This article has been accepted for publication in IEEE Transactions on Signal Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TSP.2023.3343553

Learning High-Dimensional Differential Graphs From Multi-Attribute Data

Jitendra K. Tugnait

***Abstract*—We consider the problem of estimating differences in two Gaussian graphical models (GGMs) which are known to have similar structure. The GGM structure is encoded in its precision (inverse covariance) matrix. In many applications one is interested in estimating the difference in two precision matrices to characterize underlying changes in conditional dependencies of two sets of data. Existing methods for differential graph estimation are based on single-attribute (SA) models where one associates a scalar random variable with each node. In multi-attribute (MA) graphical models, each node represents a random vector. In this paper, we analyze a group lasso penalized D-trace loss function approach for differential graph learning from multi-attribute data. An alternating direction method of multipliers (ADMM) algorithm is presented to optimize the objective function. Theoretical analysis establishing consistency in support recovery and estimation in high-dimensional settings is provided. Numerical results based on synthetic as well as real data are presented.**

***Index Terms*—Sparse graph learning; differential graph estimation; undirected graph; multi-attribute graphs.**

I. Introduction

GRAPHICAL models provide a powerful tool for analyzing multivariate data [1], [2]. In a statistical graphical model, the conditional statistical dependency structure among p random variables $x_1, x_1, \cdots, x_p$, is represented using an undirected graph $\mathcal{G} = (V, \mathcal{E})$, where $V = \{1, 2, \cdots, p\} = [p]$ is the set of p nodes corresponding to the p random variables x_is, and $\mathcal{E} \subseteq V \times V$ is the set of undirected edges describing conditional dependencies among the components of $\boldsymbol{x}$. The graph $\mathcal{G}$ then is a conditional independence graph (CIG) where there is no edge between nodes i and j (i.e., $\{i, j\} \notin \mathcal{E}$) iff x_i and x_j are conditionally independent given the remaining p-2 variables x_ℓ, $\ell \in [p]$, $\ell \neq i$, $\ell \neq j$. In particular, Gaussian graphical models (GGMs) are CIGs where $\boldsymbol{x}$ is multivariate Gaussian. Suppose $\boldsymbol{x}$ has positive-definite covariance matrix $\boldsymbol{\Sigma}$ with inverse covariance matrix $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$. Then Ω_{ij}, the (i, j)-th element of $\boldsymbol{\Omega}$, is zero iff x_i and x_j are conditionally independent. Such models for $\boldsymbol{x}$ have been extensively studied. Given n samples of $\boldsymbol{x}$, in *high-dimensional settings* where $p \gg 1$ and/or n is of the order of p, one estimates $\boldsymbol{\Omega}$ under some sparsity constraints; see [3]–[6].

More recently there has been increasing interest in differential network analysis where one is interested in estimating the difference in two inverse covariance matrices [7]–[9]. Given observations $\boldsymbol{x}$ and $\boldsymbol{y}$ from two groups of subjects, one is interested in the difference $\boldsymbol{\Delta} = \boldsymbol{\Omega}_y - \boldsymbol{\Omega}_x$, where $\boldsymbol{\Omega}_x = (E\{\boldsymbol{x}\boldsymbol{x}^\top\})^{-1}$ and $\boldsymbol{\Omega}_y = (E\{\boldsymbol{y}\boldsymbol{y}^\top\})^{-1}$. The associated differential graph is $\mathcal{G}_\Delta = (V, \mathcal{E}_\Delta)$ where $\{i, j\} \in \mathcal{E}_\Delta$ iff $\boldsymbol{\Delta}_{ij} \neq 0$. It characterizes differences between the GGMs of the two sets of data. We use the term differential graph as in [10], [11] ([7]–[9] use the term differential network). As noted in [9], in biostatistics, the differential network/graph describes the changes in conditional dependencies between components under different environmental or genetic conditions. For instance, one may be interested in the differences in the graphical models of healthy and impaired subjects, or models under different disease states, given gene expression data or functional MRI signals [3], [12], [13].

