

RESEARCH ARTICLE

Multi-Attribute Graph Estimation With Sparse-Group Non-Convex Penalties

JITENDRA K. TUGNAIT[✉], (Life Fellow, IEEE)

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

e-mail: tugnajk@auburn.edu

This work was supported by the National Science Foundation under Grant CCF-2308473.

ABSTRACT We consider the problem of inferring the conditional independence graph (CIG) of high-dimensional Gaussian vectors from multi-attribute data. Most existing methods for graph estimation are based on single-attribute models where one associates a scalar random variable with each node. In multi-attribute graphical models, each node represents a random vector. In this paper we provide a unified theoretical analysis of multi-attribute graph learning using a penalized log-likelihood objective function. We consider both convex (sparse-group lasso) and sparse-group non-convex (log-sum and smoothly clipped absolute deviation (SCAD) penalties) penalty/regularization functions. An alternating direction method of multipliers (ADMM) approach coupled with local linear approximation to non-convex penalties is presented for optimization of the objective function. For non-convex penalties, theoretical analysis establishing local consistency in support recovery, local convexity and precision matrix estimation in high-dimensional settings is provided under two sets of sufficient conditions: with and without some irrepresentability conditions. We illustrate our approaches using both synthetic and real-data numerical examples. In the synthetic data examples the sparse-group log-sum penalized objective function significantly outperformed the lasso penalized as well as SCAD penalized objective functions with F_1 -score and Hamming distance as performance metrics.

INDEX TERMS Graph learning, inverse covariance estimation, undirected graph, sparse-group lasso, multi-attribute graphs, sparse-group log-sum and SCAD penalties.

I. INTRODUCTION

Graphical models provide a powerful tool for analyzing multivariate data [1], [2]. In an undirected graphical model, the conditional dependency structure among p random variables x_1, x_2, \dots, x_p , ($\mathbf{x} = [x_1 \ x_2 \ \dots \ x_p]^\top$), is represented using an undirected graph $\mathcal{G} = (V, \mathcal{E})$, where $V = \{1, 2, \dots, p\} = [p]$ is the set of p nodes corresponding to the p random variables x_i 's, and $\mathcal{E} \subseteq [p] \times [p]$ is the set of undirected edges describing conditional dependencies among x_i 's. The graph \mathcal{G} is a conditional independence graph (CIG) where there is no edge between nodes i and j if and only if (iff) x_i and x_j are conditionally independent given the remaining $p-2$ variables.

The associate editor coordinating the review of this manuscript and approving it for publication was Sangsoo Lim[✉].

Gaussian graphical models (GGMs) are CIGs where \mathbf{x} is multivariate Gaussian. Suppose \mathbf{x} has positive-definite covariance matrix Σ with inverse covariance matrix $\Omega = \Sigma^{-1}$. Then Ω_{ij} , the (i, j) -th element of Ω , is zero iff x_i and x_j are conditionally independent. Given n samples of \mathbf{x} , in high-dimensional settings, one estimates Ω under some sparsity constraints; see, e.g., [3], [4], [5], and [6]. In these graphs each node represents a scalar random variable. In many applications, there may be more than one random variable associated with a node. This class of graphical models has been called multi-attribute graphical models in [7], [8], and [9] and vector graphs or networks in [10] and [11]. In [8], a sparse-group lasso [12], [13] based penalized log-likelihood approach for graph learning from multi-attribute data was presented whereas [7] considers only group lasso [14]. Both sparse-group lasso and group lasso are convex penalties. It is well-known that use of non-convex penalties such as

smoothly clipped absolute deviation (SCAD) [4], [15] or log-sum [16], can yield more accurate results. Such penalties can produce sparse set of solution like lasso, and approximately unbiased coefficients for large coefficients, unlike lasso.

The objective of this paper is to investigate use of sparse-group non-convex log-sum and SCAD penalties for estimation of multi-attribute graphs.

A. RELATED WORK

Although non-convex penalties have been extensively used for graph estimation (see [4], [17], [18], [19], [20], [21] and references therein), its use for multi-attribute graph estimation is almost non-existent with the exception of [22] where SCAD is investigated. In [8], a sparse-group lasso based penalized log-likelihood approach for graph learning from multi-attribute data was presented whereas [7] considers only group lasso. In sparse-group lasso there is group level lasso penalty as well as element-wise lasso penalty (see equations (6) and (7) in Sec. III in the sequel). We extend the approach of [8] to non-convex sparse-group penalties. For analysis, we rely on the proof technique of [23] as well as the primal-dual witness technique of [5], both originally used in the context of element-wise lasso penalty for single-attribute graphs. The technique of [5] was extended to group-lasso penalty for multi-attribute graphs in [7]. The SCAD penalty for multi-attribute graphs has been considered in [22] but it does not have counterparts to our Lemma 1 and Theorems 2-5. Moreover, the sparse-group SCAD penalty used in this paper is different than that in [22]. In this paper we apply the primal-dual witness technique in a sparse-group non-convex penalty setting. Some prior results in an element-wise non-convex penalty setting are in [20], [21].

For numerical optimization of the penalized log-likelihood we exploit an alternating direction method of multipliers (ADMM) approach [24] as in [8] and [22], where for non-convex penalties (SCAD or log-sum), we use a local linear approximation of the penalties ([4], [17], [22]), initialized via sparse-group lasso results of [8].

B. OUR CONTRIBUTIONS

In this paper we provide a unified theoretical analysis of multi-attribute graph learning using a penalized log-likelihood objective function where both convex (sparse-group lasso) and non-convex (sparse-group log-sum and SCAD) penalty/regularization functions are considered. We establish sufficient conditions in a high-dimensional setting for consistency (convergence of the precision matrix to true value in the Frobenius norm), local convexity when using non-convex penalties, and graph recovery. Two alternative sets of sufficient conditions are investigated, with and without some irrepresentability conditions. For the latter, we follow the proof technique of [23] (as in [8]), and for the former, we follow the primal-dual witness technique of [5] (as in [7]), both applied in a sparse-group setting. Theoretical results not relying on irrepresentability

conditions are in Theorems 1-3 and that based on some irrepresentability conditions are in Theorems 4 and 5. While the non-convex penalized log-likelihood objective function results in a non-convex optimization problem, Theorems 2 and 4 specify conditions under which it becomes a convex optimization problem (see Remark 2 in Sec. V-A). These conditions favor log-sum penalty over SCAD.

A preliminary version of parts of this paper appears in a conference paper [25]. Theorems 4-5, proof of Theorem 3, and synthetic and real data examples do not appear in [25]. Only sketches of proofs of Theorems 1 and 2 appear in [25].

C. OUTLINE

The rest of the paper is organized as follows. The system model is presented in Sec. II where we describe the multi-attribute graphical model with m random variables per node, and also an associated larger single-attribute graph. A penalized log-likelihood objective function is presented in Sec. III for estimation of multi-attribute graph using non-convex penalties. An ADMM approach coupled with a local linear approximation to non-convex penalties is presented for optimization of the objective function in Sec. IV. In Sec. V we present sufficient conditions in a high-dimensional setting for consistency, local convexity when using non-convex penalties, and graph recovery. Numerical results based on synthetic as well as real data are presented in Sec. VI to illustrate the proposed approach. Proofs of Theorems 1-5 and related technical lemmas are given in Appendices A, B and C.

D. NOTATION

For a set V , $|V|$ or $\text{card}(V)$ denotes its cardinality. The abbreviations w.p. and w.h.p. stand for with probability and with high probability, respectively. Given $A \in \mathbb{R}^{p \times p}$, we use $\phi_{\min}(A)$, $\phi_{\max}(A)$, $|A|$ and $\text{tr}(A)$ to denote the minimum eigenvalue, maximum eigenvalue, determinant and trace of A , respectively. For $B \in \mathbb{R}^{p \times q}$, we define $\|B\| = \sqrt{\phi_{\max}(B^T B)}$, $\|B\|_F = \sqrt{\text{tr}(B^T B)}$, $\|B\|_1 = \sum_{i,j} |B_{ij}|$, where B_{ij} is the (i, j) -th element of B (also denoted by $[B]_{ij}$), $\|B\|_{\infty} = \max_{i,j} |B_{ij}|$ and $\|B\|_{1,\infty} = \max_i \sum_j |B_{ij}|$. Given $A \in \mathbb{R}^{p \times p}$, $A^+ = \text{diag}(A)$ is a diagonal matrix with the same diagonal as A , and $A^- = A - A^+$ is A with all its diagonal elements set to zero. The notation $y_n = \mathcal{O}_P(x_n)$ for random vectors $y_n, x_n \in \mathbb{R}^p$ means that for any $\varepsilon > 0$, there exists $0 < M < \infty$ such that $P(\|y_n\| \leq M\|x_n\|) \geq 1 - \varepsilon \forall n \geq 1$. The symbols \otimes and \boxtimes denote Kronecker product and Tracy-Singh product [26], respectively. In particular, given block partitioned matrices $A = [A_{ij}]$ and $B = [B_{kl}]$ with submatrices A_{ij} and B_{kl} , Tracy-Singh product yields another block partitioned matrix $A \boxtimes B = [A_{ij} \boxtimes B_{kl}]_{ij} = [[A_{ij} \otimes B_{kl}]_{kl}]_{ij}$ [27]. Given $A = [A_{ij}] \in \mathbb{R}^{mp \times mp}$ with $A_{ij} \in \mathbb{R}^{m \times m}$, $\text{vec}(A) \in \mathbb{R}^{m^2 p^2}$ denotes the vectorization of A which stacks the columns of the matrix A , and

$$\text{bvec}(A) = [(\text{vec}(A_{11}))^T (\text{vec}(A_{21}))^T \cdots (\text{vec}(A_{p1}))^T \\ (\text{vec}(A_{12}))^T \cdots (\text{vec}(A_{p2}))^T \cdots (\text{vec}(A_{pp}))^T]^T.$$

Let $V = [p]$, $T \subseteq V \times V$, $A \in \mathbb{R}^{mp \times mp}$ and let $A^{(k\ell)} \in \mathbb{R}^{m \times m}$ denote an $m \times m$ submatrix of A with k and ℓ indexing some m rows and m columns, respectively, of A . Then A_T denotes the submatrix of A with rows and columns indexed by T , i.e., $A_T = [A^{(k\ell)}]_{(k,\ell) \in T}$. Suppose $\Gamma = A \boxtimes B$ given block partitioned matrices $A = [A_{ij}]$ and $B = [B_{kl}]$. For any two subsets T_1 and T_2 of $V \times V$, Γ_{T_1, T_2} denotes the submatrix of Γ with block rows and columns indexed by T_1 and T_2 , i.e., $\Gamma_{T_1, T_2} = [A_{j\ell} \otimes B_{kq}]_{(j,k) \in T_1, (\ell,q) \in T_2}$. Following [7], an operator $\mathcal{C}(\cdot)$ is used in Sec. V-B. Consider $A \in \mathbb{R}^{mp \times mp}$ with (k, l) th $m \times m$ submatrix $A^{(k\ell)}$. Then $\mathcal{C}(\cdot)$ operates on A as

$$\begin{bmatrix} A^{(11)} & \dots & A^{(1p)} \\ \vdots & \ddots & \vdots \\ A^{(p1)} & \dots & A^{(pp)} \end{bmatrix} \xrightarrow{\mathcal{C}(\cdot)} \begin{bmatrix} \|A^{(11)}\|_F & \dots & \|A^{(1p)}\|_F \\ \vdots & \ddots & \vdots \\ \|A^{(p1)}\|_F & \dots & \|A^{(pp)}\|_F \end{bmatrix}$$

with $\mathcal{C}(A^{(k\ell)}) = \|A^{(k\ell)}\|_F$ and $\mathcal{C}(A) \in \mathbb{R}^{p \times p}$. Now consider $A_1, A_2 \in \mathbb{R}^{mp \times mp}$ with (k, l) th $m \times m$ submatrices $A_1^{(k\ell)}$ and $A_2^{(k\ell)}$, respectively, and Tracy-Singh product $A_1 \boxtimes A_2 \in \mathbb{R}^{(mp)^2 \times (mp)^2}$. Then $\mathcal{C}(\cdot)$ operates on $A_1 \boxtimes A_2$ as $\mathcal{C}(A_1 \boxtimes A_2) \in \mathbb{R}^{p^2 \times p^2}$ with $\mathcal{C}(A_1^{(k_1\ell_1)} \otimes A_2^{(k_2\ell_2)}) = \|A_1^{(k_1\ell_1)} \otimes A_2^{(k_2\ell_2)}\|_F (= \|A_1^{(k_1\ell_1)}\|_F \|A_2^{(k_2\ell_2)}\|_F)$. That is, each $m^2 \times m^2$ submatrix $A_1^{(k_1\ell_1)} \otimes A_2^{(k_2\ell_2)}$ of $A_1 \boxtimes A_2$ is mapped into its Frobenius norm.

II. SYSTEM MODEL

We will call \mathcal{G} considered earlier a *single-attribute graphical model* for \mathbf{x} . Now consider p jointly Gaussian random vectors $\mathbf{z}_i \in \mathbb{R}^m$, $i \in [p]$. We associate \mathbf{z}_i with the i th node of an undirected graph $\mathcal{G} = (V, \mathcal{E})$ where $V = [p]$ and edges in \mathcal{E} describe the conditional dependencies among vectors $\{\mathbf{z}_i, i \in V\}$. As in the scalar case ($m = 1$), there is no edge between node i and node j in \mathcal{G} iff random vectors \mathbf{z}_i and \mathbf{z}_j are conditionally independent given all the remaining random vectors [7]. This is the *multi-attribute Gaussian graphical model* of interest in this paper.

Define the mp -vector

$$\mathbf{x} = [\mathbf{z}_1^\top \mathbf{z}_2^\top \dots \mathbf{z}_p^\top]^\top \in \mathbb{R}^{mp}. \quad (1)$$

Suppose we have n i.i.d. observations $\mathbf{x}(t)$, $t = 1, 2, \dots, n$, of zero-mean \mathbf{x} . Our objective is to estimate the inverse covariance matrix $(\mathbb{E}\{\mathbf{x}\mathbf{x}^\top\})^{-1}$ and to determine if edge $\{i, j\} \in \mathcal{E}$, given data $\{\mathbf{x}(t)\}_{t=1}^n$. Let us associate \mathbf{x} with an “enlarged” graph $\bar{\mathcal{G}} = (\bar{V}, \bar{\mathcal{E}})$, where $\bar{V} = [mp]$ and $\bar{\mathcal{E}} \subseteq \bar{V} \times \bar{V}$. Now $[z_j]_\ell$, the ℓ th component of \mathbf{z}_j associated with node j of $\mathcal{G} = (V, \mathcal{E})$, is the random variable $x_q = [\mathbf{x}]_q$, where $q = (j-1)m + \ell$, $j \in [p]$ and $\ell \in [m]$. The random variable x_q is associated with node q of $\bar{\mathcal{G}} = (\bar{V}, \bar{\mathcal{E}})$. Corresponding to the edge $\{j, k\} \in \mathcal{E}$ in the multi-attribute $\mathcal{G} = (V, \mathcal{E})$, there are m^2 edges $\{q, r\} \in \bar{\mathcal{E}}$ specified by $q = (j-1)m + s$ and $r = (k-1)m + t$, where $s, t \in [m]$. The graph $\bar{\mathcal{G}} = (\bar{V}, \bar{\mathcal{E}})$ is a single-attribute graph. In order for $\bar{\mathcal{G}}$ to reflect the conditional independencies encoded in \mathcal{G} , we must have the equivalence

$$\{j, k\} \notin \mathcal{E} \Leftrightarrow \bar{\mathcal{E}}^{(jk)} \cap \bar{\mathcal{E}} = \emptyset$$

where

$$\bar{\mathcal{E}}^{(jk)} = \{\{q, r\} : q = (j-1)m + s, r = (k-1)m + t, s, t \in [m]\}.$$

Let $\mathbf{R}_{xx} = \mathbb{E}\{\mathbf{x}\mathbf{x}^\top\} \succ \mathbf{0}$ and $\mathbf{\Omega} = \mathbf{R}_{xx}^{-1}$. Define the (j, k) th $m \times m$ subblock $\mathbf{\Omega}^{(jk)}$ of $\mathbf{\Omega}$ as

$$[\mathbf{\Omega}^{(jk)}]_{st} = [\mathbf{\Omega}]_{(j-1)m+s, (k-1)m+t}, \quad s, t \in [m]. \quad (2)$$

It is established in [7, Sec. 2.1] that $\mathbf{\Omega}^{(jk)} = \mathbf{0} \Leftrightarrow \{j, k\} \notin \mathcal{E}$. Since $\mathbf{\Omega}^{(jk)} = \mathbf{0}$ is equivalent to $[\mathbf{\Omega}]_{qr} = 0$ for every $\{q, r\} \in \bar{\mathcal{E}}^{(jk)}$, and since, by [2, Proposition 5.2], $[\mathbf{\Omega}]_{qr} = 0$ iff x_q and x_r are conditionally independent, hence, iff $\{q, r\} \notin \bar{\mathcal{E}}$, it follows that the aforementioned equivalence holds true.

