2024 IEEE 34th International Workshop on Machine Learning for Signal Processing (MLSP) | 979-8-3503-7225-0/24/$31.00 ©2024 IEEE | DOI: 10.1109/MLSP58920.2024.10734734

CONDITIONAL INDEPENDENCE GRAPH ESTIMATION FROM MULTI-ATTRIBUTE DEPENDENT TIME SERIES

Jitendra K. Tugnait

Department of Electrical & Computer Engineering
Auburn University, Auburn, AL 36849, USA
tugnajk@auburn.edu

ABSTRACT

Estimation of the conditional independence graph (CIG) of high-dimensional multivariate Gaussian time series from multi-attribute data is considered. All existing methods for graph estimation for such data are based on single-attribute models where one associates a scalar time series with each node. In multi-attribute graphical models, each node represents a random vector or vector time series. In this paper we provide a unified theoretical analysis of multi-attribute graph learning for dependent time series using a penalized log-likelihood objective function. We consider both convex (sparse-group lasso) and non-convex (log-sum and SCAD group penalties) penalty/regularization functions. We establish sufficient conditions in a high-dimensional setting for consistency (convergence of the inverse power spectral density to true value in the Frobenius norm), local convexity when using non-convex penalties, and graph recovery. We illustrate our approach using numerical examples utilizing both synthetic and real data.

Index Terms— Sparse graph learning, multi-attribute graphs, time series, undirected graph, inverse spectral density estimation.

1. INTRODUCTION

Graphical models are a useful tool for analyzing multivariate data where conditional independence is a central concept [1–4]. Consider a 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 $\mathcal{E} \subseteq [p] \times [p]$. Also consider a stationary (real-valued), zero-mean, $p-$dimensional multivariate Gaussian time series $\boldsymbol{x}(t)$, $t = 0, \pm1, \pm2, \cdots$, with ith component $x_i(t)$, and correlation (covariance) matrix function $\boldsymbol{R}_{xx}(\tau) = \mathbb{E}\{\boldsymbol{x}(t+\tau)\boldsymbol{x}^T(t)\}$, $\tau = 0, \pm1, \cdots$. Given $\{\boldsymbol{x}(t)\}$, in the corresponding graph $\mathcal{G}$, each component series $\{x_i(t)\}$ is represented by a node (i in V), and associations between components $\{x_i(t)\}$ and $\{x_j(t)\}$ are represented by edges between nodes i and j of $\mathcal{G}$. In a conditional independence graph (CIG), there is no edge between nodes i and j (i.e., $\{i,j\} \notin \mathcal{E}$) if and only if (iff) $x_i(t)$ and $x_j(t)$ are conditionally independent given the remaining p-2 scalar series $x_\ell(t)$, $\ell \in [p]$, $\ell \neq i$, $\ell \neq j$. (This is a generalization of CIG for random vectors where $\{i,j\} \notin \mathcal{E}$ iff $\Omega_{ij} = 0$; $\boldsymbol{\Omega} = (E\{\boldsymbol{x}(t)\boldsymbol{x}^\top(t)\})^{-1}$ is the precision matrix.)

Denote the power spectral density (PSD) matrix of $\{\boldsymbol{x}(t)\}$ by $\boldsymbol{S}_x(f)$, where $\boldsymbol{S}_x(f) = \sum_{\tau=-\infty}^{\infty} \boldsymbol{R}_{xx}(\tau) e^{-\iota 2\pi f \tau}$ and $\iota = \sqrt{-1}$. 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. Hence one can use estimated inverse PSD of observed time series to infer the associated graph. Nonparametric approaches for graphical modeling of time series in high-dimensional settings (sample size n is smaller than or of the order of p) have been formulated in frequency-domain in [6–11] using group lasso penalties. A sparse-group non-convex log-sum penalty is investigated in [12] to regularize the problem, motivated by [13].

In many applications, there may be more than one random variable (or scalar time series) associated with a node. This class of graphical models has been called multi-attribute graphical models in [14, 15]. Such models have been considered in the literature only for random vectors, not for time series graphical models. The objective of this paper is to fill this gap. In this paper we provide a unified theoretical analysis of multi-attribute graph learning for dependent time series using a penalized log-likelihood objective function. We consider both convex (sparse-group lasso [16,17]) and non-convex (log-sum [13] and Smoothly Clipped Absolute Deviation (SCAD) [18,19] group penalties) penalty functions. It is well-known that use of non-convex penalties can yield more accurate results, i.e., they can produce sparse set of solution like lasso, and approximately unbiased coefficients for large coefficients, unlike lasso [13, 18, 19].

Notation. The superscripts $*$, $\top$ and H denote the complex conjugate, transpose and Hermitian (conjugate transpose) operations, respectively, and the sets of real and complex numbers are denoted by $\mathbb{R}$ and $\mathbb{C}$, respectively. Given $\boldsymbol{A} \in \mathbb{C}^{p\times p}$, we use $\phi_{\min}(\boldsymbol{A})$, $\phi_{\max}(\boldsymbol{A})$, $|\mathbf{A}|$, tr$(\boldsymbol{A})$ and etr$(\boldsymbol{A})$ to 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 $\|\boldsymbol{B}\| = \sqrt{\phi_{\max}(\boldsymbol{B}^H\boldsymbol{B})}$, $\|\boldsymbol{B}\|_F = \sqrt{\text{tr}(\boldsymbol{B}^H\boldsymbol{B})}$ 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 $\|\boldsymbol{\theta}\|_2$. 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 $\boldsymbol{\Sigma}$.

2. SYSTEM MODEL

Consider p jointly Gaussian, zero-mean stationary, vector sequences $\{\boldsymbol{z}_i(t)\}_{t\in\mathbb{Z}}$, $\boldsymbol{z}_i(t) \in \mathbb{R}^m$, $i \in [p]$. In a multi-attribute time series graphical model, we associate $\{\boldsymbol{z}_i(t)\}_{t\in\mathbb{Z}}$ with the ith node of an undirected graph $\mathcal{G} = (V, \mathcal{E})$ where $V = [p]$ is the set of p

This work was supported by NSF Grant CCF-2308473.

nodes (vertices) and $\mathcal{E} \subseteq V \times V$ is the set of undirected edges that describe the conditional dependencies among the p sequences $\{\{\boldsymbol{z}_i(t)\}_{t\in\mathbb{Z}},\ i \in V\}$. Similar to the scalar case ($m = 1$), edge $\{i,j\} \notin \mathcal{E}$ iff the sequences $\{\boldsymbol{z}_i(t)\}$ and $\{\boldsymbol{z}_j(t)\}$ are conditionally independent given the remaining $p-2$ vector sequences $\{\boldsymbol{z}_\ell(t)\}$, $\ell \in V\backslash\{i,j\}$.