In the preceding 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 (MA) graphical models in [14]–[17] and vector graphs or networks in [18]–[21]. In a gene regulatory network, one may have different molecular profiles available for a single gene, such as protein, DNA and RNA. Since these molecular profiles are on the same set of biological samples, they constitute multi-attribute data for gene regulatory graphical models in [14], [16]. Consider p jointly Gaussian vectors $\boldsymbol{z}_i \in \mathbb{R}^m$, $i \in [p]$. We associate $\boldsymbol{z}_i$ with the ith node of graph $\mathcal{G} = (V, \mathcal{E})$, $V = [p]$, $\mathcal{E} \subseteq V \times V$. We now have m attributes per node. Now $\{i, j\} \in \mathcal{E}$ iff vectors $\boldsymbol{z}_i$ and $\boldsymbol{z}_j$ are conditionally independent given the remaining p-2 vectors $\{\boldsymbol{z}_\ell, \ell \in V \backslash \{i, j\}\}$. Let $\boldsymbol{x} = [\boldsymbol{z}_1^\top \ \boldsymbol{z}_2^\top \ \cdots \ \boldsymbol{z}_p^\top]^\top \in \mathbb{R}^{mp}$. Let $\boldsymbol{\Omega} = (E\{\boldsymbol{x}\boldsymbol{x}^\top\})^{-1}$ assuming $E\{\boldsymbol{x}\boldsymbol{x}^\top\} \succ \boldsymbol{0}$. Define the $m \times m$ subblock $\boldsymbol{\Omega}^{(ij)}$ of $\boldsymbol{\Omega}$ as $[\boldsymbol{\Omega}^{(ij)}]_{rs} = [\boldsymbol{\Omega}]_{(i-1)m+r,(j-1)m+s}$, $r, s = 1, 2, \cdots, m$. Then we have the following equivalence [16, Sec. 2.1]

$$\{i, j\} \notin \mathcal{E} \ \Leftrightarrow \ \boldsymbol{\Omega}^{(ij)} = \boldsymbol{0}\,. \tag{1}$$

This paper is concerned with estimation of differential graphs from multi-attribute data. Given independent and identically distributed (i.i.d.) samples $\boldsymbol{x}(t)$, $t = 1, 2, \cdots, n_x$, of $\boldsymbol{x} = [\boldsymbol{z}_1^\top \ \boldsymbol{z}_2^\top \ \cdots \ \boldsymbol{z}_p^\top]^\top \in \mathbb{R}^{mp}$ where $\boldsymbol{z}_i \in \mathbb{R}^m$, $i \in [p]$, are jointly Gaussian, and similarly given samples $\boldsymbol{y}(t)$, $t = 1, 2, \cdots, n_y$, of $\boldsymbol{y} \in \mathbb{R}^{mp}$, our objective is to estimate the difference $\boldsymbol{\Delta} = \boldsymbol{\Omega}_y - \boldsymbol{\Omega}_x$, and determine the differential graph $\mathcal{G}_\Delta = (V, \mathcal{E}_\Delta)$ with edgeset $\mathcal{E}_\Delta = \{\{k, \ell\} : \|\boldsymbol{\Delta}^{(k\ell)}\|_F \neq 0\}$.

A. Related Work

All prior work on high-dimensional differential graph estimation from i.i.d. samples addresses single-attribute (SA) models where each node represents a scalar random variable. One naive approach would be to estimate the two precision

J.K. Tugnait is with the Department of Electrical & Computer Engineering, 200 Broun Hall, Auburn University, Auburn, AL 36849, USA. Email: tugnajk@auburn.edu .

