{2} \sum_{k=1}^{M} \|\mathbf{\Phi}_k - \boldsymbol{W}_k + \boldsymbol{U}_k\|_F^2 \end{aligned}$$

where $\{\boldsymbol{U}\} = \{\boldsymbol{U}_k\}_{k=1}^M$ are dual variables, similarly $\{\boldsymbol{W}_k\}_{k=1}^M$ are the "split" variables, $\rho > 0$ is the penalty parameter, $\boldsymbol{U}_k, \boldsymbol{W}_k \in \mathbb{C}^{q \times q}$. Given the results $\{\tilde{\mathbf{\Phi}}^{(i)}\}, \{\boldsymbol{W}^{(i)}\}, \{\boldsymbol{U}^{(i)}\}$ of the ith iteration, in the $(i+1)$st iteration, the ADMM algorithm executes the following three updates, given in Steps (a)-(c), until convergence. To distinguish between the estimates $\mathbf{\Gamma}^{(m)}$ and $\mathbf{\Phi}_k^{(m)}$ of the mth iteration of the flip-flip optimization and the estimate of the ith iteration of the ADMM algorithm, we use $\tilde{\mathbf{\Phi}}_k^{(i)}$ for the latter.

Step (a). $\{\tilde{\mathbf{\Phi}}^{(i+1)}\} \leftarrow \arg\min_{\{\mathbf{\Phi}\}} \mathcal{L}_2^{aL}(\{\mathbf{\Phi}\}, \{\boldsymbol{W}^{(i)}\}, \{\boldsymbol{U}^{(i)}\})$. Up to some terms not dependent upon $\mathbf{\Phi}_k$'s [9]

$$\begin{aligned} &\mathcal{L}_2^{aL}(\{\mathbf{\Phi}\}, \{\boldsymbol{W}^{(i)}\}, \{\boldsymbol{U}^{(i)}) \\ &\quad = \frac{1}{2Mq} \sum_{k=1}^{M} \mathcal{L}_{2k}^{aL}(\mathbf{\Phi}_k, \boldsymbol{W}_k^{(i)}, \boldsymbol{U}_k^{(i)}) \,, \end{aligned}$$

$$\mathcal{L}_{2k}^{aL}(\boldsymbol{\Phi}_k, \boldsymbol{W}_k^{(i)}, \boldsymbol{U}_k^{(i)}) = -\ln(|\boldsymbol{\Phi}_k|) - \ln(|\boldsymbol{\Phi}_k^*|) + \mathrm{tr}\big(\tilde{\boldsymbol{\Theta}}_k \boldsymbol{\Phi}_k + \tilde{\boldsymbol{\Theta}}_k^* \boldsymbol{\Phi}_k^*\big) + Mq\rho \|\boldsymbol{\Phi}_k - \boldsymbol{W}_k^{(i)} + \boldsymbol{U}_k^{(i)}\|_F^2 ,$$

that is, the objective function is separable in k. For each k, the solution is as follows [9]. Let $\boldsymbol{P}\boldsymbol{\Delta}\boldsymbol{P}^H$ denote the eigen-decomposition of the Hermitian $\tilde{\boldsymbol{\Theta}}_k - Mq\rho\left(\boldsymbol{W}_k^{(i)} - \boldsymbol{U}_k^{(i)}\right)$ with diagonal matrix $\boldsymbol{\Delta}$ consisting of the eigenvalues and $\boldsymbol{P}\boldsymbol{P}^H = \boldsymbol{P}^H\boldsymbol{P} = \boldsymbol{I}_q$. Then $\tilde{\boldsymbol{\Phi}}_k^{(i+1)} = \boldsymbol{P}\tilde{\boldsymbol{\Delta}}\boldsymbol{P}^H$ where $\tilde{\boldsymbol{\Delta}}$ is the diagonal matrix with ℓth diagonal element

$$[\tilde{\boldsymbol{\Delta}}]_{\ell\ell} = \frac{-[\boldsymbol{\Delta}]_{\ell\ell} + \sqrt{([\boldsymbol{\Delta}]_{\ell\ell})^2 + 4Mq\rho}}{2Mq\rho} .$$

Step (b). Here we have

$$\{\mathbf{W}^{(i+1)}\} \leftarrow \arg\min_{\{\boldsymbol{W}\}} \mathcal{L}_2^{aL}(\{\tilde{\boldsymbol{\Phi}}^{(i+1)}\}, \{\boldsymbol{W}\}, \{\boldsymbol{U}^{(i)}\}).$$

We update $\{\boldsymbol{W}_k^{(i+1)}\}_{k=1}^M$ as the minimizer w.r.t. $\{\boldsymbol{W}\}_{k=1}^M$ of

$$\frac{\rho}{2}\sum_{k=1}^{M} \|\boldsymbol{W}_k - (\tilde{\boldsymbol{\Phi}}_k^{(i+1)} + \boldsymbol{U}_k^{(i)})\|_F^2 + P_q(\{\boldsymbol{W}\}).$$

The solution follows from [9, Lemma 1]. Let $\boldsymbol{G}_k = \tilde{\boldsymbol{\Phi}}_k^{(i+1)} + \boldsymbol{U}_k^{(i)} \in \mathbb{C}^{q\times q}$ and let $\boldsymbol{G}^{(j\ell)} \in \mathbb{C}^M$ be defined as in (23), but based on $\boldsymbol{G}_k$'s. Then the update of $\{\boldsymbol{W}\}$ is given by

$$[\boldsymbol{W}_k^{(i+1)}]_{j\ell} = [\boldsymbol{G}_k]_{jj} , \qquad \text{if } j = \ell$$

$$[\boldsymbol{W}_k^{(i+1)}]_{j\ell} = \left(1 - \frac{(1-\alpha)\lambda_q\sqrt{M}}{\rho\|\boldsymbol{S}_F(\boldsymbol{G}^{(j\ell)},\, \alpha\lambda_q/\rho)\|}\right)_+ \times S_F\Big([\boldsymbol{G}^{(j\ell)}]_k, \frac{\alpha\lambda_q}{\rho}\Big), \qquad \text{if } j \neq \ell ,$$

where $(b)_+ := \max(0, b)$, $S_F(b, \beta) := (1 - \beta/|b|)_+ b$ (for complex scalar $b \neq 0$) is the soft-thresholding scalar operator, and $[\boldsymbol{S}_F(\boldsymbol{a}, \beta)]_j = S(a_j, \beta)$ with $a_j = [\boldsymbol{a}]_j$, is the soft-thresholding vector operator.

Step (c). $\{\boldsymbol{U}^{(i+1)}\} \leftarrow \{\boldsymbol{U}^{(i)}\} + \Big(\{\tilde{\boldsymbol{\Phi}}^{(i+1)}\} - \{\boldsymbol{W}^{(i+1)}\}\Big)$.

C. ADMM for Estimation of $\boldsymbol{\Omega}$

Using variable splitting, consider

$$\min_{\boldsymbol{\Omega}\succ\mathbf{0}, \bar{\boldsymbol{W}}} \Big\{\frac{1}{p}\big(\mathrm{tr}(\check{\boldsymbol{\Theta}}\boldsymbol{\Omega}) - \ln(|\boldsymbol{\Omega}|)\big) + \lambda_p\|\bar{\boldsymbol{W}}^-\|_1\Big\}$$

subject to $\boldsymbol{\Omega} = \bar{\boldsymbol{W}}$. The scaled augmented Lagrangian for this problem is [49]

$$\mathcal{L}_1^{aL}(\boldsymbol{\Omega}, \bar{\boldsymbol{W}}, \bar{\boldsymbol{U}}) = (1/p)\big(\mathrm{tr}(\check{\boldsymbol{\Theta}}\boldsymbol{\Omega}) - \ln(|\boldsymbol{\Omega}|)\big) + \lambda_p\|\bar{\boldsymbol{W}}^-\|_1 + \frac{\rho}{2}\|\boldsymbol{\Omega} - \bar{\boldsymbol{W}} + \bar{\boldsymbol{U}}\|_F^2$$

where $\bar{\boldsymbol{U}}$ is the dual variable, and $\rho > 0$ is the penalty parameter. Given the results $\tilde{\boldsymbol{\Omega}}^{(i)}, \bar{\boldsymbol{W}}^{(i)}, \bar{\boldsymbol{U}}^{(i)}$ of the ith iteration, in the $(i+1)$st iteration, an ADMM algorithm executes the following three updates until convergence:

Step (a). $\tilde{\boldsymbol{\Omega}}^{(i+1)} \leftarrow \arg\min_{\boldsymbol{\Omega}} \mathcal{L}_1^{aL}(\boldsymbol{\Omega}, \bar{\boldsymbol{W}}^{(i)}, \bar{\boldsymbol{U}}^{(i)})$. We choose $\boldsymbol{\Omega}$ to minimize

$$\mathrm{tr}(\check{\boldsymbol{\Theta}}\boldsymbol{\Omega}) - \ln(|\boldsymbol{\Omega}|) + \frac{p\rho}{2}\|\boldsymbol{\Omega} - \bar{\boldsymbol{W}}^{(i)} + \bar{\boldsymbol{U}}^{(i)}\|_F^2.$$

The solution is as follows [40]. Let $\boldsymbol{Q}\boldsymbol{J}\boldsymbol{Q}^\top$ denote the eigen-decomposition of $\check{\boldsymbol{\Theta}} - p\rho\left(\bar{\boldsymbol{W}}^{(i)} - \bar{\boldsymbol{U}}^{(i)}\right)$ with diagonal matrix $\boldsymbol{J}$ consisting of the eigenvalues and $\boldsymbol{Q}\boldsymbol{Q}^\top = \boldsymbol{Q}^\top\boldsymbol{Q} = \boldsymbol{I}_q$. Then $\tilde{\boldsymbol{\Omega}}^{(i+1)} = \boldsymbol{Q}\tilde{\boldsymbol{J}}\boldsymbol{Q}^\top$ where $\tilde{\boldsymbol{J}}$ is the diagonal matrix with ℓth diagonal element

$$[\tilde{\boldsymbol{J}}]_{\ell\ell} = \frac{-[\boldsymbol{J}]_{\ell\ell} + \sqrt{([\boldsymbol{J}]_{\ell\ell})^2 + 4p\rho}}{2p\rho} .$$

Step (b). $\bar{\boldsymbol{W}}^{(i+1)} \leftarrow \arg\min_{\bar{\boldsymbol{W}}} \mathcal{L}_1^{aL}(\tilde{\boldsymbol{\Omega}}^{(i+1)}, \bar{\boldsymbol{W}}, \bar{\boldsymbol{U}}^{(i)})$. We update $\bar{\boldsymbol{W}}^{(i+1)}$ as the minimizer w.r.t. $\bar{\boldsymbol{W}}$ of

$$\lambda_p\|\bar{\boldsymbol{W}}^-\|_1 + \frac{\rho}{2}\|\tilde{\boldsymbol{\Omega}}^{(i+1)} - \bar{\boldsymbol{W}} + \bar{\boldsymbol{U}}^{(i)}\|_F^2.$$

The solution is soft thresholding given by [40]

$$[\bar{\boldsymbol{W}}]_{jk}^{(i+1)} = \begin{cases} [\tilde{\boldsymbol{\Omega}}^{(i+1)} - \bar{\boldsymbol{U}}^{(i)}]_{jj} & \text{if } j = k \\ S_F([\tilde{\boldsymbol{\Omega}}^{(i+1)} - \bar{\boldsymbol{U}}^{(i)}]_{jk}, \frac{\lambda_p}{\rho}) & \text{if } j \neq k \end{cases}$$

where $S_F()$ denotes soft-thresholding as in Sec. IV-C.

Step (c). $\bar{\boldsymbol{U}}^{(i+1)} \leftarrow \bar{\boldsymbol{U}}^{(i)} + \Big(\tilde{\boldsymbol{\Omega}}^{(i+1)} - \bar{\boldsymbol{W}}^{(i+1)}\Big)$.

D. Practical Implementation

Here we present our implementation of the algorithms of Secs. IV-A-IV-C that was used in our numerical results.

I. Parameters $\bar{\mu} = 10$, $\tau_{rel} = \tau_{abs} = 10^{-4}$, $\tau_{ff} = 10^{-5}$, $m_{\max} = 20$, $i_{\max} = 100$ and $\rho^{(0)} = 2$. Initialize $m = 1$, $\boldsymbol{\Omega}^{(0)} = \boldsymbol{I}_p$, $\boldsymbol{\Phi}_k^{(0)} = \boldsymbol{I}_q$, $k \in [M]$.

II. For $m = 1, 2, \cdots, m_{\max}$, do steps III-IV.

III. Set $\hat{\boldsymbol{\Omega}} = \boldsymbol{\Omega}^{(m-1)}$ in (27). Pick $\rho = \rho^{(0)}$, $\tilde{\boldsymbol{\Phi}}_k^{(0)} = \boldsymbol{I}_q$ for $k \in [M]$. For $i = 0, 1, \cdots, i_{\max}$, do steps 1-6 below.

1. For $k \in [M]$, update $\boldsymbol{\Phi}_k$ as $\tilde{\boldsymbol{\Phi}}_k^{(i+1)}$ as in step (a), Sec. IV-B, then update $\boldsymbol{W}_k$ as $\boldsymbol{W}_k^{(i+1)}$ as in step (b), Sec. IV-B, and then update $\boldsymbol{U}_k$ as $\boldsymbol{U}_k^{(i+1)}$ as in step (c), Sec. IV-B, all with $\rho = \rho^{(i)}$.
2. Check for convergence following [9, Sec. 4.1.5]. Define the primal residual matrix $\boldsymbol{E}_{pri}^{(i+1)} \in \mathbb{C}^{q\times(qM)}$ at the $(i+1)$st iteration as

$$\boldsymbol{E}_{pri}^{(i+1)} = \Big[\tilde{\boldsymbol{\Phi}}_1^{(i+1)} - \boldsymbol{W}_1^{(i+1)},\ \tilde{\boldsymbol{\Phi}}_2^{(i+1)} - \boldsymbol{W}_2^{(i+1)},\ \cdots,\ \tilde{\boldsymbol{\Phi}}_M^{(i+1)} - \boldsymbol{W}_M^{(i+1)}\Big]$$

and the dual residual matrix $\boldsymbol{E}_{dual}^{(i+1)} \in \mathbb{C}^{q\times(qM)}$ at the $(i+1)$st iteration as

$$\boldsymbol{E}_{dual}^{(i+1)} = \rho^{(i)}\Big[\boldsymbol{W}_1^{(i+1)} - \boldsymbol{W}_1^{(i)},\ \boldsymbol{W}_2^{(i+1)} - \boldsymbol{W}_2^{(i)},\ \cdots,\ \boldsymbol{W}_M^{(i+1)} - \boldsymbol{W}_M^{(i)}\Big] .$$

Let $e_1 = \|[\tilde{\boldsymbol{\Phi}}_1^{(i+1)}\ \ \tilde{\boldsymbol{\Phi}}_2^{(i+1)}\ \ \cdots\ \ \tilde{\boldsymbol{\Phi}}_M^{(i+1)}]\|_F$, $e_2 = \|[\boldsymbol{W}_1^{(i+1)}\ \ \boldsymbol{W}_2^{(i+1)}\ \ \cdots\ \ \boldsymbol{W}_M^{(i+1)}]\|_F$, $e_3 = \|[\boldsymbol{U}_1^{(i+1)}\ \ \boldsymbol{U}_2^{(i+1)}\ \ \cdots\ \ \boldsymbol{U}_M^{(i+1)}]\|_F$, $\tau_{pri} = q\sqrt{M}\,\tau_{abs} + \tau_{rel}\,\max(e_1, e_2)$ and $\tau_{dual} = q\sqrt{M}\,\tau_{abs} + \tau_{rel}\,e_3/\rho^{(i)}$. If $\|\boldsymbol{E}_{pri}^{(i+1)}\|_F \leq \tau_{pri}$ and $\|\boldsymbol{E}_{dual}^{(i+1)}\|_F \leq \tau_{dual}$, the convergence criterion is met. If the convergence criterion is met or if $i + 1 > i_{\max}$, exit to step IV after setting $\boldsymbol{\Phi}_k^{(m)} = \tilde{\boldsymbol{\Phi}}_k^{(i+1)}$, $k \in [M]$, and then normalizing $\boldsymbol{\Gamma}^{(m)} \leftarrow \boldsymbol{\Gamma}^{(m)}/\|\boldsymbol{\Gamma}^{(m)}\|_F$, else continue,
3. Update variable penalty parameter ρ as

$$\rho^{(i+1)} = \begin{cases} 2\rho^{(i)} & \text{if } \|\boldsymbol{E}_{pri}^{(i+1)}\|_F > \bar{\mu}\|\boldsymbol{E}_{dual}^{(i+1)}\|_F \\ \rho^{(i)}/2 & \text{if } \|\boldsymbol{E}_{dual}^{(i+1)}\|_F > \bar{\mu}\|\boldsymbol{E}_{pri}^{(i+1)}\|_F \\ \rho^{(i)} & \text{otherwise.} \end{cases}$$

For $k \in [M]$, set $\boldsymbol{U}_k^{(i+1)} = \boldsymbol{U}_k^{(i)}/2$ if $\|\boldsymbol{E}_{pri}^{(i+1)}\|_F > \bar{\mu}\|\boldsymbol{E}_{dual}^{(i+1)}\|_F$ and $\boldsymbol{U}_k^{(i+1)} = 2\boldsymbol{U}_k^{(i)}$ if $\|\boldsymbol{E}_{dual}^{(i+1)}\|_F > \bar{\mu}\|\boldsymbol{E}_{pri}^{(i+1)}\|_F$.

4. Set $i \leftarrow i+1$ and return to step 2.

IV. Set $\hat{\boldsymbol{\Gamma}} = \boldsymbol{\Gamma}^{(m)}$ in (25). Pick $\rho = \rho^{(0)}$, $\tilde{\boldsymbol{\Omega}}^{(0)} = \boldsymbol{I}_p$. For $i = 0, 1, \cdots, i_{\max}$, do steps i-v below.

i. Update $\boldsymbol{\Omega}$ as $\tilde{\boldsymbol{\Omega}}^{(i+1)}$ as in step (a), Sec. IV-C, then update $\bar{\boldsymbol{W}}$ as $\bar{\boldsymbol{W}}^{(i+1)}$ as in step (b), Sec. IV-C, and then update $\bar{\boldsymbol{U}}$ as $\bar{\boldsymbol{U}}^{(i+1)}$ as in step (c), Sec. IV-C, all with $\rho = \rho^{(i)}$.

ii. Check for convergence following [40, Sec. II-A]. Define the primal residual matrix $\boldsymbol{H}_{pri}^{(i+1)} = \tilde{\boldsymbol{\Omega}}^{(i+1)} - \bar{\boldsymbol{W}}^{(i+1)}$ and the dual residual matrix $\boldsymbol{H}_{dual}^{(i+1)} = \rho^{(i)}\big[\bar{\boldsymbol{W}}^{(i+1)} - \bar{\boldsymbol{W}}^{(i)}\big]$ where $\boldsymbol{H}_{pri}^{(i+1)}, \boldsymbol{H}_{dual}^{(i+1)} \in \mathbb{C}^{p\times p}$. Let

$$\begin{aligned}\tau_{pri} &= p\,\tau_{abs} + \tau_{rel}\,\max(\|\tilde{\boldsymbol{\Omega}}^{(i+1)}\|_F, \|\tilde{\boldsymbol{W}}^{(i+1)}\|_F)\\ \tau_{dual} &= p\,\tau_{abs} + \tau_{rel}\,\|\tilde{\boldsymbol{U}}^{(i+1)}\|_F/\rho^{(i)}\,.\end{aligned}$$

If $\|\boldsymbol{H}_{pri}^{(i+1)}\|_F \le \tau_{pri}$ and $\|\boldsymbol{H}_{dual}^{(i+1)}\|_F \le \tau_{dual}$, the convergence criterion is met. If the convergence criterion is met or if $i+1 > i_{\max}$, exit to step V after setting $\boldsymbol{\Omega}^{(m)} = \tilde{\boldsymbol{\Omega}}^{(i+1)}$, else continue.

iii. Update variable penalty parameter ρ as

$$\rho^{(i+1)} = \begin{cases} 2\rho^{(i)} & \text{if } \|\boldsymbol{H}_{pri}^{(i+1)}\|_F > \bar{\mu}\|\boldsymbol{H}_{dual}^{(i+1)}\|_F \\ \rho^{(i)}/2 & \text{if } \|\boldsymbol{H}_{dual}^{(i+1)}\|_F > \bar{\mu}\|\boldsymbol{H}_{pri}^{(i+1)}\|_F \\ \rho^{(i)} & \text{otherwise}\end{cases}$$

Set $\bar{\boldsymbol{U}}^{(i+1)} = \bar{\boldsymbol{U}}^{(i)}/2$ if $\|\boldsymbol{H}_{pri}^{(i+1)}\|_F > \bar{\mu}\|\boldsymbol{H}_{dual}^{(i+1)}\|_F$ and $\bar{\boldsymbol{U}}^{(i+1)} = 2\bar{\boldsymbol{U}}^{(i)}$ if $\|\boldsymbol{H}_{dual}^{(i+1)}\|_F > \bar{\mu}\|\boldsymbol{H}_{pri}^{(i+1)}\|_F$.

iv. Set $i \leftarrow i+1$ and return to step ii.