III. PENALIZED NEGATIVE LOG-LIKELIHOOD

Consider a finite set of data comprised of n i.i.d. zero-mean observations $\mathbf{x}(t)$, $t = 1, 2, \dots, n$. Parametrizing in terms of the precision (inverse covariance) matrix $\mathbf{\Omega}$, the negative log-likelihood, up to some irrelevant constants, is given by

$$\mathcal{L}(\mathbf{\Omega}) := -\ln(|\mathbf{\Omega}|) + tr(\hat{\mathbf{\Sigma}}\mathbf{\Omega}) \quad (3)$$

where

$$\hat{\mathbf{\Sigma}} = \frac{1}{n} \sum_{t=1}^n \mathbf{x}(t)\mathbf{x}^\top(t). \quad (4)$$

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

$$\tilde{\mathcal{L}}(\mathbf{\Omega}) = \mathcal{L}(\mathbf{\Omega}) + \alpha P_e(\mathbf{\Omega}) + (1 - \alpha) P_g(\mathbf{\Omega}), \quad (5)$$

$$P_e(\mathbf{\Omega}) = \sum_{i \neq j}^{mp} \rho_\lambda(|[\mathbf{\Omega}]_{ij}|), \quad (6)$$

$$P_g(\mathbf{\Omega}) = m \sum_{q \neq \ell}^p \rho_\lambda(\|\mathbf{\Omega}^{(q\ell)}\|_F) \quad (7)$$

where $\mathbf{\Omega}^{(q\ell)} \in \mathbb{R}^{m \times m}$ is defined as in (2), $\lambda > 0$, $\alpha \in [0, 1]$, m in (7) reflects the number of group variables [14], and for $u \in \mathbb{R}$, $\rho_\lambda(u)$ is a penalty function that is function of $|u|$. In (6), the penalty term is applied to each off-diagonal element of $\mathbf{\Omega}$ and in (7), the penalty term is applied to the off-block-diagonal group of m^2 terms via $\mathbf{\Omega}^{(q\ell)}$. The parameter $\alpha \in [0, 1]$ “balances” element-wise and group-wise penalties [8], [12], [13].

The following penalty functions are considered:

- *Lasso*. For some $\lambda > 0$, $\rho_\lambda(u) = \lambda|u|$, $u \in \mathbb{R}$.
- *Log-sum*. For some $\lambda > 0$ and $1 \gg \epsilon > 0$, $\rho_\lambda(u) = \lambda \epsilon \ln\left(1 + \frac{|u|}{\epsilon}\right)$.

- **SCAD.** For some $\lambda > 0$ and $a > 2$,

$$\rho_\lambda(u) = \begin{cases} \lambda|u|, & |u| \leq \lambda, \\ \frac{2a\lambda|u| - |u|^2 - \lambda^2}{2(a-1)}, & \lambda < |u| < a\lambda \\ \frac{\lambda^2(a+1)}{2}, & |u| \geq a\lambda. \end{cases} \quad (8)$$

In the terminology of [21], all of the above three penalties are “ μ -amenable” for some $\mu \geq 0$. As defined in [21, 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) $\lim_{u \rightarrow 0^+} \frac{d\rho_\lambda(u)}{du} = \lambda$.

It is shown in [21, 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 [21] the log-sum penalty is defined as $\rho_\lambda(u) = \ln(1 + \lambda|u|)$ whereas in [16], it is defined as $\rho_\lambda(u) = \lambda \ln\left(1 + \frac{|u|}{\epsilon}\right)$. We follow [16] 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 following properties also hold for the three penalty functions:

- (vii) For some $C_\lambda > 0$ and $\delta_\lambda > 0$, we have

$$\rho_\lambda(u) \geq C_\lambda|u| \text{ for } |u| \leq \delta_\lambda. \quad (9)$$

- (viii) $\frac{d\rho_\lambda(u)}{d|u|} \leq \lambda$ 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.

We seek $\hat{\Omega} = \arg \min_{\Omega \succ 0} \tilde{\mathcal{L}}(\Omega)$.

IV. OPTIMIZATION

The objective function $\tilde{\mathcal{L}}(\Omega)$ is non-convex for the non-convex SCAD and log-sum penalties. In this section we discuss an ADMM approach, following the ADMM approach given in [8] for sparse group lasso, to attain a local minimum of $\tilde{\mathcal{L}}(\Omega)$ w.r.t. Ω .

For non-convex $\rho_\lambda(u)$, we use a local linear approximation (LLA) to $\rho_\lambda(u)$ as in [4] and [17], to yield

$$\rho_\lambda(u) \approx \rho_\lambda(|u_0|) + \rho'_\lambda(|u_0|)(|u| - |u_0|), \quad (10)$$

where u_0 is an initial guess, and the gradient of the penalty function is

$$\rho'_\lambda(|u_0|) = \begin{cases} \frac{\lambda\epsilon}{|u_0| + \epsilon} & \text{for log-sum,} \\ \lambda, & \text{if } |u_0| \leq \lambda \\ \frac{a\lambda - |u_0|}{a-1}, & \text{if } \lambda < |u_0| \leq a\lambda \\ 0, & \text{if } a\lambda < |u_0| \\ \text{for SCAD.} \end{cases} \quad (11)$$

Therefore, with u_0 fixed, we need to consider only the term dependent upon u for optimization w.r.t. u :

$$\rho_\lambda(u) \Rightarrow \rho'_\lambda(|u_0|)|u|. \quad (12)$$

By [17, Theorem 1], the LLA provides a majorization of the non-convex penalty, thereby yielding a majorization-minimization approach. By [17, Theorem 2], the LLA is the best convex majorization of the LSP and SCAD penalties.

Thus in LSP, with some initial guess $\tilde{\Omega}$, we replace

$$\rho_\lambda(|[\Omega]_{ij}|) \rightarrow \lambda_{e,ij} := \frac{\lambda\epsilon}{|[\tilde{\Omega}]_{ij}| + \epsilon}, \quad (13)$$

$$\rho_\lambda(\|\Omega^{(k\ell)}\|_F) \rightarrow \lambda_{g,k\ell} := \frac{\lambda\epsilon}{\|\tilde{\Omega}^{(k\ell)}\|_F + \epsilon}. \quad (14)$$

The solution $\hat{\Omega}_{lasso}$ to the convex sparse group-lasso-penalized objective function may be used as an initial guess with $\tilde{\Omega} = \hat{\Omega}_{lasso}$. Similarly, for SCAD, we have

$$\lambda_{e,ij} = \begin{cases} \lambda, & \text{if } |[\tilde{\Omega}]_{ij}| \leq \lambda \\ \frac{a\lambda - |[\tilde{\Omega}]_{ij}|}{a-1}, & \text{if } \lambda < |[\tilde{\Omega}]_{ij}| \leq a\lambda \\ 0, & \text{if } a\lambda < |[\tilde{\Omega}]_{ij}| \end{cases}, \quad (15)$$

$$\lambda_{g,k\ell} = \begin{cases} \lambda, & \text{if } \|\tilde{\Omega}^{(k\ell)}\|_F \leq \lambda \\ \frac{a\lambda - \|\tilde{\Omega}^{(k\ell)}\|_F}{a-1}, & \text{if } \lambda < \|\tilde{\Omega}^{(k\ell)}\|_F \leq a\lambda \\ 0, & \text{if } a\lambda < \|\tilde{\Omega}^{(k\ell)}\|_F \end{cases}. \quad (16)$$

With LLA, the original objective function is transformed to its convex LLA approximation

$$\tilde{\mathcal{L}}(\Omega) = \mathcal{L}(\Omega) + \alpha\tilde{P}_e(\Omega) + (1-\alpha)\tilde{P}_g(\Omega), \quad (17)$$

$$\tilde{P}_e(\Omega) = \sum_{i \neq j}^{mp} \lambda_{e,ij} |[\Omega_k]_{ij}|, \quad (18)$$

$$\tilde{P}_g(\Omega) = m \sum_{q \neq \ell}^p \lambda_{g,q\ell} \|\Omega^{(q\ell)}\|_F. \quad (19)$$

For lasso, we have $\lambda_{e,ij} = \lambda \forall i, j$ and $\lambda_{g,q\ell} = \lambda \forall q, \ell$. We follow an ADMM approach, as outlined in [8], for both lasso and LLA to LSP/SCAD. Consider the scaled augmented Lagrangian [24] for this problem after variable splitting, given by

$$\begin{aligned} \tilde{\mathcal{L}}_\rho(\Omega, V, U) &= \mathcal{L}(\Omega) + \alpha\tilde{P}_e(V) \\ &+ (1-\alpha)\tilde{P}_g(V) + \frac{\rho}{2} \|\Omega - V + U\|_F^2, \end{aligned} \quad (20)$$

Algorithm 1 ADMM Algorithm for Sparse-Group Graphical Lasso

Require: Sample covariance $\hat{\Sigma}$ (see (4)), regularization and penalty parameters $\lambda_{e,ij}$ ($i, j \in [mp]$), $\lambda_{g,k\ell}$ ($k, \ell \in [p]$), α and $\rho = \bar{\rho}$, tolerances τ_{abs} and τ_{rel} , variable penalty factor ϕ , maximum number of iterations t_{max} . Initial guess $\bar{\Omega}$.

Ensure: estimated inverse covariance $\hat{\Omega}$ and edge-set $\hat{\mathcal{E}}$

- 1: Initialize: $U^{(0)} = V^{(0)} = \mathbf{0}$, $\Omega^{(0)} = \bar{\Omega}$, where $U, V \in \mathbb{R}^{(mp) \times (mp)}$, $\rho^{(0)} = \bar{\rho}$
- 2: converged = **false**, $t = 0$
- 3: **while** converged = **false** **and** $t \leq t_{max}$, **do**
- 4: Eigen-decompose $\hat{\Sigma} - \rho^{(t)}(V^{(t)} - U^{(t)})$ as $\hat{\Sigma} - \rho^{(t)}(V^{(t)} - U^{(t)}) = PDP^T$ with diagonal matrix D consisting of eigenvalues. Define diagonal matrix \tilde{D} with ℓ th diagonal element $\tilde{D}_{\ell\ell} = (-D_{\ell\ell} + \sqrt{D_{\ell\ell}^2 + 4\rho^{(t)}})/(2\rho^{(t)})$. Set $\Omega^{(t+1)} = P\tilde{D}P^T$.
- 5: Define soft thresholding scalar operator $T_{st}(a, \beta) := (1 - \beta/|a|)_+ a$. Set $A^{(k\ell)} = (\Omega^{(t+1)})^{(k\ell)} + (U^{(t)})^{(k\ell)}$. The diagonal $m \times m$ subblocks of V are updated as

$$[(V^{(t+1)})^{(kk)}]_{uv} = \begin{cases} [A^{(kk)}]_{uu} & \text{if } u = v \\ T_{st}([A_k^{(kk)}]_{uv}, \frac{\alpha\lambda_{e,ij}}{\rho^{(t)}}) & \text{if } u \neq v \end{cases}$$

$k \in [p]$, $u, v \in [m]$, $i = (k-1)m + u$, $j = (k-1)m + v$. The off-diagonal $m \times m$ subblocks of V are updated as

$$(V^{(t+1)})^{(k\ell)} = B \left(1 - \frac{(1-\alpha)m\lambda_{g,k\ell}}{\rho^{(t)}\|B\|_F} \right) +$$

where $k \neq \ell \in [p]$, $m \times m$ B has its (u, v) th element as $[B]_{uv} = T_{st}([A_k^{(k\ell)}]_{uv}, \alpha\lambda_{e,ij}/\rho^{(t)})$, $i = (k-1)m + u$, $j = (\ell-1)m + v$.

- 6: Dual update $U^{(t+1)} = U^{(t)} + (\Omega^{(t+1)} - V^{(t+1)})$.
- 7: Check convergence. Set tolerances

$$\tau_{pri} = mp \tau_{abs} + \tau_{rel} \max(\|\Omega^{(t+1)}\|_F, \|V^{(t+1)}\|_F)$$

$$\tau_{dual} = mp \tau_{abs} + \tau_{rel} \|U^{(t+1)}\|_F / \rho^{(t)}.$$

Define $d_p = \|\Omega^{(t+1)} - V^{(t+1)}\|_F$ and $d_d = \rho^{(t)}\|V^{(t+1)} - V^{(t)}\|_F$. If $(d_p \leq \tau_{pri})$ **and** $(d_d \leq \tau_{dual})$, set converged = **true**.

- 8: Update penalty parameter ρ :

$$\rho^{(t+1)} = \begin{cases} 2\rho^{(t)} & \text{if } d_p > \phi d_d \\ \rho^{(t)}/2 & \text{if } d_d > \phi d_p \\ \rho^{(t)} & \text{otherwise.} \end{cases}$$

We also need to set $U^{(t+1)} = U^{(t+1)}/2$ for $d_p > \phi d_d$ and $U^{(t+1)} = 2U^{(t+1)}$ for $d_d > \phi d_p$.

- 9: $t \leftarrow t + 1$
- 10: **end while**
- 11: For $k \neq \ell$, if $\|V^{(k\ell)}\|_F > 0$, assign edge $\{k, \ell\} \in \hat{\mathcal{E}}$, else $\{k, \ell\} \notin \hat{\mathcal{E}}$. Inverse covariance estimate $\hat{\Omega} = V$.

where $V \in \mathbb{R}^{(mp) \times (mp)}$ results from variable splitting, and in the penalties we use V instead of Ω , adding the equality

constraint $V = \Omega$, U is the dual variable, and $\rho > 0$ is the “penalty parameter” [24].

The main difference between [8] and this paper is the fact that $P_g(V)$ and $P_g(\Omega)$ are penalized slightly differently in the two papers (the factor m is missing from [8]). For non-convex penalties (not considered in [8]), we have an iterative solution: first solve with sparse group-lasso penalty, then use the LLA formulation and solve the resulting adaptive lasso type convex problem. In practice, just two iterations seem to be enough. A pseudo code for the ADMM algorithm used in this paper is given in Algorithm 1 where we use the stopping (convergence) criterion following [24, Sec. 3.3.1] and varying penalty parameter ρ following [24, Sec. 3.4.1]. See [8] for further details; note that by construction, $\Omega^{(t+1)}$ in step 4 of Algorithm 1 is positive definite. Our ADMM-based optimization algorithm is as follows.

1. Calculate sample covariance $\hat{\Sigma}$ as in (4). Initialize iteration $\tilde{m} = 1$, $\Omega^{(0)} = (\text{diag}(\hat{\Sigma}))^{-1}$, $\bar{\Omega} = \Omega^{(0)}$ and use $\bar{\Omega}$ to compute $\lambda_{e,ij}$ ’s and $\lambda_{g,k\ell}$.
2. Execute Algorithm 1 with initial guess $\bar{\Omega}$.
3. Quit if using sparse-group lasso, else set $\Omega^{(\tilde{m})} = \hat{\Omega}$ and $\bar{\Omega} = \Omega^{(\tilde{m})}$ to re-compute $\lambda_{e,ij}$ ’s and $\lambda_{g,k\ell}$ ’s via the LLA. Set $\tilde{m} \leftarrow \tilde{m} + 1$.
4. Repeat steps 2 and 3 until convergence. The converged $\hat{\Omega}$ is the final estimate of the inverse covariance. (For the numerical results shown in Sec. VI, we terminated after two iterations of steps 2 and 3, similar to [4] and [17].)

A. CONVERGENCE AND MODEL SELECTION

In the LLA approach, each approximation yields a convex objective function, therefore, convergence to a global minimum of $\tilde{\mathcal{L}}(\Omega)$ is guaranteed. Overall it is a majorization-minimization approach, hence, after repeated LLA’s, one gets a local minimum of the original non-convex objective function. In practice, two iterations seem to be enough: first run Algorithm 1 for lasso, then using lasso-based LLA, run Algorithm 1 once more.

For model selection we follow the BIC information criterion as discussed in [8]. Let $\hat{\Omega}$ and $\hat{\mathcal{E}} (= \{\{i, j\} : |V_{ij}| > 0, i \neq j\}, V \text{ as in Algorithm 1 after convergence})$ denote the estimated inverse covariance matrix and estimated enlarged edge-set, and let $|\hat{\mathcal{E}}|$ denote the cardinality (# of nonzero elements) of $\hat{\mathcal{E}}$. Noting that $\hat{\Omega}$ is symmetric with nonzero diagonal elements, the number of free nonzero elements of $\hat{\Omega}$ equal $\frac{1}{2}|\hat{\mathcal{E}}| + pm$. The BIC is then given by

$$\text{BIC}(\lambda, \alpha) = \text{tr}(\hat{\Sigma}\hat{\Omega}) - \ln(|\hat{\Omega}|) + \frac{\ln(n)}{n} \left(\frac{1}{2}|\hat{\mathcal{E}}| \right) \quad (21)$$

based on the optimized negative log-likelihood $-\ln f_X(X) \propto \frac{n}{2}(\text{tr}(\hat{\Sigma}\hat{\Omega}) - \ln|\hat{\Omega}|)$, where $X = \{x(t)\}_{t=1}^n$ and $f_X(X)$ is the joint probability density function of X . The pair (λ, α) is selected to minimize BIC. Unlike [8], in this paper to simplify computations, we fix $\alpha = 0.05$ and select λ to minimize BIC($\lambda, 0.05$) by searching over a grid of values for synthetic data. For real data, we search over both λ and α as follows.