Define the mp-dimensional sequence

$$\boldsymbol{x}(t) = \left[\boldsymbol{z}_1^\top(t),\ \boldsymbol{z}_2^\top(t),\ \cdots,\ \boldsymbol{z}_m^\top(t)\right]^\top \in \mathbb{R}^{mp}\,. \tag{1}$$

Associate $\{\boldsymbol{x}(t)\}_{t\in\mathbb{Z}}$ with an enlarged graph $\bar{\mathcal{G}} = \left(\bar{V}, \bar{\mathcal{E}}\right)$ where $\bar{V} = [mp]$ and $\bar{\mathcal{E}} \subseteq \bar{V} \times \bar{V}$. The ℓth component of $\{\boldsymbol{z}_j(t)\}$, denoted by $\{[\boldsymbol{z}_j]_\ell(t)\}$, associated with the node j of $\mathcal{G}$, is the scalar sequence $\{x_q(t)\}$, $x_q = [\boldsymbol{x}]_q$, $q = (j-1)m + \ell$, $j \in [p]$ and $\ell \in [m]$. The scalar sequence $\{x_q(t)\}$ is associated with node q of enlarged graph $\bar{\mathcal{G}}$. Corresponding to the edge $\{j,k\} \in V \times V$ in $\mathcal{G}$, there are m^2 edges $\{q,r\} \in \bar{V} \times \bar{V}$ in $\bar{\mathcal{G}}$ where $q = (j-1)m + u$ and $r = (k-1)m + v$ with $u, v \in [m]$.

As in Sec. 1, denote the power spectral density (PSD) matrix of $\{\boldsymbol{x}(t)\}$ by $\boldsymbol{S}_x(f)$, where $\boldsymbol{S}_x(f) = \sum_{\tau=-\infty}^{\infty} \boldsymbol{R}_{xx}(\tau)e^{-\iota 2\pi f\tau}$ and $\boldsymbol{R}_{xx}(\tau) = \mathbb{E}\{\boldsymbol{x}(t+\tau)\boldsymbol{x}^T(t)\}$. Here f is the normalized frequency, in Hz. Given a matrix $\boldsymbol{A} \in \mathbb{C}^{(mp)\times(mp)}$, we use $\boldsymbol{A}^{(jk)}$ to denote the $m \times m$ submatrix of $\boldsymbol{A}$ whose (u,v)th element is given by

$$[\boldsymbol{A}^{(jk)}]_{uv} = [\boldsymbol{A}]_{(j-1)m+u,(k-1)m+v}\,, \quad u, v \in [m]. \tag{2}$$

By [5, Theorem 2.4], in the conditional independence graph (CIG) $\mathcal{G} = (V, \mathcal{E})$ of the multi-attribute time series $\{\boldsymbol{x}(t)\}_{t\in\mathbb{Z}}$ originating via (1), we have

$$\{j,k\} \notin \mathcal{E} \ \Leftrightarrow\ \left(\boldsymbol{S}_x^{-1}(f)\right)^{(jk)} \equiv 0 \tag{3}$$

provided $\boldsymbol{S}_x(f) \succ \boldsymbol{0}\ \forall f$. (Note that while most of the discussion and all of the numerical results in [5] pertain to scalar time series per node, the theory is shown to apply to vector series per node also.)

2.1. Problem Formulation

Given time-domain data $\{\boldsymbol{x}(t)\}_{t=0}^{n-1}$ originating from a $mp-$dimensional stationary Gaussian sequence, our objective is to first estimate the inverse PSD $\boldsymbol{S}_x^{-1}(f)$ at distinct frequencies, and then select the edge $\{j,k\}$ in the multi-attribute time series graphical model $\mathcal{G}$ based on whether or not $\left(\boldsymbol{S}_x^{-1}(f)\right)^{(jk)} = \boldsymbol{0}$ for every f. The single attribute case ($m = 1$) has been discussed in [11] with group lasso penalty and in [12] with group log-sum penalty. Since for real-valued time series, $\boldsymbol{S}_x(f) = \boldsymbol{S}_x^H(-f)$, and $\boldsymbol{S}_x(f)$ is periodic in f with period one, knowledge of $\boldsymbol{S}_x(f)$ in the interval $[0, 0.5]$ completely specifies $\boldsymbol{S}_x(f)$ for other values of f. Hence, it is enough to check if $\left(\boldsymbol{S}_x^{-1}(f)\right)^{(jk)} = \boldsymbol{0}$ for every $f \in [0, 0.5]$.

Given $\boldsymbol{x}(t)$ for $t = 0, 1, 2, \cdots, n-1$. Define the (normalized) DFT $\boldsymbol{d}_x(f_\ell)$ of $\boldsymbol{x}(t)$, ($\iota = \sqrt{-1}$),

$$\boldsymbol{d}_x(f_\ell) = \frac{1}{\sqrt{n}} \sum_{t=0}^{n-1} \boldsymbol{x}(t) \exp\left(-\iota 2\pi f_\ell t\right) \tag{4}$$

where

$$f_\ell = \ell/n, \quad \ell = 0, 1, \cdots, n-1. \tag{5}$$

Since $\{\boldsymbol{x}(t)\}$ is Gaussian, so is $\boldsymbol{d}_x(f_\ell)$. As discussed in [11], the set of complex-valued random vectors $\{\boldsymbol{d}_x(f_\ell)\}_{\ell=0}^{n/2}$, n even, is a sufficient statistic for any statistical inference problem, including our problem of estimation of inverse PSD.

We need the following assumption in order to invoke [20, Theorem 4.4.1], used extensively later.