This work was supported by the National Science Foundation Grants ECCS-2040536 and CCF-2308473.

matrices separately by any existing estimator (see [4], [5] and references therein) and then calculate their difference to estimate the differential graph. (This approach is also applicable to MA graphs.) This approach estimates twice the number of parameters, hence needs larger sample sizes for same accuracy, and also imposes sparsity constraints on each precision matrix for the methods to work. The same comment applies to methods such as [3], [6], where the two precision matrices and their differences are jointly estimated. A recent survey is in [22]. In these approaches, given $K \geq 2$ related groups of data, each p-variate and sharing the same set of nodes V, but possibly differing in connected edgesets, the objective is to jointly estimate the K precision matrices and their pairwise differences, with sparsity constraints on each of the K precision matrices and their pairwise differences. These approaches require each of the K precision matrices to be sparse. If only the differences in the precision matrices is of interest, alternative approaches exist where no sparsity constraints are imposed on individual precision matrices. For instance, direct estimation of the difference in the two precision matrices has been considered for SA graphs in [7]–[9], [12], [23]–[28], where only the difference is required to be sparse, not the two individual precision matrices. In [7]–[9], [23], [24], [28] precision difference matrix estimators are based on a D-trace loss [29], while [12] discusses a Dantzig selector type estimator. In [25]–[27] differential graph is estimated by directly modeling the ratio of the probability densities of the random vectors under the two graphs.

Estimation of MA differential graphs has not been investigated before. The work of [10], [11] is similar to an MA formulation except that in [10], [11], $\boldsymbol{x}(t)$ and $\boldsymbol{y}(t)$ are non-stationary ("functional" modeling), and instead of a single record (sample) of $\boldsymbol{x}(t)$, $t = 1, 2, \cdots, n_x$ and $\boldsymbol{y}(t)$, $t = 1, 2, \cdots, n_y$, as in this paper, they assume multiple independent observations of $\boldsymbol{x}(t)$, $t \in \mathcal{T}$ (a closed subset of real line), and $\boldsymbol{y}(t)$, $t \in \mathcal{T}$. The objective function in [11, Eqns. (10)-(11)] is the same as our objective function (3)-(4), but consequent estimation of edges and theoretical analysis are vastly different. We estimate edges as in (6), i.e., our threshold is set at zero and this is the method analyzed in our Theorem 1(iv) for graph recovery with high probability. In [10], [11], this threshold is set at a parameter $\epsilon_n > 0$ (see [11, Eqn. (13)]) which is a function of sample size n, number of nonzero entries in true $\boldsymbol{\Delta}$, smallest eigenvalues of true covariances $\boldsymbol{\Omega}_y^{-1}$ and $\boldsymbol{\Omega}_x^{-1}$ (in our notation), and several other factors. That is, ϵ_n is unknowable for practical implementation and it is used as a theoretical construct to establish graph support recovery in [11, Theorem 10]. In simulations, [10], [11] set $\epsilon_n = 0$. That is, [10], [11] do not analyze what they implement (the proof does not hold for $\epsilon_n = 0$), and they do not implement what they analyze (ϵ_n is unknowable). There is no counterpart to our Theorem 1 in [10], [11], and the methodology of our Theorem 1 allows us to set the edge detection threshold to zero. Our Theorem 2 follows the general framework of [30] to bound the Frobenius norm of the error in estimating $\boldsymbol{\Delta}$, and [10], [11] also follow the general framework of [30] for the same purpose. But their extension of this result to graph recovery does not permit zero threshold for edge detection. We attempt no such extension.

B. Our Contributions

In this paper, we analyze a group lasso penalized D-trace loss function approach for differential graph learning from MA data, extending the SA approach of [8], [28]. A two-block ADMM algorithm is presented to optimize the objective function. The two-block ADMM is guaranteed to be convergent unlike the three-block ADMM method used in [8]. Two different approaches to theoretical analysis of the proposed approach in high-dimensional settings are presented. Theorem 1 follows the approach(es) of [8], [16], [28], [29], [31] while Theorem 2 follows the general framework of [30], not used in [8], [28]. The general method of [31] requires an irrepresentability condition (see (20)) which is also required in [8], [28] for SA graphs, but is not needed by the method of [30], hence in our Theorem 2. Numerical results based on synthetic as well as real data are presented.

Preliminary version of parts of this paper appear in a conference paper [32]. Theorem 2, proof of Theorem 1 and real data example do not appear in [32].