Fix $\alpha = 0.05$ and select the best λ by searching over a grid of values, and then with this optimized λ , select the best α by searching over a grid of values in $[0.01, 0.3]$.

In our simulations we search over $\lambda \in [\lambda_\ell, \lambda_u]$, where λ_ℓ and λ_u are selected via a heuristic as in [8]. Find the smallest λ , labeled λ_{sm} for which we get a no-edge model; then we set $\lambda_u = \lambda_{sm}/2$ and $\lambda_\ell = \lambda_u/10$ for both synthetic and real datasets. For the numerical results presented in Sec. VI, we picked $t_{\max} = 200$, $\bar{\rho} = 2$, $\phi = 10$, $\tau_{abs} = \tau_{rel} = 10^{-4}$ in Algorithm 1. For the SCAD penalty $a = 3.7$ (as in [4]) and for the log-sum penalty $\epsilon = 0.0001$.

V. THEORETICAL ANALYSIS

Here we analyze the properties of $\hat{\Omega} = \arg \min_{\Omega \succ 0} \tilde{\mathcal{L}}(\Omega)$. Since the SCAD and log-sum penalties are non-convex, the objective function is non-convex and in general, any optimization of the objective function will yield only a stationary point or a local minimum. We now allow p and λ to be functions of sample size n , denoted as p_n and λ_n , respectively. Recall that we have the original multi-attribute graph $\mathcal{G} = (V, \mathcal{E})$ with $|V| = p_n$ and the corresponding enlarged graph $\tilde{\mathcal{G}} = (\tilde{V}, \tilde{\mathcal{E}})$ with $|\tilde{V}| = mp_n$.

A. ANALYSIS WITHOUT IRREPRESENTABILITY CONDITIONS

We assume the following regarding \mathcal{G} .

- (A1) Denote the true edge set of the graph by \mathcal{E}^* , implying that $\mathcal{E}^* = \{j, k\} : \|(\Omega^*)^{(jk)}\|_F > 0, j \neq k\}$ where Ω^* denotes the true precision matrix of $\mathbf{x}(t)$. Assume that $\text{card}(\mathcal{E}^*) = |\mathcal{E}^*| \leq s_n^*$.
- (A2) The minimum and maximum eigenvalues of $(mp_n) \times (mp_n)$ true covariance $\Sigma^* \succ 0$ satisfy

$$0 < \beta_{\min} \leq \phi_{\min}(\Sigma^*) \leq \phi_{\max}(\Sigma^*) \leq \beta_{\max} < \infty.$$

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

Let $\hat{\Omega} = \arg \min_{\Omega \succ 0} \tilde{\mathcal{L}}(\Omega)$. Theorem 1 establishes local consistency of $\hat{\Omega}$ (the proof is in Appendix A).

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

$$C_0 = 40 \max_{k \in [mp]} \{(\Sigma^*)_{kk}\} \sqrt{N_1 / \ln(p_n)}, \quad (22)$$

$$R = 8(1 + m)C_0 / \beta_{\min}^2, \quad (23)$$

$$r_n = \sqrt{(mp_n + m^2 s_n^*) \ln(p_n) / n} = o(1), \quad (24)$$

$$N_1 = 2 \ln(4m^2 p_n^*), \quad (25)$$

$$N_2 = \arg \min \{n : r_n \leq 0.1 / (R\beta_{\min})\}, \quad (26)$$

$$N_3 = \arg \min \left\{ n : r_n \leq \frac{\epsilon}{R} \right\}, \quad (27)$$

$$N_4 = \arg \min \left\{ n : \lambda_n \leq \frac{\min_{(i,j): [\Omega^*]_{ij} \neq 0} \|[\Omega^*]_{ij}\|}{a + 1} \right\}, \quad (28)$$

$$\lambda_{n\ell} = 2C_0 \sqrt{\ln(p_n) / n}, \quad (29)$$

$$\lambda_{nu1} = C_0(m + 1)r_n / (m\sqrt{s_n^*}), \quad (30)$$

$$\lambda_{nu2} = \min(Rr_n, \lambda_{nu1}). \quad (31)$$

Under assumptions (A1)-(A2), there exists a local minimizer $\hat{\Omega}$ of $\tilde{\mathcal{L}}(\Omega)$ satisfying

$$\|\hat{\Omega} - \Omega^*\|_F \leq Rr_n \quad (32)$$

with probability greater than $1 - 1/p_n^{\tau-2}$ if

- (i) for the lasso penalty $n > \max\{N_1, N_2\}$ and λ_n satisfies $\lambda_{n\ell} \leq \lambda_n \leq \lambda_{nu1}$,
- (ii) for the SCAD penalty $n > \max\{N_1, N_2, N_4\}$ and λ_n satisfies $\lambda_n = \lambda_{nu2}$,
- (iii) for the log-sum penalty $n > \max\{N_1, N_2, N_3\}$ and λ_n satisfies $\lambda_{n\ell} \leq \lambda_n \leq \lambda_{nu1}$.

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

Remark 1 (Convergence Rate): In terms of the rate of convergence, $\|\hat{\Omega} - \Omega^*\|_F = \mathcal{O}_P(r_n) = \mathcal{O}_P(r_n/m)$ for fixed m . Therefore, for $\|\hat{\Omega} - \Omega^*\|_F \rightarrow 0$ as $n \rightarrow \infty$, we must have $(p_n m^{-1} + s_n^*) \ln(p_n) / n \rightarrow 0$. Notice that $mp_n + m^2 s_n^*$ is the maximum number of nonzero elements in Ω^* . □

We follow the proof technique of [21, Lemma 6] in establishing Lemma 1 (the proof is in Appendix B).

Lemma 1 (Local Convexity): The optimization problem

$$\hat{\Omega} = \arg \min_{\Omega \in \mathcal{B}} \tilde{\mathcal{L}}(\Omega), \quad (33)$$

$$\mathcal{B} = \{\Omega : \Omega \succ 0, \|\Omega\| \leq 0.99 \bar{\mu}\}, \quad (34)$$

$$\bar{\mu} = \begin{cases} \infty & : \text{lasso} \\ \sqrt{(a-1)/m} & : \text{SCAD} \\ \sqrt{\epsilon/(m\lambda_n)} & : \text{log-sum}, \end{cases} \quad (35)$$

consists of a strictly convex objective function over a convex constraint set for all three penalties where λ_n is as in Theorem 1. •

Lemma 1 and Theorem 1 lead to Theorem 2, as proved in Appendix B.

Theorem 2: Assume the conditions of Theorem 1. If $Rr_n + 1/\beta_{\min} \leq 0.99 \bar{\mu}$, then $\hat{\Omega}$ as defined in Lemma 1 is a unique minimum, satisfying all results of Theorem 1. •

Remark 2: We see from Theorem 1 that as $n \rightarrow \infty$, $\lambda_n \rightarrow 0$ (since $r_n = o(1)$), therefore, we eventually have “global” convexity for log-sum penalty by (35) for any Ω^* . But such is not the case for SCAD where one may need a to become large in which case it would behave more like lasso. □

We now turn to graph recovery. Define

$$\hat{\mathcal{E}} = \{ \{q, \ell\} : \|\hat{\Omega}^{(q\ell)}\|_F > \theta_n > 0, q \neq \ell \}, \quad (36)$$

$$\mathcal{E}^* = \{ \{q, \ell\} : \|(\Omega^*)^{(q\ell)}\|_F > 0, q \neq \ell \}, \quad (37)$$

$$\bar{\sigma}_n = Rr_n, \quad (38)$$

$$v = \min_{\{q, \ell\} \in \mathcal{E}^*} \|(\Omega^*)^{(q\ell)}\|_F, \quad (39)$$

$$N_4 = \arg \min \{ n : \bar{\sigma}_n \leq 0.4v \}, \quad (40)$$

where R and r_n are as in (23) and (24), respectively. We follow the proof technique of [28, Theorem 10] in establishing Theorem 3 whose proof is in Appendix B.

Theorem 3: For $\theta_n = 0.5\nu$ and $n \geq N_4$, $\hat{\mathcal{E}} = \mathcal{E}^*$ with probability $> 1 - 1/p_n^{\tau-2}$ under the conditions of Theorem 1.

Remark 3: In practice we do not know the value of ν , hence cannot calculate θ_n needed in (36). For the numerical results presented in Sec. VI, we used $\theta_n = 0$. Using some irrepresentability conditions (not needed in Theorem 1) and the primal-dual witness method, in Theorem 4(iv) of Sec. V-B we establish a result similar to Theorem 3 but with $\theta_n = 0$. That is, additional sufficient conditions on the system model lead to sharper results in Sec. V-B. \square

B. ANALYSIS WITH IRREPRESENTABILITY CONDITIONS

Here we impose additional conditions and obtain sharper results. Now we follow the primal-dual witness technique of [5], originally used in the context of element-wise lasso penalty for single-attribute graphs. The technique of [5] was extended to group-lasso penalty for multi-attribute graphs in [7]. In this paper we apply the primal-dual witness technique in a sparse-group non-convex penalty setting. Some prior results in an element-wise non-convex penalty setting are in [20] and [21].

Denote the true extended graph edgeset $\bar{\mathcal{E}}$ by $\bar{\mathcal{E}}^*$. Define

$$S = \mathcal{E}^* \cup \{\{k, \ell\} : k = \ell \in [p_n]\} \subseteq [p_n] \times [p_n], \quad (41)$$

$$\bar{S} = \bar{\mathcal{E}}^* \cup \{\{i, j\} : i = j \in [mp_n]\} \subseteq [mp_n] \times [mp_n], \quad (42)$$

$$\Gamma^* = (\Omega^*)^{-1} \boxtimes (\Omega^*)^{-1}, \quad (43)$$

$$\hat{\Gamma} = \hat{\Omega}^{-1} \boxtimes \hat{\Omega}^{-1}, \quad (44)$$

$$\kappa_{\Gamma^*} = \|\mathcal{C}((\Gamma_{S,S}^*)^{-1})\|_{1,\infty} \quad (45)$$

$$\bar{\kappa}_{\Gamma^*} = \|(\Gamma_{\bar{S},\bar{S}}^*)^{-1}\|_{1,\infty} \quad (46)$$

$$\kappa_{\Sigma^*} = \|\mathcal{C}(\Sigma^*)\|_{1,\infty} \quad (47)$$

$$\bar{\kappa}_{\Sigma^*} = \|\Sigma^*\|_{1,\infty} \quad (48)$$

$$d_n = \text{maximum degree of } S, \quad (49)$$

$$\bar{d}_n = \text{maximum degree of } \bar{S}. \quad (50)$$

In (49), d_n is the maximum number of non-zero elements per row of $\mathcal{C}(\Omega^*)$, and similarly \bar{d}_n in (50) is the maximum number of non-zero elements per row of Ω^* . As discussed in Sec. II, with the true graph $\mathcal{G}^* = (V, \mathcal{E}^*)$, $V = [p_n]$, we associate an enlarged graph $\bar{\mathcal{G}}^* = (\bar{V}, \bar{\mathcal{E}}^*)$, $\bar{V} = [mp_n]$, such that corresponding to an edge $f = \{k, \ell\} \in \mathcal{E}^*$, there are m^2 edges $\{i, j\} \in \bar{\mathcal{E}}^*$ specified by $i = (k-1)m + q$, $j = (\ell-1)m + r$, $q, r \in [m]$. To keep notation light, for an edge $f = \{k, \ell\} \in \mathcal{E}^*$, we use e_f to denote an edge $\{i, j\} \in \bar{\mathcal{E}}^*$ that corresponds to one of m^2 edges of $\bar{\mathcal{E}}^*$ associated with edge f . Similar notation will be used for edges in $(\mathcal{E}^*)^c$ and $(\bar{\mathcal{E}}^*)^c$, and edges in S and S^c . Using this notation, assume that for some $\gamma \in (0, 1]$, the following *irrepresentability* conditions

hold:

$$\max_{f \in S^c} \|\mathcal{C}(\Gamma_{f,S}^* (\Gamma_{S,S}^*)^{-1})\|_1 \leq 1 - \gamma, \quad (51)$$

$$\max_{e_f \in f \in S^c} \|\Gamma_{e_f,S}^* (\Gamma_{S,S}^*)^{-1}\|_1 \leq 1 - \gamma. \quad (52)$$

With C_0 and N_1 as defined in (22) and (25), respectively, define $\tilde{C}_0 = mC_0$ and

$$N_5 = 36 d_n^2 \kappa_{\Gamma^*}^2 \left(1 + \frac{4}{\gamma}\right)^2 \tilde{C}_0 \ln(p_n) \max\{\kappa_{\Sigma^*}^2, \kappa_{\Gamma^*}^2 \kappa_{\Sigma^*}^6\}, \quad (53)$$

$$N_6 = 36 d_n^2 \left(1 + \frac{4}{\gamma}\right)^4 \tilde{C}_0^2 \ln(p_n) \kappa_{\Gamma^*}^4 \kappa_{\Sigma^*}^6, \quad (54)$$

$$N_7 = 36 \bar{d}_n^2 \left(1 + \frac{4}{\gamma}\right)^4 m^2 \tilde{C}_0^2 \ln(p_n) \kappa_{\Gamma^*}^4 \kappa_{\Sigma^*}^6. \quad (55)$$

Let $\partial \tilde{\mathcal{L}}(\Omega)$ denote the sub-differential of $\tilde{\mathcal{L}}(\Omega)$ at Ω . Suppose that $\hat{\Omega}$ is a solution to

$$0 \in \partial \tilde{\mathcal{L}}(\hat{\Omega}), \quad (56)$$

which is a first-order necessary condition for a stationary point of $\tilde{\mathcal{L}}(\Omega)$. Theorem 4 addresses some properties of this $\hat{\Omega}$.

Theorem 4: For the system model of Sec. II, under the irrepresentability conditions (51)-(52) for some $\gamma \in (0, 1]$, if

$$\lambda_n = \frac{4}{\gamma} C_0 \sqrt{\frac{\ln(p_n)}{n}}, \quad (57)$$

then for $n > \max(N_1, N_5, N_6, N_7)$ and for any $\tau > 2$, there exists a stationary point $\hat{\Omega}$ of $\tilde{\mathcal{L}}(\Omega)$ satisfying with probability $> 1 - 1/p_n^{\tau-2}$,

$$(i) \quad \|\mathcal{C}(\hat{\Omega} - \Omega^*)\|_{\infty} \leq 2\kappa_{\Gamma^*} \left(1 + \frac{4}{\gamma}\right) mC_0 \sqrt{\frac{\ln(p_n)}{n}},$$

$$(ii) \quad \hat{\Omega}_{S^c} = 0.$$

$$(iii) \quad \|\mathcal{C}(\hat{\Omega} - \Omega^*)\|_F \leq \sqrt{s_n^* + p_n} \|\mathcal{C}(\hat{\Omega} - \Omega^*)\|_{\infty}.$$

$$(iv) \quad \text{Additionally, if } \min_{(k,\ell) \in S} \|(\Omega^*)^{(k\ell)}\|_F \geq 4\kappa_{\Gamma^*} \left(1 + \frac{4}{\gamma}\right) mC_0 \sqrt{\frac{\ln(p_n)}{n}}, \text{ then } P(\hat{\mathcal{E}} = \mathcal{E}^*) > 1 - 1/p_n^{\tau-2} \text{ where } \hat{\mathcal{E}} = \{\{q, \ell\} : \|\hat{\Omega}^{(q\ell)}\|_F > 0, q \neq \ell\}.$$

The proof of Theorem 4 is given in Appendix C.