(A1) The $mp-$dimensional time series $\{\boldsymbol{x}(t)\}_{t\in\mathbb{Z}}$ is zero-mean stationary and Gaussian, satisfying

$$\sum_{\tau=-\infty}^{\infty} |[\boldsymbol{R}_{xx}(\tau)]_{k\ell}| < \infty \text{ for every } k, \ell \in \bar{V}\,.$$

It follows from [20, Theorem 4.4.1] that under assumption (A1), asymptotically (as $n \to \infty$), $\boldsymbol{d}_x(f_\ell)$, $\ell = 1, 2, \cdots, (n/2) - 1$, ($n$ even), are independent proper (i.e., circularly symmetric), complex Gaussian $\mathcal{N}_c(\boldsymbol{0}, \boldsymbol{S}_x(f_\ell))$ random vectors, respectively. Also, asymptotically, $\boldsymbol{d}_x(f_0)$ and $\boldsymbol{d}_x(f_{n/2})$, ($n$ even), are independent real Gaussian $\mathcal{N}_r(\boldsymbol{0}, \boldsymbol{S}_x(f_0))$ and $\mathcal{N}_r(\boldsymbol{0}, \boldsymbol{S}_x(f_{n/2}))$ random vectors, respectively, independent of $\boldsymbol{d}_x(f_\ell)$, $\ell \in \{1, 2, \cdots, (n/2) - 1\}$. We will ignore these two frequency points f_0 and $f_{n/2}$.

Define

$$\boldsymbol{D} = \left[\boldsymbol{d}_x(f_1)\ \cdots\ \boldsymbol{d}_x(f_{(n/2)-1})\right] \in \mathbb{C}^{(mp)\times((n/2)-1)}\,. \tag{6}$$

We assume that $\boldsymbol{S}_x(f_\ell)$ is locally smooth (a standard assumption in PSD estimation [20]), so that $\boldsymbol{S}_x(f_\ell)$ is (approximately) constant over $K = 2m_t + 1$, $m_t > 0$, frequency points. Pick

$$\tilde{f}_k = \frac{(k-1)K + m_t + 1}{n}, \qquad k = 1, 2, \cdots, M\,, \tag{7}$$

$$M = \left\lfloor \frac{\frac{n}{2} - m_t - 1}{K} \right\rfloor\,, \tag{8}$$

leading to M equally spaced frequencies $\tilde{f}_k$ in the interval $(0, 0.5)$, at intervals of K/n. We state the local smoothness assumption as assumption (A2).

(A2) Assume that for $\ell = -m_t, -m_t + 1, \cdots, m_t$,

$$\boldsymbol{S}_x(\tilde{f}_{k,\ell}) = \boldsymbol{S}_x(\tilde{f}_k)\,, \tag{9}$$

$$\text{where } \tilde{f}_{k,\ell} = \big((k-1)K + m_t + 1 + \ell\big)/n\,. \tag{10}$$

Under assumptions (A1)-(A2), the joint pdf of $\boldsymbol{D}$ is given by

$$f_{\boldsymbol{D}}(\boldsymbol{D}) = \prod_{k=1}^{M} \left[\prod_{\ell=-m_t}^{m_t} \frac{\exp\left(-g_{kl} - g_{kl}^*\right)}{\pi^{mp}\,|\boldsymbol{S}_x^{-1}(\tilde{f}_k)|^{1/2}\,|\boldsymbol{S}_x^{-*}(\tilde{f}_k)|^{1/2}} \right]\,, \tag{11}$$

$$g_{kl} = \frac{1}{2}\boldsymbol{d}_x^H(\tilde{f}_{k,\ell})\boldsymbol{S}_x^{-1}(\tilde{f}_k)\boldsymbol{d}_x(\tilde{f}_{k,\ell})\,, \tag{12}$$

where $\boldsymbol{A}^{-*}$ stands for $(\boldsymbol{A}^{-1})^*$. Parameterizing in terms of the inverse PSD matrix $\boldsymbol{\Phi}_k := \boldsymbol{S}_x^{-1}(\tilde{f}_k)$, the negative log-likelihood, up to some irrelevant constants, is given by

$$-\ln f_{\boldsymbol{D}}(\boldsymbol{D}) \propto \mathcal{L}(\boldsymbol{\Omega}) \tag{13}$$

$$:= \sum_{k=1}^{M} \frac{1}{2}\left[\ln(|\boldsymbol{\Phi}_k|) + \ln(|\boldsymbol{\Phi}_k^*|) - \operatorname{tr}\left(\hat{\boldsymbol{S}}_k\boldsymbol{\Phi}_k + \hat{\boldsymbol{S}}_k^*\boldsymbol{\Phi}_k^*\right)\right] \tag{14}$$

where

$$\boldsymbol{\Omega} = [\boldsymbol{\Phi}_1\,,\ \boldsymbol{\Phi}_2\,,\ \cdots\ ,\boldsymbol{\Phi}_M] \in \mathbb{C}^{(mp)\times(mpM)}\,, \tag{15}$$

$$\hat{\boldsymbol{S}}_k = \frac{1}{K}\sum_{\ell=-m_t}^{m_t} \mathbf{d}_x(\tilde{f}_{k,\ell})\mathbf{d}_x^H(\tilde{f}_{k,\ell})\,. \tag{16}$$

Note that $\hat{\boldsymbol{S}}_k$ represents PSD estimator at frequency $\tilde{f}_k$ using unweighted frequency-domain smoothing [20].

3. PENALIZED NEGATIVE LOG-LIKELIHOOD

In the high-dimensional case of $K < p$, to enforce sparsity and to make the problem well-conditioned, we propose to minimize a penalized version $\bar{\mathcal{L}}(\mathbf{\Omega})$ of $\mathcal{L}(\mathbf{\Omega})$ where we penalize (regularize) at both element-wise and group-wise. We have

$$\bar{\mathcal{L}}(\mathbf{\Omega}) = \mathcal{L}(\mathbf{\Omega}) + \alpha P_e(\mathbf{\Omega}) + (1-\alpha) P_g(\mathbf{\Omega}), \tag{17}$$

$$P_e(\mathbf{\Omega}) = \sum_{k=1}^{M} \sum_{i \neq j}^{mp} \rho_\lambda \left(\left| [\mathbf{\Phi}_k]_{ij} \right| \right), \tag{18}$$

$$P_g(\mathbf{\Omega}) = m\sqrt{M} \sum_{q \neq \ell}^{p} \rho_\lambda \left(\| \mathbf{\Omega}^{(q\ell M)} \|_F \right) \tag{19}$$

where $\mathbf{\Omega}^{(q\ell M)} \in \mathbb{C}^{m \times (mM)}$ is defined as

$$\mathbf{\Omega}^{(q\ell M)} := [\mathbf{\Phi}_1^{(q\ell)},\ \mathbf{\Phi}_2^{(q\ell)},\ \cdots,\ \mathbf{\Phi}_M^{(q\ell)}], \tag{20}$$