C. Outline and Notation

The rest of the paper is organized as follows. A group lasso penalized D-trace loss function is presented in Sec. II for estimation of multi-attribute differential graph. An ADMM algorithm is presented in Sec. III to optimize the convex objective function. In Sec. IV we analyze the properties of the estimator of the difference $\boldsymbol{\Delta} = \boldsymbol{\Omega}_y - \boldsymbol{\Omega}_x$. Theorem 1 follows the approach(es) of [8], [16], [28], [29], [31] while Theorem 2 follows the general framework of [30]. The general method of [31] requires an irrepresentability condition (see (20)) which is not needed by the method of [30]. On the other hand, our Theorem 2 does not have a result like Theorem 1(ii), the oracle property, nor does it have a result as in Theorem 1(iv), support recovery. Numerical results based on synthetic as well as real data are presented in Sec. V to illustrate the proposed approach. Proofs of Theorems 1 and 2 are given in Appendices A and B, respectively.

For a set V, $|V|$ or card(V) denotes its cardinality. Given $\boldsymbol{A} \in \mathbb{R}^{p\times p}$, we use $\phi_{\min}(\boldsymbol{A})$, $\phi_{\max}(\boldsymbol{A})$, $|\boldsymbol{A}|$ and tr$(\boldsymbol{A})$ to denote the minimum eigenvalue, maximum eigenvalue, determinant and trace of $\boldsymbol{A}$, respectively. For $\boldsymbol{B} \in \mathbb{R}^{p\times q}$, we define $\|\boldsymbol{B}\| = \sqrt{\phi_{\max}(\boldsymbol{B}^\top \boldsymbol{B})}$, $\|\boldsymbol{B}\|_F = \sqrt{\text{tr}(\boldsymbol{B}^\top \boldsymbol{B})}$, $\|\boldsymbol{B}\|_1 = \sum_{i,j} |B_{ij}|$, where B_{ij} is the (i,j)-th element of $\boldsymbol{B}$ (also denoted by $[\boldsymbol{B}]_{ij}$), $\|\boldsymbol{B}\|_\infty = \max_{i,j} |B_{ij}|$ and $\|\boldsymbol{B}\|_{1,\infty} = \max_i \sum_j |B_{ij}|$. The symbols $\otimes$ and $\boxtimes$ denote Kronecker product and Tracy-Singh product [33], respectively. In particular, given block partitioned matrices $\boldsymbol{A} = [\boldsymbol{A}_{ij}]$ and $\boldsymbol{B} = [\boldsymbol{B}_{k\ell}]$ with submatrices $\boldsymbol{A}_{ij}$ and $\boldsymbol{B}_{k\ell}$, Tracy-Singh product yields another block partitioned matrix $\boldsymbol{A} \boxtimes \boldsymbol{B} = [\boldsymbol{A}_{ij} \boxtimes \boldsymbol{B}]_{ij} = [[\boldsymbol{A}_{ij} \otimes \boldsymbol{B}_{k\ell}]_{k\ell}]_{ij}$ [34]. Given $\boldsymbol{A} = [\boldsymbol{A}_{ij}] \in \mathbb{R}^{mp\times mp}$ with $\boldsymbol{A}_{ij} \in \mathbb{R}^{m\times m}$, vec$(\boldsymbol{A}) \in \mathbb{R}^{m^2p^2}$ denotes the vectorization of $\boldsymbol{A}$ which stacks the columns of the matrix $\boldsymbol{A}$, and

