SPARSE HIGH-DIMENSIONAL MATRIX-VALUED GRAPHICAL MODEL LEARNING FROM DEPENDENT DATA

Jitendra K. Tugnait

Department of Electrical & Computer Engineering
Auburn University, Auburn, AL 36849, USA

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 matrix graphical models assume that i.i.d. observations of 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 results also yields a rate of convergence. We illustrate our approach using numerical examples.

Keywords: Sparse graph learning; matrix graph estimation; matrix time series; undirected graph; inverse spectral density estimation.

1. 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 $x_1, x_1, \cdots, x_p$, 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 $\boldsymbol{x}$ has positive-definite covariance matrix $\boldsymbol{\Sigma}$ with precision matrix $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$. Then Ω_{ij}, the (i, j)-th element of $\boldsymbol{\Omega}$, is zero iff x_i and x_j are conditionally independent [1]. Such models for $\boldsymbol{x}$ have been extensively studied [2–4].

These models are vector graphical models. Time series (dependent data) graphical models are much less studied. 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 $\{x_\ell(t),\ \ell \in [p],\ \ell \neq i,\ \ell \neq j\}$ [5]. Vector graphical models (based only on the precision matrix $\boldsymbol{\Omega}$) do not necessarily capture the "true" series graphical model if the data originates from a time-dependent series. Denote the power spectral density (PSD) matrix of $\{\boldsymbol{x}(t)\}$ by $\boldsymbol{S}_x(f)$, where $\boldsymbol{S}_x(f) = \sum_{\tau=-\infty}^{\infty} \boldsymbol{R}_{xx}(\tau) e^{-j2\pi f\tau}$, $\boldsymbol{R}_{xx}(\tau) = E\{\boldsymbol{x}(t+\tau)\boldsymbol{x}^\top(t)\}$. In [5] it was shown that conditional independence of two time series components given all other components of the time series, is encoded by zeros in the inverse PSD, that is, $\{i, j\} \notin \mathcal{E}$ iff the (i, j)-th element of $\boldsymbol{S}_x(f)$, $[\boldsymbol{S}_x^{-1}(f)]_{ij} = 0$ for every f.

The need for matrix-valued graphical models arises in several applications [6–15]. 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$(\boldsymbol{Z})$, 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 moreover, it ignores any structural information among rows and columns of the matrix observations [6]. Prior work [6–15] all assume that i.i.d. observations of $\boldsymbol{Z}$ are available for graphical modeling. Our objective in this paper is to learn the graph associated with time-dependent matrix-valued $p \times q$ Gaussian sequence $\boldsymbol{Z}(t)$, given observations of $\boldsymbol{Z}(t)$ for $t = 0, 1, \cdots, n-1$.

Notation: $|\boldsymbol{A}|$ and $\mathrm{tr}(\boldsymbol{A})$ denote the determinant and the trace of the square matrix $\boldsymbol{A}$, respectively, and $\mathrm{etr}(\boldsymbol{A}) = \exp(\mathrm{tr}(\boldsymbol{A}))$. $[\boldsymbol{B}]_{ij}$ denotes the (i, j)-th element of $\boldsymbol{B}$, and so does B_{ij}. $\boldsymbol{I}_m$ is the $m \times m$ identity matrix. The superscripts $*$ and H denote the complex conjugate and conjugate transpose operations, respectively, $\boldsymbol{x} \sim \mathcal{N}_c(\boldsymbol{m}, \boldsymbol{\Sigma})$ denotes a random vector $\boldsymbol{x}$ that is circularly symmetric (proper) complex Gaussian with mean $\boldsymbol{m}$ and covariance $\boldsymbol{\Sigma}$, and $\otimes$ denotes the Kronecker product.

2. SYSTEM MODEL

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 given by [16, Chap. 2]

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

Equivalently,

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

Here $\boldsymbol{\Psi}$ is the row covariance matrix and $\boldsymbol{\Sigma}$ is the column covariance matrix [16] since the kth column $\boldsymbol{Z}_{\cdot k} \sim \mathcal{N}(\boldsymbol{0}, [\boldsymbol{\Psi}]_{kk}\boldsymbol{\Sigma})$ and the ith row $\boldsymbol{Z}_{i\cdot}^\top \sim \mathcal{N}(\boldsymbol{0}, [\boldsymbol{\Sigma}]_{ii}\boldsymbol{\Psi})$.