Lemma 1 and Theorem 4 lead to Theorem 5, as proved in Appendix C. First define

$$\tilde{R} = 2\kappa_{\Gamma^*} \left(1 + \frac{4}{\gamma}\right) mC_0, \quad (58)$$

$$\tilde{r}_n = \sqrt{(s_n^* + p_n) \frac{\ln(p_n)}{n}}. \quad (59)$$

Theorem 5: Assume the conditions of Theorem 4, and as in assumption (A2), suppose that $\beta_{\min} \leq \phi_{\min}(\Sigma^*)$. If $\tilde{R}\tilde{r}_n + 1/\beta_{\min} \leq 0.99 \tilde{\mu}$, then $\hat{\Omega}$ as defined in Lemma 1 is a unique minimum, satisfying all results of Theorem 4. \bullet

Remark 4: In terms of the rate of convergence, $\|\mathcal{C}(\hat{\Omega} - \Omega^*)\|_F = \mathcal{O}_P(\tilde{r}_n)$. Therefore, for $\|\mathcal{C}(\hat{\Omega} - \Omega^*)\|_F \rightarrow 0$ as $n \rightarrow \infty$, we must have $(p_n + s_n^*) \ln(p_n)/n \rightarrow 0$. This is similar to the results of Theorem 1 (see Remark 1). But unlike Theorem 1,

TABLE 1. F_1 scores, Hamming distances, normalized Frobenius norm of estimation error ($\|\hat{\Omega} - \Omega^*\|_F / \|\Omega^*\|_F$), and timing, for the synthetic data examples ($p = 100, m = 4$), averaged over 100 runs (standard deviation σ in parentheses).

n	200	400	800	200	400	800
λ 's picked to maximize F_1 score						
ER graph: F_1 score (σ)				BA graph: F_1 score (σ)		
Lasso	0.742 (0.065)	0.916 (0.032)	0.983 (0.011)	0.573 (0.058)	0.784 (0.067)	0.918 (0.048)
Log-sum	0.804 (0.046)	0.964 (0.008)	0.998 (0.002)	0.647 (0.053)	0.886 (0.044)	0.983 (0.017)
SCAD	0.752 (0.072)	0.931 (0.017)	0.988 (0.007)	0.590 (0.062)	0.800 (0.075)	0.933 (0.053)
ER graph: Hamming distance (σ)				BA graph: Hamming distance (σ)		
Lasso	113.4 (18.24)	39.60 (13.78)	08.49 (05.50)	181.4 (68.10)	80.66 (18.54)	31.38 (14.98)
Log-sum	86.91 (17.52)	18.01 (04.26)	0.880 (0.898)	128.3 (10.15)	41.32 (13.59)	06.45 (06.14)
SCAD	105.7 (21.39)	34.63 (08.74)	06.16 (03.48)	161.6 (46.10)	71.55 (19.23)	24.89 (16.26)
ER graph: Est. error (σ)				BA graph: Est. error (σ)		
Lasso	0.335 (0.008)	0.303 (0.010)	0.266 (0.010)	0.268 (0.005)	0.241 (0.003)	0.212 (0.005)
Log-sum	0.307 (0.008)	0.227 (0.008)	0.170 (0.007)	0.265 (0.004)	0.214 (0.004)	0.164 (0.006)
SCAD	0.313 (0.008)	0.222 (0.007)	0.149 (0.005)	0.304 (0.028)	0.217 (0.004)	0.152 (0.010)
ER graph: Timing (s) (σ)				BA graph: Timing (s) (σ)		
Lasso	1.897 (0.140)	1.895 (0.251)	1.891 (0.059)	2.038 (0.358)	1.932 (0.315)	1.958 (0.277)
Log-sum	7.604 (0.819)	6.251 (0.185)	5.765 (0.228)	6.902 (0.236)	6.431 (0.263)	5.857 (0.341)
SCAD	5.158 (0.455)	5.291 (0.243)	5.027 (0.237)	6.574 (0.808)	5.473 (0.437)	5.133 (0.302)
λ 's picked to minimize BIC						
ER graph: F_1 score (σ)				BA graph: F_1 score (σ)		
Lasso	0.329 (0.147)	0.813 (0.085)	0.965 (0.039)	0.414 (0.149)	0.622 (0.158)	0.901 (0.068)
Log-sum	0.766 (0.083)	0.942 (0.021)	0.996 (0.004)	0.582 (0.120)	0.859 (0.076)	0.936 (0.099)
ER graph: Hamming distance (σ)				BA graph: Hamming distance (σ)		
Lasso	1239. (538.2)	125.6 (167.3)	019.1 (023.2)	739.1 (781.3)	255.3 (204.9)	36.15 (19.49)
Log-sum	136.3 (70.81)	27.54 (10.01)	01.88 (01.64)	240.1 (185.7)	48.29 (20.05)	20.91 (30.22)

by Theorem 4(ii), we have the oracle result: $\hat{\Omega}_{Sc} = \mathbf{0}$ just as $\Omega_{Sc}^* = \mathbf{0}$, i.e., all absent edges in the true graph are absent in the estimated graph with high probability. Such a result does not exist for Theorem 1. Also, the comments made in Remark 2 apply here as well. Finally, Theorem 4 does not need a minimum amplitude condition like (28) in Theorem 1 for SCAD. \square

VI. NUMERICAL EXAMPLES

We now present numerical results for synthetic as well as real data to illustrate the proposed non-convex penalty approaches. In the synthetic data examples the ground truth is known which allows for an assessment of the efficacy of various approaches. In the real data example our goal is visualization and exploration of the conditional dependency structure underlying the data since the ground truth is unknown.

A. SYNTHETIC DATA: ERDÖS-RÉNYI AND BARABÁSI-ALBERT GRAPHS

We consider two types of graphs: Erdős-Rényi (ER) graph and Barabási-Albert (BA) graph [29], [30]. In the ER graph, $p = 100$ nodes are connected to each other with probability $p_{er} = 0.05$ and there are $m = 4$ attributes per node whereas in the BA graph, we used $p = 100$ and mean degree of 2 to generate a BA graph using the procedure given in [30]. In the upper triangular Ω , we set $[\Omega^{(jk)}]_{st} = 0.5^{|s-t|}$ for $j = k \in [p]$, $s, t \in [m]$. For $j \neq k$, if the two nodes are not connected in the graph (ER or BA), we have $\Omega^{(jk)} = \mathbf{0}$, and if nodes j and k are connected, then $[\Omega^{(jk)}]_{st}$ is uniformly distributed

over $[-0.4, -0.1] \cup [0.1, 0.4]$ for $s \neq t$, otherwise it is zero. Then add lower triangular elements to make Ω a symmetric matrix. Finally add δI to Ω and pick δ so that the minimum eigenvalue of Ω is 0.5; this is similar to the simulation example 3 in [7, Sec. 5.1]. With $\Phi\Phi^T = (\Omega + \delta I)^{-1}$, we generate $\mathbf{x} = \Phi\mathbf{w}$ with $\mathbf{w} \in \mathbb{R}^{mp}$ as zero-mean Gaussian, with identity covariance. We generate n i.i.d. observations for \mathbf{x} , with $m = 4, p = 100, n \in \{200, 400, 800\}$.

Simulation results based on 100 runs are shown in Table 1 for ER and BA graphs where the performance measures are F_1 -score and Hamming distance (between estimated and true edgesets $\hat{\mathcal{E}}$ and \mathcal{E}^*) for efficacy in edge detection, normalized estimation error $\|\hat{\Omega} - \Omega^*\|_F / \|\Omega^*\|_F$ and execution time (based on tic-toc functions in MATLAB). All simulations 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 used the ADMM algorithm (with LLA for non-convex penalties) given in Algorithm 1 with $\alpha = 0.05$ for all three regularizations: lasso, log-sum and SCAD. It is seen that log-sum penalty outperforms lasso and SCAD with F_1 score or Hamming distance as the performance metric. For $n = 800$, SCAD yields smaller estimation errors in estimating Ω but its performance in terms of F_1 score and Hamming distance metrics is, in general, poor. In practice we do not know the ground truth, hence cannot pick λ to maximize the F_1 score. In Table 1 we also show results for lasso and log-sum penalties when λ is picked based the BIC information criterion as discussed in Sec. IV-A with $\alpha = 0.05$. Here again the log-sum penalty outperforms lasso.

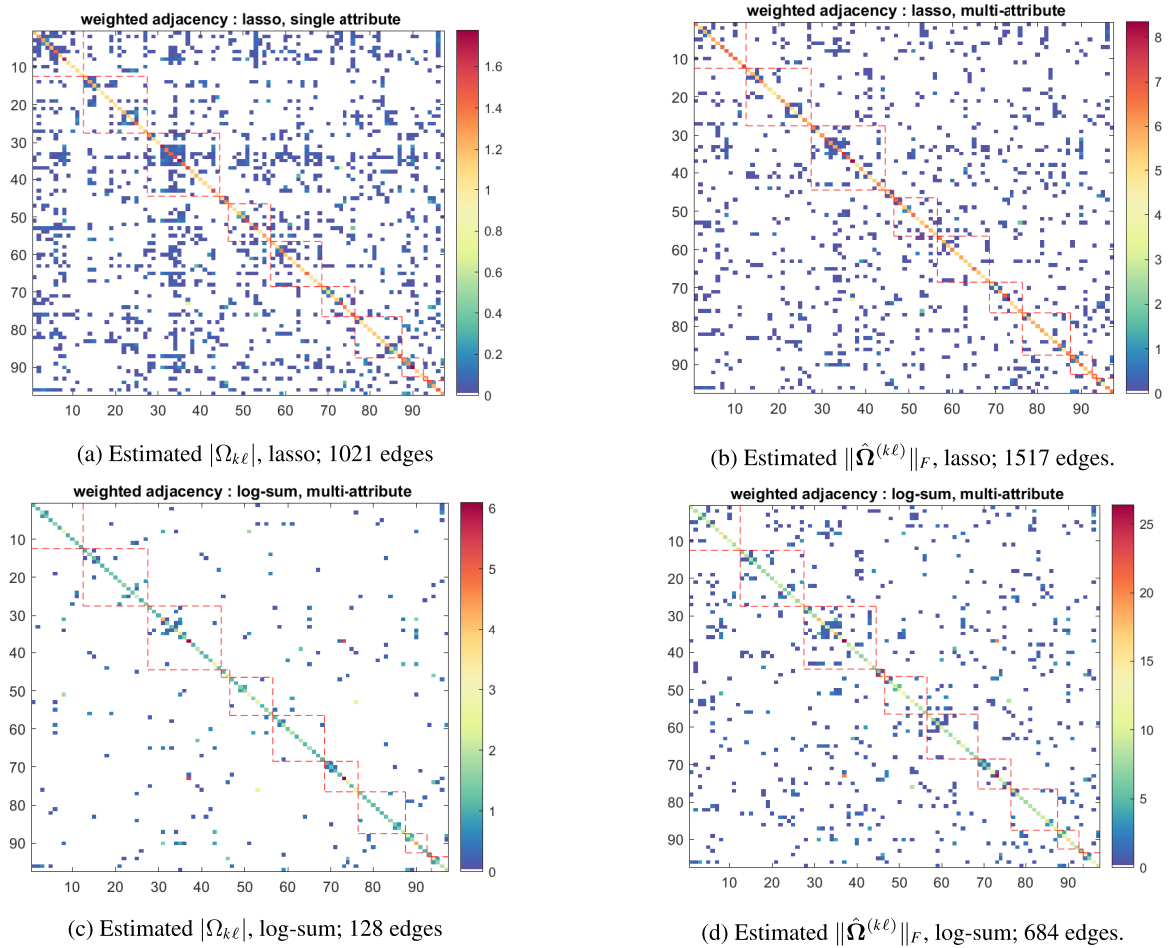


FIGURE 1. Estimated precision matrices based edge weights for financial time series using proposed log-sum penalty with BIC for λ selection. (a) Lasso, single-attribute case ($m = 1$, $p = 97$) with $|\hat{\Omega}_{k\ell}|$, $k, \ell \in [97]$ as edge weight. (b) Lasso, multi-attribute case ($m = 4$, $p = 97$) with $\|\hat{\Omega}^{(k\ell)}\|_F$, $k, \ell \in [97]$ as edge weight. (c) Log-sum, single-attribute case ($m = 1$, $p = 97$) with $|\hat{\Omega}_{k\ell}|$, $k, \ell \in [97]$ as edge weight. (d) Log-sum, multi-attribute case ($m = 4$, $p = 97$) with $\|\hat{\Omega}^{(k\ell)}\|_F$, $k, \ell \in [97]$ as edge weight. In all figures, the red squares (in dashed lines) show the 11 sectors – they are not part of the edge weights.

B. REAL DATA: FINANCIAL TIME SERIES

We consider daily high, low and close-of-the-day share prices and daily trade volume ($m = 4$) of 97 ($p = 97$) stocks in the S&P 100 index from April 1, 2015 through April 1, 2020, yielding 1259 samples. This data was gathered from the Yahoo Finance website. Let $[z_i(t)]_\ell$ denote the ℓ th feature ($\ell \in [m]$) of the i th stock on day t . Then we consider (as is conventional in such studies) $[x_i(t)]_\ell = \ln([z_i(t)]_\ell / [z_i(t)]_\ell)$ as the time series to analyze, yielding $n = 1258$, $m = 4$ and $p = 97$. These 97 stocks are classified into 11 sectors (according to the Global Industry Classification Standard (GICS)) and we order the nodes to group them as information technology (nodes 1-12), health care (nodes 13-27), financials (nodes 28-44), real estate (nodes 45-46), consumer discretionary (nodes 47-56), industrials (nodes 57-68), communication services (nodes 69-76), consumer staples (nodes 77-87), energy (nodes 88-92), materials (node 93), utilities (nodes 94-97). For each i and ℓ , $[x_i(t)]_\ell$ was centered and normalized to unit variance. We applied the BIC criterion for λ selection using log-sum penalty as well

as lasso, for two cases: $m = 4$, and $m = 1$ where only the close-of-the-day share prices were used. To apply BIC, with $\alpha = 0.05$, we selected the best λ as detailed in Sec. IV-A, and then using this optimized λ , we selected the best α by searching over a grid of values in $[0.01, 0.3]$. The resulting estimated precision matrices based edge weights are shown in Fig. 1. In the multi-attribute case ($m = 4$, Figs. 1b and 1d), we get 1517 and 684 edges for lasso and log-sum regularizations, respectively, whereas in the single-attribute case ($m = 1$, Figs. 1a and 1c) we get 1021 and 128 edges for lasso and log-sum regularizations, respectively. The multi-attribute graph exploits many more relevant features compared to the single-attribute graph which cannot use more than one feature. The log-sum penalty yields sparser graphs for both single-attribute and multi-attribute cases, compared to the lasso penalty.

VII. CONCLUSION

We investigated multi-attribute graph learning using a penalized log-likelihood objective function where both

convex (sparse-group lasso) and non-convex (sparse-group log-sum and SCAD) regularization functions were considered. An ADMM approach coupled with a local linear approximation to non-convex penalties was presented for optimization of the objective function. We established sufficient conditions in a high-dimensional setting for consistency, local convexity when using non-convex penalties, and graph recovery. Two alternative sets of sufficient conditions were investigated, with and without some irrepresentability conditions. With irrepresentability conditions we could establish sharper results such as the oracle property (Theorem 4(ii)) and maximum error bound (Theorem 4(i)). While the non-convex penalized log-likelihood objective function results in a non-convex optimization problem, Theorems 2 and 5 specify conditions under which it becomes a convex optimization problem. These conditions favor log-sum penalty over SCAD. Numerical results based on synthetic and real data were presented to illustrate the proposed approaches. In the synthetic data examples the log-sum penalized objective function significantly outperformed the lasso penalized as well as SCAD penalized objective functions with F_1 -score and Hamming distance as performance metrics.

APPENDIX A

LEMMA 2 AND PROOF OF THEOREMS 1

Lemma 2 follows from [5, Lemma 1] and [9, Lemma 1].

Lemma 2: Suppose $\hat{\Sigma} = (1/n) \sum_{t=1}^n \mathbf{x}(t)\mathbf{x}^\top(t)$, given n i.i.d. samples $\{\mathbf{x}(t)\}_{t=1}^n$ of zero-mean Gaussian $\mathbf{x} \in \mathbb{R}^{mp}$ with covariance Σ^* such that each component $x_i/\sqrt{\Sigma_{ii}^*}$ is Gaussian with unit variance. Define $\sigma_{\max} = \max_{1 \leq i \leq mp_n} \Sigma_{ii}^*$ and

$$\tilde{C}_0 = 40 m \sigma_{\max} \sqrt{2 \ln(4m^2 p_n^\tau) / \ln(p_n)}. \quad (60)$$

Then for any $\tau > 2$,

$$P\left(\|\mathcal{C}(\hat{\Sigma} - \Sigma^*)\|_\infty > \tilde{C}_0 \sqrt{\ln(p_n)/n}\right) \leq 1/p_n^{\tau-2} \quad (61)$$

and

$$P\left(\|\hat{\Sigma} - \Sigma^*\|_\infty > C_0 \sqrt{\ln(p_n)/n}\right) \leq 1/p_n^{\tau-2} \quad (62)$$

if $n > 2 \ln(4m^2 p_n^\tau)$, where $C_0 = \tilde{C}_0/m$. •

Proof: The bound (61) follows from [9, Lemma 1] when [9, Lemma 1] is applied to Gaussian distributions. Both [5, Lemma 1] and [9, Lemma 1] are stated for sub-Gaussian distributions, and the latter is based on the former. The bound (62) follows similar to (61). ■

Proof of Theorem 1: Let $\Omega = \Omega^* + \Delta$ with both Ω , $\Omega^* > 0$, and

$$Q(\Omega) := \tilde{\mathcal{L}}(\Omega) - \tilde{\mathcal{L}}(\Omega^*). \quad (63)$$

The estimate $\hat{\Omega}$ minimizes $Q(\Omega)$, or equivalently, $\hat{\Delta} = \hat{\Omega} - \Omega^*$ minimizes $G(\Delta) := Q(\Omega^* + \Delta)$. We will follow the method of proof of [8, Theorem 1], which, in turn, for the most part, follows the method of proof of [23, Theorem 1] pertaining to lasso penalty. Consider the set

$$\Theta_n(R) := \left\{ \Delta : \Delta = \Delta^\top, \|\Delta\|_F = Rr_n \right\} \quad (64)$$

where R and r_n are as in (23) and (24), respectively. Since $G(\hat{\Delta}) \leq G(0) = 0$, if we can show that $\inf_{\Delta \in \Theta_n(R)} \{G(\Delta) : \Delta \in \Theta_n(R)\} > 0$, then the minimizer $\hat{\Delta}$ must be inside $\Theta_n(R)$, and hence $\|\hat{\Delta}\|_F \leq Rr_n$. It is shown in [23, (9)] that

$$\ln(|\Omega^* \mathcal{C} \Delta|) - \ln(|\Omega^*|) = \text{tr}(\Sigma^* \Delta) - A_1 \quad (65)$$

where, with $H(\Omega^*, \Delta, v) = (\Omega^* + v\Delta)^{-1} \otimes (\Omega^* + v\Delta)^{-1}$ and v denoting a scalar,

$$A_1 = \text{vec}(\Delta)^\top \left(\int_0^1 (1-v) H(\Omega^*, \Delta, v) dv \right) \text{vec}(\Delta). \quad (66)$$