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

Let $S = \mathcal{E}_\Delta = \{\{k,\ell\} : \|\boldsymbol{\Delta}^{(k\ell)}\|_F \neq 0\}$ where $\boldsymbol{\Delta} = [\boldsymbol{\Delta}^{(k\ell)}] \in \mathbb{R}^{mp\times mp}$ with $\boldsymbol{\Delta}^{(k\ell)} \in \mathbb{R}^{m\times m}$ denoting the (k,l)th $m\times m$ submatrix of $\boldsymbol{\Delta}$. Then $\boldsymbol{\Delta}_S$ denotes the submatrix of $\boldsymbol{\Delta}$ with block rows and columns indexed by S, i.e., $\boldsymbol{\Delta}_S = [\boldsymbol{\Delta}^{(k\ell)}]_{(k,\ell)\in S}$. Suppose $\boldsymbol{\Gamma} = \boldsymbol{A} \boxtimes \boldsymbol{B}$ given block partitioned matrices $\boldsymbol{A} = [\boldsymbol{A}_{ij}]$ and $\boldsymbol{B} = [\boldsymbol{B}_{k\ell}]$. For any two subsets T_1 and T_2 of $V \times V$, $\boldsymbol{\Gamma}_{T_1,T_2}$ denotes the submatrix of $\boldsymbol{\Gamma}$ with block rows and columns indexed by T_1 and T_2, i.e., $\boldsymbol{\Gamma}_{T_1,T_2} = [\boldsymbol{A}_{j\ell} \otimes \boldsymbol{B}_{kq}]_{(j,k)\in T_1,(\ell,q)\in T_2}$. Following [16], an operator $\boldsymbol{\mathcal{C}}(\cdot)$ is used in Sec. IV. Consider $\boldsymbol{A} \in \mathbb{R}^{mp\times mp}$ with (k,l)th $m\times m$ submatrix $\boldsymbol{A}^{(k\ell)}$. Then $\boldsymbol{\mathcal{C}}(\cdot)$ operates on $\boldsymbol{A}$ as

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

with $\boldsymbol{\mathcal{C}}(\boldsymbol{A}^{(k\ell)}) = \|\boldsymbol{A}^{(k\ell)}\|_F$ and $\boldsymbol{\mathcal{C}}(\boldsymbol{A}) \in \mathbb{R}^{p\times p}$. Now consider $\boldsymbol{A}_1, \boldsymbol{A}_2 \in \mathbb{R}^{mp\times mp}$ with (k,l)th $m\times m$ submatrices $\boldsymbol{A}_1^{(k\ell)}$ and $\boldsymbol{A}_2^{(k\ell)}$, respectively, and Tracy-Singh product $\boldsymbol{A}_1 \boxtimes \boldsymbol{A}_2 \in \mathbb{R}^{(mp)^2\times(mp)^2}$. Then $\boldsymbol{\mathcal{C}}(\cdot)$ operates on $\boldsymbol{A}_1 \boxtimes \boldsymbol{A}_2$ as $\boldsymbol{\mathcal{C}}(\boldsymbol{A}_1 \boxtimes \boldsymbol{A}_2) \in \mathbb{R}^{p^2\times p^2}$ with $\boldsymbol{\mathcal{C}}(\boldsymbol{A}_1^{(k_1\ell_1)} \otimes \boldsymbol{A}_2^{(k_2\ell_2)}) = \|\boldsymbol{A}_1^{(k_1\ell_1)} \otimes \boldsymbol{A}_2^{(k_2\ell_2)}\|_F$ (=$\|\boldsymbol{A}_1^{(k_1\ell_1)}\|_F \, \|\boldsymbol{A}_2^{(k_2\ell_2)}\|_F$). That is, each $m^2 \times m^2$ submatrix $\boldsymbol{A}_1^{(k_1\ell_1)} \otimes \boldsymbol{A}_2^{(k_2\ell_2)}$ of $\boldsymbol{A}_1 \boxtimes \boldsymbol{A}_2$ is mapped into its Frobenius norm.

II. Group Lasso Penalized D-Trace Loss

Let $\boldsymbol{x} = [\boldsymbol{z}_{1x}^\top \; \boldsymbol{z}_{2x}^\top \; \cdots \; \boldsymbol{z}_{px}^\top]^\top \in \mathbb{R}^{mp}$ where $\boldsymbol{z}_{ix} \in \mathbb{R}^m$, $i \in [p]$, are zero-mean, jointly Gaussian. Similarly, let $\boldsymbol{y} = [\boldsymbol{z}_{1y}^\top \; \boldsymbol{z}_{2y}^\top \; \cdots \; \boldsymbol{z}_{py}^\top]^\top \in \mathbb{R}^{mp}$ where $\boldsymbol{z}_{iy} \in \mathbb{R}^m$, $i \in [p]$, are zero-mean, jointly Gaussian. Given i.i.d. samples $\boldsymbol{x}(t)$, $t = 1, 2, \cdots, n_x$, of $\boldsymbol{x}$, and similarly given i.i.d. samples $\boldsymbol{y}(t)$, $t = 1, 2, \cdots, n_y$, of $\boldsymbol{y} \in \mathbb{R}^{mp}$, form the sample covariance estimates