Graphical modeling of random vectors to characterize conditional dependence of its components [1, 3] has been extended to matrix data with structured information [6–9, 13]. With $\boldsymbol{Z} \in \mathbb{R}^{p\times q}$ modeled as a zero-mean matrix normal vector and $\boldsymbol{z} = \mathrm{vec}(\boldsymbol{Z})$, [6] assumes

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

which could be interpreted as follows. Let $\boldsymbol{x} \in \mathbb{R}^p$, $\boldsymbol{y} \in \mathbb{R}^q$, $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$ and $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Psi})$. Express $\boldsymbol{Z}$ as

$$\boldsymbol{Z} = \boldsymbol{x} \otimes \boldsymbol{y}^\top, \text{ or } \boldsymbol{z} = \mathrm{vec}(\boldsymbol{Z}) = \mathrm{vec}(\boldsymbol{x}\boldsymbol{y}^\top) = \boldsymbol{y} \otimes \boldsymbol{x} \tag{4}$$

This work was supported by NSF Grant ECCS-2040536. Author's email: tugnajk@auburn.edu

such that

$$E\{\boldsymbol{z}\boldsymbol{z}^\top\} = E\{(\boldsymbol{y}\otimes\boldsymbol{x})(\boldsymbol{y}\otimes\boldsymbol{x})^\top\} = E\{(\boldsymbol{y}\boldsymbol{y}^\top)\otimes(\boldsymbol{x}\boldsymbol{x})^\top)\}$$
$$= E\{\boldsymbol{y}\boldsymbol{y}^\top\}\otimes E\{\boldsymbol{x}\boldsymbol{x}^\top\} = \boldsymbol{\Psi}\otimes\boldsymbol{\Sigma}\,, \tag{5}$$

implying a separable covariance structure [17]. Let $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ and $\boldsymbol{\Gamma} = \boldsymbol{\Psi}^{-1}$ 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 $\boldsymbol{\Omega}_{ij}$ and $\boldsymbol{\Gamma}_{k\ell}$ is zero when $i\neq k$, $j\neq\ell$, (ii) $\boldsymbol{\Omega}_{ij}=0$ when $i\neq k$, $j=\ell$, and (iii) $\boldsymbol{\Gamma}_{k\ell}=0$ when $i=k$, $j\neq\ell$ [6]. Prior work [6–9, 13] all assume that i.i.d. observations of $\boldsymbol{Z}$ are available for graphical modeling.

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

$$E\{\boldsymbol{z}(t+\tau)\boldsymbol{z}^\top(t)\} = \boldsymbol{\Psi}(\tau)\otimes\boldsymbol{\Sigma} \tag{6}$$

where $\boldsymbol{\Psi}(\tau)$, $\tau = 0,\pm1,\cdots$ models time-dependence while $\boldsymbol{\Sigma}\succ\boldsymbol{0}$ is fixed. With $\{\boldsymbol{e}(t)\}$ i.i.d., $\boldsymbol{e}(t)\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I})$, a generative model for $\boldsymbol{z}(t)$ is given by

$$\boldsymbol{z}(t) = \sum_{i=0}^{L}(\boldsymbol{B}_i\otimes\boldsymbol{F})\boldsymbol{e}(t-i)\,,\ \boldsymbol{B}_i\in\mathbb{R}^q\,,\ \boldsymbol{F}\in\mathbb{R}^p \tag{7}$$

$$\Rightarrow\ E\{\boldsymbol{z}(t+\tau)\boldsymbol{z}^\top(t)\} = \underbrace{\big(\sum_{i=0}^{L}\boldsymbol{B}_i\boldsymbol{B}_{i-\tau}^\top\big)}_{=\boldsymbol{\Psi}(\tau)}\otimes\underbrace{(\boldsymbol{F}\boldsymbol{F}^\top)}_{=\boldsymbol{\Sigma}}\,. \tag{8}$$

An example considered in [6] is that of a United States Department of Agriculture (USDA) dataset reporting itemized annual export to major trading partners. The dataset with 40 years U.S. export is collected for 13 trading partners and 36 items. Each observation in the dataset can be denoted by a 13×36 matrix where the trading partners and items, as the rows and columns, respectively, of this matrix, are used as structural information for the observations. The basic idea in matrix-valued graphs is to model the covariance of $\text{vec}(\boldsymbol{Z})$ as $\boldsymbol{\Psi}\otimes\boldsymbol{\Sigma}$ reducing the number of unknowns from $\mathcal{O}(p^2q^2)$ in the precision matrix for the "full" vectorized model to $\mathcal{O}(p^2+q^2)$ for the matrix model, while also preserving the structural information. Given data, one estimates two precision matrices $\boldsymbol{\Omega}=\boldsymbol{\Sigma}^{-1}$ and $\boldsymbol{\Gamma}=\boldsymbol{\Psi}^{-1}$. In the matrix graph, conditional independence between $\boldsymbol{Z}_{ij}$ and $\boldsymbol{Z}_{k\ell}$ is determined by zeros in $\boldsymbol{\Omega}$ and $\boldsymbol{\Gamma}$ [6]. While [6] and others ([8–11]) all consider only i.i.d. observations, we allow possible temporal dependence in matrix observations via $\boldsymbol{\Psi}(\tau)$.