$\mathbf{\Phi}_i^{(q\ell)}$, $i \in [M]$, is defined as in (2), $\lambda > 0$, $\alpha \in [0,1]$, $m\sqrt{M}$ in (19) reflects the number of group variables [21], and for $u \in \mathbb{R}$, $\rho_\lambda(u)$ is a penalty function that is function of $|u|$. In (18), the penalty term is applied to each off-diagonal element of $\mathbf{\Phi}_k$ and in (19), the penalty term is applied to the off-block-diagonal group of $m^2 M$ terms via $\mathbf{\Omega}^{(q\ell M)}$, defined in (20). The parameter $\alpha \in [0,1]$ "balances" element-wise and group-wise penalties [11, 16]

The following penalty functions are considered:

- *Lasso*. For some $\lambda > 0$,

$$\rho_\lambda(u) = \lambda |u|, \quad u \in \mathbb{R}. \tag{21}$$

- *Log-sum*. For some $\lambda > 0$ and $1 \gg \epsilon > 0$,

$$\rho_\lambda(u) = \lambda\epsilon \ln\left(1 + \frac{|u|}{\epsilon}\right). \tag{22}$$

- *Smoothly Clipped Absolute Deviation (SCAD)*. For some $\lambda > 0$ and $a > 2$,

$$\rho_\lambda(u) = \begin{cases} \lambda|u| & \text{for } |u| \leq \lambda \\ \frac{2a\lambda|u| - |u|^2 - \lambda^2}{2(a-1)} & \text{for } \lambda < |u| < a\lambda \\ \frac{\lambda^2(a+1)}{2} & \text{for } |u| \geq a\lambda. \end{cases} \tag{23}$$

In the terminology of [22], all of the above three penalties are "μ-amenable" for some $\mu \geq 0$. As defined in [22, Sec. 2.2], $\rho_\lambda(u)$ is μ-amenable for some $\mu \geq 0$ if
(i) The function $\rho_\lambda(u)$ is symmetric around zero, i.e., $\rho_\lambda(u) = \rho_\lambda(-u)$ and $\rho_\lambda(0) = 0$. **(ii)** The function $\rho_\lambda(u)$ is nondecreasing on $\mathbb{R}_+$. **(iii)** The function $\rho_\lambda(u)/u$ is nonincreasing on $\mathbb{R}_+$. **(iv)** The function $\rho_\lambda(u)$ is differentiable for $u \neq 0$. **(v)** The function $\rho_\lambda(u) + \frac{\mu}{2}u^2$ is convex, for some $\mu \geq 0$. **(vi)** $\lim_{u \to 0^+} \frac{d\rho_\lambda(u)}{du} = \lambda$.

It is shown in [22, Appendix A.1], that all of the above three penalties are μ-amenable with $\mu = 0$ for Lasso and $\mu = 1/(a-1)$ for SCAD. In [22] the log-sum penalty is defined as $\rho_\lambda(u) = \ln(1 + \lambda|u|)$ whereas in [13], it is defined as $\rho_\lambda(u) = \lambda \ln\left(1 + \frac{|u|}{\epsilon}\right)$. We follow [13] but modify it so that property (vi) in the definition of μ-amenable penalties holds. In our case $\mu = \frac{\lambda}{\epsilon}$ for the log-sum penalty since $\frac{d^2\rho_\lambda(u)}{du^2} = -\lambda\epsilon/(\epsilon + |u|)^2$ for $u \neq 0$. The above three penalty functions also have the following properties: **(vii)** For some $C_\lambda > 0$ and $\delta_\lambda > 0$, the function $\rho_\lambda(u)$ has a lower bound $\rho_\lambda(u) \geq C_\lambda |u|$ for $|u| \leq \delta_\lambda$. **(viii)** $\rho'_\lambda(|u|) := \frac{d\rho_\lambda(u)}{d|u|} \leq \lambda$ for $u \neq 0$.

Property (viii) is straightforward to verify. For Lasso, $C_\lambda = \lambda$ and $\delta_\lambda = \infty$. For SCAD, $C_\lambda = \lambda$ and $\delta_\lambda = \lambda$. Since $\ln(1+x) \geq x/(1+x)$ for $x > -1$, we have $\ln(1+x) \geq x/C_1$ for $0 \leq x \leq C_1 - 1$, $C_1 > 1$. Take $C_1 = 2$. Then log-sum $\rho_\lambda(u) \geq \frac{\lambda}{2}|u|$ for any $|u| \leq \epsilon$, leading to $C_\lambda = \frac{\lambda}{2}$ and $\delta_\lambda = \epsilon$. We may and will take $C_\lambda = \frac{\lambda}{2}$ for lasso and SCAD penalties as well.

4. OPTIMIZATION

Consider the scaled augmented Lagrangian for this problem [23] after variable splitting, given by

$$\bar{\mathcal{L}}_\rho(\{\mathbf{\Omega}\}, \{\boldsymbol{W}\}, \{\boldsymbol{U}\}) = \mathcal{L}(\{\mathbf{\Omega}\} + \alpha P_e(\boldsymbol{W}) + (1-\alpha) P_g(\boldsymbol{W}) + \frac{\rho}{2} \sum_{k=1}^{M} \|\mathbf{\Phi}_k - \boldsymbol{W}_k + \boldsymbol{U}_k\|_F^2, \tag{24}$$

$$P_e(\boldsymbol{W}) = \sum_{k=1}^{M} \sum_{i \neq j}^{mp} \rho_\lambda \left(\left| [\boldsymbol{W}_k]_{ij} \right| \right), \tag{25}$$

$$P_g(\boldsymbol{W}) = m\sqrt{M} \sum_{q \neq \ell}^{p} \rho_\lambda \left(\| \boldsymbol{W}^{(q\ell M)} \|_F \right) \tag{26}$$

where $\{\boldsymbol{W}\} = \{\boldsymbol{W}_k,\ k = 1, 2, \cdots, M\}$ results from variable splitting where in the penalties we use $\boldsymbol{W}_k$'s instead of $\mathbf{\Phi}_k$'s, adding the equality constraint $\boldsymbol{W}_k = \mathbf{\Phi}_k$, $\{\boldsymbol{U}\} = \{\boldsymbol{U}_k,\ k = 1, 2, \cdots, M\}$ are dual variables, and $\rho > 0$ is the "penalty parameter" [23]. For non-convex $\rho_\lambda(u)$, we use a local linear approximation (LLA) (as in [19, 24]), to yield