$$\hat{\boldsymbol{\Sigma}}_x = \frac{1}{n_x}\sum_{t=1}^{n_x} \boldsymbol{x}(t)\boldsymbol{x}^\top(t)\,, \quad \hat{\boldsymbol{\Sigma}}_y = \frac{1}{n_y}\sum_{t=1}^{n_y} \boldsymbol{y}(t)\boldsymbol{y}^\top(t)\,. \tag{2}$$

and denote their true values as $\boldsymbol{\Sigma}_x^* = \boldsymbol{\Omega}_x^{-*} (= (\boldsymbol{\Omega}_x^*)^{-1})$ and $\boldsymbol{\Sigma}_y^* = \boldsymbol{\Omega}_y^{-*}$. Assume that $\{\boldsymbol{x}(t)\}$ and $\{\boldsymbol{y}(t)\}$ are mutually independent sequences. Assume $\boldsymbol{\Sigma}_x^*$ and $\boldsymbol{\Sigma}_y^*$ are positive definite. We wish to estimate $\boldsymbol{\Delta} = \boldsymbol{\Omega}_y^* - \boldsymbol{\Omega}_x^*$ and graph $\mathcal{G}_\Delta = (V, \mathcal{E}_\Delta)$, based on $\hat{\boldsymbol{\Sigma}}_x$ and $\hat{\boldsymbol{\Sigma}}_y$. Following the SA formulation of [8] (see also [28, Sec. 2.1]), we will use a convex D-trace loss function given by

$$L(\boldsymbol{\Delta}, \hat{\boldsymbol{\Sigma}}_x, \hat{\boldsymbol{\Sigma}}_y) = \frac{1}{2}\mathrm{tr}(\hat{\boldsymbol{\Sigma}}_x\boldsymbol{\Delta}\hat{\boldsymbol{\Sigma}}_y\boldsymbol{\Delta}^\top) - \mathrm{tr}(\boldsymbol{\Delta}(\hat{\boldsymbol{\Sigma}}_x - \hat{\boldsymbol{\Sigma}}_y)) \tag{3}$$

where D-trace refers to difference-in-trace loss function, a term coined in [29] in the context of graphical model estimation. The function $L(\boldsymbol{\Delta}, \boldsymbol{\Sigma}_x^*, \boldsymbol{\Sigma}_y^*)$ is strictly convex in $\boldsymbol{\Delta}$ (its Hessian w.r.t. vec($\boldsymbol{\Delta}$) is $\boldsymbol{\Sigma}_y^* \otimes \boldsymbol{\Sigma}_x^*$), and has a unique minimum at $\boldsymbol{\Delta}^* = \boldsymbol{\Omega}_y^* - \boldsymbol{\Omega}_x^*$ [8], [28]. When we use sample covariances, we propose to estimate $\boldsymbol{\Delta}$ by minimizing the group-lasso penalized loss function

$$L_\lambda(\boldsymbol{\Delta}, \hat{\boldsymbol{\Sigma}}_x, \hat{\boldsymbol{\Sigma}}_y) = L(\boldsymbol{\Delta}, \hat{\boldsymbol{\Sigma}}_x, \hat{\boldsymbol{\Sigma}}_y) + \lambda \sum_{k,\ell=1}^{p} \|\boldsymbol{\Delta}^{(k\ell)}\|_F \tag{4}$$

where $\lambda > 0$ is a tuning parameter and $\|\boldsymbol{\Delta}^{(k\ell)}\|_F$ promotes blockwise sparsity in $\boldsymbol{\Delta}$ [35]–[37] where, if we partition $\boldsymbol{\Delta}$ into $m\times m$ submatrices, $\boldsymbol{\Delta}^{(k\ell)}$ denotes its (k,ℓ)th submatrix, associated with edge $\{k,\ell\}$ of the differential graph $\mathcal{G}_\Delta = (V, \mathcal{E}_\Delta)$.