V. Check for convergence of the flip-flop algorithm. If $\|\boldsymbol{\Gamma}^{(m)} - \boldsymbol{\Gamma}^{(m-1)}\|_F/\|\boldsymbol{\Gamma}^{(m-1)}\|_F \le \tau_{ff}$ and $\|\boldsymbol{\Omega}^{(m)} - \boldsymbol{\Omega}^{(m-1)}\|_F/\|\boldsymbol{\Omega}^{(m-1)}\|_F \le \tau_{ff}$, or $m > m_{\max}$, go to step VI, else set $m \leftarrow m+1$ and return to step III.

VI. The final estimates are given by $\hat{\boldsymbol{\Omega}} = \boldsymbol{\Omega}^{(m)}$ and $\hat{\boldsymbol{\Gamma}} = \boldsymbol{\Gamma}^{(m)}$, and $\mathcal{E}_{\hat{\boldsymbol{\Omega}}} = \{(i,j) : |[\hat{\boldsymbol{\Omega}}]_{ij}| > 0\}$ and $\mathcal{E}_{\hat{\boldsymbol{\Gamma}}} = \{(i,j) : \|\hat{\boldsymbol{\Phi}}^{(ij)}\| > 0\}$ are the estimated edgesets for $\boldsymbol{\Omega}$ and $\boldsymbol{\Gamma}$ respectively.

Remark 1. We terminate the flip-flop optimization (step V) when relative improvements in new updates of both $\boldsymbol{\Omega}^{(m)}$ and $\boldsymbol{\Gamma}^{(m)}$ are below the threshold τ_{ff}, or the maximum number of iterations in m is reached. The ADMM algorithms are terminated when both primary and dual residuals are below the respective tolerances, or the maximum number of iterations in i is reached; here we follow [49, Sec. 3.3.1] (see also [40] and [9]). The variable penalty $\rho^{(i)}$ follows the recommendations in [49, Sec. 3.4.1]. The most expensive computation in Sec. IV-B is in step (a) requiring the eigen-decomposition of M $q \times q$ matrices, with computational complexity $\mathcal{O}(Mq^3)$. Similarly, the most expensive computation in Sec. IV-C is in step (a) requiring the eigen-decomposition of a $p\times p$ matrix, with computational complexity $\mathcal{O}(p^3)$. Thus the overall computational complexity of our proposed approach is $\mathcal{O}(Mq^3 + p^3)$. □

E. BIC for selection of λ_p, λ_q (and α)

Given n, K and M, the Bayesian information criterion (BIC) is given by (see also [9])

$$\begin{aligned}\mathrm{BIC}(\lambda_p, \lambda_q, \alpha) &= -2KMq\ln(|\hat{\boldsymbol{\Omega}}|)\\ &\quad + 2Kp\sum_{k=1}^{M}\Big(-\ln(|\hat{\boldsymbol{\Phi}}_k|) + p^{-1}\,\mathrm{Re}\big(\mathrm{tr}(\hat{\boldsymbol{A}}_k)\big)\Big)\\ &\quad + \ln(2KM)\big(|\hat{\boldsymbol{\Omega}}|_0/2 + \sum_{k=1}^{M}|\hat{\boldsymbol{\Phi}}_k|_0\big)\end{aligned} \tag{28}$$

where $\hat{\boldsymbol{A}}_k$ is given by (19) with $\boldsymbol{\Omega}$ and $\boldsymbol{\Phi}_k$ therein replaced with $\hat{\boldsymbol{\Omega}}$ and $\hat{\boldsymbol{\Phi}}_k$, respectively, $|\boldsymbol{H}|_0$ denotes number of nonzero elements in $\boldsymbol{H}$, $2KM$ is total number of real-valued measurements in frequency-domain and $2K$ is the number of real-valued measurements per frequency point, with total M frequencies in $(0, 0.5)$. A general expression for BIC is $-2\log$-likelihood+(number of model parameters)$\times\log$(number of data points). The expression in (28) follows by using $\{\boldsymbol{d}_z(f_m)\}_{m=1}^{(n/2)-1)}$ as complex-valued data in frequency-domain whose log-likelihood is given by (18). We count each complex value as two real values, both for data points and for parameters (entries of $\hat{\boldsymbol{\Phi}}_k$), and also use the fact that $\hat{\boldsymbol{\Omega}}$ is symmetric and $\hat{\boldsymbol{\Phi}}_k$ is Hermitian.

Pick α, λ_q and λ_p to minimize BIC. In our simulations we fixed $\alpha = 0.05$ and then picked λ_q and λ_p over a grid of values, as follows. We search over $\lambda_q \in [\lambda_{q\ell}, \lambda_{qu}]$ and $\lambda_p \in [\lambda_{p\ell}, \lambda_{pu}]$ selected via a heuristic as in [40]. Find the smallest λ_q and λ_p, labeled λ_{qsm} and λ_{psm}, for which we get a no-edge model; then we set $\lambda_{qu} = \lambda_{qsm}/2$ and $\lambda_{q\ell} = \lambda_{qu}/10$; similarly for λ_{pu} and $\lambda_{p\ell}$.

V. Theoretical Analysis

Now we provide sufficient conditions for local convergence in the Frobenius norm of the Kronecker-decomposable inverse PSD estimators to the true value or a scaled version of it. First some notation. The true values of $\boldsymbol{\Omega}$, $\boldsymbol{\Sigma}$ and $\bar{\boldsymbol{S}}(f)$ will be denoted as $\boldsymbol{\Omega}^\diamond$, $\boldsymbol{\Sigma}^\diamond$ and $\bar{\boldsymbol{S}}^\diamond(f)$, respectively. Therefore, $\boldsymbol{\Omega}^\diamond = (\boldsymbol{\Sigma}^\diamond)^{-1}$. Since we use $\boldsymbol{\Phi}_k := \bar{\boldsymbol{S}}^{-1}(\tilde{f}_k)$, we have $\boldsymbol{\Phi}_k^\diamond := \bar{\boldsymbol{S}}^{-\diamond}(\tilde{f}_k)$ (where $\boldsymbol{A}^{-\diamond} = (\boldsymbol{A}^\diamond)^{-1}$). Therefore, in this notation, $\boldsymbol{d}_z(f_m) \sim \mathcal{N}_c(\boldsymbol{0}, \boldsymbol{S}_z^\diamond(f_m))$ and $\boldsymbol{S}_z^\diamond(f_m) = \bar{\boldsymbol{S}}^\diamond(f_m) \otimes \boldsymbol{\Sigma}^\diamond$. Also in this section, we replace $\hat{\boldsymbol{\Phi}}_k$'s in (25) with $\boldsymbol{\Phi}_k$'s and still use the notation $\tilde{\boldsymbol{\Theta}}$ for the sum (25) and the notation $\mathcal{L}_1(\boldsymbol{\Omega})$ for (24), and similarly, we replace $\hat{\boldsymbol{\Omega}}$ in (27) with $\boldsymbol{\Omega}$ and still use the notation $\tilde{\boldsymbol{\Theta}}$ for the sum (27) and $\mathcal{L}_2(\boldsymbol{\Gamma})$ for (26).

We follow [22] in first considering the solution to the unpenalized population objective function (i.e., expectation of $G(\boldsymbol{\Omega}, \{\boldsymbol{\Phi}\}, \{\boldsymbol{\Phi}^*\})$ given by 18). We have

$$\begin{aligned}\bar{G}(\boldsymbol{\Omega}, \{\boldsymbol{\Phi}\}, \{\boldsymbol{\Phi}^*\}) &= E\{G(\boldsymbol{\Omega}, \{\boldsymbol{\Phi}\}, \{\boldsymbol{\Phi}^*\})\}\\ &= -\frac{1}{p}\ln(|\boldsymbol{\Omega}|) - \frac{1}{2Mq}\sum_{k=1}^{M}\Big[\ln(|\boldsymbol{\Phi}_k|) + \ln(|\boldsymbol{\Phi}_k^*|)\\ &\quad - \frac{1}{p}\big(\mathrm{tr}(\bar{\boldsymbol{S}}_k^\diamond\boldsymbol{\Phi}_k) + \mathrm{tr}(\bar{\boldsymbol{S}}_k^\diamond\boldsymbol{\Phi}_k)^*\big)\mathrm{tr}(\boldsymbol{\Sigma}^\diamond\boldsymbol{\Omega})\Big],\end{aligned} \tag{29}$$

where we have used the facts $\bar{\boldsymbol{S}}_k^\diamond = \bar{\boldsymbol{S}}^\diamond(\tilde{f}_k)$,

$$E\{\mathrm{tr}(\boldsymbol{A}_k)\} = \frac{1}{Kp}\sum_{\ell=-m_t}^{m_t} \mathrm{tr}(E\{\boldsymbol{d}_z(\tilde{f}_{k,\ell})\boldsymbol{d}_z^H(\tilde{f}_{k,\ell})\}(\boldsymbol{\Phi}_k \otimes \boldsymbol{\Omega}))$$
$$= \frac{1}{p}\mathrm{tr}((\bar{\boldsymbol{S}}_k^\diamond \otimes \boldsymbol{\Sigma}^\diamond)(\boldsymbol{\Phi}_k \otimes \boldsymbol{\Omega})) = \frac{1}{p}\mathrm{tr}((\bar{\boldsymbol{S}}_k^\diamond \boldsymbol{\Phi}_k) \otimes (\boldsymbol{\Sigma}^\diamond \boldsymbol{\Omega}))$$
$$= p^{-1}\mathrm{tr}(\bar{\boldsymbol{S}}_k^\diamond \boldsymbol{\Phi}_k)\mathrm{tr}(\boldsymbol{\Sigma}^\diamond \boldsymbol{\Omega})\,. \tag{30}$$

Define

$$\bar{\boldsymbol{\Omega}}(\boldsymbol{\Gamma}) = \arg\min_{\boldsymbol{\Omega}} \bar{G}(\boldsymbol{\Omega}, \{\boldsymbol{\Phi}\}, \{\boldsymbol{\Phi}^*\})\,, \tag{31}$$
$$\bar{\boldsymbol{\Gamma}}(\boldsymbol{\Omega}) = \arg\min_{\boldsymbol{\Gamma}} \bar{G}(\boldsymbol{\Omega}, \{\boldsymbol{\Phi}\}, \{\boldsymbol{\Phi}^*\}) \tag{32}$$

where $\bar{\boldsymbol{\Gamma}}(\boldsymbol{\Omega}) = [\bar{\boldsymbol{\Phi}}_1(\boldsymbol{\Omega})\ \bar{\boldsymbol{\Phi}}_2(\boldsymbol{\Omega})\ \cdots\ \bar{\boldsymbol{\Phi}}_M(\boldsymbol{\Omega})]$.

Theorem 1. If $\sum_{k=1}^M \left(\mathrm{tr}(\bar{\boldsymbol{S}}_k^\diamond \boldsymbol{\Phi}_k) + \mathrm{tr}(\bar{\boldsymbol{S}}_k^\diamond \boldsymbol{\Phi}_k)^*\right) \neq 0$, then

$$\bar{\boldsymbol{\Omega}}(\boldsymbol{\Gamma}) = \frac{2Mq}{\sum_{k=1}^M \left(\mathrm{tr}(\bar{\boldsymbol{S}}_k^\diamond \boldsymbol{\Phi}_k) + \mathrm{tr}(\bar{\boldsymbol{S}}_k^\diamond \boldsymbol{\Phi}_k)^*\right)}\,\boldsymbol{\Omega}^\diamond\,, \tag{33}$$

and if $\mathrm{tr}(\boldsymbol{\Sigma}^\diamond \boldsymbol{\Omega}) \neq 0$, then for $k \in [M]$,

$$\bar{\boldsymbol{\Phi}}_k(\boldsymbol{\Omega}) = \frac{p}{\mathrm{tr}(\boldsymbol{\Sigma}^\diamond \boldsymbol{\Omega})}\,(\bar{\boldsymbol{S}}_k^\diamond)^{-1} = \frac{p}{\mathrm{tr}(\boldsymbol{\Sigma}^\diamond \boldsymbol{\Omega})}\,\boldsymbol{\Phi}_k^\diamond \quad \bullet \tag{34}$$

Theorem 1 shows that the unpenalized population objective function yields true values up to a constant scalar. Notice that $\bar{\boldsymbol{\Omega}}(\boldsymbol{\Gamma}) = \boldsymbol{\Omega}^\diamond$ if $\boldsymbol{\Gamma} = \boldsymbol{\Gamma}^\diamond$, and similarly, $\bar{\boldsymbol{\Phi}}_k(\boldsymbol{\Omega}) = \boldsymbol{\Phi}_k^\diamond$, $k = 1, 2, \cdots, M$, if $\boldsymbol{\Omega} = \boldsymbol{\Omega}^\diamond$.

We now turn to penalized data-based objective function $\mathcal{L}(\boldsymbol{\Omega}, \boldsymbol{\Gamma})$ which is minimized alternatingly as $\mathcal{L}_2(\boldsymbol{\Gamma})$ w.r.t. $\boldsymbol{\Phi}_k$'s and as $\mathcal{L}_1(\boldsymbol{\Omega})$ w.r.t. $\boldsymbol{\Omega}$. Here in addition to assumptions (A1) and (A2), we assume

(A3) Define the true edgesets $\mathcal{S}_q = \{\{i,j\} : [(\bar{\boldsymbol{S}}^\diamond)^{-1}(f)]_{ij} \not\equiv 0,\ i \neq j,\ 0 \leq f \leq 0.5,\ i,j \in [q]\}$ and $\mathcal{S}_p = \{\{i,j\} : [\boldsymbol{\Omega}^\diamond]_{ij} \neq 0,\ i \neq j,\ i,j \in [p]\}$, where $\bar{\boldsymbol{S}}^\diamond(f)$ denotes DTFT of $\boldsymbol{\Psi}(\tau)$ and $\boldsymbol{\Omega}^\diamond = (\boldsymbol{\Sigma}^\diamond)^{-1}$ denotes the true value of $\boldsymbol{\Omega}$. Assume that number of nonzero elements in the true edgesets $\mathcal{S}_q$ and $\mathcal{S}_p$ are upperbounded as $|\mathcal{S}_q| \leq s_q$ and $|\mathcal{S}_p| \leq s_p$.

(A4) The minimum and maximum eigenvalues of $q \times q$ PSD $\bar{\boldsymbol{S}}^\diamond(f) \succ \boldsymbol{0}$ satisfy $0 < \beta_{q,\min} \leq \min_{f\in[0,0.5]} \phi_{\min}(\bar{\boldsymbol{S}}^\diamond(f))$ and $\max_{f\in[0,0.5]} \phi_{\max}(\bar{\boldsymbol{S}}^\diamond(f)) \leq \beta_{q,\max} < \infty$. Similarly, $0 < \beta_{p,\min} \leq \phi_{\min}(\boldsymbol{\Sigma}^\diamond) \leq \phi_{\max}(\boldsymbol{\Sigma}^\diamond) \leq \beta_{p,\max} < \infty$. Here $\beta_{\cdot,\min}$ and $\beta_{\cdot,\max}$ are not functions of n, p, q.

Theorem 2 establishes bounds on estimation errors of local minimizers $\hat{\boldsymbol{\Omega}}(\boldsymbol{\Gamma})$ and $\hat{\boldsymbol{\Gamma}}(\boldsymbol{\Omega})$ of $\mathcal{L}_1(\boldsymbol{\Omega})$ and $\mathcal{L}_2(\boldsymbol{\Gamma})$, respectively. We now explicitly allow p, q, M, K, s_p, s_q, λ_p and λ_q to be functions of sample size n, denoted as p_n, q_n, M_n, K_n, s_{pn}, s_{qn}, λ_{pn} and λ_{qn}, respectively. (In the appendices we do not do so to keep the notation simple.) First we define some variables. For $\tau > 2$, define

$$\gamma_p = 0.1/\beta_{p,\max}\,, \tag{35}$$
$$C_{1q} = \frac{2}{\sqrt{\ln(M_n^{1/\tau} q_n)}} + \sqrt{2\tau + \frac{2\ln(16)}{\ln(M_n^{1/\tau} q_n)}}\,, \tag{36}$$
$$C_{0q} = 16C_{1q}(1 + \gamma_p \beta_{p,\max})\beta_{q,\max}\,, \tag{37}$$
$$\gamma_q = 0.1/\beta_{q,\max}\,, \tag{38}$$
$$C_{1p} = \sqrt{\frac{2}{\ln(p_n)}} + \sqrt{\tau + \frac{\ln(4)}{\ln(p_n)}}\,, \tag{39}$$
$$C_{0p} = 8C_{1p}\,(2 + \gamma_q \beta_{q,\max})\beta_{p,\max}\,, \tag{40}$$
$$r_{qn} = \sqrt{M_n(q_n + s_{qn})\,\ln(M_n^{1/\tau} q_n)/(K_n p_n)} = o(1)\,, \tag{41}$$
$$r_{pn} = \sqrt{(p_n + s_{pn})\,\ln(p_n)/(M_n K_n q_n)} = o(1)\,. \tag{42}$$

Theorem 2. Let $\tau > 2$.