The PSD of $\{\boldsymbol{z}(t)\}$ is $\boldsymbol{S}_z(f) = \bar{\boldsymbol{S}}(f)\otimes\boldsymbol{\Sigma}$ where $\bar{\boldsymbol{S}}(f) = \sum_\tau\boldsymbol{\Psi}(\tau)e^{-j2\pi f\tau}$. Then $\boldsymbol{S}_z^{-1}(f) = \bar{\boldsymbol{S}}^{-1}(f)\otimes\boldsymbol{\Sigma}^{-1}$, and by [5], in the $pq-$node graph $\mathcal{G}=(V,\mathcal{E})$, $|V|=pq$, associated with $\{\boldsymbol{z}(t)\}$, edge $\{i,j\}\in\mathcal{E}$ 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 $\Gamma=\boldsymbol{\Psi}^{-1}$, using [5, 6], 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 $\boldsymbol{\Omega}_{ij}$ and $[\bar{\boldsymbol{S}}^{-1}(f)]_{k\ell}$, $f\in[0,0.5]$ is zero when $i\neq k$, $j\neq\ell$, (ii) $\boldsymbol{\Omega}_{ij}=0$ when $i\neq k$, $j=\ell$, and (iii) $[\bar{\boldsymbol{S}}^{-1}(f)]_{k\ell}=0$ for $f\in[0,0.5]$ when $i=k$, $j\neq\ell$.

As an example, consider $\boldsymbol{x}(t)=\sum_{i=0}^{L}\boldsymbol{B}_i\boldsymbol{\epsilon}(t-i)$, $L\geq1$, where $\{\boldsymbol{\epsilon}(t)\}$ is zero-mean, i.i.d. Gaussian, with covariance$=\boldsymbol{I}$,

$$\boldsymbol{B}_0 = \begin{bmatrix}1&0&0\\0&\sqrt{1-b^2}&0\\a&0&\sqrt{1-a^2}\end{bmatrix}\,,\quad \boldsymbol{B}_L = \begin{bmatrix}0&0&0\\b&0&0\\0&0&0\end{bmatrix}\,,$$
$$\boldsymbol{B}_i = \boldsymbol{0}\ \forall i\neq0,\ i\neq L,\ \ |a|<1,\ |b|<1. \tag{9}$$

Straightforward calculations yield

$$\boldsymbol{\Gamma}_x = \boldsymbol{R}_{xx}^{-1}(0) = \frac{1}{1-a^2}\begin{bmatrix}1&0&-a\\0&1-a^2&0\\-a&0&1\end{bmatrix}\,, \tag{10}$$

$$c\,\boldsymbol{S}_x^{-1}(f) = \begin{bmatrix}1-(ab)^2&d_1&-a(1-b^2)\\d_1^*&1-a^2&0\\-a(1-b^2)&0&1-b^2\end{bmatrix} \tag{11}$$

where $c=(1-a^2)(1-b^2)$ and $d_1=-b(1-a^2)e^{j2\pi fL}$. Notice that in $\boldsymbol{\Gamma}_x$, edges $\{1,2\}$ and $\{2,1\}$ are missing whereas they are present in $\boldsymbol{S}_x^{-1}(f)$, that is, $\boldsymbol{\Gamma}_x$ does not capture the true dependencies among various components of the dependent series.

Our objective is to learn the graph associated with time-dependent sequence $\{\boldsymbol{Z}(t)\}$, given observations $t=0,1,\cdots,n-1$, 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 will take $\Omega_{11}=1$.

3. PENALIZED NEGATIVE LOG-LIKELIHOOD

Given $\boldsymbol{z}(t)$ for $t=0,1,2,\cdots,n-1$. Define the (normalized) DFT $\boldsymbol{d}_z(f_m)$ of $\boldsymbol{z}(t)$, ($j=\sqrt{-1}$, $f_m=m/n$), over $m=0,1,\cdots,n-1$ as $\boldsymbol{d}_z(f_m) = \frac{1}{\sqrt{n}}\sum_{t=0}^{n-1}\boldsymbol{z}(t)\exp(-j2\pi f_mt)$. Let $\boldsymbol{D}_z(f_m) = \frac{1}{\sqrt{n}}\sum_{t=0}^{n-1}\boldsymbol{Z}(t)\exp(-j2\pi f_mt)$, then $\boldsymbol{d}_z(f_m) = \text{vec}(\boldsymbol{D}_z(f_m))$. It is established in [18] (see also [19]) that 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; 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

$$\tilde{f}_k = \frac{(k-1)K+m_t+1}{n},\qquad k=1,2,\cdots,M, \tag{12}$$

yielding M equally spaced frequencies $\tilde{f}_k$ in the interval $(0,0.5)$. By local smoothness

$$\boldsymbol{S}_z(\tilde{f}_{k,\ell}) = \boldsymbol{S}_z(\tilde{f}_k)\ \text{ for } \ell=-m_t,-m_t+1,\cdots,m_t, \tag{13}$$
$$\text{where }\ \tilde{f}_{k,\ell} = \frac{(k-1)K+m_t+1+\ell}{n}. \tag{14}$$