Noting that $\Omega^{-1} = \Sigma$, we can rewrite $G(\Delta)$ as

$$G(\Delta) = A_1 + A_2 + A_3 + A_4, \quad (67)$$

where

$$A_2 = \text{tr} \left((\hat{\Sigma} - \Sigma^*) \Delta \right), \quad (68)$$

$$A_3 = \alpha \sum_{i,j=1; i \neq j}^{mp_n} \left(\rho_\lambda(|\Omega_{ij}^* + \Delta_{ij}|) - \rho_\lambda(|\Omega_{ij}^*|) \right), \quad (69)$$

$$A_4 = (1-\alpha)m \sum_{q,\ell=1; q \neq \ell}^{p_n} \left(\rho_\lambda(\|(\Omega^*)^{(q\ell)} + \Delta^{(q\ell)}\|_F) - \rho_\lambda(\|(\Omega^*)^{(q\ell)}\|_F) \right). \quad (70)$$

Following [23, p. 502], we have

$$A_1 \geq \frac{\|\Delta\|_F^2}{2(\|\Omega^*\| + \|\Delta\|)^2} \geq \frac{\|\Delta\|_F^2}{2(\beta_{\min}^{-1} + Rr_n)^2} \quad (71)$$

where we have used the fact that $\|\Omega^*\| = \|(\Sigma^*)^{-1}\| = \phi_{\max}((\Sigma^*)^{-1}) = (\phi_{\min}(\Sigma^*))^{-1} \leq \beta_{\min}^{-1}$ and $\|\Delta\| \leq \|\Delta\|_F = Rr_n$. We now consider A_2 in (68). We have

$$A_2 = \underbrace{\sum_{i,j=1; i \neq j}^{mp_n} [\hat{\Sigma} - \Sigma^*]_{ij} \Delta_{ji}}_{L_1} + \underbrace{\sum_{i=1}^{mp_n} [\hat{\Sigma} - \Sigma^*]_{ii} \Delta_{ii}}_{L_2} \quad (72)$$

By Lemma 2, the sample covariance $\hat{\Sigma}$ satisfies the tail bound

$$P\left(\max_{k,\ell} |[\hat{\Sigma} - \Sigma^*]_{kl}| > C_0 \sqrt{\frac{\ln(p_n)}{n}}\right) \leq \frac{1}{(p_n)^{\tau-2}} \quad (73)$$

for $\tau > 2$, if the sample size $n > N_1$ (N_1 is defined in (25)). To bound L_1 , using Lemma 2, with probability $> 1 - 1/p_n^{\tau-2}$,

$$|L_1| \leq \|\Delta^-\|_1 \max_{i,j} |[\hat{\Sigma} - \Sigma^*]_{ij}| \leq \|\Delta^-\|_1 C_0 \sqrt{\frac{\ln(p_n)}{n}}. \quad (74)$$

Similarly, by Cauchy-Schwarz inequality, Lemma 2 and (24),

$$\begin{aligned} |L_2| &\leq \|\Delta^+\|_1 C_0 \sqrt{\frac{\ln(p_n)}{n}} \\ &\leq C_0 \sqrt{\frac{\ln(p_n)}{n}} \sqrt{mp_n} \|\Delta^+\|_F \leq \|\Delta^+\|_F C_0 r_n. \end{aligned} \quad (75)$$

Therefore, with probability $> 1 - 1/p_n^{\tau-2}$,

$$|A_2| \leq \|\Delta^-\|_1 C_0 \sqrt{\frac{\ln(p_n)}{n}} + \|\Delta^+\|_F C_0 r_n. \quad (76)$$

We now derive a different bound on A_2 . Define $\tilde{\Delta} \in \mathbb{R}^{p_n \times p_n}$ with (i, j) -th element $\tilde{\Delta}_{ij} = \|\Delta^{(ij)}\|_F$, where $\Delta^{(ij)}$ is defined from Δ similar to (2). By Cauchy-Schwarz inequality,

$$\begin{aligned} \|\Delta^-\|_1 &= \sum_{i,j=1; i \neq j}^{mp_n} |\Delta_{ij}| \leq m \|\tilde{\Delta}^-\|_1 \\ &+ \underbrace{\left(\sum_{k=1}^{p_n} \|\Delta^{(kk)}\|_1 - \|\Delta^+\|_1 \right)}_{=:B}. \end{aligned} \quad (77)$$

Then using $\sum_k \|\Delta^{(kk)}\|_1 \leq m \sum_k \tilde{\Delta}_{kk} \leq m \sqrt{p_n} \|\tilde{\Delta}^+\|_F$, we have

$$\begin{aligned} |L_2| + C_0 \sqrt{\frac{\ln(p_n)}{n}} B &\leq C_0 \sqrt{\frac{\ln(p_n)}{n}} \left(\sum_{k=1}^{p_n} \|\Delta^{(kk)}\|_1 \right) \\ &\leq \|\tilde{\Delta}^+\|_F \sqrt{m} C_0 r_n. \end{aligned}$$

Therefore, an alternative bound is

$$|A_2| \leq m \|\tilde{\Delta}^-\|_1 C_0 \sqrt{\frac{\ln(p_n)}{n}} + \sqrt{m} \|\tilde{\Delta}^+\|_F C_0 r_n. \quad (78)$$

For the rest of the proof we have two different approaches, one for lasso and log-sum and the other for SCAD penalty.

For Lasso and Log-Sum Penalties: We now bound A_3 in (69). Let \mathcal{E}^* denote the true enlarged edge-set corresponding to \mathcal{E}^* when one interprets multi-attribute model as a single-attribute model. Let $(\mathcal{E}^*)^c$ denote its complement. Using the mean-value theorem, we have $(\rho'_\lambda(u) = \frac{d\rho_\lambda(u)}{du})$

$$\begin{aligned} \rho_\lambda(|\Omega_{ij}^* + \Delta_{ij}|) &= \rho_\lambda(|\Omega_{ij}^*|) \\ &+ \rho'_\lambda(|\tilde{\Omega}_{ij}|)(|\Omega_{ij}^* + \Delta_{ij}| - |\Omega_{ij}^*|) \end{aligned} \quad (79)$$

where $|\tilde{\Omega}_{ij}| = |\Omega_{ij}^*| + \gamma(|\Omega_{ij}^* + \Delta_{ij}| - |\Omega_{ij}^*|)$ for some $\gamma \in [0, 1]$. We have

$$\begin{aligned} A_3 &= \alpha \sum_{(i,j) \in \mathcal{E}^*} \rho'_\lambda(|\tilde{\Omega}_{ij}|)(|\Omega_{ij}^* + \Delta_{ij}| - |\Omega_{ij}^*|) \\ &+ \alpha \sum_{(i,j) \in (\mathcal{E}^*)^c} \rho_\lambda(|\Delta_{ij}|) \end{aligned} \quad (80)$$

$$\begin{aligned} &\geq -\alpha \sum_{(i,j) \in \mathcal{E}^*} \rho'_\lambda(|\tilde{\Omega}_{ij}|) |\Delta_{ij}| + \alpha \sum_{(i,j) \in (\mathcal{E}^*)^c} C_\lambda |\Delta_{ij}| \\ &\text{for } |\Delta_{ij}| \leq \delta_\lambda, \end{aligned} \quad (81)$$

using the triangle inequality and (9) in the last step above. Now use property (viii) of the penalty functions and $C_\lambda = \lambda/2$ to conclude that

$$A_3 \geq -\alpha \lambda_n \sum_{(i,j) \in \mathcal{E}^*} |\Delta_{ij}| + \alpha (\lambda_n/2) \sum_{(i,j) \in (\mathcal{E}^*)^c} |\Delta_{ij}|. \quad (82)$$

Next we bound A_4 in (70). Considering the true edge-set \mathcal{E}^* for the multi-attribute graph, let $(\mathcal{E}^*)^c$ denote its complement.

If the edge $\{i, j\} \in (\mathcal{E}^*)^c$, then $(\|\Omega^*\|^{(ij)}) = \mathbf{0}$, therefore, $(\|\Omega^*\|^{(ij)} + \Delta^{(ij)})_F - \|\Omega^*\|^{(ij)}_F = \|\Delta^{(ij)}\|_F$. For $\{i, j\} \in \mathcal{E}^*$, by the triangle inequality, $\|(\Omega^*)^{(ij)} + \Delta^{(ij)}\|_F - \|(\Omega^*)^{(ij)}\|_F \geq -\|\Delta^{(ij)}\|_F$. Thus, mimicking the steps for bounding A_3 , we have

$$\begin{aligned} A_4 &\geq -(1 - \alpha) m \lambda_n \sum_{(i,j) \in \mathcal{E}^*} \|\Delta^{(ij)}\|_F \\ &+ (1 - \alpha) m (\lambda_n/2) \sum_{(i,j) \in ((\mathcal{E}^*)^c)} \|\Delta^{(ij)}\|_F. \end{aligned} \quad (83)$$

Split A_2 as $A_2 = \alpha A_2 + (1 - \alpha) A_2$, apply bound (76) to αA_2 and (78) to $(1 - \alpha) A_2$, use $\|\Delta^-\|_1 = \|\Delta_{\mathcal{E}^*}^-\|_1 + \|\Delta_{(\mathcal{E}^*)^c}^-\|_1$ and $\|\tilde{\Delta}^-\|_1 = \|\tilde{\Delta}_{\mathcal{E}^*}^-\|_1 + \|\tilde{\Delta}_{(\mathcal{E}^*)^c}^-\|_1$. Define $d_1 = \frac{\ln(p_n)}{n}$, then $r_n = \sqrt{mp_n + m^2 s_n^*} d_1$. We have

$$\begin{aligned} \alpha A_2 + A_3 &\geq -\alpha |A_2| + \alpha \lambda_n (0.5 \|\Delta_{(\mathcal{E}^*)^c}^-\|_1 - \|\Delta_{\mathcal{E}^*}^-\|_1) \\ &\geq \alpha (0.5 \lambda_n - C_0 d_1) \sum_{(i,j) \in (\mathcal{E}^*)^c} |\Delta_{ij}| \\ &- \alpha (\lambda_n + C_0 d_1) \sum_{(i,j) \in \mathcal{E}^*} |\Delta_{ij}| - \alpha C_0 r_n \|\Delta^+\|_F. \end{aligned} \quad (84)$$

Since we pick $\lambda_n \geq \lambda_{n\ell}$ in Theorem 1, $0.5 \lambda_n - C_0 d_1 \geq 0$ and therefore, the first term above can be neglected. Now $\sum_{(i,j) \in \mathcal{E}^*} |\Delta_{ij}| \leq \sqrt{m^2 s_n^*} \|\Delta\|_F$, by the Cauchy-Schwarz inequality, and $\|\Delta^+\|_F \leq \|\Delta\|_F$. We then have

$$\alpha A_2 + A_3 \geq -\alpha \left((\lambda_n + C_0 d_1) \sqrt{m^2 s_n^*} + C_0 r_n \right) \|\Delta\|_F. \quad (85)$$

Similarly, we have

$$\begin{aligned} (1 - \alpha) A_2 + A_4 &\geq -(1 - \alpha) m \\ &\times \left((\lambda_n + C_0 d_1) \sqrt{s_n^*} + C_0 r_n \right) \|\Delta\|_F. \end{aligned} \quad (86)$$

From (85) and (86) we have

$$\begin{aligned} A_2 + A_3 + A_4 &\geq -\|\Delta\|_F \left(\lambda_n m \sqrt{s_n^*} + C_0 d_1 m \sqrt{s_n^*} + m C_0 r_n \right) \\ &\geq -\|\Delta\|_F \left(\lambda_n m \sqrt{s_n^*} + (1 + m) C_0 r_n \right) \\ &\geq -2(1 + m) C_0 r_n \|\Delta\|_F \end{aligned} \quad (87)$$

where we used the fact that since $\lambda_n \leq \lambda_{nu1}$, $\lambda_n m \sqrt{s_n^*} \leq C_0 (1 + m) r_n$. Using (67), the bound (71) on A_1 , bound (96) on $A_2 + A_3 + A_4$, and $\|\Delta\|_F = R r_n$, we have with probability $> 1 - 1/p_n^{\tau-2}$,

$$G(\Delta) \geq \|\Delta\|_F^2 \left[\frac{1}{2(\beta_{\min}^{-1} + R r_n)^2} - \frac{2C_0(1 + m)}{R} \right]. \quad (88)$$

For the given choice of N_2 , $R r_n \leq R r_{N_2} \leq 0.1/\beta_{\min}$ for $n \geq N_2$. Also, $2C_0(1 + m)/R = \beta_{\min}^2/4$ by (23). Then for $n \geq N_2$,

$$\frac{1}{2(\beta_{\min}^{-1} + R r_n)^2} - \frac{2C_0(1 + m)}{R} \geq \beta_{\min}^2 \left(\frac{1}{2.42} - \frac{1}{4} \right) > 0,$$

implying $G(\Delta) > 0$. This proves (32). The choice of N_3 for log-sum penalty ensures that $|\Delta_{ij}| \leq \delta_\lambda = \epsilon$ needed in (81) is satisfied w.h.p.: if $R r_n \leq \epsilon$, then $|\Delta_{ij}| \leq \|\Delta\|_F \leq R r_n \leq \epsilon$.

For SCAD Penalty: Here we address (79) differently. Using triangle inequality, we have

$$\begin{aligned} |\tilde{\Omega}_{ij}| &\geq |\Omega_{ij}^*| + \gamma(|\Omega_{ij}^*| - |\Delta_{ij}| - |\Omega_{ij}^*|) \\ &\geq |\Omega_{ij}^*| - |\Delta_{ij}|. \end{aligned} \quad (89)$$

Since $|\Delta_{ij}| \leq \|\Delta\|_F \leq Rr_n$, the choice $\lambda_n = \lambda_{nu2}$ implies that $\lambda_n \geq Rr_n$, satisfying $|\Delta_{ij}| \leq \lambda_n$. Therefore, $|\tilde{\Omega}_{ij}| \geq |\Omega_{ij}^*| - \lambda_n$. For $n \geq N_4$, $\rho'_\lambda(|\tilde{\Omega}_{ij}|) = 0$ (see (28)) if $\{i, j\} \in \tilde{\mathcal{E}}^*$, i.e., $|\Omega_{ij}^*| \neq 0$, since in this case $|\tilde{\Omega}_{ij}| \geq (a+1)\lambda_n - \lambda_n = a\lambda_n$. As in (80), we have

$$\begin{aligned} A_3 &= \alpha \sum_{(i,j) \in \tilde{\mathcal{E}}^*} \rho'_\lambda(|\tilde{\Omega}_{ij}|)(|\Omega_{ij}^*| + |\Delta_{ij}| - |\Omega_{ij}^*|) \\ &\quad + \alpha \sum_{(i,j) \in (\tilde{\mathcal{E}}^*)^c} \rho_\lambda(|\Delta_{ij}|) \\ &\geq \alpha \sum_{(i,j) \in (\tilde{\mathcal{E}}^*)^c} C_\lambda |\Delta_{ij}| \quad \text{for } |\Delta_{ij}| \leq \delta_\lambda, \end{aligned} \quad (90)$$

$$= \alpha(\lambda_n/2) \sum_{(i,j) \in (\tilde{\mathcal{E}}^*)^c} |\Delta_{ij}|. \quad (91)$$

Mimicking the steps for bounding A_3 above and under same conditions, we have

$$A_4 \geq (1 - \alpha)m(\lambda_n/2) \sum_{(i,j) \in (\tilde{\mathcal{E}}^*)^c} \|\Delta^{(ij)}\|_F. \quad (92)$$

Thus

$$\begin{aligned} \alpha A_2 + A_3 &\geq -\alpha|A_2| + 0.5\alpha\lambda_n \|\Delta_{(\tilde{\mathcal{E}}^*)^c}^-\|_1 \\ &\geq \alpha(0.5\lambda_n - C_0 d_1) \sum_{(i,j) \in (\tilde{\mathcal{E}}^*)^c} |\Delta_{ij}| \\ &\quad - \alpha C_0 d_1 \sum_{(i,j) \in \tilde{\mathcal{E}}^*} |\Delta_{ij}| - \alpha C_0 r_n \|\Delta^+\|_F. \end{aligned} \quad (93)$$

Since we pick $\lambda_n = \max(Rr_n, \lambda_{nu1})$ in Theorem 1, $0.5\lambda_n - C_0 d_1 \geq 0$ and therefore, the first term above can be neglected. Now $\sum_{(i,j) \in \tilde{\mathcal{E}}^*} |\Delta_{ij}| \leq \sqrt{m^2 s_n^*} \|\Delta\|_F$, by the Cauchy-Schwarz inequality, and $\|\Delta^+\|_F \leq \|\Delta\|_F$. We then have

$$\alpha A_2 + A_3 \geq -\alpha \left(C_0 d_1 \sqrt{m^2 s_n^*} + C_0 r_n \right) \|\Delta\|_F. \quad (94)$$