(i) Let $\mathcal{B}(\boldsymbol{\Gamma}^\diamond) = \{\boldsymbol{\Gamma} :\ \|\boldsymbol{\Gamma} - \boldsymbol{\Gamma}^\diamond\|_F \leq \gamma_q,\ \boldsymbol{\Phi}_k = \boldsymbol{\Phi}_k^H \succ \boldsymbol{0}\}$ and $\hat{\boldsymbol{\Omega}}(\boldsymbol{\Gamma}) = \arg\min_{\{\boldsymbol{\Omega}:\,\boldsymbol{\Gamma}\in\mathcal{B}(\boldsymbol{\Gamma}^\diamond)\}} \mathcal{L}_1(\boldsymbol{\Omega})$. Suppose λ_{pn} satisfies

$$\frac{C_{0p}}{p_n}\sqrt{\frac{\ln(p_n)}{M_n K_n q_n}} \leq \lambda_{pn} \leq \frac{C_{0p}}{p_n}\sqrt{\left(1 + \frac{p_n}{s_{pn}}\right)\frac{\ln(p_n)}{M_n K_n q_n}}\,. \tag{43}$$

Let $N_p := \arg\min_n \left\{n :\ r_{pn} \leq \beta_{p,\min}/(34C_{0p})\right\}$. Then under assumptions (A1)-(A4), for $n > N_p$, $\hat{\boldsymbol{\Omega}}(\boldsymbol{\Gamma})$ satisfies

$$\|\hat{\boldsymbol{\Omega}}(\boldsymbol{\Gamma}) - \bar{\boldsymbol{\Omega}}(\boldsymbol{\Gamma})\|_F \leq \frac{17C_{0p}}{\beta_{p,\min}^2}\,r_{pn} \tag{44}$$

with probability greater than $1 - \frac{1}{p^{\tau-2}} - 4p^2 e^{-KqM}$.

(ii) Let $\mathcal{B}(\boldsymbol{\Omega}^\diamond) = \{\boldsymbol{\Omega} :\ \|\boldsymbol{\Omega} - \boldsymbol{\Omega}^\diamond\|_F \leq \gamma_p,\ \boldsymbol{\Omega} = \boldsymbol{\Omega}^\top \succ \boldsymbol{0}\}$ and $\hat{\boldsymbol{\Gamma}}(\boldsymbol{\Omega}) = \arg\min_{\{\boldsymbol{\Gamma}:\,\boldsymbol{\Omega}\in\mathcal{B}(\boldsymbol{\Omega}^\diamond)\}} \mathcal{L}_2(\boldsymbol{\Gamma})$. Suppose λ_{qn} satisfies

$$\frac{C_{0q}}{M_n q_n}\sqrt{d_n} \leq \lambda_{qn} \leq \frac{C_{0q}}{M_n q_n}\sqrt{\left(1 + \frac{q_n}{s_{qn}}\right)d_n}\,, \tag{45}$$

where $d_n = \ln(M_n^{1/\tau} q_n)/(K_n p_n)$. Let $N_q := \arg\min_n \left\{n :\ r_{qn} \leq \beta_{q,\min}/(34C_{0q})\right\}$. Then under assumptions (A1)-(A4), for $n > N_q$ and $\alpha \in [0,1]$, $\hat{\boldsymbol{\Gamma}}(\boldsymbol{\Omega})$ satisfies

$$\|\hat{\boldsymbol{\Gamma}}(\boldsymbol{\Omega}) - \bar{\boldsymbol{\Gamma}}(\boldsymbol{\Omega})\|_F \leq \frac{17C_{0q}}{\beta_{q,\min}^2}\,r_{qn} \tag{46}$$

with probability greater than $1 - \frac{1}{q^{\tau-2}} - 16Mq^2 e^{-Kp/2}$ $\bullet$

Remark 2. Theorem 2 helps determine how to choose M_n and K_n so that for given n, s_{pn}, s_{qn}, q_n and p_n, $\lim_{n\to\infty} r_{pn} = 0$ and $\lim_{n\to\infty} r_{qn} = 0$, and moreover, how fast can s_{pn} and s_{qn} grow with n and still have r_{qn} and $r_{pn} \downarrow 0$. Since $K_n M_n \approx n/2$, if one picks $K_n = \mathcal{O}(n^\mu)$, then $M_n = \mathcal{O}(n^{1-\mu})$ for some $0 < \mu < 1$. We assume p_n and q_n are of the same order. (i) First consider the case $\mathcal{O}(p_n) = \mathcal{O}(p_n + s_{pn}) = \mathcal{O}(q_n) = \mathcal{O}(q_n + s_{qn})$, which, for example, is true for chain graphs. Also, take $\mathcal{O}(p_n) \propto n^\nu$ for some $\nu > 0$. Then $r_{pn} = \mathcal{O}(\sqrt{\ln(n)/n}) \to 0$ as $n \to \infty$, and $r_{qn} = \mathcal{O}(\sqrt{\ln(n)/n^{2\mu-1}}) \to 0$ as $n \to \infty$ if $\mu > 0.5$. This holds for any $\nu > 0$. If $\mu = \frac{3}{4}$, then $r_{qn} = \mathcal{O}(\sqrt{\ln(n)}/n^{1/4}) > r_{pn}$. If $\mu = \frac{2}{3}$, then $r_{qn} = \mathcal{O}(\sqrt{\ln(n)}/n^{1/6}) > r_{pn}$. (ii) Now suppose $\mathcal{O}(p_n) = \mathcal{O}(q_n) \propto n^\nu$ for some $\nu > 0$, but $\mathcal{O}(s_{pn}) = \mathcal{O}(s_{qn}) \propto n^{2\nu} = \mathcal{O}(p_n^2)$, which is true for Erdös-Rènyi graphs, e.g. Then $r_{pn} = \mathcal{O}(\sqrt{\ln(n)/n^{1-\nu}}) \to 0$ as $n \to \infty$ if $\nu < 1$, and $r_{qn} = \mathcal{O}(\sqrt{\ln(n)/n^{2\mu-1-\nu}}) \to 0$ as $n \to \infty$ if $2\mu - \nu > 1$. Clearly $\nu = 1$ does not work. Suppose $\nu = 0.25$ and $\mu = 0.75$. Then $r_{pn} = \mathcal{O}(\sqrt{\ln(n)}/n^{0.375})$ and $r_{qn} = \mathcal{O}(\sqrt{\ln(n)}/n^{1/8})$. $\square$

Remark 3. The values of γ_p and γ_q specified in (35) and (38), respectively, are used in the proofs of Theorem 2(ii) (see after (116)) and Theorem 2(i) (see (101)), respectively. One can enlarge γ_p and γ_q to $\gamma_p = 0.1\sqrt{p_n}/\beta_{p,\min}$ and $\gamma_q = 0.1\sqrt{M_n q_n}/\beta_{q,\min}$, respectively, and the proofs and the other results remain unchanged and valid. Enlarging these values implies that the balls $\mathcal{B}(\mathbf{\Gamma}^\diamond)$ and $\mathcal{B}(\mathbf{\Omega}^\diamond)$ specified in Theorem 2 are larger, signifying larger convergence regions for initialization of $\mathbf{\Gamma}$ and $\mathbf{\Omega}$. However, this would slow the convergence rates from $\|\hat{\bar{\mathbf{\Omega}}}(\mathbf{\Gamma}) - \bar{\mathbf{\Omega}}(\mathbf{\Gamma})\|_F = \mathcal{O}_P(r_{pn})$ and $\|\hat{\bar{\mathbf{\Gamma}}}(\mathbf{\Omega}) - \bar{\mathbf{\Gamma}}(\mathbf{\Omega})\|_F = \mathcal{O}_P(r_{qn})$ to $\|\hat{\bar{\mathbf{\Omega}}}(\mathbf{\Gamma}) - \bar{\mathbf{\Omega}}(\mathbf{\Gamma})\|_F = \mathcal{O}_P(\sqrt{p_n}\, r_{pn})$ and $\|\hat{\bar{\mathbf{\Gamma}}}(\mathbf{\Omega}) - \bar{\mathbf{\Gamma}}(\mathbf{\Omega})\|_F = \mathcal{O}_P(\sqrt{M_n q_n}\, r_{qn})$, respectively. □

Theorem 3. Assume $\|\mathbf{\Omega}^\diamond\|_F = 1$.
(i) Define $\hat{\mathbf{\Omega}} = \hat{\bar{\mathbf{\Omega}}}(\mathbf{\Gamma})/\|\hat{\bar{\mathbf{\Omega}}}(\mathbf{\Gamma})\|_F$. Let $N_{2p} := \arg\min_n \{n : r_{pn} \le \beta^2_{p,\min}\|\bar{\mathbf{\Omega}}(\mathbf{\Gamma})\|_F/(34C_{0p})\}$ and $\gamma_r = (\beta_{q,\max} + \beta_{q,\min})/\beta_{q,\min}$. Under the assumptions of Theorem 2(i), for $n > \max\{N_p, N_{2p}\}$, $\hat{\mathbf{\Omega}}$ satisfies

$$\|\hat{\mathbf{\Omega}} - \mathbf{\Omega}^\diamond\|_F \le 4\gamma_r\|\hat{\bar{\mathbf{\Omega}}}(\mathbf{\Gamma}) - \bar{\mathbf{\Omega}}(\mathbf{\Gamma})\|_F \le \frac{68\gamma_r C_{0p}}{\beta^2_{p,\min}}\, r_{pn} \quad (47)$$

with probability greater than $1 - \frac{1}{p^{\tau-2}} - 4p^2e^{-KqM}$.
(ii) Let $C_{2p} = 68\gamma_r\sqrt{p}\beta_{p,\max}C_{0p}/\beta^2_{p,\min}$, $U_{1p} = p/(2C_{2p})$, $U_{2p} = 0.1\beta^2_{p,\min}/(68\gamma_r C_{0p}\beta_{p,\max})$, $N_{3p} := \arg\min_n \{n : r_{pn} \le \max\{U_{1p}, U_{2p}\}\}$ and $C_{2q} = 2C_{2p}\|\mathbf{\Gamma}^\diamond\|_F/p$. Let $\hat{\mathbf{\Gamma}}(\hat{\mathbf{\Omega}}) = \arg\min_{\{\mathbf{\Gamma}:\, \mathbf{\Omega}=\hat{\mathbf{\Omega}}\in\mathcal{B}(\mathbf{\Omega}^\diamond)\}} \mathcal{L}_2(\mathbf{\Gamma})$ where $\hat{\mathbf{\Omega}}$ is as in Theorem 3(i). Under the assumptions of Theorem 2, for $n > \max\{N_p, N_{2p}, N_q, N_{3p}\}$ and $\alpha \in [0,1]$, $\hat{\mathbf{\Gamma}}(\hat{\mathbf{\Omega}})$ satisfies

$$\|\hat{\mathbf{\Gamma}}(\hat{\mathbf{\Omega}}) - \mathbf{\Gamma}^\diamond\|_F \le \frac{17C_{0q}}{\beta^2_{q,\min}}\, r_{qn} + C_{2q}r_{pn} \quad (48)$$

with probability greater than $1 - \frac{1}{p^{\tau-2}} - 4p^2e^{-KqM} - \frac{1}{q^{\tau-2}} - 16Mq^2e^{-Kp/2}$ •

VI. Numerical Results

We now present numerical results for both synthetic and real data to illustrate the proposed approach. In synthetic data examples the ground truth is known and this allows for assessment of the efficacy of various approaches. In real data examples where the ground truth is unknown, our goal is visualization and exploration of the linear conditional dependency structures underlying the data.

A. Synthetic Data

We use model (5)-(6) to generate synthetic data where $\mathbf{\Psi}(\tau)$ is controlled via a vector autoregressive (VAR) model impulse response and $\mathbf{\Sigma}$ is determined via an Erdös-Rènyi graph. We take $p = q = 15$. Consider the impulse response $\boldsymbol{H}_i^{(r)} \in \mathbb{R}^{5\times5}$ generated as $\boldsymbol{H}_i^{(r)} = \sum_{k=1}^3 \boldsymbol{A}_k^{(r)}\boldsymbol{H}_{i-k}^{(r)} + \boldsymbol{I}_5\delta_i$, where $\boldsymbol{H}_i^{(r)} = 0$ for $i < 0$, δ_i is the Kronecker delta, $r = 1, 2, 3$, and only 5% of entries of $\boldsymbol{A}_i^{(r)}$'s are nonzero and the nonzero elements are independently and uniformly distributed over $[-0.8, 0.8]$. We then check if the VAR(3) model is stable with all eigenvalues of the companion matrix ≤ 0.95 in magnitude; if not, we re-draw randomly till this condition is fulfilled. The impulse response $\boldsymbol{B}_i \in \mathbb{R}^{15\times15}$ in (5) is given by $\boldsymbol{B}_i = \text{block-diag}\{\boldsymbol{H}_i^{(1)}, \boldsymbol{H}_i^{(2)}, \boldsymbol{H}_i^{(3)}\}$, for $0 \le i \le L = 40$, otherwise it is set to zero. Thus $\boldsymbol{B}_i$'s in (5) have a block-diagonal structure with 3 blocks, each block is 5×5. In the Erdös-Rènyi graph with $p = 15$ nodes, the nodes are connected with probability $p_{er} = 0.05$. In the upper triangular $\bar{\mathbf{\Omega}}$, $\bar{\mathbf{\Omega}}_{ij} = 0$ if $\{i,j\} \notin \mathcal{S}_p$, $\bar{\mathbf{\Omega}}_{ij}$ is uniformly distributed over $[-0.4, -0.1] \cup [0.1, 0.4]$ if $\{i,j\} \in \mathcal{S}_p$, and $\bar{\mathbf{\Omega}}_{ii} = 0.5$. With $\bar{\mathbf{\Omega}} = \bar{\mathbf{\Omega}}^\top$, add $\kappa\boldsymbol{I}_p$ to $\bar{\mathbf{\Omega}}$ with κ picked to make minimum eigenvalue of $\mathbf{\Omega} = \bar{\mathbf{\Omega}} + \kappa\boldsymbol{I}_p$ equal to 0.5. Let $\mathbf{\Omega} = \tilde{\boldsymbol{F}}\tilde{\boldsymbol{F}}$ (matrix square-root), then $\boldsymbol{F} = \tilde{\boldsymbol{F}}^{-1}$ in (5).

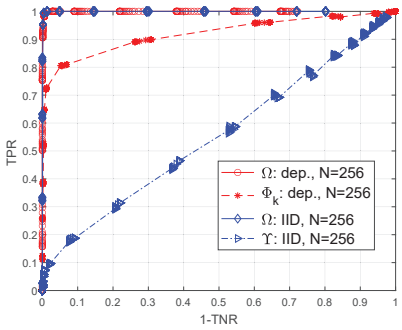


Fig. 1: ROC curves: plots labeled "IID" are from the approach of [17], [18], [25], and the plots labeled "dep." are from our proposed approach. TPR=true positive rate, TNR=true negative rate

We applied our proposed approach with $n = 256$, $M = 2$, $K = 63$ and compared with the approach of [17] (which is also the approach of [18], [25], all of whom assume i.i.d. observations and have two lasso penalties one each on $\mathbf{\Omega}$ and $\mathbf{\Upsilon}$, counterpart to our $\mathbf{\Phi}_k$). In our approach, we fix $\alpha = 0.05$ for all simulations and real data results. For fixed values of λ_q and λ_p, using our proposed approach of Sec. IV-D, we calculated the true positive rate (TPR) and false positive rate 1-TNR (where TNR is the true negative rate) over 100 runs, separately for $\mathbf{\Omega}$ and $\{\mathbf{\Phi}_k\}/\mathbf{\Upsilon}$, based on the estimated edges. As we vary λ_q and λ_p over a wide range of values, we can compute the corresponding pairs of estimated (1-TNR, TPR). The receiver operating characteristic (ROC) is shown in Fig. 1 based on 100 runs, using the estimated (1-TNR, TPR). We repeat this method for the i.i.d. modeling approach of [17], [18], [25]. Fig. 1 shows that the i.i.d. modeling of [17], [18], [25] is unable to capture the "dependent" edges (cf. (4)) via $\mathbf{\Upsilon}$ whereas it has no issues with $\mathbf{\Omega}$. Our approach works well for both components of the Kronecker product graph.

In Table I we show the results based on 100 runs under different parameter settings and samples sizes. Here we show the F_1score, TPR, 1-TNR and timing values for the overall graph (not the two Kronecker product components separately) along with the $\pm\sigma$ errors. All algorithms were run on a Window 10 Pro operating system with processor Intel(R) Core(TM) i7-10700 CPU @2.90 GHz with 32 GB RAM, using MATLAB R2023a. We take $n = 64, 128, 256, 512, 1024, 2048$, and for our proposed approach, the corresponding m_t values leading to different M values are $m_t = 7, 15, 31, 63, 127, 255$ ($M = 2$), $m_t = 4, 9, 20, 41, 84, 169$ ($M = 3$), $m_t = 3, 7, 14, 31, 63, 127$ ($M = 4$), $m_t = 2, 5, 12, 24, 50, 101$ ($M = 5$), $m_t = **, 4, 10, 20, 42, 84$ ($M = 6$), $m_t = **, **, **, 15, 31, 63$ ($M = 8$), and $m_t = **, **, **, 12, 25, 50$ ($M = 10$). Here $**$ denotes that no simulation were performed for the corresponding sample size n (since $K = 2m_t + 1$ is too

This article has been accepted for publication in IEEE Transactions on Signal Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TSP.2024.3395897

TABLE I: *F_1 scores, TPR, 1-TNR and timing per run for fixed tuning parameters, for the synthetic data example, averaged over 100 runs. The entries $**$ denote no simulations done for these parameters.*