It is known ([20, Theorem 4.4.1]) that 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, provided all elements of $\boldsymbol{R}_{zz}(\tau)=E\{\boldsymbol{z}(t+\tau)\boldsymbol{z}^\top(t)\}$ are absolutely summable. Denote the joint probability density function of $\boldsymbol{d}_z(f_m)$, $m=1,2,\cdots,(n/2)-1$, as $f_{\boldsymbol{\mathcal{D}}}(\boldsymbol{\mathcal{D}})$. Then we have [18, 19]

$$f_{\boldsymbol{\mathcal{D}}}(\boldsymbol{\mathcal{D}}) = \prod_{k=1}^{M}\Bigg[\prod_{\ell=-m_t}^{m_t}\frac{\exp(-g-g^*)}{\pi^{pq}\,|\bar{\boldsymbol{S}}(\tilde{f}_k)\otimes\boldsymbol{\Sigma}|^{1/2}|\bar{\boldsymbol{S}}^*(\tilde{f}_k)\otimes\boldsymbol{\Sigma}|^{1/2}}\Bigg] \tag{15}$$

$$\text{where }\ g = \frac{1}{2}\boldsymbol{d}_z^H(\tilde{f}_{k,\ell})\big(\bar{\boldsymbol{S}}^{-1}(\tilde{f}_k)\otimes\boldsymbol{\Sigma}^{-1}\big)\boldsymbol{d}_z(\tilde{f}_{k,\ell})\,. \tag{16}$$

Using $\text{tr}\big(\boldsymbol{A}^\top\boldsymbol{B}\boldsymbol{C}\boldsymbol{D}^\top\big) = (\text{vec}(\boldsymbol{A}))^\top(\boldsymbol{D}\otimes\boldsymbol{B})\text{vec}(\boldsymbol{C})$ and $|\bar{\boldsymbol{S}}(\tilde{f}_k)\otimes\boldsymbol{\Sigma}| = |\bar{\boldsymbol{S}}(\tilde{f}_k)|^p\,|\boldsymbol{\Sigma}|^q$, and parametrizing in terms of $\boldsymbol{\Phi}_k := \bar{\boldsymbol{S}}^{-1}(\tilde{f}_k)$ and $\boldsymbol{\Omega}=\boldsymbol{\Sigma}^{-1}$, up to some constants the negative log-likelihood follows from (15) as ($\{\boldsymbol{\Phi}\}$ denotes $\{\boldsymbol{\Phi}_k,\ k=$

$1, \cdots, M\}$)

$$-\frac{1}{KMpq}\ln f_{\boldsymbol{\mathcal{D}}}(\boldsymbol{\mathcal{D}}) \propto G(\boldsymbol{\Omega}, \{\boldsymbol{\Phi}\}, \{\boldsymbol{\Phi}^*\}) := -\frac{1}{p}\ln(|\boldsymbol{\Omega}|) - \frac{1}{2Mq}\sum_{k=1}^{M}\big(\ln|\boldsymbol{\Phi}_k| + \ln|\boldsymbol{\Phi}_k^*| - \mathrm{tr}(\boldsymbol{A}_k + \boldsymbol{A}_k^*)\big) \quad (17)$$

$$\text{where } \boldsymbol{A}_k = \frac{1}{Kp}\sum_{\ell=-m_t}^{m_t} \boldsymbol{D}_z^H(\tilde{f}_{k,\ell})\boldsymbol{\Omega}\boldsymbol{D}_z(\tilde{f}_{k,\ell})\boldsymbol{\Phi}_k^*\,. \quad (18)$$

In the high-dimension case, one needs to use penalty terms to enforce sparsity and to make the problem well-conditioned. Imposing a sparse-group lasso sparsity constraint on $\{\boldsymbol{\Phi}\}$ (cf. [2, 21, 22]) and a lasso constraint on $\boldsymbol{\Omega}$, we propose to minimize a penalized version of negative log-likelihood w.r.t. $\boldsymbol{\Omega}$ and $\{\boldsymbol{\Phi}\}$,

$$\mathcal{L}(\boldsymbol{\Omega}, \{\boldsymbol{\Phi}\}) = G(\boldsymbol{\Omega}, \{\boldsymbol{\Phi}\}, \{\boldsymbol{\Phi}^*\}) + P_p(\boldsymbol{\Omega}) + P_q(\{\boldsymbol{\Phi}\}), \quad (19)$$

$$P_p(\boldsymbol{\Omega}) = \lambda_p \sum_{i\neq j}^{p} |\boldsymbol{\Omega}_{ij}| \quad (20)$$

$$P_q(\{\boldsymbol{\Phi}\}) = \alpha\lambda_q \sum_{k=1}^{M}\sum_{i\neq j}^{p}\big|[\boldsymbol{\Phi}_k]_{ij}\big| + (1-\alpha)\sqrt{M}\lambda_q \sum_{i\neq j}^{p}\|\boldsymbol{\Phi}^{(ij)}\| \quad (21)$$

$$\text{where } \boldsymbol{\Phi}^{(ij)} := [[\boldsymbol{\Phi}_1]_{ij}\ [\boldsymbol{\Phi}_2]_{ij}\ \cdots\ [\boldsymbol{\Phi}_M]_{ij}]^\top \in \mathbb{C}^M\,, \quad (22)$$

and $\alpha \in [0,1]$ and $\lambda_p, \lambda_q > 0$ are tuning parameters.