For SA models ($m = 1$), [28] has used the lasso-penalized loss function $L_J(\boldsymbol{\Delta}) = L(\boldsymbol{\Delta}, \hat{\boldsymbol{\Sigma}}_x, \hat{\boldsymbol{\Sigma}}_y) + \lambda \sum_{k,\ell=1}^{p} \big|\boldsymbol{\Delta}_{k\ell}\big|$. The cost $L_J(\boldsymbol{\Delta})$ is optimized in [28] using a two-block ADMM approach which is known to be convergent. The resulting estimator $\hat{\boldsymbol{\Delta}}$ that minimizes the above cost is not necessarily symmetric. To obtain a symmetric estimator for SA models, [8] proposes the lasso-penalized loss function $L_Y(\boldsymbol{\Delta}) = \frac{1}{4}\mathrm{tr}(\hat{\boldsymbol{\Sigma}}_x\boldsymbol{\Delta}\hat{\boldsymbol{\Sigma}}_y\boldsymbol{\Delta}^\top + \hat{\boldsymbol{\Sigma}}_y\boldsymbol{\Delta}\hat{\boldsymbol{\Sigma}}_x\boldsymbol{\Delta}^\top) - \mathrm{tr}(\boldsymbol{\Delta}(\hat{\boldsymbol{\Sigma}}_x - \hat{\boldsymbol{\Sigma}}_y)) + \lambda \sum_{k,\ell=1}^{p} \big|\boldsymbol{\Delta}_{k\ell}\big|$. In [8], cost $L_Y(\boldsymbol{\Delta})$ is optimized using a three-block ADMM method which is not necessarily convergent.

Suppose

$$\hat{\boldsymbol{\Delta}} = \arg\min_{\boldsymbol{\Delta}} L_\lambda(\boldsymbol{\Delta}, \hat{\boldsymbol{\Sigma}}_x, \hat{\boldsymbol{\Sigma}}_y)\,. \tag{5}$$

Even though $\boldsymbol{\Delta}$ is symmetric, $\hat{\boldsymbol{\Delta}}$ is not. We can symmetrize it by setting $\hat{\boldsymbol{\Delta}}_{sym} = \frac{1}{2}(\hat{\boldsymbol{\Delta}} + \hat{\boldsymbol{\Delta}}^\top)$, after obtaining $\hat{\boldsymbol{\Delta}}$. Then the differential graph edges are estimated as

$$\hat{\mathcal{E}}_\Delta = \Big\{\{k,\ell\} : \|\hat{\boldsymbol{\Delta}}_{sym}^{(k\ell)}\|_F > 0\Big\}\,. \tag{6}$$

III. Optimization

The objective function $L_\lambda(\boldsymbol{\Delta}, \hat{\boldsymbol{\Sigma}}_x, \hat{\boldsymbol{\Sigma}}_y)$, given by (4), is strictly convex. Several existing approaches such as an alternating direction method of multipliers (ADMM) [38] or proximal gradient descent (PGD) methods [39], can be followed to minimize (4). Note that [8], [28] use ADMM while [9] uses a proximal gradient method, all for SA graphs. It is stated in [8, Sec. 2.2] that in their simulation example, ADMM approach yielded a slightly smaller value of the objective function compared to the PGD approach. In [10], [11], similar to [9], a proximal gradient method is used for an objective function similar to our (4). In this paper, motivated by [8], we will develop an ADMM method. In a simulation example (Sec. V-A) we compare our ADMM approach with ADMM and PGD approaches of [28] and [9], [11], respectively.