n	64	128	256	512	1024	2048
Proposed Approach: F_1 scores $\pm\sigma$ when λ's are selected to minimize BIC						
M=2	0.5163 ± 0.1530	0.5660 ± 0.1580	0.6440 ± 0.1709	0.7061 ± 0.1147	0.7018 ± 0.1176	0.7190 ± 0.1217
M=3	0.5111 ± 0.1705	0.5876 ± 0.1560	0.6969 ± 0.1421	0.7266 ± 0.1159	0.7322 ± 0.1031	0.7474 ± 0.1000
M=4	0.5454 ± 0.1852	0.6470 ± 0.1489	0.7106 ± 0.1465	0.7376 ± 0.1011	0.7446 ± 0.1069	0.7524 ± 0.1047
M=5	0.5977 ± 0.1717	0.6609 ± 0.1465	0.7049 ± 0.1367	0.7253 ± 0.1104	0.7401 ± 0.1028	0.7467 ± 0.1002
M=6	**	0.6277 ± 0.1353	0.6773 ± 0.1379	0.7115 ± 0.1172	0.7343 ± 0.1025	0.7369 ± 0.0971
M=8	**	**	**	0.7016 ± 0.1071	0.7366 ± 0.1013	0.7365 ± 0.0974
M=10	**	**	**	0.7123 ± 0.1117	0.7319 ± 0.1044	0.7367 ± 0.1022
Proposed Approach: F_1 scores $\pm\sigma$ when λ's are selected to maximize F_1 score						
M=2	0.6826 ± 0.1440	0.6954 ± 0.1588	0.7485 ± 0.1632	0.8026 ± 0.1139	0.8032 ± 0.1588	0.8440 ± 0.1184
M=3	0.6984 ± 0.1383	0.7322 ± 0.1730	0.8055 ± 0.1383	0.8293 ± 0.1190	0.8372 ± 0.1442	0.8670 ± 0.1295
M=4	0.7041 ± 0.1355	0.7364 ± 0.1646	0.8074 ± 0.1434	0.8282 ± 0.1169	0.8401 ± 0.1197	0.8633 ± 0.1397
M=5	0.6652 ± 0.1664	0.7309 ± 0.1431	0.8072 ± 0.1466	0.8411 ± 0.1158	0.8451 ± 0.1251	0.8637 ± 0.1314
M=6	**	0.7218 ± 0.1490	0.8089 ± 0.1324	0.8252 ± 0.1206	0.8433 ± 0.1282	0.8583 ± 0.1396
M=8	**	**	**	0.8329 ± 0.1130	0.8382 ± 0.1221	0.8601 ± 0.1404
M=10	**	**	**	0.8187 ± 0.1216	0.8286 ± 0.1525	0.8496 ± 0.1406
IID modeling [17], [18], [25]: λ's are selected to maximize F_1 score						
F_1 scores $\pm\sigma$	0.4329 ± 0.1244	0.4230 ± 0.1208	0.4368 ± 0.1228	0.4746 ± 0.1367	0.4483 ± 0.1180	0.4709 ± 0.1104
timing (s) per run $\pm\sigma$	0.0051 ± 0.0011	0.0073 ± 0.0014	0.0111 ± 0.0020	0.0195 ± 0.0031	0.0342 ± 0.0035	0.0640 ± 0.0050
Proposed approach under model mismatch – non-Gaussian $\boldsymbol{e}$ in (5) : F_1 scores $\pm\sigma$ when λ's are selected to minimize BIC						
Exponential $\boldsymbol{e}$, M=4	0.5518 ± 0.1853	0.6565 ± 0.1728	0.7098 ± 0.1349	0.7355 ± 0.0976	0.7514 ± 0.1141	0.7555 ± 0.0888
Uniform $\boldsymbol{e}$, M=4	0.5434 ± 0.1772	0.6510 ± 0.1693	0.7137 ± 0.1364	0.7400 ± 0.1043	0.7494 ± 0.1146	0.7537 ± 0.0982
Proposed Approach: TPR $\pm\sigma$ when λ's are selected to maximize F_1 score						
M=2	0.6312 ± 0.1675	0.6420 ± 0.1541	0.6937 ± 0.1852	0.7533 ± 0.1332	0.8146 ± 0.1187	0.8249 ± 0.1199
M=4	0.6793 ± 0.1493	0.7120 ± 0.1477	0.7595 ± 0.1529	0.7919 ± 0.1307	0.8142 ± 0.1229	0.8836 ± 0.1021
M=6	**	0.6711 ± 0.1529	0.7459 ± 0.1608	0.8024 ± 0.1287	0.8162 ± 0.1215	0.8275 ± 0.1290
M=10	**	**	**	0.7867 ± 0.1269	0.8278 ± 0.1199	0.8504 ± 0.1177
Proposed Approach: 1-TNR $\pm\sigma$ when λ's are selected to maximize F_1 score						
M=2	0.0032 ± 0.0092	0.0033 ± 0.0090	0.0022 ± 0.0074	0.0018 ± 0.0061	0.0049 ± 0.0157	0.0025 ± 0.0096
M=4	0.0041 ± 0.0118	0.0044 ± 0.0127	0.0020 ± 0.0074	0.0021 ± 0.0097	0.0023 ± 0.0086	0.0043 ± 0.0174
M=6	**	0.0035 ± 0.0116	0.0013 ± 0.0050	0.0026 ± 0.0113	0.0027 ± 0.0120	0.0030 ± 0.0161
M=10	**	**	**	0.0025 ± 0.0113	0.0046 ± 0.0173	0.0040 ± 0.0173
Proposed Approach: timing (s) per run $\pm\sigma$ when λ's are selected to minimize BIC						
M=2	0.1687 ± 0.0400	0.1688 ± 0.0418	0.1791 ± 0.1005	0.1777 ± 0.0289	0.2166 ± 0.0322	0.3131 ± 0.1051
M=4	0.2294 ± 0.1650	0.2470 ± 0.1026	0.2284 ± 0.0846	0.2278 ± 0.0392	0.2890 ± 0.1338	0.3627 ± 0.0494
M=6	**	0.2738 ± 0.0903	0.2426 ± 0.0633	0.2507 ± 0.0537	0.2944 ± 0.0902	0.3842 ± 0.0471
M=10	**	**	**	0.3040 ± 0.1400	0.3230 ± 0.0733	0.4285 ± 0.0890

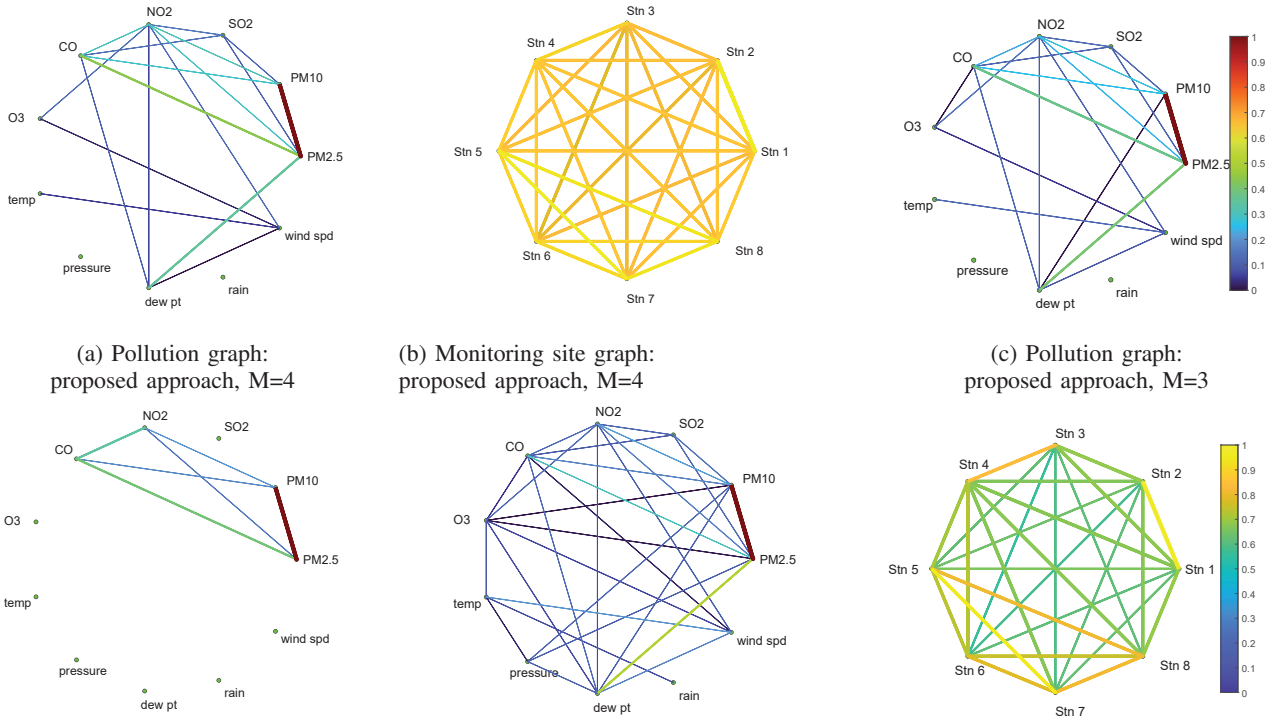
small). We show the resulting F_1 scores under two different scenarios: when we use the proposed BIC parameter selection method (Sec. IV-E) and when F_1 score was selected based on λ values that maximize the F_1 score. While the latter approach is not practical, it is presented to illustrate what is possible using the proposed approach and what may be "lost" when there are errors in the BIC parameter selection method. The number of unknown parameters being estimated are $\mathcal{O}(p^2+Mq^2)$, with $\mathcal{O}(p^2)$ for $\mathbf{\Omega}$ and $\mathcal{O}(Mq^2)$ for M $\mathbf{\Phi}_k$'s. We see that for a fixed n, at first the performance improves with increasing M, then it slowly declines as more parameters need to be estimated with increasing M. Increasing M also reduces $K = 2m_t + 1$ since $KM \approx \frac{n}{2}$, which reduces the number of frequency-domain samples (K) for averaging for the kth model for $\mathbf{\Phi}_k$, $k \in [M]$ (see assumption (A1) in Sec. III). Note also that by (46) of Theorem 2(ii), the error in estimating $\mathbf{\Phi}_k$'s $\propto r_{qn} \propto \sqrt{(Mq)/(Kp)}$. For a fixed M, the performance improves, in general, with increasing n but more slowly for higher n's. Higher n values implies higher resolution in the frequency-domain and for fixed M, higher n implies higher K (and m_t), in which case assumption (A1) in Sec. III may not hold. The TPR, 1-TNR and timing values are shown for selected M's for the proposed approach where timing per run is for the λ values picked by the BIC criterion. It is seen that increasing M and/or n leads to only a small increase in timing.

In Table I we also show the performance of i.i.d. modeling approach of [17], [18], [25], in terms of the F_1 score and timing. The i.i.d. modeling approach is significantly faster but the accuracy in edge detection in terms of the F_1 score is much poorer. Finally, to assess sensitivity to modeling errors such as violation of the Gaussianity assumption, we used either exponential or uniform $\boldsymbol{e}(t)$ in (5), both with zero-mean unit variance, instead of the assumed Gaussian $\boldsymbol{e}(t)$ in our model. The results are shown for $M = 4$ and we see that the performance is robust w.r.t. violation of this assumption.

B. Real Data: Beijing air-quality dataset [50]

Here we consider Beijing air-quality dataset [50], [51], downloaded from https://archive.ics.uci.edu/dataset/501/beijing+multi+site+air+quality+data. This data set includes hourly air pollutants data from 12 nationally-controlled air-quality monitoring sites in the Beijing area. The time period is from March 1st, 2013 to February 28th, 2017. The six air pollutants are $PM_{2.5}$, PM_{10}, SO_2, NO_2, CO, and O_3, and the meteorological data is comprised of five features: temperature, atmospheric pressure, dew point, wind speed, and rain; we did not use wind direction. Thus we have eleven ($= q$) features (pollutants and weather variables). We used data from 8 ($= p$) sites: 4 rural/suburban sites Changping, Dingling, Huairou,

This article has been accepted for publication in IEEE Transactions on Signal Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TSP.2024.3395897



(a) Pollution graph: proposed approach, M=4

(b) Monitoring site graph: proposed approach, M=4

(c) Pollution graph: proposed approach, M=3

(d) Pollution graph: proposed approach, M=5

(e) Pollution graph: i.i.d. modeling approach [17], [18], [25]

(f) Monitoring site graph: i.i.d. modeling approach [17], [18], [25]

Fig. 2: Pollution and site graphs for the Beijing air-quality dataset [50] for year 2013-14: 8 monitoring sites and 11 features ($p = 8$, $q = 11$, $n = 364$). Number of distinct edges $= 18, 28, 20, 6, 30, 28$ in graphs (a), (b), (c), (d), (e) and (f), respectively. Monitoring sites labeled Stn. 1-4 are the rural/suburban sites and those labeled Stn. 5-8 are the urban sites (see the text). For the pollution graph, estimated $\hat{\mathbf{\Phi}}^{(ij)}$ is the edge weight (normalized to have $\max_{i\neq j} \hat{\mathbf{\Phi}}^{(ij)} = 1$) and for the site graph, estimated $|\hat{\Omega}_{ij}|$ is the edge weight (normalized to have $\max_{i\neq j} |\hat{\Omega}_{ij}| = 1$). The edge weights are color coded (all pollution graphs share the same color legend, and similarly for the site graphs), in addition to the edges with higher weights being drawn thicker.

Shunyi, and 4 urban sites Aotizhongxin, Dongsi, Guanyuan, Gucheng (labeled Stn 1 through 8 in Fig. 2). The data are averaged over 24 hour period to yield daily averages. We used one year 2013-14 of daily data resulting in $n = 365$ days. Arranging stations as rows and features as columns, we have $\boldsymbol{Z}(t) \in \mathbb{R}^{8\times 11}$, $t = 1, 2, \cdots, 365$. We pre-processed the data as follows. Given jth feature data $\boldsymbol{Z}_{ij}(t)$ at ith station, we transform it to $\bar{\boldsymbol{Z}}_{ij}(t) = \ln(\boldsymbol{Z}_{ij}(t)/\boldsymbol{Z}_{ij}(t-1))$ for each i and j, and then detrend it (i.e., remove the best straight-line fit). Finally, we scale the detrended scalar sequence to have a mean-square value of one. All temperatures were converted from Celsius to Kelvin to avoid negative numbers. If a value of a feature is zero (e.g., wind speed), we added a small positive number to it so that the log transformation is well-defined.

We applied our proposed approach with $M = 4$, $K = 45$ and $n = 364$ ($p = 8$, $q = 11$) and compared it with the i.i.d. modeling approach of [17], [18], [25]. The objective here is to compare the two approaches in estimation of the pollution (feature) graph and the site graph. The spatio-temporal data has a matrix structure and one is interested in learning two aspects of conditional dependencies: the relationship among the features via the pollution graph and the relationship among the sites via the site graph. We have not yet tested if our model assumptions apply to this dataset (this needs further theoretical analysis to devise suitable statistical tests, particularly in a high-dimensional setting), but it still seems to be useful to compare the results of our proposed approach and that of [17], [18], [25]. Fig. 2(a) shows the resulting graph for the air quality and environmental variables where $\{i, j\} \in \mathcal{S}_q$ iff $\hat{\mathbf{\Phi}}^{(ij)} = (\sum_{k=1}^{M} |[\hat{\mathbf{\Phi}}_k]_{ij}|^2)^{1/2} > 0$ for $i \neq j$, and Fig. 2(b) shows the resulting graph for the sites around the Beijing area where $\{i, j\} \in \mathcal{S}_p$ iff $|\hat{\Omega}_{ij}| > 0$ for $i \neq j$. Since all the sites are physically close to one another, it is not surprising that the site graph in Fig. 2(b) is fully connected. But we do see that the rural/suburban sites stn. 1 through stn. 4 have higher weight edges among the group and the urban sites stn. 5 through stn. 8 have higher weight edges among the urban group, with inter-group edge weights being slightly weaker (but fully connected). Automobile exhaust is the main cause of NO_2 which is likely to undergo a chemical reaction with Ozone O_3, thereby, lowering its concentration [51]. This fact is captured by the edge between NO_2 and Ozone O_3 in Fig. 2(a). Cold, dry air from the north of Beijing reduces both dew point and $PM_{2.5}$ particle concentration in suburban areas while southerly wind brings warmer and more humid air from the more polluted south that elevates both dew point and $PM_{2.5}$ concentration [50]. This fact is captured by the edge between dew point and $PM_{2.5}$ in Fig. 2(a). The counterparts to Figs. 2(a) and 2(b) when using the i.i.d. modeling approach of [17], [18], [25], are shown in Figs. 2(e) and 2(f), respectively. While the site graph in Fig. 2(f) is fully connected and quite similar to the proposed approach's site graph in Fig. 2(b), the pollution graph in Fig. 2(e) far denser than the proposed approach's

pollution graph in Fig. 2(a).

We do not have any systematic approach for selection of M for a given sample size n. Since $KM \approx \frac{n}{2}$, fixing M fixes $K = 2m_t + 1$, and vice-versa. Using BIC to pick M does not work as BIC always picks the smallest M. The synthetic data results presented in Table I show that the performance is not unduly sensitive to the choice of M. To illustrate the sensitivity of the proposed approach in Beijing data case, we show the pollution graphs in Figs. 2(c) and 2(d) for the choice $M = 3$ (K=59) and $M = 5$ ($K = 35$), respectively. There is not much difference between pollution graphs for $M = 4$ and $M = 3$, but that for $M = 5$ is much sparser. This is consistent with the results of Sec. VI-A on synthetic data.

VII. Conclusions

Sparse-group lasso penalized log-likelihood approach in frequency-domain with a Kronecker-decomposable PSD was investigated for matrix CIG learning for dependent time series. An ADMM-based flip-flop approach for iterative optimization of the bi-convex problem was presented. We provided sufficient conditions for consistency of a local estimator of inverse PSD. We illustrated our approach using numerical examples utilizing both synthetic and real data. Lasso and related approaches are known to yield biased estimates [52]. To remedy this, various non-convex penalties have been suggested [52] and typically, lasso-based approaches provide the initial guess for iterative optimization. In the context of this paper, adaptive lasso has been used in [40] (the basis of the ADMM method of Sec. IV-C), and a log-sum penalty has been used in [53] (which modifies [9], the basis for Sec. IV-B). Investigation of such non-convex penalties is left for future research.

Appendix A
Proof of Theorem 1

With fixed $\boldsymbol{\Gamma}$, let $\bar{G}_1(\boldsymbol{\Omega})$ denote $\bar{G}(\boldsymbol{\Omega}, \{\boldsymbol{\Phi}\}, \{\boldsymbol{\Phi}^*\})$ up to some irrelevant constants. Then

$$\bar{G}_1(\boldsymbol{\Omega}) = -\frac{1}{p}\ln(|\boldsymbol{\Omega}|) + B\,\mathrm{tr}(\boldsymbol{\Sigma}^\diamond\boldsymbol{\Omega})\,, \tag{49}$$

where $B = \mathrm{tr}(\bar{\boldsymbol{S}}_k^\diamond\boldsymbol{\Phi}_k)^*/(2Mqp)$. We have

$$\mathbf{0} = \frac{\partial\bar{G}_1(\boldsymbol{\Omega})}{\partial\boldsymbol{\Omega}} = -\frac{1}{p}\boldsymbol{\Omega}^{-1} + B\,\boldsymbol{\Sigma}^\diamond\,, \tag{50}$$

establishing (33) if $B \neq 0$. The solution is unique since the Hessian of $\bar{G}_1(\boldsymbol{\Omega})$, given by $\frac{1}{p}\boldsymbol{\Omega}^{-1}\otimes\boldsymbol{\Omega}^{-1}$, is positive definite at $\boldsymbol{\Omega} = \bar{\boldsymbol{\Omega}}(\boldsymbol{\Gamma})$. Similarly, with fixed $\boldsymbol{\Omega}$, let $\bar{G}_2(\boldsymbol{\Gamma})$ denote $\bar{G}(\boldsymbol{\Omega}, \{\boldsymbol{\Phi}\}, \{\boldsymbol{\Phi}^*\})$ up to some irrelevant constants. Then

$$\bar{G}_2(\boldsymbol{\Gamma}) = \sum_{k=1}^{M}\bar{G}_{2k}(\boldsymbol{\Phi}_k)\,, \tag{51}$$

$$\begin{aligned}\bar{G}_{2k}(\boldsymbol{\Phi}_k) =& -\ln(|\boldsymbol{\Phi}_k|) - \ln(|\boldsymbol{\Phi}_k^*|)\\ &+ \frac{1}{p}\big(\mathrm{tr}(\bar{\boldsymbol{S}}_k^\diamond\boldsymbol{\Phi}_k) + \mathrm{tr}(\bar{\boldsymbol{S}}_k^\diamond\boldsymbol{\Phi}_k)^*\big)\mathrm{tr}(\boldsymbol{\Sigma}^\diamond\boldsymbol{\Omega})\,.\end{aligned} \tag{52}$$

The cost $\bar{G}_2(\boldsymbol{\Gamma})$ is separable in k, $\boldsymbol{\Phi}_k$. We have

$$\mathbf{0} = \frac{\partial\bar{G}_{2k}(\boldsymbol{\Phi}_k)}{\partial\boldsymbol{\Phi}_k^*} = -\boldsymbol{\Phi}_k^{-1} + \frac{1}{p}\bar{\boldsymbol{S}}_k^\diamond\mathrm{tr}(\boldsymbol{\Sigma}^\diamond\boldsymbol{\Omega})\,, \tag{53}$$

establishing (34) if $\mathrm{tr}(\boldsymbol{\Sigma}^\diamond\boldsymbol{\Omega}) \neq 0$. Similar to [9, Lemma 4], the Hessian of $\bar{G}_{2k}(\boldsymbol{\Phi}_k)$ is positive definite at $\boldsymbol{\Phi}_k = \bar{\boldsymbol{\Phi}}_k(\boldsymbol{\Omega})$. Therefore, the solution is unique. □

Appendix B
Technical Lemmas and Proof of Theorem 2

Lemma 1 is a restatement of [22, Lemma S.1, Supplementary].