4. OPTIMIZATION

The objective function $\mathcal{L}(\boldsymbol{\Omega}, \{\boldsymbol{\Phi}\})$ in (19) is biconvex: (strictly) convex in $\{\boldsymbol{\Phi}\}$, $\boldsymbol{\Phi}_k \succ \mathbf{0}$, for fixed $\boldsymbol{\Omega}$, and (strictly) convex in $\boldsymbol{\Omega}$, $\boldsymbol{\Omega} \succ \mathbf{0}$, for fixed $\{\boldsymbol{\Phi}\}$. As in [6, 7] (and others) pertaining to the i.i.d. observations case, and as is a general approach for biconvex function optimization [23], we will use an iterative and alternating minimization approach where we optimize w.r.t. $\boldsymbol{\Omega}$ with $\{\boldsymbol{\Phi}\}$ fixed, and then optimize w.r.t. $\{\boldsymbol{\Phi}\}$ with $\boldsymbol{\Omega}$ fixed at the last optimized value, and repeat the two optimizations (flip-flop). There is no guarantee that the algorithm converges to the global minimum, however, the algorithm converges to a local stationary point of $\mathcal{L}(\boldsymbol{\Omega}, \{\boldsymbol{\Phi}\})$ [23].

With $\{\hat{\boldsymbol{\Phi}}\}$ denoting the estimate of $\{\boldsymbol{\Phi}\}$, fix $\{\boldsymbol{\Phi}\} = \{\hat{\boldsymbol{\Phi}}\}$ and let $\mathcal{L}_1(\boldsymbol{\Omega})$ denote $\mathcal{L}(\boldsymbol{\Omega}, \{\hat{\boldsymbol{\Phi}}\})$ up to some irrelevant constants. We minimize $\mathcal{L}_1(\boldsymbol{\Omega})$ w.r.t. $\boldsymbol{\Omega}$ to obtain estimate $\hat{\boldsymbol{\Omega}}$, where

$$\mathcal{L}_1(\boldsymbol{\Omega}) = -\frac{1}{p}\ln(|\boldsymbol{\Omega}|) + \frac{1}{p}\mathrm{tr}\left(\boldsymbol{\Omega}\bar{\boldsymbol{S}}\right) + P_p(\boldsymbol{\Omega})\,, \quad (23)$$

$$\bar{\boldsymbol{S}} = \frac{1}{MKq}\sum_{k=1}^{M}\sum_{\ell=-m_t}^{m_t}\mathcal{R}e\big\{\boldsymbol{D}_z(\tilde{f}_{k,\ell})\hat{\boldsymbol{\Phi}}_k^*\boldsymbol{D}_z^H(\tilde{f}_{k,\ell})\big\}\,. \quad (24)$$

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

$$\mathcal{L}_2(\{\boldsymbol{\Phi}\}) = -\frac{1}{2Mq}\sum_{k=1}^{M}(\ln|\boldsymbol{\Phi}_k| + \ln|\boldsymbol{\Phi}_k^*|) + \frac{1}{2Mq}\sum_{k=1}^{M}\mathrm{tr}\left(\tilde{\boldsymbol{S}}_k\boldsymbol{\Phi}_k + \tilde{\boldsymbol{S}}_k^*\boldsymbol{\Phi}_k^*\right) + P_q(\{\boldsymbol{\Phi}\})\,, \quad (25)$$

$$\tilde{\boldsymbol{S}}_k = \frac{1}{Kp}\sum_{\ell=-m_t}^{m_t}\boldsymbol{D}_z^\top(\tilde{f}_{k,\ell})\hat{\boldsymbol{\Omega}}\boldsymbol{D}_z^*(\tilde{f}_{k,\ell})\,. \quad (26)$$

Our optimization algorithm (used in our simulations) is as follows.

1. Initialize $m = 1$, $\boldsymbol{\Omega}^{(0)} = \boldsymbol{I}_p$, $\boldsymbol{\Phi}_k^{(0)} = \boldsymbol{I}_q$, $k = 1, 2, \cdots, M$.
2. Set $\hat{\boldsymbol{\Omega}} = \boldsymbol{\Omega}^{(m-1)}$ in (26). Use the iterative alternating direction method of multipliers (ADMM) algorithm [24], as outlined in [19, Sec. 4], to minimize $\mathcal{L}_2(\{\boldsymbol{\Phi}\})$ (given by (25)) w.r.t. $\{\boldsymbol{\Phi}\}$ to obtain estimates $\boldsymbol{\Phi}_k^{(m)}$, $k = 1, 2, \cdots, M$. Cost (40) in [19] corresponds to (25) of this paper.
3. Set $\{\hat{\boldsymbol{\Phi}}\} = \{\hat{\boldsymbol{\Phi}}^{(m)}\}$ in (24). Use the ADMM algorithm of [25, Sec. III] (with $\alpha = 1$ therein, no group-lasso penalty) to minimize $\mathcal{L}_1(\boldsymbol{\Omega})$ w.r.t. $\boldsymbol{\Omega}$, to obtain estimate $\boldsymbol{\Omega}^{(m)}$. Cost (7) in [25] (after setting $\alpha = 1$) corresponds to (23) of this paper. Normalize $\hat{\Omega}_{11}^{(m)} = 1$ to resolve the scaling ambiguity. Let $m \leftarrow m + 1$.
4. Repeat steps 2 and 3 until convergence.