By very similar arguments we also have

$$\begin{aligned} (1 - \alpha)A_2 + A_4 &\geq -(1 - \alpha)m \\ &\quad \times \left(C_0 d_1 \sqrt{s_n^*} + C_0 r_n \right) \|\Delta\|_F. \end{aligned} \quad (95)$$

From (94) and (95) we have

$$\begin{aligned} A_2 + A_3 + A_4 &\geq -\|\Delta\|_F \left(C_0 d_1 m \sqrt{s_n^*} + m C_0 r_n \right) \\ &\geq -(1 + m) C_0 r_n \|\Delta\|_F \end{aligned} \quad (96)$$

where we used the fact that $C_0 d_1 m \sqrt{s_n^*} \leq C_0 r_n$. Mimicking (88), with probability $> 1 - 1/p_n^{\tau-2}$, we have

$$\begin{aligned} G(\Delta) &\geq \|\Delta\|_F^2 \left[\frac{1}{2(\beta_{\min}^{-1} + Rr_n)^2} - \frac{(1+m)C_0}{R} \right] \\ &\geq \beta_{\min}^2 \left(\frac{1}{2.42} - \frac{1}{8} \right) > 0, \end{aligned} \quad (97)$$

implying $G(\Delta) > 0$. This proves (32). For the SCAD penalty, we need $|\Delta_{ij}| \leq \delta_\lambda = \lambda_n$ in (91). Since $|\Delta_{ij}| \leq \|\Delta\|_F \leq Rr_n$, the choice $\lambda_n = \lambda_{nu2}$ implies that $\lambda_n \geq Rr_n$, satisfying $|\Delta_{ij}| \leq \lambda_n$. This completes the proof. ■

APPENDIX B

PROOFS OF LEMMA 2 AND THEOREMS 2 AND 3

Proof of Lemma 1: Consider $h(\Omega) = \mathcal{L}(\Omega) - \frac{\mu}{2} \|\Omega\|_F^2$ for some $\mu \geq 0$. The Hessian of $\mathcal{L}(\Omega)$ w.r.t. $\text{vec}(\Omega)$ is $\nabla^2 \mathcal{L}(\Omega) = \Omega^{-1} \otimes \Omega^{-1}$ with

$$\phi_{\min}(\nabla^2 \mathcal{L}(\Omega)) = \phi_{\min}^2(\Omega^{-1}) = 1/\phi_{\max}^2(\Omega) = 1/\|\Omega\|^2. \quad (98)$$

Since $\nabla^2 h(\Omega) = \Omega^{-1} \otimes \Omega^{-1} - \mu I_{(mp)^2}$, it follows that $h(\Omega)$ is positive semi-definite, hence convex, if

$$\|\Omega\| \leq \sqrt{\frac{1}{\mu}}. \quad (99)$$

By property (v) of the penalty functions, $g(u) := \rho_\lambda(u) + \frac{\mu}{2} u^2$ is convex, for some $\mu \geq 0$, and by property (ii), it is non-decreasing on \mathbb{R}_+ . Therefore, by the composition rules [31, Sec. 3.2.4], $g(|[\Omega]_{ij}|)$ and $g(\|\Omega^{(q\ell)}\|_F)$ are convex. Hence,

$$P_e(\Omega) + \frac{\mu_e}{2} \|\Omega\|_F^2 = \sum_{i \neq j}^{mp_n} \left(\rho_\lambda(|[\Omega]_{ij}|) + \frac{\mu_e}{2} |[\Omega]_{ij}|^2 \right) \quad (100)$$

is convex for $\mu_e = \mu \geq 0$, and similarly,

$$P_g(\Omega) + \frac{\mu_g}{2} \|\Omega\|_F^2 = m \sum_{q \neq \ell}^{p_n} \left(\rho_\lambda(\|\Omega^{(q\ell)}\|_F) + \frac{\mu_g}{2m} \|\Omega^{(q\ell)}\|_F^2 \right) \quad (101)$$

is convex for $\mu_g = m\mu$, where μ is the value that renders $\rho_\lambda(u) + \frac{\mu}{2} u^2$ convex. Express $\tilde{\mathcal{L}}(\Omega)$ as

$$\tilde{\mathcal{L}}(\Omega) = \alpha \tilde{\mathcal{L}}_e(\Omega) + (1 - \alpha) \tilde{\mathcal{L}}_g(\Omega), \quad (102)$$

$$\tilde{\mathcal{L}}_e(\Omega) = \mathcal{L}(\Omega) - \frac{\mu}{2} \|\Omega\|_F^2 + P_e(\Omega) + \frac{\mu}{2} \|\Omega\|_F^2, \quad (103)$$

$$\tilde{\mathcal{L}}_g(\Omega) = \mathcal{L}(\Omega) - \frac{\mu}{2} \|\Omega\|_F^2 + P_g(\Omega) + \frac{\mu}{2} \|\Omega\|_F^2. \quad (104)$$

Now $\tilde{\mathcal{L}}_e(\Omega)$ is convex function of Ω if $\|\Omega\| \leq \sqrt{\frac{1}{\mu}}$, and $\tilde{\mathcal{L}}_g(\Omega)$ is convex in Ω if $\|\Omega\| \leq \sqrt{\frac{1}{\mu_g}} = \sqrt{\frac{1}{m\mu}}$. Thus, for $\tilde{\mathcal{L}}(\Omega)$ to

be strictly convex, using the (minimum) values of μ to make $\rho_\lambda(u) + \frac{\mu}{2}u^2$ convex, we require

$$\begin{aligned} \|\Omega\| &< \bar{\mu} \\ &= \begin{cases} \infty & : \text{lasso} \\ \sqrt{(a-1)/m} & : \text{SCAD} \\ \sqrt{\epsilon/(m\lambda_n)} & : \text{log-sum}, \end{cases} \end{aligned} \quad (105)$$

The choice $\|\Omega\| < \bar{\mu}$ makes $\mathcal{L}(\Omega) - \frac{\mu}{2}\|\Omega\|_F^2$ positive definite, hence strictly convex. We take $\|\Omega\| = 0.99\bar{\mu}$, completing the proof. ■

Proof of Theorem 2: If $1/\beta_{\min} \leq 0.99\bar{\mu}$, then $\Omega^* \in \mathcal{B}$ since $\|\Omega^*\| \leq 1/\beta_{\min}$ by assumption (A2). Now we establish that $\hat{\Omega} \in \mathcal{B}$. To this end, consider

$$\begin{aligned} \|\hat{\Omega}\| &\leq \|\hat{\Omega} - \Omega^*\| + \|\Omega^*\| \\ &\leq \|\hat{\Omega} - \Omega^*\|_F + \|\Omega^*\| \\ &\leq Rr_n + 1/\beta_{\min}. \end{aligned} \quad (106)$$

Therefore, $\hat{\Omega} \in \mathcal{B}$. Thus, both $\hat{\Omega}$ and Ω^* are feasible. The desired result then follows from Theorem 1 and (local) strict convexity of $\tilde{\mathcal{L}}(\Omega)$ over \mathcal{B} implied by Lemma 1. ■

Proof of Theorem 3: We have $\|\hat{\Omega}^{(q\ell)} - (\Omega^*)^{(q\ell)}\|_F \leq \|\hat{\Omega} - \Omega^*\|_F \leq \bar{\sigma}_n$ w.h.p. For the edge $\{q, \ell\} \in \mathcal{E}^*$, we have

$$\begin{aligned} \|\hat{\Omega}^{(q\ell)}\|_F &= \|(\Omega^*)^{(q\ell)} + \hat{\Omega}^{(q\ell)} - (\Omega^*)^{(q\ell)}\|_F \\ &\geq \|(\Omega^*)^{(q\ell)}\|_F - \|\hat{\Omega}^{(q\ell)} - (\Omega^*)^{(q\ell)}\|_F \\ &\geq \nu - \bar{\sigma}_n \geq 0.6\nu \quad \text{for } n \geq N_4 \\ &> \theta_n. \end{aligned} \quad (107)$$

Thus, $\mathcal{E}^* \subseteq \hat{\mathcal{E}}$. Now consider the set complements $(\mathcal{E}^*)^c$ and $\hat{\mathcal{E}}^c$. For the edge $\{q, \ell\} \in (\mathcal{E}^*)^c$, $\|(\Omega^*)^{(q\ell)}\|_F = 0$. For $n \geq N_4$, w.h.p. we have

$$\begin{aligned} \|\hat{\Omega}^{(q\ell)}\|_F &\leq \|(\Omega^*)^{(q\ell)}\|_F + \|\hat{\Omega}^{(q\ell)} - (\Omega^*)^{(q\ell)}\|_F \\ &\leq 0 + \bar{\sigma}_n \leq 0.4\nu < \theta_n, \end{aligned} \quad (108)$$

implying that $\{q, \ell\} \in (\hat{\mathcal{E}}^*)^c$. Thus, $(\mathcal{E}^*)^c \subseteq \hat{\mathcal{E}}^c$, hence $\hat{\mathcal{E}} \subseteq \mathcal{E}^*$, establishing $\hat{\mathcal{E}} = \mathcal{E}^*$. ■

APPENDIX C

TECHNICAL LEMMAS AND PROOFS OF THEOREMS 4 AND 5

In this Appendix, we prove Theorems 4 and 5. A first-order necessary condition for minimization of non-convex $\tilde{\mathcal{L}}(\Omega)$, given by (5), w.r.t. $\Omega \in \mathbb{R}^{mp_n \times mp_n}$ is that the zero matrix belongs to the sub-differential of $\tilde{\mathcal{L}}(\Omega)$ at the solution $\hat{\Omega}$. That is, at $\Omega = \hat{\Omega}$,

$$\begin{aligned} 0 \in \partial \tilde{\mathcal{L}}(\Omega) &= \frac{\partial \mathcal{L}(\Omega)}{\partial \Omega} + \alpha \partial P_e(\Omega) + (1-\alpha) \partial P_g(\Omega) \\ &= \hat{\Sigma} - \Omega^{-1} + \alpha \lambda_n \mathbf{Z}(\Omega) + (1-\alpha) m \lambda_n \mathbf{Y}(\Omega) \end{aligned} \quad (109)$$

where $\lambda_n \mathbf{Z}(\Omega) \in \partial \sum_{i \neq j}^{mp_n} \rho_\lambda(|\Omega_{ij}|) \in \mathbb{R}^{mp_n \times mp_n}$, the sub-differential of (possibly non-convex) element-wise

penalty term for $i \neq j$, is given by

$$[\mathbf{Z}(\Omega)]_{ij} = \begin{cases} v \in [-1, 1], & \text{if } \Omega_{ij} = 0 \\ \frac{\Omega_{ij}}{|\Omega_{ij}|} & \text{if } \Omega_{ij} \neq 0 : \text{lasso} \\ C_{eij} & \text{if } \Omega_{ij} \neq 0 : \text{log-sum} \\ D_{eij} & \text{if } \Omega_{ij} \neq 0 : \text{SCAD}, \end{cases} \quad (110)$$

$$C_{eij} = \frac{\epsilon}{\epsilon + |\Omega_{ij}|} \frac{\Omega_{ij}}{|\Omega_{ij}|}, \quad (111)$$

$$D_{eij} = \begin{cases} \frac{\Omega_{ij}}{|\Omega_{ij}|} & \text{if } 0 < |\Omega_{ij}| \leq \lambda_n \\ \frac{a - |\Omega_{ij}|/\lambda_n}{a-1} \frac{\Omega_{ij}}{|\Omega_{ij}|} & \text{if } \lambda_n < |\Omega_{ij}| \leq a\lambda_n \\ 0 & \text{if } a\lambda_n < |\Omega_{ij}|, \end{cases} \quad (112)$$

and $\lambda_n \mathbf{Y}(\Omega) \in m^{-1} \partial \sum_{k \neq \ell}^{p_n} \rho_\lambda(\|\Omega^{(k\ell)}\|_F) \in \mathbb{R}^{mp_n \times mp_n}$, the sub-differential of (possibly non-convex) group penalty term for $k \neq \ell$, is given by

$$(\mathbf{Y}(\Omega))^{(k\ell)} = \begin{cases} V \in \mathbb{R}^{m \times m}, \|V\|_F \leq 1, \\ \quad \text{if } \|\Omega^{(k\ell)}\|_F = 0 \\ \frac{\Omega^{(k\ell)}}{\|\Omega^{(k\ell)}\|_F} & \text{if } \|\Omega^{(k\ell)}\|_F \neq 0 : \text{lasso} \\ C_g^{(k\ell)} & \text{if } \|\Omega^{(k\ell)}\|_F \neq 0 : \text{log-sum} \\ D_g^{(k\ell)} & \text{if } \|\Omega^{(k\ell)}\|_F \neq 0 : \text{SCAD}, \end{cases} \quad (113)$$

$$C_g^{(k\ell)} = \frac{\epsilon}{\epsilon + \|\Omega^{(k\ell)}\|_F} \frac{\Omega^{(k\ell)}}{\|\Omega^{(k\ell)}\|_F}, \quad (114)$$

$$D_g^{(k\ell)} = \begin{cases} \frac{\Omega^{(k\ell)}}{\|\Omega^{(k\ell)}\|_F} & \text{if } 0 < \|\Omega^{(k\ell)}\|_F \leq \lambda_n \\ \frac{a - \|\Omega^{(k\ell)}\|_F/\lambda_n}{a-1} \frac{\Omega^{(k\ell)}}{\|\Omega^{(k\ell)}\|_F} & \text{if } \lambda_n < \|\Omega^{(k\ell)}\|_F \leq a\lambda_n \\ 0 & \text{if } a\lambda_n < \|\Omega^{(k\ell)}\|_F. \end{cases} \quad (115)$$

We have $|\mathbf{Z}(\Omega)_{ij}| \leq 1$ and $\|(\mathbf{Y}(\Omega))^{(k\ell)}\|_F = \|\text{vec}((\mathbf{Y}(\Omega))^{(k\ell)})\|_2 \leq 1$ for all three penalties; note that $[\mathbf{Z}(\Omega)]_{ij} = 0$ for $i = j$ and $(\mathbf{Y}(\Omega))^{(k\ell)} = 0$ for $k = \ell$. Suppose that $\hat{\Omega}$ is a solution to (109) which is a first-order necessary condition for a stationary point of $\tilde{\mathcal{L}}(\Omega)$. Theorem 4 addresses some properties of this $\hat{\Omega}$.

Let $\tilde{\Omega}$ be a stationary point of $\tilde{\mathcal{L}}(\Omega)$ under the constraint $\Omega_{\text{sc}} = 0$, i.e., $\tilde{\Omega}$ is a solution to $0 \in \partial(\tilde{\mathcal{L}}(\Omega)|_{\Omega_{\text{sc}}=0})$. Define

$$\Delta = \tilde{\Omega} - \Omega^*, \quad (116)$$

$$\mathbf{R}(\Delta) = \tilde{\Omega}^{-1} - (\Omega^*)^{-1} + (\Omega^*)^{-1} \Delta (\Omega^*)^{-1}, \quad (117)$$

$$\mathbf{W} = \hat{\Sigma} - \Sigma^*. \quad (118)$$

Lemma 3: If $\|\Delta\|_\infty < 1/(3\bar{\kappa}_{\Sigma^*} \bar{d}_n)$ then

$$\|\mathbf{R}(\Delta)\|_\infty \leq \frac{3}{2} \bar{d}_n \|\Delta\|_\infty^2 \bar{\kappa}_{\Sigma^*}^3. \quad (119)$$

If $\|\mathbf{C}(\Delta)\|_\infty < 1/(3\kappa_{\Sigma^*} d_n)$ then

$$\|\mathbf{C}(\mathbf{R}(\Delta))\|_\infty \leq \frac{3}{2} d_n \|\mathbf{C}(\Delta)\|_\infty^2 \kappa_{\Sigma^*}^3. \quad (120)$$

Proof: The bound (119) is proved in [5, Lemma 5] and the bound (120) is proved in [7, Lemma 9]. ■

Lemma 4 establishes sufficient conditions under which $\tilde{\Omega}$ is also a solution to $\mathbf{0} \in \partial \tilde{\mathcal{L}}(\Omega)$.