$$\rho_\lambda(u) \approx \rho_\lambda(|u_0|) + \rho'_\lambda(|u_0|)(|u| - |u_0|) \Rightarrow \rho'_\lambda(|u_0|)|u|, \tag{27}$$

where u_0 is an initial guess, $\rho'_\lambda(|u_0|) = \lambda\epsilon/(|u_0| + \epsilon)$ for LSP, and for SCAD, $\rho'_\lambda(|u_0|) = \lambda$ for $|u| \leq \lambda$, $= \frac{a\lambda - |u|}{a-1}$ for $\lambda < |u| < a\lambda$, and $= 0$ for $|u| \geq a\lambda$. Therefore, with u_0 fixed, we consider only the last term above for optimization w.r.t. u. By [24, Theorem 1], the LLA provides a majorization of non-convex penalty, thereby yielding a majorization-minimization approach. Thus in LSP, with initial guess $\hat{\boldsymbol{W}}_k$, we replace $\rho_\lambda(|[\boldsymbol{W}_k]_{ij}|) \to \lambda\epsilon/(|[\hat{\boldsymbol{W}}_k]_{ij}| + \epsilon) =: \lambda_{kij}$ and $\rho_\lambda(\|\boldsymbol{W}^{(q\ell M)}\|_F) \to \lambda\epsilon/(\|\hat{\boldsymbol{W}}^{(q\ell M)}\|_F + \epsilon) =: \lambda_{q\ell M}$, leading an adaptive sparse-group lasso convex problem. The initial guess follows from the solution to lasso-penalized objective function.

We follow an ADMM (alternating direction method of multipliers) approach, as outlined in [11], for both lasso and LLA to LSP/SCAD. The main difference between [11] and this paper is that in [11], $\boldsymbol{W}_k$ and $\mathbf{\Phi}_k$ are $p \times p$ whereas in this paper we have $\boldsymbol{W}_k$ and $\mathbf{\Phi}_k$ as $(mp) \times (mp)$ matrices. Therefore, the approach of [11] is applicable after we account for the dimension difference, and additionally for that fact that $P_g(\boldsymbol{W})$ and $P_g(\mathbf{\Omega})$ are penalized slightly differently in the two papers (the factor $m\sqrt{M}$ is missing from [11]). See [11] for further details.For non-convex penalties, we have an iterative solution: first solve with lasso penalty, then use the solution for LLA and solve again the adaptive lasso type convex problem. In practice, just two iterations seem to be enough.

4.1. BIC for Tuning Parameter Selection

Given n and choice of K and M, we follow the Bayesian information criterion (BIC) as given in [11], to select λ (with $\alpha = 0.05$ fixed), for all penalty functions.

5. THEORETICAL ANALYSIS

We now allow p, M, K (see (7), (8)), and λ to be functions of sample size n, denoted as p_n, M_n, K_n and λ_n, respectively. We take p_n to be a non-decreasing function of n, as is typical in high-dimensional settings. Note that $K_n M_n \approx n/2$. Pick $K_n = a_1 n^\gamma$ and $M_n = a_2 n^{1-\gamma}$ for some $0.5 < \gamma < 1$, $0 < a_1, a_2 < \infty$, so that both $M_n/K_n \to 0$ and $K_n/n \to 0$ as $n \to \infty$ (cf. [11, Remark 1]).

Recall that we have the original multi-attribute graph $\mathcal{G} = (V, \mathcal{E})$ with $|V| = p_n$ and the enlarged graph $\bar{\mathcal{G}} = (\bar{V}, \bar{\mathcal{E}})$ with $|\bar{V}| = mp_n$. We assume the following regarding $\mathcal{G}$.

(A3) Denote the true edge set of the graph by $\mathcal{E}_0$, implying that $\mathcal{E}_0 = \{\{j,k\} \,:\, (\boldsymbol{S}_0^{-1}(f))^{(jk)} \not\equiv 0,\ j \neq k,\ 0 \le f \le 0.5\}$ where $\boldsymbol{S}_0(f)$ denotes the true PSD of $\boldsymbol{x}(t)$. (We also use $\boldsymbol{\Phi}_{0k}$ for $\boldsymbol{S}_0^{-1}(\tilde{f}_k)$ where $\tilde{f}_k$ is as in (7), and use $\boldsymbol{\Omega}_0$ to denote the true value of $\boldsymbol{\Omega}$). Assume that $\text{card}(\mathcal{E}_0) = |(\mathcal{E}_0)| \le s_{n0}$.

(A4) The minimum and maximum eigenvalues of $mp_n \times mp_n$ PSD $\boldsymbol{S}_0(f) \succ \boldsymbol{0}$ satisfy

$$0 < \beta_{\min} \le \min_{f\in[0,0.5]} \phi_{\min}(\boldsymbol{S}_0(f)) \le \max_{f\in[0,0.5]} \phi_{\max}(\boldsymbol{S}_0(f)) \le \beta_{\max} < \infty\,.$$

Here $\beta_{\min}$ and $\beta_{\max}$ are not functions of n (or p_n).

Let $\hat{\boldsymbol{\Omega}}_\lambda = \arg\min_{\boldsymbol{\Omega}:\boldsymbol{\Phi}_k \succ \boldsymbol{0}} \bar{\mathcal{L}}(\boldsymbol{\Omega})$. Theorem 1 establishes local consistency of $\hat{\boldsymbol{\Omega}}_\lambda$.