4.1. 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 [19]) $\mathrm{BIC}(\lambda_p, \lambda_q, \alpha) = -2KMq\ln|\hat{\boldsymbol{\Omega}}| + 2Kp\sum_{k=1}^{M}\big(-\ln|\hat{\boldsymbol{\Phi}}_k| + p^{-1}\mathrm{tr}\big(\hat{\boldsymbol{A}}_k\big)\big) + \ln(2KM)\big(|\hat{\boldsymbol{\Omega}}|_0/2 + \sum_{k=1}^{M}|\hat{\boldsymbol{\Phi}}_k|_0\big)$ where $\hat{\boldsymbol{A}}_k$ is given by (18) with $\boldsymbol{\Omega}$ and $\boldsymbol{\Phi}_k$ therein replaced with $\hat{\boldsymbol{\Omega}}$ and $\hat{\boldsymbol{\Phi}}_k$, respectively, $|\boldsymbol{J}|_0$ denotes number of nonzero elements in $\boldsymbol{J}$, $2KM$ 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 $(0, \pi)$. 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 [25]. 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 $\lambda_{p\ell}$.

5. CONSISTENCY

Now we provide sufficient conditions for local convergence in the Frobenius norm of the Kronecker-decomposable inverse PSD estimators to the true value. Define $q \times (qM)$ matrix $\tilde{\boldsymbol{\Omega}}$ as

$$\tilde{\boldsymbol{\Omega}} = [\boldsymbol{\Phi}_1\ \boldsymbol{\Phi}_2\ \cdots\ \boldsymbol{\Phi}_M]\,. \quad (27)$$

We now allow p, q, M, K (see (12)), λ_p and λ_q to be functions of sample size n, denoted as p_n, q_n, M_n, K_n, λ_{pn} and λ_{qn}, respectively. Assume

(A1) 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 $k, \ell \in [q]$.

(A2) Define the true edgesets $\mathcal{S}_q = \{\{i,j\} : [\bar{\boldsymbol{S}}_0^{-1}(f)]_{ij} \not\equiv 0,\ i \neq j,\ 0 \leq f \leq 0.5,\ i,j \in [q]\}$ and $\mathcal{S}_p = \{\{i,j\} : \boldsymbol{\Omega}_{ij} \not/ 0,\ i \neq j,\ i,j \in [p]\}$, where $\bar{\boldsymbol{S}}_0(f)$ denotes DTFT of $\boldsymbol{\Psi}(\tau)$ and $\boldsymbol{\Omega}_0 = \boldsymbol{\Sigma}_0^{-1}$ denotes the true value of $\boldsymbol{\Omega}$. Assume that $|\mathcal{S}_q| \leq s_{qn}$ and $|\mathcal{S}_p| \leq s_{pn}$.

(A3) The minimum and maximum eigenvalues of $q_n \times q_n$ PSD $\bar{\boldsymbol{S}}_0(f) \succ \mathbf{0}$ satisfy $0 < \beta_{q,\min} \leq \min_{f\in[0,0.5]}\phi_{\min}(\bar{\boldsymbol{S}}_0(f))$ and $\max_{f\in[0,0.5]}\phi_{\max}(\bar{\boldsymbol{S}}_0(f)) \leq \beta_{q,\max} < \infty$. Similarly, $0 < \beta_{p,\min} \leq \phi_{\min}(\boldsymbol{\Sigma}_0) \leq \phi_{\min}(\boldsymbol{\Sigma}_0) \leq \beta_{p,\max} < \infty$. Here $\beta_{\min}$ and $\beta_{\max}$ are not functions of n (or p_n, q_n).