Lemma 1. Assume that i.i.d. data $\boldsymbol{X}_i \in \mathbb{R}^{p\times q}$, $i = 1, 2, \cdots, n$, follows the matrix-valued normal distribution $\mathcal{MVN}(\mathbf{0}, \boldsymbol{\Sigma}^\diamond, \boldsymbol{\Psi}^\diamond)$, with $\boldsymbol{\Sigma}^\diamond \in \mathbb{R}^{p\times p}$, $\boldsymbol{\Psi}^\diamond \in \mathbb{R}^{q\times q}$, $\boldsymbol{\Sigma}^\diamond \succ \mathbf{0}$ and $\boldsymbol{\Psi}^\diamond \succ \mathbf{0}$, i.e., $\mathrm{vec}(\boldsymbol{X}_i) \sim \mathcal{N}_r\big(\mathbf{0}, \boldsymbol{\Psi}^\diamond\otimes\boldsymbol{\Sigma}^\diamond\big)$. Assume that $\phi_{max}(\boldsymbol{\Sigma}^\diamond) \leq C_{1h} < \infty$ and $\phi_{max}(\boldsymbol{\Psi}^\diamond) \leq C_{2h} < \infty$ for some positive constants C_{1h} and C_{2h}. For any symmetric positive-definite $\boldsymbol{\Omega} \in \mathbb{R}^{p\times p}$ such that $\|\boldsymbol{\Omega} - \boldsymbol{\Omega}^\diamond\|_F \leq \gamma$, $\boldsymbol{\Omega}^\diamond = (\boldsymbol{\Sigma}^\diamond)^{-1}$, we have

$$\begin{aligned}P\Big(\max_{i,j}\Big|\big[\frac{1}{np}\sum_{i=1}^{n}\boldsymbol{X}_i^\top\boldsymbol{\Omega}\boldsymbol{X}_i - \frac{1}{p}E\{\boldsymbol{X}_i^\top\boldsymbol{\Omega}\boldsymbol{X}_i\}\big]_{ij}\Big| \geq \delta\Big)\\ \leq 4q^2\Big[\exp\big\{-\frac{np}{2}\big[\frac{\delta}{8(1+\gamma C_{1h})C_{2h}} - \frac{2}{\sqrt{np}}\big]^2\big\}\\ + \exp\big\{-\frac{np}{2}\big\}\Big]\end{aligned} \tag{54}$$

for any $\delta > 16(1+\gamma C_{1h})C_{2h}/\sqrt{np}$ •

The lower bound on δ follows from [22, Lemma S.12, Supplementary] and (54) is [22, Eqn. (S.26), Supplementary] in our notation.

Lemma 2 collects some useful results from [42, Theorem 2.3.5].

Lemma 2. Suppose $\boldsymbol{X} \sim \mathcal{MVN}(\mathbf{0}, \boldsymbol{\Sigma}, \boldsymbol{\Psi})$ where $\boldsymbol{X} \in \mathbb{R}^{p\times q}$, $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$, $\boldsymbol{\Psi} \in \mathbb{R}^{q\times q}$, i.e., $\mathrm{vec}(\boldsymbol{X}) \sim \mathcal{N}_r\big(\mathbf{0}, \boldsymbol{\Psi}\otimes\boldsymbol{\Sigma}\big)$. Then

(i) $\boldsymbol{X}^\top \sim \mathcal{MVN}(\mathbf{0}, \boldsymbol{\Psi}, \boldsymbol{\Sigma})$, i.e., $\mathrm{vec}(\boldsymbol{X}^\top) \sim \mathcal{N}_r\big(\mathbf{0}, \boldsymbol{\Sigma}\otimes\boldsymbol{\Psi}\big)$.
(ii) For any $\boldsymbol{A} \in \mathbb{R}^{q\times q}$, $E\{\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^\top\} = \mathrm{tr}(\boldsymbol{A}^\top\boldsymbol{\Psi})\,\boldsymbol{\Sigma}$.
(iii) For any $\boldsymbol{B} \in \mathbb{R}^{p\times p}$, $E\{\boldsymbol{X}^\top\boldsymbol{B}\boldsymbol{X}\} = \mathrm{tr}(\boldsymbol{B}^\top\boldsymbol{\Sigma})\,\boldsymbol{\Psi}$.
(iv) For any $\boldsymbol{C} \in \mathbb{R}^{q\times p}$, $E\{\boldsymbol{X}\boldsymbol{C}\boldsymbol{X}\} = \boldsymbol{\Sigma}\boldsymbol{C}^\top\boldsymbol{\Psi}$ •

Lemma 3. Suppose $\boldsymbol{X} \in \mathbb{C}^{p\times q}$, $\mathrm{vec}(\boldsymbol{X}) \sim \mathcal{N}_c\big(\mathbf{0}, \boldsymbol{S}\otimes\boldsymbol{\Sigma}\big)$ where $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$, $\boldsymbol{S} \in \mathbb{C}^{q\times q}$, $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^\top \succ \mathbf{0}$, $\boldsymbol{S} = \boldsymbol{S}^H \succ \mathbf{0}$ and $\boldsymbol{S} = \boldsymbol{S}_r + j\boldsymbol{S}_i$ with $\boldsymbol{S}_r, \boldsymbol{S}_i \in \mathbb{R}^{q\times q}$.

(i) Let $\boldsymbol{X} = \boldsymbol{X}_r + j\boldsymbol{X}_i$, $\boldsymbol{X}_r, \boldsymbol{X}_i \in \mathbb{R}^{p\times q}$. Then

$$\tilde{\boldsymbol{X}} = [\boldsymbol{X}_r\ \boldsymbol{X}_i] \sim \mathcal{MVN}(\mathbf{0}, \boldsymbol{\Sigma}, \tilde{\boldsymbol{S}}) \tag{55}$$

i.e., $\mathrm{vec}(\tilde{\boldsymbol{X}}) \sim \mathcal{N}_r\big(\mathbf{0}, \tilde{\boldsymbol{S}}\otimes\boldsymbol{\Sigma}\big)$, where

$$\tilde{\boldsymbol{S}} = \frac{1}{2}\begin{bmatrix}\boldsymbol{S}_r & -\boldsymbol{S}_i\\ \boldsymbol{S}_i & \boldsymbol{S}_r\end{bmatrix} \in \mathbb{R}^{2q\times 2q}\,. \tag{56}$$

(ii) For any $\boldsymbol{\Omega} \in \mathbb{R}^{p\times p}$, $E\{\tilde{\boldsymbol{X}}^\top\boldsymbol{\Omega}\tilde{\boldsymbol{X}}\} = \mathrm{tr}(\boldsymbol{\Omega}^\top\boldsymbol{\Sigma})\,\tilde{\boldsymbol{S}}$.
(iii) For any $\boldsymbol{\Phi} \in \mathbb{C}^{q\times q}$, $\boldsymbol{\Phi} = \boldsymbol{\Phi}_r + j\boldsymbol{\Phi}_i = \boldsymbol{\Phi}^H$, $\boldsymbol{\Phi}_r, \boldsymbol{\Phi}_i \in \mathbb{R}^{q\times q}$,

$$E\{\mathrm{Re}\big(\boldsymbol{X}\boldsymbol{\Phi}^*\boldsymbol{X}^H\big)\} = E\{\tilde{\boldsymbol{X}}\tilde{\boldsymbol{\Phi}}\tilde{\boldsymbol{X}}^\top\big)\} = \mathrm{tr}(\tilde{\boldsymbol{\Phi}}^\top\tilde{\boldsymbol{S}})\,\boldsymbol{\Sigma} \tag{57}$$

where

$$\tilde{\boldsymbol{\Phi}} = \begin{bmatrix}\boldsymbol{\Phi}_r & -\boldsymbol{\Phi}_i\\ \boldsymbol{\Phi}_i & \boldsymbol{\Phi}_r\end{bmatrix} \in \mathbb{R}^{2q\times 2q} \quad \bullet \tag{58}$$

Proof.

(i) If $\boldsymbol{x} = \mathrm{vec}(\boldsymbol{X}) \sim \mathcal{N}_c\big(\mathbf{0}, \boldsymbol{S}\otimes\boldsymbol{\Sigma}\big)$, then by [38, Sec. 2.3],

$$\tilde{\boldsymbol{x}} = \mathrm{vec}(\tilde{\boldsymbol{X}}) \sim \mathcal{N}_r\big(\mathbf{0}, \boldsymbol{R}\big) \tag{59}$$

where, with $\boldsymbol{x} = \boldsymbol{x}_r + j\boldsymbol{x}_i$, $\boldsymbol{x}_r, \boldsymbol{x}_i \in \mathbb{R}^{pq}$,

$$\boldsymbol{R} = = \begin{bmatrix} E\{\boldsymbol{x}_r\boldsymbol{x}_r^\top\} & E\{\boldsymbol{x}_i\boldsymbol{x}_r^\top\} \\ E\{\boldsymbol{x}_r\boldsymbol{x}_i^\top\} & E\{\boldsymbol{x}_i\boldsymbol{x}_i^\top\} \end{bmatrix} = \begin{bmatrix} \boldsymbol{R}_{rr} & \boldsymbol{R}_{ir} \\ \boldsymbol{R}_{ri} & \boldsymbol{R}_{ii} \end{bmatrix} \tag{60}$$

$$\boldsymbol{R}_{rr} = \boldsymbol{R}_{ii}\,, \quad \boldsymbol{R}_{ri} = -\boldsymbol{R}_{ri}^\top = \boldsymbol{R}_{ir}^\top\,. \tag{61}$$

Now $\boldsymbol{R}_{rr} = \frac{1}{2}\boldsymbol{S}_r \otimes \boldsymbol{\Sigma} = \boldsymbol{R}_{ii}$ and $\boldsymbol{R}_{ri} = -\frac{1}{2}\boldsymbol{S}_i^\top \otimes \boldsymbol{\Sigma}$. Therefore, $\boldsymbol{R} = \tilde{\boldsymbol{S}} \otimes \boldsymbol{\Sigma}$, yielding the desired result.

(ii) It follows from Lemma 2(iii) and Lemma 3(i).

(iii) Since $\boldsymbol{\Phi} = \boldsymbol{\Phi}^H$, it follows that $\boldsymbol{\Phi}_r = \boldsymbol{\Phi}_r^\top$ and $\boldsymbol{\Phi}_i = -\boldsymbol{\Phi}_i^\top$. We have $\mathrm{Re}\big(\boldsymbol{X}\boldsymbol{\Phi}^*\boldsymbol{X}^H\big) = \tilde{\boldsymbol{X}}\tilde{\boldsymbol{\Phi}}\tilde{\boldsymbol{X}}^\top$. Then the given expression for $E\{\tilde{\boldsymbol{X}}\tilde{\boldsymbol{\Phi}}\tilde{\boldsymbol{X}}^\top)\}$ follows from Lemma 2(ii). □

We now consider a tail bound on $\tilde{\boldsymbol{\Theta}}_k$ defined in (27). First we need Lemma 4.

Lemma 4. Given $\boldsymbol{S} \in \mathbb{C}^{q\times q}$ and $\tilde{\boldsymbol{S}} \in \mathbb{R}^{2q\times 2q}$ as in Lemma 3. Then $\tilde{\boldsymbol{S}} \succ \boldsymbol{0}$ and $\phi_{max}(\tilde{\boldsymbol{S}}) = \frac{1}{2}\phi_{max}(\boldsymbol{S})$.

Proof. If λ is a an eigenvalue of $\boldsymbol{S}$, then for some $\boldsymbol{v} = \boldsymbol{v}_r + j\boldsymbol{v}_i \in \mathbb{C}^q$, $\boldsymbol{v}_r, \boldsymbol{v}_i \in \mathbb{R}^q$, we have $\boldsymbol{S}\boldsymbol{v} = \lambda\boldsymbol{v}$, where λ is real positive since $\boldsymbol{S}$ is Hermitian, positive-definite. It then follows that

$$\tilde{\boldsymbol{S}}\begin{bmatrix}\boldsymbol{v}_r \\ \boldsymbol{v}_i\end{bmatrix} = \frac{1}{2}\lambda\begin{bmatrix}\boldsymbol{v}_r \\ \boldsymbol{v}_i\end{bmatrix}\,, \quad \tilde{\boldsymbol{S}}\begin{bmatrix}-\boldsymbol{v}_i \\ \boldsymbol{v}_r\end{bmatrix} = \frac{1}{2}\lambda\begin{bmatrix}-\boldsymbol{v}_i \\ \boldsymbol{v}_r\end{bmatrix}\,. \tag{62}$$

That is, each eigenvalue of $\boldsymbol{S}$ is also an eigenvalue of $2\tilde{\boldsymbol{S}}$ with multiplicity two. This proves the desired result. □

Lemma 5. Under assumptions (A1) and (A2), for any symmetric positive-definite $\hat{\boldsymbol{\Omega}} \in \mathbb{R}^{p\times p}$ such that $\|\hat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}^\diamond\|_F \le \gamma_p$, $\boldsymbol{\Omega}^\diamond = (\boldsymbol{\Sigma}^\diamond)^{-1}$, and $\tau > 2$, we have

$$\begin{aligned} P\Big(\max_{k,i,j}\big|\big[\tilde{\boldsymbol{\Theta}}_k^* - E\{\tilde{\boldsymbol{\Theta}}_k^*\}\big]_{ij}\big| \ge C_{0q}\sqrt{\frac{\ln(M^{1/\tau}q)}{Kp}}\,\Big) \\ \le \frac{1}{q^{\tau-2}} + 16Mq^2e^{-Kp/2} \end{aligned} \tag{63}$$

for any $q \ge 1$, where C_{0q} is given by (37) and

$$E\{\tilde{\boldsymbol{\Theta}}_k^*\} = \frac{1}{p}\,\mathrm{tr}\big(\hat{\boldsymbol{\Omega}}\boldsymbol{\Sigma}^\diamond\big)\,(\bar{\boldsymbol{S}}_k^\diamond)^*\,. \tag{64}$$

Proof. Let $\boldsymbol{D}_z(\tilde{f}_{k,\ell}) = \boldsymbol{D}_{r,kl} + j\boldsymbol{D}_{i,kl}$, $\boldsymbol{D}_{r,kl}, \boldsymbol{D}_{i,kl} \in \mathbb{R}^{p\times q}$. Define

$$\boldsymbol{X}_{kl} = \big[\boldsymbol{D}_{r,kl} \quad \boldsymbol{D}_{i,kl}\big] \in \mathbb{R}^{p\times(2q)}\,, \tag{65}$$

$$\boldsymbol{B}_{kl} = \boldsymbol{X}_{kl}^\top\hat{\boldsymbol{\Omega}}\boldsymbol{X}_{kl}\,, \quad \boldsymbol{F}_k = \frac{1}{Kp}\sum_{\ell=-m_t}^{m_t}\boldsymbol{B}_{kl}\,. \tag{66}$$

Since

$$\begin{aligned} \boldsymbol{D}_z^H(\tilde{f}_{k,\ell})\hat{\boldsymbol{\Omega}}\boldsymbol{D}_z(\tilde{f}_{k,\ell}) = \boldsymbol{D}_{r,kl}^\top\hat{\boldsymbol{\Omega}}\boldsymbol{D}_{r,kl} + \boldsymbol{D}_{i,kl}^\top\hat{\boldsymbol{\Omega}}\boldsymbol{D}_{i,kl} \\ + j\big[\boldsymbol{D}_{r,kl}^\top\hat{\boldsymbol{\Omega}}\boldsymbol{D}_{i,kl} - \boldsymbol{D}_{i,kl}^\top\hat{\boldsymbol{\Omega}}\boldsymbol{D}_{r,kl}\big]\,, \end{aligned} \tag{67}$$

it follows that

$$\max_{k,i,j}\big|\big[\tilde{\boldsymbol{\Theta}}_k^* - E\{\tilde{\boldsymbol{\Theta}}_k^*\}\big]_{ij}\big| \le 4\max_{k,i,j}\big|\big[\boldsymbol{F}_k - E\{\boldsymbol{F}_k\}\big]_{ij}\big|\,. \tag{68}$$

Therefore,

$$\begin{aligned} \Big\{\max_{k,i,j}\big|\big[\boldsymbol{F}_k - E\{\boldsymbol{F}_k\}\big]_{ij}\big| < \frac{\delta}{4}\Big\} \\ \subseteq \Big\{\max_{k,i,j}\big|\big[\tilde{\boldsymbol{\Theta}}_k^* - E\{\tilde{\boldsymbol{\Theta}}_k^*\}\big]_{ij}\big| < \delta\Big\}\,, \end{aligned} \tag{69}$$

implying

$$\begin{aligned} P\Big(\max_{k,i,j}\big|\big[\tilde{\boldsymbol{\Theta}}_k^* - E\{\tilde{\boldsymbol{\Theta}}_k^*\}\big]_{ij}\big| \ge \delta\Big) \\ \le P\Big(\max_{k,i,j}\big|\big[\boldsymbol{F}_k - E\{\boldsymbol{F}_k\}\big]_{ij}\big| \ge \frac{\delta}{4}\Big)\,. \end{aligned} \tag{70}$$

Since $\boldsymbol{d}_z(\tilde{f}_{k,\ell}) = \mathrm{vec}(\boldsymbol{D}_z(\tilde{f}_{k,\ell})) \sim \mathcal{N}_c(\boldsymbol{0}, \bar{\boldsymbol{S}}^\diamond(\tilde{f}_k) \otimes \boldsymbol{\Sigma}^\diamond)$, it follows from Lemma 3(i) that

$$\boldsymbol{X}_{kl} \sim \mathcal{MVN}(\boldsymbol{0}, \boldsymbol{\Sigma}^\diamond, \tilde{\boldsymbol{S}}_k^\diamond)\,, \tag{71}$$

$$\tilde{\boldsymbol{S}}_k^\diamond = \frac{1}{2}\begin{bmatrix}\bar{\boldsymbol{S}}_{rk}^\diamond & -\bar{\boldsymbol{S}}_{ik}^\diamond \\ \bar{\boldsymbol{S}}_{ik}^\diamond & \bar{\boldsymbol{S}}_{rk}^\diamond\end{bmatrix}\,, \quad \bar{\boldsymbol{S}}^\diamond(\tilde{f}_k) = \bar{\boldsymbol{S}}_{rk}^\diamond + j\bar{\boldsymbol{S}}_{ik}^\diamond\,. \tag{72}$$

By assumption (A4), $\phi_{\max}(\boldsymbol{\Sigma}^\diamond) \le \beta_{p,\max}$ and additionally, by Lemma 4, $\phi_{\max}(\tilde{\boldsymbol{S}}_k^\diamond) \le \beta_{q,\max}/2$ for every k. With $a = 4(1+\gamma_p\beta_{p,\max})\beta_{q,\max}$, invoking Lemma 1 for the sum $\boldsymbol{F}_k$, we have

$$\begin{aligned} &P\Big(\max_{i,j}\big|\big[\boldsymbol{F}_k - E\{\boldsymbol{F}_k\}\big]_{ij}\big| \ge \frac{\delta}{4}\Big) \\ &\le 4(2q)^2\Big[\exp\Big\{-\frac{Kp}{2}\big[\frac{\delta/4}{a} - \frac{2}{\sqrt{Kp}}\big]^2\Big\} + e^{-Kp/2}\Big] = P_{qtb}\,. \end{aligned} \tag{73}$$

Maximizing over all $k = 1, 2, \cdots, M$, and using the union bound, we obtain

$$P\Big(\max_{k,i,j}\big|\big[\boldsymbol{F}_k - E\{\boldsymbol{F}_k\}\big]_{ij}\big| \ge \frac{\delta}{4}\Big) \le MP_{qtb}\,. \tag{74}$$

For $\tau > 2$, pick $\delta = 4a(\sqrt{2\ln(16Mq^\tau)/(Kp)} + 2/\sqrt{Kp}\,)$, leading to $\delta = C_{0q}\sqrt{\ln(M^{1/\tau}q)/(Kp)}$ and $(Kp/2)[(\delta/(4a)) - 2/\sqrt{Kp}]^2 = \ln(16M\,q^\tau)$. Thus

$$\begin{aligned} MP_{qtb} &= 16Mq^2\Big[e^{-\ln(16Mq^\tau)} + e^{-Kp/2}\Big] \\ &= \frac{1}{q^{\tau-2}} + 16Mq^2e^{-Kp/2}\,. \end{aligned} \tag{75}$$

Thus we have established (63). The lower bound on $\delta/4$ specified in Lemma 1 is satisfied if $(\delta/(4a) > 2/\sqrt{Kp}$, which is true for any $q \ge 1$. Turning to (64), by (66), (71) and Lemma 2(iii), we have

$$E\{\boldsymbol{B}_{kl}\} = \mathrm{tr}(\hat{\boldsymbol{\Omega}}\boldsymbol{\Sigma}^\diamond)\frac{1}{2}\begin{bmatrix}\bar{\boldsymbol{S}}_{rk}^\diamond & -\bar{\boldsymbol{S}}_{ik}^\diamond \\ \bar{\boldsymbol{S}}_{ik}^\diamond & \bar{\boldsymbol{S}}_{rk}^\diamond\end{bmatrix}\,. \tag{76}$$

By assumption (A1), (66), (67) and (76),

$$\begin{aligned} E\{\boldsymbol{D}_z^H(\tilde{f}_{k,\ell})\hat{\boldsymbol{\Omega}}\boldsymbol{D}_z(\tilde{f}_{k,\ell})\} &= \frac{1}{2}\mathrm{tr}(\hat{\boldsymbol{\Omega}}\boldsymbol{\Sigma}^\diamond)\Big(2\bar{\boldsymbol{S}}_{rk}^\diamond - j2\bar{\boldsymbol{S}}_{ik}^\diamond\Big) \\ &= \mathrm{tr}(\hat{\boldsymbol{\Omega}}\boldsymbol{\Sigma}^\diamond)(\bar{\boldsymbol{S}}_k^\diamond)^*\,. \end{aligned} \tag{77}$$

By (27) and (77), we obtain (64). □

Now we consider a tail bound on $\check{\boldsymbol{\Theta}}$ defined in (25).