Theorem 1 (Local Consistency). For $\tau > 2$, let

$$C_0 = 80 \max_{\ell,f}([\boldsymbol{S}_0(f)]_{\ell\ell})\sqrt{N_0/\ln(mp_n)} \tag{28}$$

where

$$N_0 = 2\ln(16(mp_n)^\tau M_n)\,. \tag{29}$$

Define

$$R = 8(1+m)C_0/\beta_{\min}^2\,, \tag{30}$$

$$r_n = \sqrt{M_n(mp_n + m^2 s_{n0})\ln(mp_n)/K_n} = o(1)\,, \tag{31}$$

$$N_1 = \arg\min\left\{n \,:\, K_n > N_0\right\}, \tag{32}$$

$$N_2 = \arg\min\left\{n \,:\, r_n \le 0.1/(R\beta_{\min})\right\}, \tag{33}$$

$$N_3 = \arg\min\left\{n \,:\, r_n \le \epsilon/R\right\}, \tag{34}$$

$$\lambda_{n\ell} = 2C_0\sqrt{\ln(mp_n)/K_n}\,, \tag{35}$$

$$\lambda_{nu1} = C_0(1+\frac{1}{m})\sqrt{(m^2 + \frac{mp_n}{s_{n0}})\frac{\ln(mp_n)}{K_n}}\,, \tag{36}$$

$$\lambda_{nu2} = \min\left(Rr_n, \lambda_{nu1}\right)\,. \tag{37}$$

Under assumptions (A1)-(A4), there exists a local minimizer $\hat{\boldsymbol{\Omega}}_\lambda$ of $\bar{\mathcal{L}}(\boldsymbol{\Omega})$ in the neighborhood of $\boldsymbol{\Omega}_0$, satisfying

$$\|\hat{\boldsymbol{\Omega}}_\lambda - \boldsymbol{\Omega}_0\|_F \le Rr_n \tag{38}$$

with probability greater than $1 - 1/(mp_n)^{\tau-2}$ if

(i) for the lasso penalty $\rho_\lambda(t) = \lambda|t|$, sample size $n > \max\{N_1, N_2\}$ and λ_n satisfies $\lambda_{n\ell} \le \lambda_n \le \lambda_{nu1}$,

(ii) for the SCAD penalty $\rho_\lambda(t)$, sample size $n > \max\{N_1, N_2\}$ and the regularization parameter satisfies $\lambda_{n\ell} \le \lambda_n \le \lambda_{nu2}$,

(iii) sample size $n > \max\{N_1, N_2, N_3\}$ and λ_n satisfies $\lambda_{n\ell} \le \lambda_n \le \lambda_{nu1}$ for the log-sum penalty $\rho_\lambda(t)$.

For the lasso penalty, $\hat{\boldsymbol{\Omega}}_\lambda$ is a global minimizer whereas for the other two penalties, it is a local minimizer. •

The proof of Theorem 1 follows for most part from [11, Theorem 1] which is based on the proof technique of [25].

We follow the proof technique of [22, Lemma 6] in establishing Lemma 1.

Lemma 1 (Local Convexity). The optimization problem

$$\hat{\boldsymbol{\Omega}}_\lambda = \arg\min_{\boldsymbol{\Omega}\,:\,\boldsymbol{\Phi}_k \in \mathcal{B}_k} \bar{\mathcal{L}}(\boldsymbol{\Omega})\,, \tag{39}$$

$$\mathcal{B}_k = \left\{\boldsymbol{\Phi}_k \,:\, \boldsymbol{\Phi}_k \succ \boldsymbol{0},\ \|\boldsymbol{\Phi}_k\| \le 0.99\sqrt{2/(m\mu\sqrt{M_n})}\right\},$$

$$\sqrt{2/(m\mu\sqrt{M_n})} = \begin{cases} \infty & : \text{ Lasso} \\ \sqrt{\frac{2(a-1)}{m\sqrt{M_n}}} & : \text{ SCAD} \\ \sqrt{\frac{2\epsilon}{m\sqrt{M_n}\lambda_n}} & : \text{ log-sum,} \end{cases} \tag{40}$$

consists of a strictly convex objective function over a convex constraint set, for all three penalties, where C_0 and λ_n are as defined in Theorem 1. •

Lemma 1 and Theorem 1 lead to Theorem 2.

Theorem 2. Assume the conditions of Theorem 1. Then $\hat{\boldsymbol{\Omega}}_\lambda$ as defined in Lemma 1 is unique, satisfying $\|\hat{\boldsymbol{\Omega}}_\lambda - \boldsymbol{\Omega}_0\|_F \le Rr_n$ with probability $> 1 - 1/(mp_n)^{\tau-2}$ if $Rr_n + 1/\beta_{\min} \le 0.99\sqrt{2/(m\mu\sqrt{M_n})}$, as defined in Lemma 1. •

Sketch of Proof. If $1/\beta_{\min} \le \frac{0.99}{\sqrt{\mu}}$, then $\boldsymbol{\Phi}_{0k} \in \mathcal{B}_k$ since $\|\boldsymbol{\Phi}_{0k}\| \le 1/\beta_{\min}$, and also $\hat{\boldsymbol{\Phi}}_k \in \mathcal{B}_k$ since $\|\hat{\boldsymbol{\Phi}}_k\| \le Rr_n + 1/\beta_{\min}$. Thus, both $\hat{\boldsymbol{\Phi}}_k$ and $\boldsymbol{\Phi}_{0k}$, hence $\hat{\boldsymbol{\Omega}}_\lambda$ and $\boldsymbol{\Omega}_0$, respectively, are feasible. □

We now turn to graph recovery. Define

$$\hat{\mathcal{E}} = \left\{\{q,\ell\} \,:\, \|\hat{\boldsymbol{\Omega}}^{(q\ell)}\|_F > \gamma_n > 0, q \neq \ell\right\}, \tag{41}$$

$$\mathcal{E}_0 = \left\{\{q,\ell\} \,:\, \|\boldsymbol{\Omega}_0^{(q\ell)}\|_F > 0, q \neq \ell\right\}, \tag{42}$$

$$\bar{\sigma}_n = Rr_n\,, \tag{43}$$

$$\nu = \min_{\{q,\ell\}\in\mathcal{E}_0} \|\boldsymbol{\Omega}_0^{(q\ell)}\|_F\,, \tag{44}$$

$$N_4 = \arg\min\left\{n \,:\, \bar{\sigma}_n \le 0.4\nu\right\}, \tag{45}$$

where R and r_n are as in (30) and (31), respectively.

Theorem 3. For $\gamma_n = 0.5\nu$ and $n \ge N_4$, $\hat{\mathcal{E}} = \mathcal{E}_0$ with probability greater than $1 - 1/(mp_n)^{\tau-2}$ under the conditions of Theorem 1. □

The proof of Theorem 3 is omitted for lack of space.