Theorem 1 whose proof is omitted for lack of space, establishes consistency of a local minimizer $(\hat{\boldsymbol{\Omega}}, \hat{\tilde{\boldsymbol{\Omega}}})$ of $\mathcal{L}(\boldsymbol{\Omega}, \{\boldsymbol{\Phi}\}) = \mathcal{L}(\boldsymbol{\Omega}, \tilde{\boldsymbol{\Omega}})$

under assumptions (A1)-(A3). First we define some variables. For $\tau > 2$, define

$$C_{0q} = 80 \max_{\ell,f}([\bar{\boldsymbol{S}}_0(f)]_{\ell\ell})\sqrt{2\ln(16q_n^{\tau}M_n)/\ln(q_n)}\,, \quad (28)$$

$$C_{0p} = 40 \max_{k}([\boldsymbol{\Sigma}_0]_{\ell\ell})\sqrt{2\ln(4p_n^{\tau})/\ln(p_n)}\,, \quad (29)$$

$$r_{qn} = \sqrt{M_n(q_n + s_{qn})\,\ln(q_n)/(K_n p_n)} = o(1)\,, \quad (30)$$

$$r_{pn} = \sqrt{(p_n + s_{pn})\,\ln(p_n)/(M_n K_n q_n)} = o(1)\,. \quad (31)$$

Recall that for random vectors $\boldsymbol{y}_n$ and $\boldsymbol{x}_n$, the notation $\boldsymbol{y}_n = \mathcal{O}_P(\boldsymbol{x}_n)$ means that for any $\varepsilon > 0$, there exist real R and integer N, $0 < R < \infty$ and $0 < N < \infty$, such that $P(\|\boldsymbol{y}_n\| \leq R\|\boldsymbol{x}_n\|) \geq 1 - \varepsilon\ \forall n \geq N$.

Theorem 1 (Consistency). For $\alpha \in [0,1]$, any $C_1 \geq 1$, and C_{0p} and C_{0q} as defined in (28) and (29), respectively, suppose λ_{pn} and λ_{qn} satisfy

$$\frac{C_{0p}}{p_n}\sqrt{\frac{\ln(p_n)}{M_n K_n q_n}} \leq \lambda_{pn} \leq \frac{C_1 C_{0p}}{p_n}\sqrt{\left(1 + \frac{p_n}{s_{pn}}\right)\frac{\ln(p_n)}{M_n K_n q_n}}\,,$$

$$\frac{C_{0q}}{M_n q_n}\sqrt{\frac{\ln(q_n)}{K_n p_n}} \leq \lambda_{qn} \leq \frac{C_1 C_{0q}}{M_n q_n}\sqrt{\left(1 + \frac{q_n}{s_{qn}}\right)\frac{\ln(q_n)}{K_n p_n}}\,.$$

Then under assumptions (A1)-(A3), there exists a local minimizer $(\hat{\boldsymbol{\Omega}}, \hat{\tilde{\boldsymbol{\Omega}}})$ of $\mathcal{L}(\boldsymbol{\Omega}, \tilde{\boldsymbol{\Omega}})$ such that

$$\|\hat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}_0\|_F = \mathcal{O}_P\left(r_{pn}\right)\,, \quad \|\hat{\tilde{\boldsymbol{\Omega}}} - \tilde{\boldsymbol{\Omega}}_0\|_F = \mathcal{O}_P\left(r_{qn}\right) \quad (32)$$

where r_{pn} and r_{qn} are as in (30) and (31), respectively. •

Remark 1. Proof of Theorem 1 is patterned after [11] pertaining to matrix graphs, exploiting the results in [19] for dependent vector time series and in [25] for multi-attribute graphical models with i.i.d. data; in turn, all these results are based on the proof technique of [26]. Theorem 1 helps determine how to choose M_n and K_n so that for given n, q_n and p_n, $\lim_{n\to\infty} r_{pn} = 0$ and $\lim_{n\to\infty} r_{qn} = 0$. See also [19, Remark 2]. □

6. NUMERICAL RESULTS

We use model (7)-(8) to generate synthetic data where $\boldsymbol{\Psi}(\tau)$ is controlled via $\boldsymbol{B}_i$'s as in (9) and $\boldsymbol{\Sigma}$ is determined via an Erdös-Rènyi graph. We take $p = q = 15$ and for $\boldsymbol{\Psi}(\tau)$, $\boldsymbol{B}_i$'s in (7) have a block-diagonal structure with 5 blocks, each block as in (9), where in each 3×3 block, a, b are uniform over $[-0.3, 0.7]$, L is uniform over $\{1, 2, 3, 4, 5\}$. In the Erdös-Rènyi graph with $p = 15$ nodes, the nodes are connected with probability $p_{er} = 0.05$. In the upper triangular $\bar{\boldsymbol{\Omega}}$, $\bar{\boldsymbol{\Omega}}_{ij} = 0$ if $\{i,j\} \notin \mathcal{S}_p$, $\bar{\boldsymbol{\Omega}}_{ij}$ is uniformly distributed over $[-0.4, -0.1] \cup [0.1, 0.4]$ if $\{i,j\} \in \mathcal{S}_p$, and $\bar{\boldsymbol{\Omega}}_{ii} = 0.5$. With $\bar{\boldsymbol{\Omega}} = \bar{\boldsymbol{\Omega}}^{\top}$, add $\kappa\boldsymbol{I}$ to $\bar{\boldsymbol{\Omega}}$ with κ picked to make minimum eigenvalue of $\boldsymbol{\Omega} = \bar{\boldsymbol{\Omega}} + \kappa\boldsymbol{I}$ equal to 0.5. Let $\boldsymbol{\Omega} = \tilde{\boldsymbol{F}}\tilde{\boldsymbol{F}}$ (matrix square-root), then $\boldsymbol{F} = \tilde{\boldsymbol{F}}^{-1}$ in (7).