Lemma 6. Under assumptions (A1) and (A2), for any Hermitian positive-definite $\hat{\boldsymbol{\Phi}}_k \in \mathbb{C}^{q\times q}$, $k = 1, 2, \cdots, M$, such that $\|\hat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}^\diamond\|_F \le \gamma_q$, $\boldsymbol{\Gamma}^\diamond = [\boldsymbol{\Phi}_1^\diamond, \cdots, \boldsymbol{\Phi}_M^\diamond]$, $\hat{\boldsymbol{\Gamma}} = [\hat{\boldsymbol{\Phi}}_1, \cdots, \hat{\boldsymbol{\Phi}}_M]$, and $\tau > 2$, we have

$$\begin{aligned} P\Big(\max_{i,j}\big|\big[\check{\boldsymbol{\Theta}} - E\{\check{\boldsymbol{\Theta}}\}\big]_{ij}\big| \ge C_{0p}\sqrt{\frac{\ln(p)}{KqM}}\,\Big) \\ \le \frac{1}{p^{\tau-2}} + 4p^2e^{-KqM} \end{aligned} \tag{78}$$

for any $p \geq 1$ where, where C_{0p} is given by (40), and

$$E\{\check{\boldsymbol{\Theta}}\} = \Big[\frac{1}{2Mq}\sum_{k=1}^{M} \mathrm{tr}\big(\bar{\boldsymbol{S}}_k^{\diamond}\hat{\boldsymbol{\Phi}}_k + (\bar{\boldsymbol{S}}_k^{\diamond}\hat{\boldsymbol{\Phi}}_k)^*\big)\Big]\boldsymbol{\Sigma}^{\diamond}\,. \tag{79}$$

Proof. We have

$$\mathrm{Re}\big(\boldsymbol{D}_z(\tilde{f}_{k,\ell})\hat{\boldsymbol{\Phi}}_k\boldsymbol{D}_z^H(\tilde{f}_{k,\ell})\big) = \boldsymbol{X}_{kl}\tilde{\boldsymbol{\Phi}}_k\boldsymbol{X}_{kl}^{\top}\,, \tag{80}$$

where $\boldsymbol{X}_{kl}$ is as in (65) and

$$\tilde{\boldsymbol{\Phi}}_k = \begin{bmatrix}\hat{\boldsymbol{\Phi}}_{rk} & -\hat{\boldsymbol{\Phi}}_{ik}\\ \hat{\boldsymbol{\Phi}}_{ik} & \hat{\boldsymbol{\Phi}}_{rk}\end{bmatrix} \in \mathbb{R}^{2q\times 2q}\,. \tag{81}$$

Define

$$\check{\boldsymbol{\Phi}} = \begin{bmatrix}\tilde{\boldsymbol{\Phi}}_1 & \boldsymbol{0} & \cdots & \boldsymbol{0}\\ \boldsymbol{0} & \tilde{\boldsymbol{\Phi}}_2 & \cdots & \boldsymbol{0}\\ \vdots & \vdots & \ddots & \vdots\\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \tilde{\boldsymbol{\Phi}}_M\end{bmatrix} \in \mathbb{R}^{(2qM)\times(2qM)}\,, \tag{82}$$

$$\check{\boldsymbol{X}}_\ell = \begin{bmatrix}\boldsymbol{X}_{1l} & \boldsymbol{X}_{2l} & \cdots & \boldsymbol{X}_{Ml}\end{bmatrix}^{\top}\,. \tag{83}$$

Then we can express $\check{\boldsymbol{\Theta}}$ as

$$\check{\boldsymbol{\Theta}} = \frac{1}{MKq}\sum_{\ell=-m_t}^{m_t}\check{\boldsymbol{X}}_\ell^{\top}\check{\boldsymbol{\Phi}}\check{\boldsymbol{X}}_\ell\,. \tag{84}$$

Since $\boldsymbol{X}_{kl} \sim \mathcal{MVN}(\boldsymbol{0}, \boldsymbol{\Sigma}^{\diamond}, \tilde{\boldsymbol{S}}_k^{\diamond})$, and $\boldsymbol{X}_{k_1 l}$ and $\boldsymbol{X}_{k_2 l}$ are independent for $k_1 \neq k_2$, we have

$$\check{\boldsymbol{X}}_\ell^{\top} \sim \mathcal{MVN}(\boldsymbol{0}, \boldsymbol{\Sigma}^{\diamond}, \check{\boldsymbol{S}}^{\diamond})\,,\ \check{\boldsymbol{X}}_\ell \sim \mathcal{MVN}(\boldsymbol{0}, \check{\boldsymbol{S}}^{\diamond}, \boldsymbol{\Sigma}^{\diamond})\,, \tag{85}$$

where

$$\check{\boldsymbol{S}}^{\diamond} = \begin{bmatrix}\tilde{\boldsymbol{S}}_1^{\diamond} & \boldsymbol{0} & \cdots & \boldsymbol{0}\\ \boldsymbol{0} & \tilde{\boldsymbol{S}}_2^{\diamond} & \cdots & \boldsymbol{0}\\ \vdots & \vdots & \ddots & \vdots\\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \tilde{\boldsymbol{S}}_M^{\diamond}\end{bmatrix} \in \mathbb{R}^{(2qM)\times(2qM)}\,. \tag{86}$$

By assumption (A4), $\phi_{\max}(\boldsymbol{\Sigma}^{\diamond}) \leq \beta_{p,\max}$ and additionally, by Lemma 4, $\phi_{\max}(\check{\boldsymbol{S}}^{\diamond}) \leq \beta_{q,\max}/2$. With $b = 8(1 + \gamma_q\beta_{q,\max}/2)\beta_{p,\max}$, apply Lemma 1 to the sum $\frac{1}{2}\check{\boldsymbol{\Theta}}$ to obtain

$$\begin{aligned} &P\Big(\max_{i,j}\big|\frac{1}{2}\big[\check{\boldsymbol{\Theta}} - E\{\check{\boldsymbol{\Theta}}\}\big]_{ij}\big| \geq \delta\Big)\\ &\leq 4p^2\Big[\exp\Big\{-\frac{2qMK}{2}\Big[\frac{\delta}{b} - \frac{2}{\sqrt{2qMK}}\Big]^2\Big\} + e^{-2qMK/2}\Big]\\ &= P_{ptb}\,. \end{aligned} \tag{87}$$

For $\tau > 2$, pick $\delta = b(\sqrt{\ln(4p^\tau)/(KqM)} + \sqrt{2/(KqM)}\,)$, leading to $(2qMK/2)[(\delta/b) - \sqrt{2/(KqM)}]^2 = \ln(4p^\tau)$. Thus

$$P_{ptb} = 4p^2\Big[e^{-\ln(4p^\tau)} + e^{-qMK}\Big] = \frac{1}{p^{\tau-2}} + 4p^2e^{-qMK}\,. \tag{88}$$

The lower bound on δ specified in Lemma 1 is satisfied if $(\delta/b) > \sqrt{2/(KqM)}$, which is true for any $p \geq 1$. With our choice of δ, we have $2\delta = C_{0p}\sqrt{\frac{\ln(p)}{KqM}}$, establishing (78). Turning to (79), by (85) and Lemma 2(iii), we have

$$E\{\check{\boldsymbol{X}}_\ell^{\top}\check{\boldsymbol{\Phi}}\check{\boldsymbol{X}}_\ell\} = \mathrm{tr}\big(\check{\boldsymbol{\Phi}}^{\top}\check{\boldsymbol{S}}^{\diamond}\big)\boldsymbol{\Sigma}^{\diamond} = \mathrm{tr}\big(\sum_{k=1}^{M}\tilde{\boldsymbol{\Phi}}_k^{\top}\tilde{\boldsymbol{S}}_k^{\diamond}\big)\boldsymbol{\Sigma}^{\diamond}\,. \tag{89}$$

By (72) and (81),

$$\begin{aligned}\mathrm{tr}\big(\tilde{\boldsymbol{\Phi}}_k^{\top}\tilde{\boldsymbol{S}}_k^{\diamond}\big) &= \mathrm{tr}\big(\hat{\boldsymbol{\Phi}}_{rk}\bar{\boldsymbol{S}}_{rk}^{\diamond} + \hat{\boldsymbol{\Phi}}_{ik}\bar{\boldsymbol{S}}_{ik}^{\diamond}\big)\\ &= \frac{1}{2}\mathrm{tr}\big(\bar{\boldsymbol{S}}_k^{\diamond}\hat{\boldsymbol{\Phi}}_k + (\bar{\boldsymbol{S}}_k^{\diamond}\hat{\boldsymbol{\Phi}}_k)^*\big)\,.\end{aligned} \tag{90}$$

Using (84), (89) and (90), we have (79). □

Proof of Theorem 2(i). Let $\boldsymbol{\Omega} = \bar{\boldsymbol{\Omega}}(\boldsymbol{\Gamma}) + \boldsymbol{\Delta}$ with $\boldsymbol{\Omega}, \bar{\boldsymbol{\Omega}}(\boldsymbol{\Gamma}) \succ \boldsymbol{0}$, and denote $Q(\boldsymbol{\Omega}) = L_1(\boldsymbol{\Omega}) - L_1(\bar{\boldsymbol{\Omega}}(\boldsymbol{\Gamma}))$. For the rest of the proof, we will denote $\bar{\boldsymbol{\Omega}}(\boldsymbol{\Gamma})$ by $\bar{\boldsymbol{\Omega}}$. Then $\hat{\boldsymbol{\Omega}}(\boldsymbol{\Gamma})$ minimizes $Q(\boldsymbol{\Omega})$, or equivalently, $\hat{\boldsymbol{\Delta}} = \hat{\boldsymbol{\Omega}}(\boldsymbol{\Gamma}) - \bar{\boldsymbol{\Omega}}$ minimizes $J(\boldsymbol{\Delta}) = Q(\bar{\boldsymbol{\Omega}} + \boldsymbol{\Delta})$. Consider the set

$$\Psi_p(R_p) := \big\{\boldsymbol{\Delta} : \boldsymbol{\Delta} = \boldsymbol{\Delta}^{\top}\,,\ \|\boldsymbol{\Delta}\|_F = R_p r_{pn}\big\} \tag{91}$$

where $R_p = 17C_{0p}/\beta_{p,\min}^2$ and r_{pn} is as in (42). Since $J(\hat{\boldsymbol{\Delta}}) \leq J(\boldsymbol{0}) = 0$, if we can show that $\inf_{\boldsymbol{\Delta}}\{J(\boldsymbol{\Delta}) : \boldsymbol{\Delta} \in \Psi_p(R_p)\} > 0$, then the minimizer $\hat{\boldsymbol{\Delta}}$ must be inside the sphere defined by $\Psi_p(R_p)$, and hence, $\|\hat{\boldsymbol{\Delta}}\|_F \leq R_p r_{pn}$. It is shown in [41, (9)] that $\ln(|\bar{\boldsymbol{\Omega}}+\boldsymbol{\Delta}|) - \ln(|\bar{\boldsymbol{\Omega}}|) = \mathrm{tr}(\bar{\boldsymbol{\Omega}}^{-1}\boldsymbol{\Delta}) - \tilde{B}_1$ where, with $\boldsymbol{H}(\bar{\boldsymbol{\Omega}}, \boldsymbol{\Delta}, v) = (\bar{\boldsymbol{\Omega}} + v\boldsymbol{\Delta})^{-1} \otimes (\bar{\boldsymbol{\Omega}} + v\boldsymbol{\Delta})^{-1}$ and v denoting a real scalar,

$$\tilde{B}_1 := \mathrm{vec}(\boldsymbol{\Delta})^{\top}\left(\int_0^1 (1-v)\boldsymbol{H}(\bar{\boldsymbol{\Omega}}, \boldsymbol{\Delta}, v)\,dv\right)\mathrm{vec}(\boldsymbol{\Delta})\,. \tag{92}$$

We have

$$J(\boldsymbol{\Delta}) = \sum_{i=1}^{3} B_i\,,\quad B_1 = \frac{1}{p}\tilde{B}_1\,, \tag{93}$$

$$B_2 := \frac{1}{p}\mathrm{tr}\big((\check{\boldsymbol{\Theta}} - \bar{\boldsymbol{\Omega}}^{-1})\boldsymbol{\Delta}\big)\,, \tag{94}$$

$$B_3 := \lambda_p\big(\|\bar{\boldsymbol{\Omega}}^{-} + \boldsymbol{\Delta}^{-}\|_1 - \|\bar{\boldsymbol{\Omega}}^{-}\|_1\big)\,. \tag{95}$$

By (33) and (79), $\bar{\boldsymbol{\Omega}}^{-1} = E\{\check{\boldsymbol{\Theta}}\} = \Big(\frac{1}{2Mq}\sum_{k=1}^{M}\mathrm{tr}\big(\bar{\boldsymbol{S}}_k^{\diamond}\boldsymbol{\Phi}_k + (\bar{\boldsymbol{S}}_k^{\diamond}\boldsymbol{\Phi}_k)^*\big)\Big)\boldsymbol{\Sigma}^{\diamond}$ (where we replaced $\hat{\boldsymbol{\Phi}}_k$ with $\boldsymbol{\Phi}_k$). By Lemma 6, $\max_{i,j}\big|\big[\check{\boldsymbol{\Theta}} - E\{\check{\boldsymbol{\Theta}}\}\big]_{ij}\big| \geq C_{0p}\sqrt{\frac{\ln(p)}{KqM}}$ w.h.p. (which refers to $1 - \frac{1}{p^{\tau-2}} - 4p^2e^{-KqM}$, cf. (78)). Following [41, p. 502], we have

$$\tilde{B}_1 \geq \|\boldsymbol{\Delta}\|_F^2/\big(2(\|\bar{\boldsymbol{\Omega}}\| + \|\boldsymbol{\Delta}\|)^2\big)\,. \tag{96}$$

Turning to $E\{\check{\boldsymbol{\Theta}}\}$, we have

$$\mathrm{tr}\big(\bar{\boldsymbol{S}}_k^{\diamond}\boldsymbol{\Phi}_k + (\bar{\boldsymbol{S}}_k^{\diamond}\boldsymbol{\Phi}_k)^*\big) = 2\mathrm{Re}\,\mathrm{tr}\big(\bar{\boldsymbol{S}}_k^{\diamond}(\boldsymbol{\Phi}_k - \boldsymbol{\Phi}_k^{\diamond} + \boldsymbol{\Phi}_k^{\diamond})\big) \tag{97}$$

$$= 2\mathrm{Re}\,\mathrm{tr}\big(\bar{\boldsymbol{S}}_k^{\diamond}(\boldsymbol{\Phi}_k - \boldsymbol{\Phi}_k^{\diamond})\big) + 2\mathrm{tr}(\boldsymbol{I}_q) \tag{98}$$

$$\geq 2q - 2\,\|\bar{\boldsymbol{S}}_k^{\diamond}\|_F\,\|\boldsymbol{\Phi}_k - \boldsymbol{\Phi}_k^{\diamond}\|_F \tag{99}$$

where we used $|\mathrm{tr}(\boldsymbol{B}\boldsymbol{C}^H)| \leq \|\boldsymbol{B}\|_F\,\|\boldsymbol{C}\|_F$ (Cauchy-Schwarz inequality). Since $\|\bar{\boldsymbol{S}}_k^{\diamond}\|_F \leq \sqrt{q}\,\|\bar{\boldsymbol{S}}_k^{\diamond}\| \leq \sqrt{q}\,\beta_{q,\max}$ and $\sum_{k=1}^{M}\|\boldsymbol{\Phi}_k - \boldsymbol{\Phi}_k^{\diamond}\|_F \leq \sqrt{M}\|\boldsymbol{\Gamma} - \boldsymbol{\Gamma}^{\diamond}\|_F \leq \sqrt{M}\,\gamma_q$, we have

$$A = 2\mathrm{Re}\sum_{k=1}^{M}\mathrm{tr}\big(\bar{\boldsymbol{S}}_k^{\diamond}\boldsymbol{\Phi}_k\big) \geq 2Mq - 2\sqrt{Mq}\,\beta_{q,\max}\gamma_q \tag{100}$$

$$\geq 2Mq - 2Mq\,\beta_{q,\max}\gamma_q = 1.8Mq\,, \tag{101}$$

where we have used the facts that $\sqrt{Mq} \leq Mq$ and $\beta_{q,\max}\gamma_q = 0.1$, as defined in (38). Therefore, $\|\bar{\boldsymbol{\Omega}}^{-1}\| = \|E\{\check{\boldsymbol{\Theta}}\}\| \geq 0.9\,\|\boldsymbol{\Sigma}^{\diamond}\|$, implying $\|\bar{\boldsymbol{\Omega}}\| \leq 10/(9\,\|\boldsymbol{\Sigma}^{\diamond}\|) \leq 10/(9\,\beta_{p,\min}) \leq 1.5/\beta_{p,\min}$. Using (93), (96), and the facts

$\|\bar{\mathbf{\Omega}}\| \leq 1.5/\beta_{p,\min}$ and $\|\mathbf{\Delta}\| \leq \|\mathbf{\Delta}\|_F = R_p r_{pn}$, we obtain w.h.p.

$$B_1 \geq \|\mathbf{\Delta}\|_F^2\, \beta_{p,\min}^2/(8p)\,, \tag{102}$$

for $n > N_p$, since $r_{pn} \leq \beta_{p,\min}/(34C_{0p})$ for $n > N_p$ and $R_p r_{pn} \leq 0.5/\beta_{p,\min}$.