6. NUMERICAL EXAMPLES

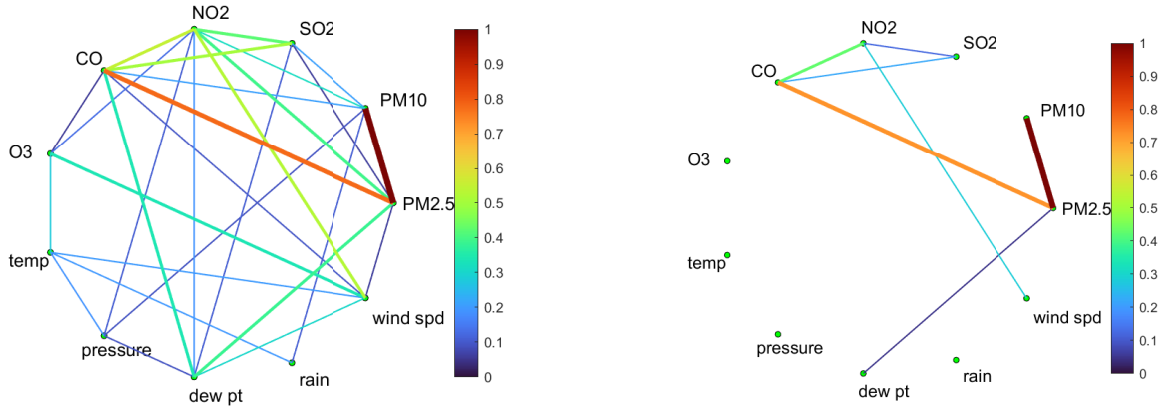
We now present numerical results for both synthetic and real data to illustrate the proposed approach.

6.1. Synthetic Data

Consider a graph with $p = 64$ nodes, each node with $m = 4$ attributes. The time series data $\{\boldsymbol{x}(t)\}$ is generated using a vector autoregressive model of order 3 (VAR(3)):

$$\boldsymbol{x}(t) = \sum_{i=1}^{3} \boldsymbol{A}_i \boldsymbol{x}(t-i) + \boldsymbol{w}(t)\,, \quad \boldsymbol{x}(t) \in \mathbb{R}^{mp}\,,$$

where $\boldsymbol{w}(t)$ is i.i.d. zero-mean Gaussian with precision matrix $\tilde{\boldsymbol{\Omega}}$. We create 8 clusters (communities) of 8 nodes each, each



(a) M=4: Lasso penalty

(b) M=4: Log-sum penalty

Fig. 1: Pollution graphs for the Beijing air-quality dataset [27] for year 2013-14: 8 monitoring sites and 11 features ($m = 8, p = 11, M = 4$, $n = 364$). Number of distinct edges $= 29$ and 7 in graphs (a) and (b), respectively. Estimated $\|\hat{\boldsymbol{\Omega}}^{(ijM)}\|_F$ is the edge weight (normalized to have $\max_{i\neq j} \|\hat{\boldsymbol{\Omega}}^{(ijM)}\|_F = 1$), see (20). The edge weights are color coded , in addition to the edges with higher weights being drawn thicker.

node with $m = 4$ attributes, where nodes within a community are not connected to any node in other communities. We set $[\tilde{\boldsymbol{\Omega}}^{(q\ell)}]_{uv} = 0.5^{|u-v|}$ for $q = \ell \in [8]$, $u \neq v$, $u, v \in [m]$ (notation as in 2), and it is zero otherwise. For $q \neq \ell$, we have $\tilde{\boldsymbol{\Omega}}^{(q\ell)} = \mathbf{0}$. We add $\gamma \boldsymbol{I}_{mp}$ to $\tilde{\boldsymbol{\Omega}}_1$ and choose γ to make the minimum eigenvalue of $\tilde{\boldsymbol{\Omega}}_1 + \gamma \boldsymbol{I}_{mp}$ equal to 0.5 . The parameters of VAR(3) model are generated similarly by having $\boldsymbol{A}_i^{(q\ell)} = \mathbf{0}$ for $q \neq \ell$, and only 10% of the entries of $\boldsymbol{A}_i^{(qq)}$'s are nonzero with the nonzero elements independently and uniformly distributed over $[-0.6, 0.6]$. We then check if the VAR(3) model is stable, a necessary and sufficient condition for which is that the roots of $a(z) = |\boldsymbol{I}_{mp} - \sum_{i=1}^{3} \boldsymbol{A}_i z^{-i}| = 0$ should all have modulus < 1; this condition is equivalent to having all eigenvalues of the corresponding $(3mp) \times (3mp)$ companion matrix to have modulus < 1 [26, Sec. 8.2.3]. Additionally, in order to avoid a "long" impulse response, we require the roots of $a(z)$ to have modulus ≤ 0.95. Suppose this condition is violated with $|z_{\max}| > 0.95$ where $|z_{\max}| = \arg\max_{\ell \in [3mp]} \{|z_\ell| : a(z_\ell) = 0\}$. In this case, we scale $\boldsymbol{A}_i$'s to $\bar{\boldsymbol{A}}_i = \gamma^i \boldsymbol{A}_i, \gamma = 0.95/|z_{\max}|$. It is easy to see that the roots of $\bar{a}(z) = |\boldsymbol{I}_{mp} - \sum_{i=1}^{3} \bar{\boldsymbol{A}}_i z^{-i}| = a(z/\gamma) = 0$ now all have modulus ≤ 0.95. First 100 samples are discarded to eliminate transients. This set-up leads to approximately 11% connected edges. In each run, we calculated the true PSD $\boldsymbol{S}(f)$ for $f \in [0, 0.5]$ at intervals of 0.01, and then take $\{q, \ell\} \in \mathcal{E}$ if $\sqrt{\sum_f \|(\boldsymbol{S}^{-1}(f))^{(q\ell)}\|_F^2} > 10^{-2}(\max_{q,\ell\in[p]} \sqrt{\sum_f \|(\boldsymbol{S}^{-1}(f))^{(q\ell)}\|_F^2})$, else $\{q, \ell\} \notin \mathcal{E}$.