Lemma 4: If $\max(\|\mathbf{W}\|_\infty, \|\mathbf{R}(\Delta)\|_\infty) \leq \gamma\lambda_n/4$ and $\max(\|\mathbf{C}(\mathbf{W})\|_\infty, \|\mathbf{C}(\mathbf{R}(\Delta))\|_\infty) \leq \gamma m\lambda_n/4$, then $\mathbf{0} \in \partial \tilde{\mathcal{L}}(\Omega)|_{\Omega=\tilde{\Omega}}$. •

Proof: This is a key step in the primal-dual witness approach of [5] for single-attribute graphs and that of [7] for multi-attribute graphs. With

$$\mathbf{X}(\tilde{\Omega}) = \alpha\lambda_n\mathbf{Z}(\tilde{\Omega}) + (1-\alpha)m\lambda_n\mathbf{Y}(\tilde{\Omega}), \quad (121)$$

(109) can be expressed as $\hat{\Sigma} - \tilde{\Omega}^{-1} + \mathbf{X}(\tilde{\Omega}) = \mathbf{0}$. By construction of $\tilde{\Omega}$, $\tilde{\Omega}_{S^c} = \mathbf{0}$ and $\hat{\Sigma}_S - (\tilde{\Omega}^{-1})_S + \mathbf{X}(\tilde{\Omega}_S) = \mathbf{0}$. For the unconstrained problem, we need to show that (109) holds for $\Omega = \tilde{\Omega}$, equivalently,

$$\hat{\Sigma}_S - (\tilde{\Omega}^{-1})_S + \mathbf{X}(\tilde{\Omega}_S) = \mathbf{0}, \quad (122)$$

$$\hat{\Sigma}_{S^c} - (\tilde{\Omega}^{-1})_{S^c} + \mathbf{X}(\tilde{\Omega}_{S^c}) = \mathbf{0}. \quad (123)$$

Now (122) is true for the constrained problem. It remains to show that $\mathbf{X}(\tilde{\Omega}_{S^c}) = (\tilde{\Omega}^{-1})_{S^c} - \hat{\Sigma}_{S^c}$ is a valid solution to (123) with $\mathbf{Z}(\tilde{\Omega}_{S^c})$ and $\mathbf{Y}(\tilde{\Omega}_{S^c})$ satisfying the sub-differential conditions (110) and (113) for every $e_f \in f \in S^c$, so that $\tilde{\Omega}$ qualifies as a stationary point of $\tilde{\mathcal{L}}(\Omega)$. To this end, we first rewrite (109) as

$$\hat{\Sigma} - (\Omega^*)^{-1} + (\Omega^*)^{-1}\Delta(\Omega^*)^{-1} - \mathbf{R}(\Delta) + \mathbf{X}(\tilde{\Omega}) = \mathbf{0}. \quad (124)$$

In terms of $m \times m$ submatrices of Δ , $\hat{\Sigma}$, Ω^* and $\mathbf{X}(\tilde{\Omega})$ corresponding to various graph edges, using $\text{bvec}(\mathbf{ADB}) = (\mathbf{B}^\top \boxtimes \mathbf{A})\text{bvec}(\mathbf{D})$ [26, Lemma 1], we may rewrite (124) as

$$\Gamma^*\text{bvec}(\Delta) + \text{bvec}(\mathbf{W} - \mathbf{R}(\Delta)) + \text{bvec}(\mathbf{X}(\tilde{\Omega})) = \mathbf{0}, \quad (125)$$

which then can be rewritten as

$$\begin{bmatrix} \Gamma_{S,S}^* & \Gamma_{S,S^c}^* \\ \Gamma_{S^c,S}^* & \Gamma_{S^c,S^c}^* \end{bmatrix} \begin{bmatrix} \text{bvec}(\Delta_S) \\ \text{bvec}(\Delta_{S^c}) \end{bmatrix} - \begin{bmatrix} \text{bvec}((\mathbf{W} - \mathbf{R}(\Delta))_S) \\ \text{bvec}((\mathbf{W} - \mathbf{R}(\Delta))_{S^c}) \end{bmatrix} + \begin{bmatrix} \text{bvec}(\mathbf{X}(\tilde{\Omega}_S)) \\ \text{bvec}(\mathbf{X}(\tilde{\Omega}_{S^c})) \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}. \quad (126)$$

Since $\tilde{\Delta}_{S^c} = \tilde{\Omega}_{S^c} - \Omega_{S^c}^* = \mathbf{0}$ by construction, (126) reduces to

$$\Gamma_{S,S}^* \text{bvec}(\Delta_S) + \text{bvec}((\mathbf{W} - \mathbf{R}(\Delta))_S) + \text{bvec}(\mathbf{X}(\tilde{\Omega}_S)) = \mathbf{0}, \quad (127)$$

$$\Gamma_{S^c,S}^* \text{bvec}(\Delta_S) + \text{bvec}((\mathbf{W} - \mathbf{R}(\Delta))_{S^c}) + \text{bvec}(\mathbf{X}(\tilde{\Omega}_{S^c})) = \mathbf{0}. \quad (128)$$

By construction of $\tilde{\Omega}$ as a stationary point of $\tilde{\mathcal{L}}(\Omega)$ under the constraint $\Delta_{S^c} = \mathbf{0}$, (127) is satisfied. It remains to show that (128) is true. Substituting for $\text{bvec}(\Delta_S)$ from (127) into (128), we have

$$\begin{aligned} \text{bvec}(\mathbf{X}(\tilde{\Omega}_{S^c})) &= \Gamma_{S^c,S}^* (\Gamma_{S,S}^*)^{-1} \left(\text{bvec}((\mathbf{W} - \mathbf{R}(\Delta))_S) \right. \\ &\quad \left. + \text{bvec}(\mathbf{X}(\tilde{\Omega}_S)) \right) - \text{bvec}((\mathbf{W} - \mathbf{R}(\Delta))_{S^c}). \end{aligned} \quad (129)$$

Using (121), we split (129) as

$$\begin{aligned} \alpha\lambda_n \text{bvec}(\mathbf{Z}(\tilde{\Omega}_{S^c})) &= \alpha\Gamma_{S^c,S}^* (\Gamma_{S,S}^*)^{-1} \\ &\quad \times \left(\text{bvec}((\mathbf{W} - \mathbf{R}(\Delta))_S) + \lambda_n \text{bvec}(\mathbf{Z}(\tilde{\Omega}_S)) \right) \\ &\quad - \alpha \text{bvec}((\mathbf{W} - \mathbf{R}(\Delta))_{S^c}), \end{aligned} \quad (130)$$

$$\begin{aligned} (1-\alpha)m\lambda_n \text{bvec}(\mathbf{Y}(\tilde{\Omega}_{S^c})) &= (1-\alpha)\Gamma_{S^c,S}^* (\Gamma_{S,S}^*)^{-1} \\ &\quad \times \left(\text{bvec}((\mathbf{W} - \mathbf{R}(\Delta))_S) + m\lambda_n \text{bvec}(\mathbf{Y}(\tilde{\Omega}_S)) \right) \\ &\quad - (1-\alpha) \text{bvec}((\mathbf{W} - \mathbf{R}(\Delta))_{S^c}). \end{aligned} \quad (131)$$

A solution for $\mathbf{Z}(\tilde{\Omega}_{S^c})$ is obtained via (130) and a solution for $\mathbf{Y}(\tilde{\Omega}_{S^c})$ is obtained via (131). Using the notation as explained in Sec. V-B, consider an edge $f \in S^c$ (implying $\|\Omega_f^*\|_F = 0$) with e_f denoting one of the corresponding m^2 edges in the corresponding enlarged graph (implying $|\Omega_{e_f}^*| = 0$). By (130), with $\mathbf{A} = \Gamma_{e_f,S}^* (\Gamma_{S,S}^*)^{-1} \in \mathbb{R}^{1 \times (ms)}$, $s = |S|$, we have

$$\begin{aligned} \lambda_n |Z(\tilde{\Omega}_{e_f})| &\leq |\mathbf{A} \text{bvec}(\mathbf{W}_S)| + |\mathbf{A} \text{bvec}((\mathbf{R}(\Delta))_S)| \\ &\quad + \lambda_n |\mathbf{A} \text{bvec}(\mathbf{Z}(\tilde{\Omega}_S))| + |\mathbf{W}_{e_f}| + |(\mathbf{R}(\Delta))_{e_f}| \end{aligned} \quad (132)$$

$$\begin{aligned} &\leq \|\mathbf{A}\|_1 \left(\|\mathbf{W}\|_\infty + \|\mathbf{R}(\Delta)\|_\infty + \lambda_n \|\mathbf{Z}(\tilde{\Omega}_S)\|_\infty \right) \\ &\quad + \|\mathbf{W}\|_\infty + \|\mathbf{R}(\Delta)\|_\infty \end{aligned} \quad (133)$$

Using (52) and the fact that $|Z(\tilde{\Omega}_{e_g})| \leq 1$ for any $e_g \in g \in S$, we have

$$\begin{aligned} \lambda_n |Z(\tilde{\Omega}_{e_f})| &\leq (2-\gamma)(\|\mathbf{W}\|_\infty + \|\mathbf{R}(\Delta)\|_\infty) + \lambda_n(1-\gamma) \\ &\leq (2-\gamma)\gamma\lambda_n/2 + \lambda_n(1-\gamma). \end{aligned} \quad (134)$$

Thus

$$|Z(\tilde{\Omega}_{e_f})| \leq \gamma - \frac{\gamma^2}{2} + 1 - \gamma = 1 - \frac{\gamma^2}{2} < 1, \quad (135)$$

establishing that (130) holds for some $Z(\tilde{\Omega}_{e_f})$ with $|Z(\tilde{\Omega}_{e_f})| < 1$ (strict feasibility) for any $e_f \in f \in S^c$. We now turn to (131) where we need to show $\|\text{vec}(\mathbf{Y}(\tilde{\Omega}_f))\|_2 < 1$, $\mathbf{Y}(\tilde{\Omega}_f) \in \mathbb{R}^{m \times m}$, for any $f \in S^c$. By (131), with $\mathbf{B} = \Gamma_{f,S}^* (\Gamma_{S,S}^*)^{-1} \in \mathbb{R}^{m \times (ms)}$, $s = |S|$, we have

$$\begin{aligned} m\lambda_n \|\text{vec}(\mathbf{Y}(\tilde{\Omega}_f))\|_2 &\leq \|\mathbf{B} \text{bvec}(\mathbf{W}_S)\|_2 \\ &\quad + \|\mathbf{B} \text{bvec}((\mathbf{R}(\Delta))_S)\|_2 + m\lambda_n \|\mathbf{B} \text{bvec}(\mathbf{Y}(\tilde{\Omega}_S))\|_2 \\ &\quad + \|\text{vec}(\mathbf{W}_f)\|_2 + \|\text{vec}((\mathbf{R}(\Delta))_f)\|_2 \\ &\leq \|\mathbf{C}(\mathbf{B})\|_1 \left(\|\mathbf{C}(\mathbf{W})\|_\infty + \|\mathbf{C}(\mathbf{R}(\Delta))\|_\infty \right. \\ &\quad \left. + m\lambda_n \|\mathbf{C}(\mathbf{Y}(\tilde{\Omega}_S))\|_\infty \right) + \|\mathbf{C}(\mathbf{W})\|_\infty \\ &\quad + \|\mathbf{C}(\mathbf{R}(\Delta))\|_\infty \end{aligned} \quad (137)$$

where we used the fact that $\|\mathbf{B} \text{bvec}(\mathbf{W}_S)\|_2 \leq \|\mathbf{C}(\mathbf{B})\|_1 \|\mathbf{C}(\mathbf{W}_S)\|_\infty$ following [9, Eqn. (80)] (see also [7, Lemma 13]). Using (51), the fact that $\|\text{vec}(\mathbf{Y}(\tilde{\Omega}_g))\|_2 \leq 1$ for any $g \in S$, and the bounds on $\|\mathbf{C}(\mathbf{W})\|_\infty$ and

$\|\mathcal{C}(\mathbf{R}(\Delta))\|_\infty$, we have

$$\begin{aligned} m\lambda_n \|\text{vec}(\mathbf{Y}(\tilde{\Omega}_f))\|_2 &\leq (2 - \gamma) \left(\|\mathcal{C}(\mathbf{W})\|_\infty \right. \\ &\quad \left. + \|\mathcal{C}(\mathbf{R}(\Delta))\|_\infty \right) + m\lambda_n (1 - \gamma) \\ &\leq (2 - \gamma)\gamma m\lambda_n/2 + m\lambda_n (1 - \gamma). \end{aligned} \quad (138)$$

Thus

$$\|\text{vec}(\mathbf{Y}(\tilde{\Omega}_f))\|_2 \leq \gamma - \frac{\gamma^2}{2} + 1 - \gamma = 1 - \frac{\gamma^2}{2} < 1, \quad (139)$$

proving that for some $\mathbf{Y}(\tilde{\Omega}_f)$ with $\|\text{vec}(\mathbf{Y}(\tilde{\Omega}_f))\|_2 = \|\mathbf{Y}(\tilde{\Omega}_f)\|_F < 1$ for any $f \in S^c$, (131) holds. Satisfaction of (130) and (131) implies that of (129), and hence, that of (125) and (124), yielding the desired result. ■

Lemma 5: Suppose that

$$\begin{aligned} r &:= 2\kappa_{\Gamma^*} (\|\mathcal{C}(\mathbf{W})\|_\infty + m\lambda_n) \\ &\leq \min \left(\frac{1}{3\kappa_{\Sigma^*} d_n}, \frac{1}{3\kappa_{\Gamma^*} \kappa_{\Sigma^*}^3 d_n} \right). \end{aligned} \quad (140)$$

Then $\tilde{\Omega} = \Omega^* + \Delta$ of Lemma 4 satisfies $\|\mathcal{C}(\Delta)\|_\infty \leq r$. •

Proof: Define the closed ball

$$\mathcal{B}(r) := \{\Delta_S : \|\mathcal{C}(\Delta_S)\|_\infty \leq r\} \quad (141)$$

and the gradient mapping (109)

$$G(\Omega) := \hat{\Sigma} - \Omega^{-1} + X(\Omega). \quad (142)$$

By construction $G(\tilde{\Omega}_S) := \hat{\Sigma}_S - (\Omega^{-1})_S + X(\Omega_S) = \mathbf{0}$ and $\tilde{\Omega}_{S^c} = \mathbf{0}$. As in (124),

$$\begin{aligned} G(\Omega^* + \Delta) &= (\Omega^*)^{-1} \Delta (\Omega^*)^{-1} - \mathbf{R}(\Delta) \\ &\quad + \mathbf{W} + X(\Omega^* + \Delta). \end{aligned} \quad (143)$$

Since $\tilde{\Omega}_{S^c} = \mathbf{0}$, we have $\Delta_{S^c} = \mathbf{0}$. Vectorizing and using decomposition as in (127), we have

$$\begin{aligned} \text{bvec}(G(\Omega_S^* + \Delta_S)) &= \Gamma_{S,S}^* \text{bvec}(\Delta_S) \\ &\quad + \text{bvec}((\mathbf{W} - \mathbf{R}(\Delta))_S) + \text{bvec}(X(\tilde{\Omega}_S)). \end{aligned} \quad (144)$$

Define a mapping $F(\Delta_S)$ on $\mathcal{B}(r)$ as

$$\begin{aligned} F(\Delta_S) &:= -(\Gamma_{S,S}^*)^{-1} \text{bvec}(G(\Omega_S^* + \Delta_S)) + \text{bvec}(\Delta_S) \\ &= -(\Gamma_{S,S}^*)^{-1} \left(\text{bvec}((\mathbf{W} - \mathbf{R}(\Delta))_S) + \text{bvec}(X(\tilde{\Omega}_S)) \right). \end{aligned} \quad (145)$$

The proof technique of [5] (see also [7]) is to show that $F(\mathcal{B}(r)) \subseteq \mathcal{B}(r)$ (i.e., $F(\Delta_S)$ maps $\Delta_S \in \mathcal{B}(r)$ to $F(\Delta_S) \in \mathcal{B}(r)$). Since the mapping F is continuous and $\mathcal{B}(r)$ is compact, by the Brouwer's fixed point theorem [32, Theorem 9.2] $F(\mathcal{B}(r)) \subseteq \mathcal{B}(r)$ implies that there exists a fixed point $\Delta_S \in \mathcal{B}(r)$ such that $F(\Delta_S) = \text{bvec}(\Delta_S)$, which, in turn, leads to $G(\Omega_S^* + \Delta_S) = G(\tilde{\Omega}_S) = \mathbf{0}$ establishing that the fixed point is a constrained stationary point of $\tilde{\mathcal{L}}(\Omega)$ with $\|\mathcal{C}(\Delta)\|_\infty \leq r$ since $\Delta_{S^c} = \mathbf{0}$. It remains to show that

$F(\mathcal{B}(r)) \subseteq \mathcal{B}(r)$. By (145), in a manner similar to (137), we have

$$\begin{aligned} \|\mathcal{C}(F(\Delta_S))\|_\infty &\leq \|\mathcal{C}(\Gamma_{S,S}^*)^{-1}\|_{1,\infty} \left(\|\mathcal{C}(\mathbf{W})\|_\infty \right. \\ &\quad \left. + \|\mathcal{C}(\mathbf{R}(\Delta))\|_\infty + \|\mathcal{C}(X(\tilde{\Omega}_S))\|_\infty \right). \end{aligned} \quad (146)$$

Since $|Z(\tilde{\Omega}_{e_g})| \leq 1$ for any $e_g \in g \in S$, $\|\text{vec}(Z(\tilde{\Omega}_g))\|_2 \leq m$ using the Cauchy-Schwarz inequality. Also, $\|\text{vec}(\mathbf{Y}(\tilde{\Omega}_g))\|_2 \leq 1$ for any $g \in S$. Using these two facts and (121), we have

$$\begin{aligned} \|\mathcal{C}(X(\tilde{\Omega}_S))\|_\infty &\leq \alpha\lambda_n \|\mathcal{C}(Z(\tilde{\Omega}_S))\|_\infty \\ &\quad + (1 - \alpha)m\lambda_n \|\mathcal{C}(\mathbf{Y}(\tilde{\Omega}_S))\|_\infty \\ &\leq \alpha\lambda_n m + (1 - \alpha)m\lambda_n = m\lambda_n. \end{aligned} \quad (147)$$

Therefore,

$$\|\mathcal{C}(F(\Delta_S))\|_\infty \leq \kappa_{\Gamma^*} (\|\mathcal{C}(\mathbf{W})\|_\infty + \|\mathcal{C}(\mathbf{R}(\Delta))\|_\infty + m\lambda_n). \quad (148)$$

Since $r \leq 1/(3\kappa_{\Sigma^*} d_n)$ and $\|\mathcal{C}(\Delta_S)\|_\infty \leq r$, by Lemma 3, $\|\mathcal{C}(\mathbf{R}(\Delta))\|_\infty \leq (3d_n/2) \|\mathcal{C}(\Delta)\|_\infty^2 \kappa_{\Sigma^*}^3$. Hence

$$\kappa_{\Gamma^*} \|\mathcal{C}(\mathbf{R}(\Delta))\|_\infty \leq 3\kappa_{\Gamma^*} d_n \kappa_{\Sigma^*}^3 r^2/2 \leq \frac{1}{r} \frac{r^2}{2} = \frac{r}{2}. \quad (149)$$

Thus

$$\begin{aligned} \|\mathcal{C}(F(\Delta_S))\|_\infty &\leq \frac{r}{2} + \kappa_{\Gamma^*} (\|\mathcal{C}(\mathbf{W})\|_\infty + m\lambda_n) \\ &= \frac{r}{2} + \frac{r}{2} = r. \end{aligned} \quad (150)$$

Therefore, $F(\Delta_S) \in \mathcal{B}(r)$, yielding the desired result. ■

We now turn to the proof of Theorem 4.