We applied our proposed approach with $n = 256$, $M = 2$, $K = 63$ and compared with the approach of [6] (which is also the approach of [7, 14], all of whom assume i.i.d. observations and have two lasso penalties one each on $\boldsymbol{\Omega}$ and $\boldsymbol{\Gamma}$, counterpart to our $\boldsymbol{\Phi}_k$). By changing the penalty parameters and determining the resulting edges, 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 $\boldsymbol{\Omega}$ and $\{\boldsymbol{\Phi}_k\}/\boldsymbol{\Gamma}$. The receiver operating characteristic (ROC) is shown in Fig. 1 based on 100 runs. Fig. 1 shows that the i.i.d. modeling of [6,7,14] is unable to capture the "dependent" edges (cf. (7)) via $\boldsymbol{\Gamma}$ whereas it has no issues with $\boldsymbol{\Omega}$. Our approach works well for both components of the graph Kronecker product.

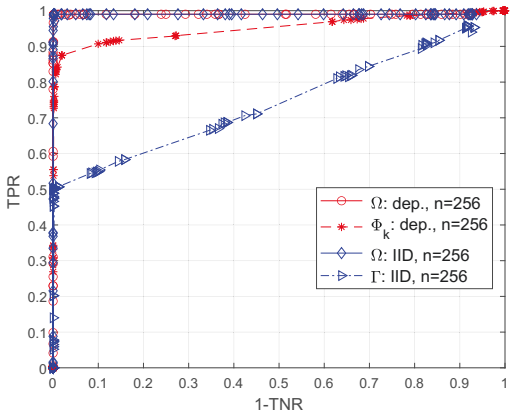

Fig. 1: ROC curves: plots labeled "IID" are from the approach of [6, 7, 14], and the plots labeled "dep." are from our proposed approach. TPR=true positive rate, TNR=true negative rate

In Fig. 2 we show the results based on 50 runs for our approach when BIC parameter selection method (Sec. 4.1) is applied. We take $n = 64, 128, 256, 512, 1024$ with corresponding m_t values as either $m_t = 7, 15, 31, 63, 127$ ($M = 2$), or $m_t = 3, 7, 14, 31, 63$ ($M = 4$); note $K = 2m_t + 1$. Here we show the TPR, 1-TNR and F_1score values for the overall graph (not the two Kronecker product components separately) along with the $\pm\sigma$ error bars. The proposed approach works well both in terms of F_1 score and TPR vs 1-TNR.

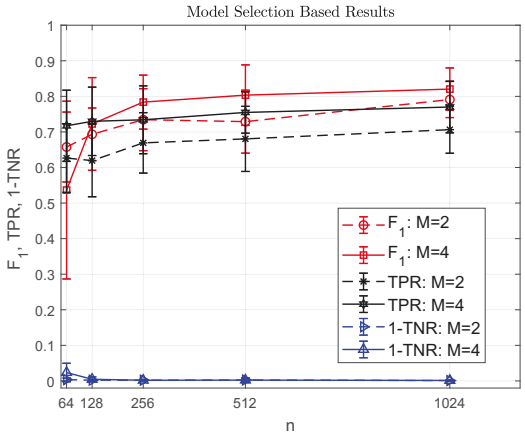

Fig. 2: BIC based results of the proposed approach: F_1-scores, TPR and 1-TNR

7. CONCLUSIONS

Sparse-group lasso penalized log-likelihood approach in frequency-domain with a Kronecker-decomposable PSD was presented for matrix graph 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 a numerical example where our approach significantly outperformed an existing i.i.d. modeling-based approach [6, 7, 14] in correctly detecting the graph edges with ROC as the performance metric.

8. REFERENCES

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

[2] P. Danaher, P. Wang and D.M. Witten, "The joint graphical lasso for inverse covariance estimation across multiple classes," *J. Royal Statistical Society, Series B (Methodological)*, vol. 76, pp. 373-397, 2014.

[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] K. Mohan, P. London, M. Fazel, D. Witten and S.I. Lee, "Node-based learning of multiple Gaussian graphical models," *J. Machine Learning Research*, vol. 15, pp. 445-488, 2014.

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

[6] C. Leng and C.Y. Tang, "Sparse matrix graphical models," *J. American Statistical Association*, vol. 107, pp. 1187-1200, Sep. 2012.

[7] 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.

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

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

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

[11] 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 and Machine Intelligence*, vol. 42, no. 8, pp. 2024-2037, 1 Aug. 2020.

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

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

[14] 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.

[15] 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.

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

[17] 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.

[18] 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.

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

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

[21] 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.

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

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

[24] 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.

[25] 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.)

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