Simulation results based on 100 runs are shown in Table 1 where the performance measure are F_1-score and Hamming distance for efficacy in edge detection. The F_1-score is defined as $F_1 = 2 \times$ precision $\times$ recall/(precision + recall) where precision $= |\hat{\mathcal{E}} \cap \mathcal{E}_0|/|\hat{\mathcal{E}}|$, recall $= |\hat{\mathcal{E}} \cap \mathcal{E}_0|/|\mathcal{E}_0|$, and $\mathcal{E}_0$ and $\hat{\mathcal{E}}$ denote the true and estimated edge sets, respectively. The Hamming distance is between $\hat{\mathcal{E}}$ and $\mathcal{E}_0$, scaled by 0.5 to count only distinct edges. For our proposed approach, we consider $M = 4$ for three samples sizes $n \in \{128, 256, 1024\}$. For $M = 4$, we used $K = 15, 31, 127$ for $n = 128, 256, 1024$, respectively. We fixed $\alpha = 0.05$ and λ was selected by searching over a grid of values to maximize the F_1-score (over 100 runs), or via BIC as in Sec. 4.1 ([11]). We used lasso (convex) or log-sum (non-convex, $\epsilon = 0.0001$) penalties. It is seen that the non-convex penalty outperforms the convex penalty.

Table 1: *F_1 scores and Hamming distances for fixed tuning parameters, for the synthetic data example, averaged over 100 runs.*

n	128	256	1024
M=4: F_1 score $\pm\sigma$: λ's picked to maximize F_1			
Lasso	0.579 ± 0.141	0.765 ± 0.131	0.968 ± 0.035
Log-sum	0.707 ± 0.052	0.868 ± 0.026	0.990 ± 0.008
M=4: Hamming distance $\pm\sigma$: λ's picked to maximize F_1			
Lasso	168.5 ± 040.3	097.4 ± 044.0	013.9 ± 014.7
Log-sum	113.3 ± 012.4	057.7 ± 011.1	004.5 ± 003.3
M=4: F_1 score $\pm\sigma$: λ's picked to minimize BIC			
Log-sum	0.439 ± 0.011	0.663 ± 0.050	0.958 ± 0.053
M=4: Hamming distance $\pm\sigma$: λ's picked to minimize BIC			
Log-sum	500.0 ± 015.9	214.1 ± 050.7	017.2 ± 020.2

6.2. Real Data: Beijing air-quality dataset [27]

Here we consider Beijing air-quality dataset [27, 28], 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 ($= p$) features (pollutants and weather variables). We used data from 8 ($= m$) sites: Changping, Dingling, Huairou, Shunyi Aotizhongxin, Dongsi, Guanyuan, Gucheng. The data are averaged over 24 hour period to yield daily averages $x_i(t)$, $i \in [88]$. We used one year 2013-14 of daily data resulting in $n = 365$ days. We pre-processed the data as follows. Given $x_i(t)$, we transform it to $\bar{x}_i(t) = \ln(x_i(t)/x_i(t-1))$ for each i (leads

to $n = 364$), 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. Fig. 1 shows the CIGs for lasso and log-sum penalties for $M = 4$ where with $\alpha = 0.05$, λ was selected via BIC: an edge exists iff $\|\hat{\mathbf{\Omega}}^{(ijM)}\|_F > 0$. It is seen that lasso yields a much denser graph (29 edges) while the graph resulting from the log-sum penalty is much sparser (7 edges). 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 [27]. This fact is captured by the edge between dew point and $PM_{2.5}$ in Fig. 1.

7. CONCLUSIONS

Estimation of the CIG of high-dimensional multivariate Gaussian time series from multi-attribute data was considered. We provided a unified theoretical analysis of multi-attribute graph learning for dependent time series using a penalized log-likelihood objective function. Both convex and non-convex regularization functions were considered. We established sufficient conditions for consistency, local convexity when using non-convex penalties, and graph recovery. Our approach was illustrated using numerical examples utilizing both synthetic and real (Beijing air-quality dataset) data. Non-convex log-sum regularization yielded more accurate results compared to convex sparse-group lasso regularization for synthetic data, and sparser graph for real data.

8. REFERENCES

[1] J. Whittaker, *Graphical Models in Applied Multivariate Statistics*. New York: Wiley, 1990.

[2] S.L. Lauritzen, *Graphical models*. Oxford, UK: Oxford Univ. Press, 1996.

[3] P. Bühlmann and S. van de Geer, *Statistics for High-Dimensional data*. Berlin: Springer, 2011.

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

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

[6] A. Jung, R. Heckel, H. Bölcskei, and F. Hlawatsch, "Compressive nonparametric graphical model selection for time series," in *Proc. IEEE ICASSP-2014*, Florence, Italy, May 2014.

[7] A. Jung, "Learning the conditional independence structure of stationary time series: A multitask learning approach," *IEEE Trans. Signal Process.*, vol. 63, no. 21, pp. 5677-5690, Nov. 1, 2015.

[8] 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.

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

[10] J.K. Tugnait, "Consistency of sparse-group lasso graphical model selection for time series," in *Proc. 54th Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, CA, Nov. 1-4, 2020, pp. 589-593.

[11] 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.

[12] J.K. Tugnait, "Sparse-group log-sum penalized graphical model learning for time series," in *Proc. ICASSP 2022*, pp. 5822-5826, Singapore, May 22-27, 2022.

[13] E.J. Candès, M.B. Wakin and S.P. Boyd, "Enhancing sparsity by reweighted ℓ_1 minimization," *J. Fourier Anal. Appl.*, vol. 14, pp. 877-905, 2008.

[14] M. Kolar, H. Liu and E.P. Xing, "Graph estimation from multi-attribute data," *J. Machine Learning Research*, vol. 15, pp. 1713-1750, 2014.

[15] 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.)

[16] 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.

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

[18] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *J. American Statistical Assoc.*, vol. 96, pp. 1348-1360, Dec. 2001.

[19] C. Lam and J. Fan, "Sparsistency and rates of convergence in large covariance matrix estimation," *Ann. Statist.*, vol. 37, no. 6B, pp. 4254-4278, 2009.

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

[21] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *J. Royal Statistical Society: Statistical Methodology, Series B*, vol. 68, no. 1, pp. 49-67, 2006. '

[22] P.-L. Loh and M.J. Wainwright, "Support recovery without incoherence: A case for nonconvex regularization," *Annals of Statistics*, vol. 45, pp. 2455-2482, 2017.

[23] 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.

[24] H. Zou and R. Li, "One-step sparse estimates in nonconcave penalized likelihood models," *Ann. Statist.*, vol. 36, no. 4, pp. 1509-1533, 2008.

[25] 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.

[26] R.S. Tsay, *Analysis of Financial Time Series*, 3rd Ed., Hoboken, NJ: John Wiley, 2010.

[27] 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.

[28] 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.