We now consider B_2 given by (94). Define $\bar{\mathcal{S}}_p = \mathcal{S}_p \cup \{\{i,j\} : i = j\}$ so that $|\bar{\mathcal{S}}_p| = s_p + p$. We have

$$|B_2| \leq B_{12} + B_{22},\;\; pB_{12} = \Big| \sum_{\{i,j\}\in\bar{\mathcal{S}}_p} [\check{\mathbf{\Theta}} - \bar{\mathbf{\Omega}}^{-1}]_{ij}\Delta_{ji} \Big|,$$
$$pB_{22} = \Big| \sum_{\{i,j\}\in\bar{\mathcal{S}}_p^c} [\check{\mathbf{\Theta}} - \bar{\mathbf{\Omega}}^{-1}]_{ij}\,\Delta_{ji} \Big|,$$

where $\bar{\mathcal{S}}_p^c$ denotes the complement of set $\bar{\mathcal{S}}_p$. For an index set $\boldsymbol{B}$ and a matrix $\boldsymbol{C} \in \mathbb{R}^{p\times p}$, we write $\boldsymbol{C}_{\boldsymbol{B}}$ to denote a matrix in $\mathbb{R}^{p\times p}$ such that $[\boldsymbol{C}_{\boldsymbol{B}}]_{ij} = C_{ij}$ if $(i,j) \in \boldsymbol{B}$, and $[\boldsymbol{C}_{\boldsymbol{B}}]_{ij} = 0$ if $(i,j) \notin \boldsymbol{B}$. Using $\big|\sum_{\{i,j\}\in\bar{\mathcal{S}}_p} \Delta_{ij}\big| \leq \sqrt{s_p+p}\,\|\mathbf{\Delta}\|_F$ (by Cauchy-Schwarz inequality),

$$\begin{aligned} pB_{12} &\leq \max_{i,j} [\check{\mathbf{\Theta}} - \bar{\mathbf{\Omega}}^{-1}]_{ij} \,\Big|\sum_{\{i,j\}\in\bar{\mathcal{S}}_p} \Delta_{ij}\Big|\,, \\ &\leq C_{0p}\sqrt{\ln(p)/(KqM)}\,\sqrt{s_p+p}\,\|\mathbf{\Delta}\|_F = C_{0p} r_{pn}\|\mathbf{\Delta}\|_F\,. \end{aligned} \tag{103}$$

We will combine B_{22} with B_3. By (95),

$$\begin{aligned} B_3 =& \lambda_p\big(\|\bar{\mathbf{\Omega}}^- + \mathbf{\Delta}^-_{\mathcal{S}_p}\|_1 + \|\mathbf{\Delta}^-_{\mathcal{S}_p^c}\|_1 - \|\bar{\mathbf{\Omega}}^-\|_1\big) \\ \geq& \lambda_p\big(\|\mathbf{\Delta}^-_{\mathcal{S}_p^c}\|_1 - \|\mathbf{\Delta}^-_{\mathcal{S}_p}\|_1\big)\,, \end{aligned} \tag{104}$$

using the triangle inequality $\|\bar{\mathbf{\Omega}}^- + \mathbf{\Delta}^-_{\mathcal{S}_p}\|_1 \geq \|\bar{\mathbf{\Omega}}^-\|_1 - \|\mathbf{\Delta}^-_{\mathcal{S}_p}\|_1$ and the fact $\bar{\mathbf{\Omega}}^-_{\mathcal{S}_p^c} = \bar{\mathbf{\Omega}}_{\mathcal{S}_p^c} = \mathbf{0}$. Hence, $B_2 + B_3 \geq -B_{12} - B_{22} + \lambda_p\big(\|\mathbf{\Delta}^-_{\mathcal{S}_p^c}\|_1 - \|\mathbf{\Delta}^-_{\mathcal{S}_p}\|_1\big)$. But $pB_{22} \leq C_{0p}\sqrt{\ln(p)/(KqM)}\,\|\mathbf{\Delta}^-_{\mathcal{S}_p^c}\|_1$ w.h.p., therefore,

$$\begin{aligned} B_2 + B_3 \geq& \big(\lambda_p - C_{0p}\sqrt{\ln(p)/(p^2KqM)}\,\big)\|\mathbf{\Delta}^-_{\mathcal{S}_p^c}\|_1 \\ &- \lambda_p\|\mathbf{\Delta}^-_{\mathcal{S}_p}\|_1 - C_{0p} r_{pn}\|\mathbf{\Delta}\|_F/p\,. \end{aligned} \tag{105}$$

Using the fact that by (43), the first term on right side of (105) is nonnegative, and $\|\mathbf{\Delta}^-_{\mathcal{S}_p}\|_1 \leq \sqrt{s_p}\,\|\mathbf{\Delta}\|_F$ by the Cauchy-Schwarz inequality, we obtain $B_2 + B_3 \geq -\big(\lambda_p\sqrt{s_p} + r_{pn}/p\big)\|\mathbf{\Delta}\|_F$. Thus, by (43), (93) and (102)

$$\begin{aligned} J(\mathbf{\Delta}) \geq& \frac{\|\mathbf{\Delta}\|_F^2\,\beta_{p,\min}^2}{8p} - \big(\lambda_p\sqrt{s_p} + C_{0p}r_{pn}/p\big)\|\mathbf{\Delta}\|_F \\ \geq& \frac{\|\mathbf{\Delta}\|_F^2\,\beta_{p,\min}^2}{8p} - \frac{2C_{0p}r_{pn}\|\mathbf{\Delta}\|_F}{p} \\ =& \frac{\|\mathbf{\Delta}\|_F^2\,\beta_{p,\min}^2}{8p}\Big(1 - \frac{16}{17}\Big) > 0 \end{aligned} \tag{106}$$

using $\|\mathbf{\Delta}\|_F = R_p r_{pn}$ and $R_p = 17C_{0p}/\beta_{p,\min}^2$. This proves Theorem 2(i). □

Proof of Theorem 2(ii). With $\mathbf{\Gamma}$ as in (17), let $\mathbf{\Gamma} = \bar{\mathbf{\Gamma}}(\mathbf{\Omega}) + \mathbf{\Lambda}$ with $\mathbf{\Phi}_k = \mathbf{\Phi}_k^H \succ \mathbf{0}$, and denote $Q(\mathbf{\Gamma}) = L_2(\mathbf{\Gamma}) - L_2(\bar{\mathbf{\Gamma}}(\mathbf{\Omega}))$. For the rest of the proof, we will denote $\bar{\mathbf{\Gamma}}(\mathbf{\Omega})$ by $\bar{\mathbf{\Gamma}}$. Then $\hat{\mathbf{\Gamma}}(\mathbf{\Omega})$ minimizes $Q(\mathbf{\Gamma})$, or equivalently, $\hat{\mathbf{\Lambda}} = \hat{\mathbf{\Gamma}}(\mathbf{\Omega}) - \bar{\mathbf{\Gamma}}$ minimizes $J(\mathbf{\Lambda}) = Q(\bar{\mathbf{\Gamma}} + \mathbf{\Lambda})$. Note that $\mathbf{\Lambda} = [\mathbf{\Lambda}_1,\, \cdots,\, \mathbf{\Lambda}_M] \in \mathbb{C}^{q\times(qM)}$ and $\mathbf{\Lambda}_k = \mathbf{\Phi}_k - \bar{\mathbf{\Phi}}_k$, $k = 1, \cdots, M$, where $\bar{\mathbf{\Phi}}_k = \bar{\mathbf{\Phi}}_k(\mathbf{\Omega}) = p\mathbf{\Phi}_k^\diamond/\mathrm{tr}(\mathbf{\Sigma}^\diamond\mathbf{\Omega})$ by (34). Consider the set

$$\Psi_q(R_q) := \big\{\mathbf{\Lambda} :\, \mathbf{\Lambda}_k = \mathbf{\Lambda}_k^H,\, k = 1, \cdots M,\, \|\mathbf{\Lambda}\|_F = R_q r_{qn}\big\} \tag{107}$$

where $R_q = 17C_{0q}/\beta_{q,\min}^2$ and r_{qn} is as in (41). Similar to the proof of Theorem 2(i), our objective is to show that $\inf_{\mathbf{\Lambda}}\{J(\mathbf{\Lambda}) : \mathbf{\Lambda} \in \Psi_q(R_q)\} > 0$, which would ensure $\|\hat{\mathbf{\Lambda}}\|_F \leq R_q r_{qn}$ w.h.p. It is shown in [9, Lemma 5] that $\ln(|\bar{\mathbf{\Phi}}_k + \mathbf{\Lambda}_k|) - \ln(|\mathbf{\Phi}_k|) + \ln(|\bar{\mathbf{\Phi}}_k^* + \mathbf{\Lambda}_k^*|) - \ln(|\mathbf{\Phi}_k^*|) = \mathrm{tr}\big(\bar{\mathbf{\Phi}}_k^{-1}\mathbf{\Lambda}_k + (\bar{\mathbf{\Phi}}_k^{-1}\mathbf{\Lambda}_k)^*\big) - \tilde{B}_{1k}$ where

$$\tilde{B}_{1k} = \boldsymbol{g}^H(\mathbf{\Lambda}_k)\left(\int_0^1 (1-v)\boldsymbol{H}_k(\bar{\mathbf{\Phi}}_k, \mathbf{\Lambda}_k, v)\,dv\right)\boldsymbol{g}(\mathbf{\Lambda}_k)\,, \tag{108}$$

$$\boldsymbol{g}(\mathbf{\Lambda}_k) = \begin{bmatrix}\mathrm{vec}(\mathbf{\Lambda}_k)\\ \mathrm{vec}(\mathbf{\Lambda}_k^*)\end{bmatrix},\; \boldsymbol{H}_k(\bar{\mathbf{\Phi}}_k, \mathbf{\Lambda}_k, v) = \begin{bmatrix}\boldsymbol{H}_{11k} & \mathbf{0}\\ \mathbf{0} & \boldsymbol{H}_{22k}\end{bmatrix}, \tag{109}$$

$$\boldsymbol{H}_{11k} = (\bar{\mathbf{\Phi}}_k + v\mathbf{\Lambda}_k)^{-*} \otimes (\bar{\mathbf{\Phi}}_k + v\mathbf{\Lambda}_k)^{-1}\,, \tag{110}$$

$$\boldsymbol{H}_{22k} = (\bar{\mathbf{\Phi}}_k + v\mathbf{\Lambda}_k)^{-1} \otimes (\bar{\mathbf{\Phi}}_k + v\mathbf{\Lambda}_k)^{-*}\,, \tag{111}$$

and v is a real scalar. Therefore,

$$J(\mathbf{\Lambda}) = \sum_{k=1}^M\sum_{i=1}^3 B_{ik} + B_4\,, \quad B_{ik} = \frac{1}{2Mq}\tilde{B}_{1k}\,, \tag{112}$$

$$B_{2k} = \frac{1}{2Mq}\mathrm{tr}\big(\tilde{B}_{2k} + \tilde{B}_{2k}^*\big)\,, \quad \tilde{B}_{2k} = (\tilde{\mathbf{\Theta}} - \bar{\mathbf{\Phi}}_k^{-1})\mathbf{\Lambda}_k\,, \tag{113}$$

$$B_{3k} = \alpha\lambda_q\big(\|\bar{\mathbf{\Phi}}_k^- + \mathbf{\Lambda}_k^-\|_1 - \|\bar{\mathbf{\Phi}}_k^-\|_1\big)\,, \tag{114}$$

$$B_4 = (1-\alpha)\sqrt{M}\lambda_q\sum_{i\neq j}^p \big(\|\mathbf{\Phi}^{(ij)} + \mathbf{\Lambda}^{(ij)}\| - \|\mathbf{\Phi}^{(ij)}\|\big)\,. \tag{115}$$

By [9, Eqn. (B.43)], we have

$$B_{1k} \geq \frac{1}{2Mq}\,\frac{\|\mathbf{\Lambda}_k\|_F^2}{(\|\bar{\mathbf{\Phi}}_k\| + \|\mathbf{\Lambda}_k\|)^2}\,. \tag{116}$$

Now $\mathrm{tr}(\mathbf{\Sigma}^\diamond\mathbf{\Omega}) = \mathrm{tr}(\mathbf{\Sigma}^\diamond(\mathbf{\Omega} - \mathbf{\Omega}^\diamond + \mathbf{\Omega}^\diamond)) = \mathrm{tr}(\mathbf{\Sigma}^\diamond(\mathbf{\Omega} - \mathbf{\Omega}^\diamond)) + p$. Since $|\mathrm{tr}(\mathbf{\Sigma}^\diamond(\mathbf{\Omega} - \mathbf{\Omega}^\diamond))| \leq \|\mathbf{\Sigma}^\diamond\|_F\,\|\mathbf{\Omega} - \mathbf{\Omega}^\diamond\|_F \leq \sqrt{p}\,\beta_{p,\max}\gamma_p$, we have $|\mathrm{tr}(\mathbf{\Sigma}^\diamond\mathbf{\Omega})| \geq p - \sqrt{p}\,\beta_{p,\min}\gamma_p \geq p - p\,\beta_{p,\min}\gamma_p = 0.9p$ since $\gamma_p = 0.1/\beta_{p,\min}$. Therefore, $\|\bar{\mathbf{\Phi}}_k\| \leq p\|\mathbf{\Phi}_k^\diamond\|/(0.9p) \leq 1.5/\beta_{q,\min}$. Also, $\|\mathbf{\Lambda}_k\| \leq \|\mathbf{\Lambda}_k\|_F \leq \|\mathbf{\Lambda}\|_F = R_q r_{qn}$. Therefore,

$$\begin{aligned} \sum_{k=1}^M B_{1k} \geq& \frac{1}{2Mq}\,\frac{\sum_{k=1}^M\|\mathbf{\Lambda}_k\|_F^2}{(1.5/\beta_{q,\min} + R_q r_{qn})^2} \\ \geq& \frac{\|\mathbf{\Lambda}\|_F^2\,\beta_{q,\min}^2}{8Mq} \end{aligned} \tag{117}$$

w.h.p. for $n > N_q$, since $r_{qn} \leq \beta_{q,\min}/(34C_{0q})$ for $n > N_q$ and $R_q r_{qn} \leq 0.5/\beta_{q,\min}$.

We now bound B_{2k} noting that $|B_{2k}| \leq L_{1k} + L_{2k}$ where

$$L_{1k} = \frac{2}{2Mq}\Big|\sum_{\{i,j\}\in\bar{\mathcal{S}}_q}[\tilde{\mathbf{\Theta}} - \bar{\mathbf{\Phi}}_k^{-1}]_{ij}\,[\mathbf{\Lambda}_k]_{ji}\Big|,$$

$$L_{2k} = \frac{2}{2Mq}\Big|\sum_{\{i,j\}\in\bar{\mathcal{S}}_q^c}[\tilde{\mathbf{\Theta}} - \bar{\mathbf{\Phi}}_k^{-1}]_{ij}\,[\mathbf{\Lambda}_k]_{ji}\Big|$$

where $\bar{\mathcal{S}}_q = \mathcal{S}_q \cup \{\{i,j\} : i = j\}$ so that $|\bar{\mathcal{S}}_q| = s_q + q$. Using Lemma 5 and $\big|\sum_{\{i,j\}\in\bar{\mathcal{S}}_q}[\mathbf{\Lambda}_k]_{ij}\big| \le \sqrt{s_q+q}\,\|\mathbf{\Lambda}_k\|_F$ (by Cauchy-Schwarz inequality), we have

$$L_{1k} \le \frac{1}{Mq} C_{0q}\sqrt{\frac{\ln(M^{1/\tau}q)}{Kp}}\Big|\sum_{\{i,j\}\in\bar{\mathcal{S}}_q}[\mathbf{\Lambda}_k]_{ij}\Big| \le \frac{C_{0q}}{M^{3/2}q}\, r_{qn}\,\|\mathbf{\Lambda}_k\|_F\,, \tag{118}$$

$$L_{2k} \le \frac{C_{0q}}{Mq}\sqrt{\frac{\ln(M^{1/\tau}q)}{Kp}}\,\|\mathbf{\Lambda}^-_{k\mathcal{S}_q^c}\|_1\,. \tag{119}$$

Alternatively, as in [9, Eqn. (B.56)], with $B_2 = \sum_{k=1}^M B_{2k}$,

$$|B_2| \le \frac{2}{2Mq}\sum_{i,j=1}^p\sum_{k=1}^M \big|[\tilde{\mathbf{\Theta}} - \bar{\mathbf{\Phi}}_k^{-1}]_{ij}\big|\,\big|[\mathbf{\Lambda}_k]_{ji}\big| \le \frac{C_{0q}}{Mq}\sqrt{\frac{\ln(M^{1/\tau}q)}{Kp}}\sum_{i,j=1}^p\sum_{k=1}^M\big|[\mathbf{\Lambda}_k]_{ij}\big|\,. \tag{120}$$

Define $\check{\mathbf{\Lambda}} \in \mathbb{R}^{q\times q}$ with $[\check{\mathbf{\Lambda}}]_{ij} = \|\mathbf{\Lambda}^{(ij)}\|_F$ and as in (23), $\mathbf{\Lambda}^{(ij)} := [[\mathbf{\Lambda}_1]_{ij}\ \cdots\ [\mathbf{\Lambda}_M]_{ij}]^\top \in \mathbb{C}^M$. Using $\sum_{k=1}^M\big|[\mathbf{\Lambda}_k]_{ij}\big| \le \sqrt{M}\,\|\mathbf{\Lambda}^{(ij)}\|_F$, we have

$$|B_2| \le \frac{C_{0q}}{\sqrt{M}\,q}\sqrt{\frac{\ln(M^{1/\tau}q)}{Kp}}\,\|\check{\mathbf{\Lambda}}\|_1\,. \tag{121}$$

Mimicking [9, Eqns. (B.56)-(B.58)], we have $B_{3k} \ge \alpha\lambda_q(\|\mathbf{\Lambda}^-_{k\mathcal{S}_q^c}\|_1 - \|\mathbf{\Lambda}^-_{k\mathcal{S}_q}\|_1)$ and $B_4 \ge (1-\alpha)\sqrt{M}\,\lambda_q(\|\check{\mathbf{\Lambda}}^-_{k\mathcal{S}_q^c}\|_1 - \|\check{\mathbf{\Lambda}}^-_{k\mathcal{S}_q}\|_1)$. With $B_3 = \sum_{k=1}^M B_{3k}$ and using (118) and (119), similar to [9, Eqns. (B.60)], we have

$$\alpha B_2 + B_3 \ge -\alpha|B_2| + B_3 \ge -\alpha\lambda_q\sum_{k=1}^M\|\mathbf{\Lambda}^-_{k\mathcal{S}_q}\|_1 - \alpha\frac{C_{0q}}{M^{3/2}q}\,r_{qn}\sum_{k=1}^M\|\mathbf{\Lambda}_k\|_F \tag{122}$$

where we also used the first inequality in (45). Using $\|\mathbf{\Lambda}^-_{k\mathcal{S}_q}\|_1 \le \sqrt{s_q}\,\|\mathbf{\Lambda}_k\|_F$, $\sum_{k=1}^M\|\mathbf{\Lambda}_k\|_F \le \sqrt{M}\,\|\mathbf{\Lambda}\|_F$ and the second inequality in (45), we can simplify (122) as

$$\alpha B_2 + B_3 \ge -2\alpha\|\mathbf{\Lambda}\|_F\,\frac{C_{0q}}{Mq}\,r_{qn}\,. \tag{123}$$