Proof of Theorem 4: Here we first verify the conditions in Lemmas 3-5. Pick λ_n as in (57). By Lemma 2, this choice ensures that $\|\mathcal{C}(\mathbf{W})\|_\infty \leq \gamma\lambda_n m/4$ and $\|\mathbf{W}\|_\infty \leq \gamma\lambda_n/4$ (needed in Lemma 4) with probability $> 1 - 1/p_n^{\tau-2}$ provided the sample size $n > N_1 = 2 \ln(4m^2 p_n^\tau)$. Now consider

$$\begin{aligned} r &= 2\kappa_{\Gamma^*} (\|\mathcal{C}(\mathbf{W})\|_\infty + m\lambda_n) \leq 2\kappa_{\Gamma^*} \left(1 + \frac{\gamma}{4}\right) m\lambda_n \\ &= 2\kappa_{\Gamma^*} \left(1 + \frac{4}{\gamma}\right) \tilde{C}_0 \sqrt{\ln(p_n)/n}. \end{aligned} \quad (151)$$

In Lemma 5 to satisfy (140), we pick $n > N_5$ which ensures that with probability $> 1 - 1/p_n^{\tau-2}$,

$$2\kappa_{\Gamma^*} \left(1 + \frac{4}{\gamma}\right) \tilde{C}_0 \sqrt{\ln(p_n)/n} \leq \min \left(\frac{1}{3\kappa_{\Sigma^*} d_n}, \frac{1}{3\kappa_{\Gamma^*} \kappa_{\Sigma^*}^3 d_n} \right). \quad (152)$$

By Lemma 5, we have $\|\mathcal{C}(\Delta)\|_\infty \leq r \leq 1/(3\kappa_{\Sigma^*} d_n)$, which invoking Lemma 3 implies that $\|\mathcal{C}(\mathbf{R}(\Delta))\|_\infty \leq (3/2)d_n \kappa_{\Sigma^*}^3 \|\mathcal{C}(\Delta)\|_\infty^2 \leq (3/2)d_n \kappa_{\Sigma^*}^3 r^2$. Therefore, we have

$$\begin{aligned} \|\mathcal{C}(\mathbf{R}(\Delta))\|_\infty &\leq \left(6 d_n \kappa_{\Sigma^*}^3 \kappa_{\Gamma^*}^2 \left(1 + \frac{4}{\gamma}\right)^2 \tilde{C}_0^2 \sqrt{\ln(p_n)/n}\right) \gamma\lambda_n m/4. \end{aligned} \quad (153)$$

We pick $n > N_6$ so that $\|\mathcal{C}(\mathbf{R}(\Delta))\|_\infty \leq \gamma\lambda_n m/4$ with probability $> 1 - 1/p_n^{\tau-2}$. It remains to show that in Lemma 4,

the condition $\|\mathbf{R}(\Delta)\|_\infty \leq \gamma\lambda_n/4$ holds. To this end we impose an additional condition on (151) as

$$r \leq 2\kappa_{\Gamma^*} \left(1 + \frac{4}{\gamma}\right) \tilde{C}_0 \sqrt{\ln(p_n)/n} \leq \frac{1}{3\bar{\kappa}_{\Sigma^*} \bar{d}_n}, \quad (154)$$

i.e., modify the bounds on r given in Lemma 5 as

$$r \leq \min \left(\frac{1}{3\kappa_{\Sigma^*} d_n}, \frac{1}{3\kappa_{\Gamma^*} \kappa_{\Sigma^*}^3 d_n}, \frac{1}{3\bar{\kappa}_{\Sigma^*} \bar{d}_n} \right). \quad (155)$$

By Lemma 5 and (155), we have $\|\Delta\|_\infty \leq \|\mathcal{C}(\Delta)\|_\infty \leq r \leq 1/(3\bar{\kappa}_{\Sigma^*} \bar{d}_n)$, which invoking Lemma 3 implies that $\|\mathbf{R}(\Delta)\|_\infty \leq (3/2)\bar{d}_n \bar{\kappa}_{\Sigma^*}^3 \|\Delta\|_\infty^2 \leq (3/2)\bar{d}_n \bar{\kappa}_{\Sigma^*}^3 r^2$. Therefore, we have

$$\begin{aligned} \|\mathbf{R}(\Delta)\|_\infty &\leq \left(6\bar{d}_n \bar{\kappa}_{\Sigma^*}^3 \kappa_{\Gamma^*}^2 \left(1 + \frac{4}{\gamma}\right)^2 m \tilde{C}_0 \sqrt{\ln(p_n)/n} \right) \gamma \frac{\lambda_n}{4}. \end{aligned} \quad (156)$$

We pick $n > N_7$ so that $\|\mathbf{R}(\Delta)\|_\infty \leq \gamma\lambda_n/4$ with probability $> 1 - 1/p_n^{\tau-2}$. Thus, we have proved Theorem 4(i). Theorem 4(ii) follows from Lemma 4. To prove part (iii), consider

$$\begin{aligned} \|\mathcal{C}(\hat{\Omega} - \Omega^*)\|_F &= \sqrt{\sum_{\{k,\ell\} \in S} \|\hat{\Omega}^{(k,\ell)} - (\Omega^*)^{(k,\ell)}\|_F^2} \\ &\leq \sqrt{s_n^* + p_n} \|\mathcal{C}(\hat{\Omega} - \Omega^*)\|_\infty, \end{aligned} \quad (157)$$

where in the last step above we used the Cauchy-Schwarz inequality. Finally, to establish part (iv), note that parts (i)-(iii) hold with probability $> 1 - 1/p_n^{\tau-2}$ (with high probability (w.h.p.)). If $\min_{(k,\ell) \in S} \|(\Omega^*)^{(k,\ell)}\|_F \geq 2\|\mathcal{C}(\hat{\Omega} - \Omega^*)\|_\infty$,

$$\begin{aligned} \|\mathcal{C}(\hat{\Omega} - \Omega^*)\|_\infty &= \|\mathcal{C}((\hat{\Omega} - \Omega^*)_S)\|_\infty \\ &\leq (1/2) \min_{(k,\ell) \in S} \|(\Omega^*)^{(k,\ell)}\|_F. \end{aligned} \quad (158)$$

For any edge $\{k, \ell\} = f \in S$, using the Cauchy-Schwarz inequality, we have

$$\begin{aligned} (1/2) \min_{(k,\ell) \in S} \|(\Omega^*)^{(k,\ell)}\|_F &\geq \|(\hat{\Omega} - \Omega^*)_f\|_F \\ &\geq \|\hat{\Omega}_f^* - \|\hat{\Omega}_f\|_F, \end{aligned} \quad (159)$$

therefore,

$$\begin{aligned} \|\hat{\Omega}_f\|_F &\geq \|\hat{\Omega}_f^*\|_F - (1/2) \min_{(k,\ell) \in S} \|(\Omega^*)^{(k,\ell)}\|_F \\ &\geq (1/2) \min_{(k,\ell) \in S} \|(\Omega^*)^{(k,\ell)}\|_F > 0, \end{aligned} \quad (160)$$

while $\hat{\Omega}_{S^c} = \mathbf{0}$ w.h.p. ■

Proof of Theorem 5: We note that in terms of \tilde{R} and \tilde{r}_n , Theorem 4 implies that

$$\|\mathcal{C}(\hat{\Omega} - \Omega^*)\|_F \leq \tilde{R} \tilde{r}_n. \quad (161)$$

If $1/\beta_{\min} \leq 0.99\tilde{\mu}$, then $\Omega^* \in \mathcal{B}$ since $\|\Omega^*\| \leq 1/\beta_{\min}$ by the stated assumption $\beta_{\min} \leq \phi_{\min}(\Sigma^*)$. Now we establish

that $\hat{\Omega} \in \mathcal{B}$. To this end, as in the proof of Theorem 2, consider

$$\begin{aligned} \|\hat{\Omega}\| &\leq \|\hat{\Omega} - \Omega^*\| + \|\Omega^*\| \\ &\leq \|\hat{\Omega} - \Omega^*\|_F + \|\Omega^*\| \\ &= \|\mathcal{C}(\hat{\Omega} - \Omega^*)\|_F + \|\Omega^*\| \\ &\leq \tilde{R} \tilde{r}_n + 1/\beta_{\min}. \end{aligned} \quad (162)$$

Therefore, $\hat{\Omega} \in \mathcal{B}$. Thus, both $\hat{\Omega}$ and Ω^* are feasible. The desired result then follows from Theorem 4 and (local) strict convexity of $\tilde{\mathcal{L}}(\Omega)$ over \mathcal{B} implied by Lemma 1. ■

REFERENCES

- [1] J. Whittaker, *Graphical Models in Applied Multivariate Statistics*. New York, NY, USA: Wiley, 1990.
- [2] S. L. Lauritzen, *Graphical Models*. Oxford, U.K.: Oxford Univ. Press, 1996.
- [3] P. Bühlmann and S. van de Geer, *Statistics for High-Dimensional Data*. Berlin, Germany: Springer, 2011.
- [4] C. Lam and J. Fan, "Sparsistency and rates of convergence in large covariance matrix estimation," *Ann. Statist.*, vol. 37, no. 6B, pp. 4254–4278, Dec. 2009.
- [5] P. Ravikumar, M. J. Wainwright, G. Raskutti, and B. Yu, "High-dimensional covariance estimation by minimizing ℓ_1 -penalized log-determinant divergence," *Electron. J. Statist.*, vol. 5, pp. 935–980, Jan. 2011.
- [6] M. J. Wainwright, *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge, U.K.: Cambridge Univ. Press, 2019.
- [7] M. Kolar, H. Liu, and E. P. Xing, "Graph estimation from multi-attribute data," *J. Mach. Learn. Res.*, vol. 15, pp. 1713–1750, Jan. 2012.
- [8] J. K. Tugnait, "Sparse-group lasso for graph learning from multi-attribute data," *IEEE Trans. Signal Process.*, vol. 69, pp. 1771–1786, 2021.
- [9] J. K. Tugnait, "Learning high-dimensional differential graphs from multiattribute data," *IEEE Trans. Signal Process.*, vol. 72, pp. 415–431, 2024.
- [10] G. Marjanovic and V. Solo, "Vector l0 sparse conditional independence graphs," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 2731–2735.
- [11] P. Sundaram, M. Luessi, M. Bianciardi, S. Stufflebeam, M. Hämmäläinen, and V. Solo, "Individual resting-state brain networks enabled by massive multivariate conditional mutual information," *IEEE Trans. Med. Imag.*, vol. 39, no. 6, pp. 1957–1966, Jun. 2020.
- [12] J. Friedman, T. Hastie, and R. Tibshirani, "A note on the group lasso and a sparse group lasso," 2010, *arXiv:1001.0736*.
- [13] N. Simon, J. H. Friedman, T. Hastie, and R. Tibshirani, "A sparse-group lasso," *J. Comput. Graph. Statist.*, vol. 22, no. 2, pp. 231–245, Sep. 2012.
- [14] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *J. Roy. Stat. Soc. Ser. B: Stat. Methodol.*, vol. 68, no. 1, pp. 49–67, Feb. 2006.
- [15] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its Oracle properties," *J. Amer. Stat. Assoc.*, vol. 96, no. 456, pp. 1348–1360, Dec. 2001.
- [16] 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, Jan. 2008.
- [17] H. Zou and R. Li, "One-step sparse estimates in nonconcave penalized likelihood models," *Ann. Statist.*, vol. 36, no. 4, pp. 1509–1533, Aug. 2008.
- [18] J. K. Tugnait, "Sparse graph learning under Laplacian-related constraints," *IEEE Access*, vol. 9, pp. 151067–151079, 2021.
- [19] J. K. Tugnait, "Sparse-group log-sum penalized graphical model learning for time series," in *Proc. ICASSP - IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Singapore, May 2022, pp. 5822–5826.
- [20] P.-L. Loh and M. J. Wainwright, "Regularized M-estimators with non-convexity: Statistical and algorithmic theory for local optima," *J. Mach. Learn. Res.*, vol. 16, pp. 559–616, Jan. 2015.
- [21] P.-L. Loh and M. J. Wainwright, "Support recovery without incoherence: A case for nonconvex regularization," *Ann. Statist.*, vol. 45, no. 6, pp. 2455–2482, Dec. 2017.

- [22] J. K. Tugnait, "Sparse-group non-convex penalized multi-attribute graphical model selection," in *Proc. 29th Eur. Signal Process. Conf. (EUSIPCO)*, Dublin, Ireland, Aug. 2021, pp. 1850–1854.
- [23] A. J. Rothman, P. J. Bickel, E. Levina, and J. Zhu, "Sparse permutation invariant covariance estimation," *Electron. J. Statist.*, vol. 2, no. none, pp. 494–515, Jan. 2008.
- [24] S. Boyd, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends[®] Mach. Learn.*, vol. 3, no. 1, pp. 1–122, 2010.
- [25] J. K. Tugnait, "On sparse high-dimensional graph estimation from multi-attribute data," in *Proc. 58th Asilomar Conf. Signals, Syst., Comput.*, Pacific Grove, CA, USA, Oct. 2024, pp. 1053–1057.
- [26] D. S. Tracy and K. G. Jinadasa, "Partitioned Kronecker products of matrices and applications," *Can. J. Statist.*, vol. 17, no. 1, pp. 107–120, Mar. 1989.
- [27] S. Liu, "Matrix results on the Khatri–Rao and Tracy–Singh products," *Linear Algebra Appl.*, vol. 289, nos. 1–3, pp. 267–277, Mar. 1999.
- [28] B. Zhao, Y. S. Wang, and M. Kolar, "FuDGE: A method to estimate a functional differential graph in a high-dimensional setting," *J. Mach. Learn. Res.*, vol. 23, pp. 1–82, Jan. 2020.
- [29] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, Oct. 1999.
- [30] S. Lu, J. Kang, W. Gong, and D. Towsley, "Complex network comparison using random walks," in *Proc. 23rd Int. Conf. World Wide Web*, Apr. 2014, pp. 727–730.
- [31] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.
- [32] V. Pata, *Fixed Point Theorems and Applications*. New York, NY, USA: Springer-Verlag, 2019.



JITENDRA K. TUGNAIT (Life Fellow, IEEE) received the B.Sc. degree (Hons.) in electronics and electrical communication engineering from Punjab Engineering College, Chandigarh, India, in 1971, the M.S. and E.E. degrees in electrical engineering from Syracuse University, Syracuse, NY, USA, in 1973 and 1974, respectively, and the Ph.D. degree in electrical engineering from the University of Illinois at Urbana–Champaign, Champaign, IL, USA, in 1978.

From 1978 to 1982, he was an Assistant Professor of electrical and computer engineering with the University of Iowa, Iowa City, IA, USA. He was with the Long Range Research Division, Exxon Production Research Company, Houston, TX, USA, from June 1982 to September 1989. In September 1989, he joined the Department of Electrical and Computer Engineering, Auburn University, Auburn, AL, USA, as a Professor, where he is currently a James B. Davis Professor. His current research interests include statistical signal processing and machine learning for signal processing.

Dr. Tugnait has served as an Associate Editor for IEEE TRANSACTIONS ON AUTOMATIC CONTROL, IEEE TRANSACTIONS ON SIGNAL PROCESSING, IEEE SIGNAL PROCESSING LETTERS, and IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS; a Senior Area Editor for IEEE TRANSACTIONS ON SIGNAL PROCESSING; and a Senior Editor for IEEE WIRELESS COMMUNICATIONS LETTERS.

• • •