In a similar manner (see also [9, Eqns. (B.61)]) using (121), we have

$$(1-\alpha)B_2 + B_4 \ge -2(1-\alpha)\|\mathbf{\Lambda}\|_F\,\frac{C_{0q}}{Mq}\,r_{qn} \tag{124}$$

under the upperbound on λ_{qn} specified in (45). Thus, by (112), (117), (123) and (124), we obtain

$$J(\mathbf{\Lambda}) \ge \frac{\|\mathbf{\Lambda}\|_F^2\,\beta_{q,\min}^2}{8Mq} - \frac{2C_{0q}r_{qn}\|\mathbf{\Lambda}\|_F}{Mq} = \frac{\|\mathbf{\Lambda}\|_F^2\,\beta_{q,\min}^2}{8Mq}\Big(1-\frac{16}{17}\Big) > 0 \tag{125}$$

using $\|\mathbf{\Lambda}\|_F = R_q r_{qn}$ and $R_q = 17C_{0q}/\beta_{q,\min}^2$. This proves Theorem 2(ii). □

Appendix C
Proof of Theorem 3

Proof of Theorem 3(i). Since $\|\mathbf{\Omega}^\diamond\|_F = 1$, we have $\bar{\mathbf{\Omega}}(\mathbf{\Gamma})/\|\bar{\mathbf{\Omega}}(\mathbf{\Gamma})\|_F = \mathbf{\Omega}^\diamond$. We have

$$\begin{aligned}\|\hat{\mathbf{\Omega}} - \mathbf{\Omega}^\diamond\|_F &= \Big\|\hat{\mathbf{\Omega}}(\mathbf{\Gamma})/\|\hat{\mathbf{\Omega}}(\mathbf{\Gamma})\|_F - \bar{\mathbf{\Omega}}(\mathbf{\Gamma})/\|\bar{\mathbf{\Omega}}(\mathbf{\Gamma})\|_F\Big\|_F\\ &= \Big\|\frac{\hat{\mathbf{\Omega}}(\mathbf{\Gamma})}{\|\hat{\mathbf{\Omega}}(\mathbf{\Gamma})\|_F} - \frac{\bar{\mathbf{\Omega}}(\mathbf{\Gamma})}{\|\hat{\mathbf{\Omega}}(\mathbf{\Gamma})\|_F} + \frac{\bar{\mathbf{\Omega}}(\mathbf{\Gamma})}{\|\hat{\mathbf{\Omega}}(\mathbf{\Gamma})\|_F} - \frac{\bar{\mathbf{\Omega}}(\mathbf{\Gamma})}{\|\bar{\mathbf{\Omega}}(\mathbf{\Gamma})\|_F}\Big\|_F\\ &\le \frac{\|\hat{\mathbf{\Omega}}(\mathbf{\Gamma}) - \bar{\mathbf{\Omega}}(\mathbf{\Gamma})\|_F}{\|\hat{\mathbf{\Omega}}(\mathbf{\Gamma})\|_F} + \|\bar{\mathbf{\Omega}}(\mathbf{\Gamma})\|_F\,\Big|\frac{1}{\|\hat{\mathbf{\Omega}}(\mathbf{\Gamma})\|_F} - \frac{1}{\|\bar{\mathbf{\Omega}}(\mathbf{\Gamma})\|_F}\Big|\\ &\le \frac{2}{\|\hat{\mathbf{\Omega}}(\mathbf{\Gamma})\|_F}\,\|\hat{\mathbf{\Omega}}(\mathbf{\Gamma}) - \bar{\mathbf{\Omega}}(\mathbf{\Gamma})\|_F\end{aligned} \tag{126}$$

using $\big|\|\bar{\mathbf{\Omega}}(\mathbf{\Gamma})\|_F - \|\hat{\mathbf{\Omega}}(\mathbf{\Gamma})\|_F\big| \le \|\hat{\mathbf{\Omega}}(\mathbf{\Gamma}) - \bar{\mathbf{\Omega}}(\mathbf{\Gamma})\|_F$ (by triangle inequality). Now $\|\hat{\mathbf{\Omega}}(\mathbf{\Gamma})\|_F = \|\hat{\mathbf{\Omega}}(\mathbf{\Gamma}) - \bar{\mathbf{\Omega}}(\mathbf{\Gamma}) + \bar{\mathbf{\Omega}}(\mathbf{\Gamma})\|_F \ge \|\bar{\mathbf{\Omega}}(\mathbf{\Gamma})\|_F - \|\hat{\mathbf{\Omega}}(\mathbf{\Gamma}) - \bar{\mathbf{\Omega}}(\mathbf{\Gamma})\|_F$. For $n > N_{2p}$, we have $\|\hat{\mathbf{\Omega}}(\mathbf{\Gamma}) - \bar{\mathbf{\Omega}}(\mathbf{\Gamma})\|_F \le 0.5\|\bar{\mathbf{\Omega}}(\mathbf{\Gamma})\|_F$, and therefore, $\|\hat{\mathbf{\Omega}}(\mathbf{\Gamma})\|_F \ge 0.5\|\bar{\mathbf{\Omega}}(\mathbf{\Gamma})\|_F$. Hence,

$$\|\hat{\mathbf{\Omega}} - \mathbf{\Omega}^\diamond\|_F \le 4\,\|\hat{\mathbf{\Omega}}(\mathbf{\Gamma}) - \bar{\mathbf{\Omega}}(\mathbf{\Gamma})\|_F/\|\bar{\mathbf{\Omega}}(\mathbf{\Gamma})\|_F\,. \tag{127}$$

We now characterize $\|\bar{\mathbf{\Omega}}(\mathbf{\Gamma})\|_F$. We have

$$\begin{aligned}A &= \Big|\sum_{k=1}^M\big(\mathrm{tr}(\bar{\boldsymbol{S}}_k^\diamond\mathbf{\Phi}_k) + \mathrm{tr}(\bar{\boldsymbol{S}}_k^\diamond\mathbf{\Phi}_k)^*\big)\Big| \le 2\sum_{k=1}^M\big|\mathrm{tr}(\bar{\boldsymbol{S}}_k^\diamond\mathbf{\Phi}_k)\big|\\ &\le 2\sum_{k=1}^M\|\bar{\boldsymbol{S}}_k^\diamond\|_F\,\|\mathbf{\Phi}_k\|_F \le 2\sqrt{q}\,\beta_{q,\max}\sum_{k=1}^M\|\mathbf{\Phi}_k\|_F\,.\end{aligned}$$

Since $\sum_{k=1}^M\|\mathbf{\Phi}_k\|_F \le \sum_{k=1}^M\|\mathbf{\Phi}_k - \mathbf{\Phi}_k^\diamond\|_F + \sum_{k=1}^M\|\mathbf{\Phi}_k^\diamond\|_F \le \sqrt{M}\,\gamma_q + \sqrt{q}\,M/\beta_{q,\min}$, we have $A \le 0.2\sqrt{qM} + 2qM\beta_{q,\max}/\beta_{q,\min} \le 2qM(1+\beta_{q,\max}/\beta_{q,\min})$. By (33) and the fact $\|\mathbf{\Omega}^\diamond\|_F = 1$, we infer $\|\bar{\mathbf{\Omega}}(\mathbf{\Gamma})\|_F \ge \beta_{q,\min}/(\beta_{q,\max} + \beta_{q,\min}) = 1/\gamma_r$, which combined with (127) and (44) yields (47). □

Proof of Theorem 3(ii). For $n > N_{3p}$, $\hat{\mathbf{\Omega}} \in \mathcal{B}(\mathbf{\Omega}^\diamond)$ (cf. Theorem 3(i)), and $C_{2p}r_{pn} \le (p/2)$ w.h.p. We have

$$\|\hat{\mathbf{\Gamma}}(\hat{\mathbf{\Omega}}) - \mathbf{\Gamma}^\diamond\|_F \le \|\hat{\mathbf{\Gamma}}(\hat{\mathbf{\Omega}}) - \bar{\mathbf{\Gamma}}(\hat{\mathbf{\Omega}})\|_F + \|\bar{\mathbf{\Gamma}}(\hat{\mathbf{\Omega}}) - \mathbf{\Gamma}^\diamond\|_F$$

where Theorem 2(ii) applies to $\|\hat{\mathbf{\Gamma}}(\hat{\mathbf{\Omega}}) - \bar{\mathbf{\Gamma}}(\hat{\mathbf{\Omega}})\|_F$. By (34),

$$\bar{\mathbf{\Gamma}}(\hat{\mathbf{\Omega}}) - \mathbf{\Gamma}^\diamond = \big(\frac{p}{\mathrm{tr}(\mathbf{\Sigma}^\diamond\hat{\mathbf{\Omega}})} - 1\big)\mathbf{\Gamma}^\diamond\,. \tag{128}$$

As in the proof of Theorem 2(ii) (following (116)), we have $\mathrm{tr}(\mathbf{\Sigma}^\diamond\hat{\mathbf{\Omega}}) = \mathrm{tr}(\mathbf{\Sigma}^\diamond(\hat{\mathbf{\Omega}} - \mathbf{\Omega}^\diamond + \mathbf{\Omega}^\diamond)) = \mathrm{tr}(\mathbf{\Sigma}^\diamond(\hat{\mathbf{\Omega}} - \mathbf{\Omega}^\diamond)) + p$ and $|\mathrm{tr}(\mathbf{\Sigma}^\diamond(\hat{\mathbf{\Omega}} - \mathbf{\Omega}^\diamond))| \le \|\mathbf{\Sigma}^\diamond\|_F\,\|\hat{\mathbf{\Omega}} - \mathbf{\Omega}^\diamond\|_F \le C_{2p}r_{pn}$ (using (47)). Therefore, $p - C_{2p}r_{pn} \le \mathrm{tr}(\mathbf{\Sigma}^\diamond\hat{\mathbf{\Omega}}) \le p + C_{2p}r_{pn}$ and $|p - \mathrm{tr}(\mathbf{\Sigma}^\diamond\hat{\mathbf{\Omega}})| \le C_{2p}r_{pn}$. Since $0 < C_{2p}r_{pn} \le (p/2)$ w.h.p., $|\mathrm{tr}(\mathbf{\Sigma}^\diamond\hat{\mathbf{\Omega}})|^{-1} \le 2/p$. Thus we have $\|\bar{\mathbf{\Gamma}}(\hat{\mathbf{\Omega}}) - \mathbf{\Gamma}^\diamond\|_F \le C_{2q}r_{pn}$, which yields (48). The given probability bound is the result of the bounds in Theorem 2 (both (44) and (46) must hold) and an application of the union bound. □

References

[1] S.L. Lauritzen, *Graphical Models*. Oxford, UK: Oxford Univ. Press, 1996.

[2] M.J. Wainwright, *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge, UK: Cambridge Univ. Press, 2019.

[3] N. Meinshausen and P. Bühlmann, "High-dimensional graphs and variable selection with the Lasso," *Ann. Statist.*, vol. 34, no. 3, pp. 1436-1462, 2006.

[4] O. Banerjee, L.E. Ghaoui and A. d'Aspremont, "Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data," *J. Mach. Learn. Res.*, vol. 9, pp. 485-516, 2008.

[5] J. Friedman, T. Hastie and R. Tibshirani, "Sparse inverse covariance estimation with the graphical lasso," *Biostatistics*, vol. 9, no. 3, pp. 432-441, July 2008.

[6] R. Dahlhaus, "Graphical interaction models for multivariate time series," *Metrika*, vol. 51, pp. 157-172, 2000.

[7] A. Jung, G. Hannak and N. Goertz, "Graphical LASSO based model selection for time series," *IEEE Signal Process. Lett.*, vol. 22, no. 10, pp. 1781-1785, Oct. 2015.

[8] J.K. Tugnait, "Graphical modeling of high-dimensional time series," in *Proc. 52nd Asilomar Conf. Signals, Systems, Computers*, pp. 840-844, Pacific Grove, CA, Oct. 29 - Oct. 31, 2018.

[9] J.K. Tugnait, "On sparse high-dimensional graphical model learning for dependent time series," *Signal Process.*, vol. 197, pp. 1-18, Aug. 2022, Article 108539.

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

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

[12] D. Alpago, M. Zorzi and A. Ferrante, "Identification of sparse reciprocal graphical models," *IEEE Control Sys. Lett.*, vol. 22, no. 4, pp. 659-664, 2018.

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

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

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

[16] S. Basu and G. Michailidis, "Regularized estimation in sparse high-dimensional time series models," *Annals Statist.*, vol. 43, no. 4, pp. 1535-1567, 2015.

[17] C. Leng and C.Y. Tang, "Sparse matrix graphical models," *J. Amer. Statist. Assoc.*, vol. 107, pp. 1187-1200, Sep. 2012.

[18] T. Tsiligkaridis, A.O. Hero, III, and S. Zhou, "On convergence of Kronecker graphical lasso algorithms," *IEEE Trans. Signal Process.*, vol. 61, no. 7, pp. 1743-1755, April 2013.

[19] Y. Zhu and L. Li, "Multiple matrix Gaussian graphs estimation," *J. Royal Statistical Soc., Series B*, vol. 80, pp. 927-950, 2018.

[20] X. Chen and W. Liu, "Graph estimation for matrix-variate Gaussian data," *Statistica Sinica*, vol. 29, pp. 479-504, 2019.

[21] K. Greenewald, S. Zhou and A. Hero III, "Tensor graphical lasso (teralasso)," *J. Royal Statistical Soc., Series B*, vol. 81, no. 5, pp. 901-931, 2019.

[22] X. Lyu, W.W. Sun, Z. Wang, H. Liu, J. Yang and G. Cheng, "Tensor graphical model: Non-convex optimization and statistical inference," *IEEE Trans. Pattern Analysis Mach. Intell.*, vol. 42, no. 8, pp. 2024-2037, 1 Aug. 2020.

[23] F. Huang and S. Chen, "Joint learning of multiple sparse matrix Gaussian graphical models," *IEEE Trans. Neural Netw. Learning Sys.*, vol. 26, no. 11, pp. 2606-2620, Nov. 2015.

[24] S. Zhou, "Gemini: Graph estimation with matrix variate normal instances," *Annals Statist.*, vol. 42, no. 2, pp. 532-562, 2014.

[25] J. Yin and H. Li, "Model selection and estimation in the matrix normal graphical model," *J. Multivariate Analysis*, vol. 107, pp. 119-140, May 2012.

[26] S. He, J. Yin, H. Li and X. Wang, "Graphical model selection and estimation for high dimensional tensor data," *J. Multivariate Analysis*, vol. 128, pp. 165-185, 2014.

[27] K. Min, Q. Mai and X. Zhang, "Fast and separable estimation in high-dimensional tensor Gaussian graphical models," *J. Comp. Graphical Statistics*, vol. 31, pp. 294-300, 2022.

[28] K. Werner, M. Jansson and P. Stoica, "On estimation of covariance matrices with Kronecker product structure," *IEEE Trans. Signal Process.*, vol. 56, no. 2, pp. 478-491, Feb. 2008.

[29] P.M. Weichsel, "The Kronecker product of graphs," *Proc. American Math. Soc.*, vol. 13, no. 1, pp. 37-52, 1962.

[30] J. Leskovec, D. Chakrabarti, J. Kleinberg, C. Faloutsos, and Z. Ghahramani, "Kronecker graphs: An approach to modeling networks," *J. Mach. Learn. Res.*, vol. 11, pp. 985-1042, Feb. 2010.

[31] C.M. Carvalho and M. West, "Dynamic matrix-variate graphical models," *Bayesian Analysis*, vol. 2, no. 1, pp. 69-98, 2007.

[32] H. Wang and M. West, "Bayesian analysis of matrix normal graphical models," *Biometrika*, vol. 96, no. 4, pp. 821-834, Dec. 2009.

[33] Y. Jiang, J. Bigot and S. Maabout, "Online graph topology learning from matrix-valued time series," arXiv:2107.08020v1 [stat.ML], July 2021.

[34] M. Zorzi, "Nonparametric identification of Kronecker networks," *Automatica*, vol. 145, no. 9, p. 110518, Nov. 2022.

[35] B. Sinquin and M. Verhaegen, "Quarks: Identification of large-scale Kronecker vector-autoregressive models," *IEEE Trans. Autom. Control*, vol. 64, no. 3, pp. 448-463, 2019.

[36] M. Zorzi, "Autoregressive identification of Kronecker graphical models," *Automatica*, vol. 119, no. 9, p. 109053, Sep. 2020.

[37] R. Chen, H. Xiao and D. Yang, "Autoregressive models for matrix-valued time series," *J. Econometrics*, vol. 222, pp. 539-560, 2021.

[38] P.J. Schreier and L.L. Scharf, *Statistical Signal Processing of Complex-Valued Data*, Cambridge, UK: Cambridge Univ. Press, 2010.

[39] J.K. Tugnait, "Sparse high-dimensional matrix-valued graphical model learning from dependent data," in *Proc. 22nd IEEE Statistical Signal Proc. Workshop (SSP-2023)*, pp. 344-348, Hanoi, Vietnam, July 2-5, 2023.

[40] J.K. Tugnait, "Sparse-group lasso for graph learning from multi-attribute data," *IEEE Trans. Signal Process.*, vol. 69, pp. 1771-1786, 2021. (Corrections: vol. 69, p. 4758, 2021.)

[41] A.J. Rothman, P.J. Bickel, E. Levina and J. Zhu, "Sparse permutation invariant covariance estimation," *Elec. J. Statistics*, vol. 2, pp. 494-515, 2008.

[42] A.K. Gupta and D.K. Nagar, *Matrix Variate Distributions*. Boca Raton, FL: Chapman and Hall/CRC Press, 1999.

[43] J.K. Tugnait, "Edge exclusion tests for graphical model selection: Complex Gaussian vectors and time series," *IEEE Trans. Signal Process.*, vol. 67, no. 19, pp. 5062-5077, Oct. 1, 2019.

[44] D.R. Brillinger, *Time Series: Data Analysis and Theory*, Expanded edition. New York: McGraw Hill, 1981.

[45] J. Friedman, T. Hastie and R. Tibshirani, "A note on the group lasso and a sparse group lasso," *arXiv:1001.0736v1 [math.ST]*, 5 Jan 2010.

[46] N. Simon, J. Friedman, T. Hastie and R. Tibshirani, "A sparse-group lasso," *J. Comp. Graphical Statistics*, vol. 22, pp. 231-245, 2013.

[47] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *J. Royal Statistical Soc., Series B*, vol. 68, pp. 49-67, 2006.

[48] J. Gorski, F. Pfeuffer and K. Klamroth, "Biconvex sets and optimization with biconvex functions: A survey and extensions," *Math. Methods Operations Res.*, vol. 66, pp. 373-408, 2007.

[49] S. Boyd, N. Parikh, E. Chu, B. Peleato and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends in Machine Learning*, vol. 3, no. 1, pp. 1-122, 2010.

[50] S. Zhang, B. Guo, A. Dong, J. He, Z. Xu and S.X. Chen, "Cautionary tales on air-quality improvement in Beijing," *Proc. Royal Soc. A*, vol. 473, p. 20170457, 2017.

[51] W. Chen, F. Wang, G. Xiao, J. Wu and S. Zhang, "Air quality of Beijing and impacts of the new ambient air quality standard," *Atmosphere*, vol. 6, pp. 1243-1258, 2015.

[52] J. Fan, Y. Feng and Y. Wu, "Network exploration via the adaptive lasso and SCAD penalties," *Annals Applied Statistics*, vol. 3, no. 2, pp. 521-541, 2009.

[53] J.K. Tugnait, "Sparse-group log-sum penalized graphical model learning for time series," in *Proc. 2022 IEEE Intern. Conf. Acoustics, Speech, Signal Process. (ICASSP 2022)*, pp. 5822-5826, Singapore, May 22-27, 2022.