A. ADMM Approach

Similar to [28] (also [8]), we use an ADMM approach [38] with variable splitting. Using variable splitting, consider

$$\min_{\boldsymbol{\Delta},\boldsymbol{W}} \Big\{L(\boldsymbol{\Delta}, \hat{\boldsymbol{\Sigma}}_x, \hat{\boldsymbol{\Sigma}}_y) + \lambda \sum_{k,\ell=1}^{p} \|\boldsymbol{W}^{(k\ell)}\|_F\Big\} \quad \text{subject to } \boldsymbol{\Delta} = \boldsymbol{W}\,. \tag{7}$$

The scaled augmented Lagrangian for this problem is [38]

$$L_\rho = L(\boldsymbol{\Delta}, \hat{\boldsymbol{\Sigma}}_x, \hat{\boldsymbol{\Sigma}}_y) + \lambda \sum_{k,\ell=1}^{p} \|\boldsymbol{W}^{(k\ell)}\|_F + \frac{\rho}{2}\|\boldsymbol{\Delta} - \boldsymbol{W} + \boldsymbol{U}\|_F^2 \tag{8}$$

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

(a) $\boldsymbol{\Delta}^{(i+1)} \leftarrow \arg\min_{\boldsymbol{\Delta}} L_a(\boldsymbol{\Delta})$, $L_a(\boldsymbol{\Delta}) := L(\boldsymbol{\Delta}, \hat{\boldsymbol{\Sigma}}_x, \hat{\boldsymbol{\Sigma}}_y) + \frac{\rho}{2}\|\boldsymbol{\Delta} - \boldsymbol{W}^{(i)} + \boldsymbol{U}^{(i)}\|_F^2$.
(b) $\boldsymbol{W}^{(i+1)} \leftarrow \arg\min_{\boldsymbol{W}} L_b(\boldsymbol{W})$, $L_b(\boldsymbol{W}) := \lambda \sum_{k,\ell=1}^{p} \|\boldsymbol{W}^{(k\ell)}\|_F + \frac{\rho}{2}\|\boldsymbol{\Delta}^{(i+1)} - \boldsymbol{W} + \boldsymbol{U}^{(i)}\|_F^2$.
(c) $\boldsymbol{U}^{(i+1)} \leftarrow \boldsymbol{U}^{(i)} + \big(\boldsymbol{\Delta}^{(i+1)} - \boldsymbol{W}^{(i+1)}\big)$.

Update (a): Differentiate $L_a(\boldsymbol{\Delta})$ w.r.t. $\boldsymbol{\Delta}$ to obtain

$$\boldsymbol{0} = \frac{\partial L_a(\boldsymbol{\Delta})}{\partial \boldsymbol{\Delta}} = \hat{\boldsymbol{\Sigma}}_x \boldsymbol{\Delta} \hat{\boldsymbol{\Sigma}}_y - (\hat{\boldsymbol{\Sigma}}_x - \hat{\boldsymbol{\Sigma}}_y) + \rho(\boldsymbol{\Delta} - \boldsymbol{W} + \boldsymbol{U}) \tag{9}$$

$$\Rightarrow (\hat{\boldsymbol{\Sigma}}_y \otimes \hat{\boldsymbol{\Sigma}}_x + \rho \boldsymbol{I})\mathrm{vec}(\boldsymbol{\Delta}) = \mathrm{vec}(\hat{\boldsymbol{\Sigma}}_x - \hat{\boldsymbol{\Sigma}}_y + \rho(\boldsymbol{W} - \boldsymbol{U})) \tag{10}$$

Direct matrix inversion solution of (10) requires inversion of a $(mp)^2 \times (mp)^2$ matrix. A computationally cheaper solution is given in [8], [28], as follows. Carry out eigendecomposition of $\hat{\boldsymbol{\Sigma}}_x$ and $\hat{\boldsymbol{\Sigma}}_y$ as $\hat{\boldsymbol{\Sigma}}_x = \boldsymbol{Q}_x \boldsymbol{D}_x \boldsymbol{Q}_x^\top$, $\boldsymbol{Q}_x \boldsymbol{Q}_x^\top = \boldsymbol{I}$ and $\hat{\boldsymbol{\Sigma}}_y = \boldsymbol{Q}_y \boldsymbol{D}_y \boldsymbol{Q}_y^\top$, $\boldsymbol{Q}_y \boldsymbol{Q}_y^\top = \boldsymbol{I}$, where $\boldsymbol{D}_x$ and $\boldsymbol{D}_y$ are diagonal matrices of the respective eigenvalues. Then $\hat{\boldsymbol{\Delta}}$ that minimizes $L_a(\boldsymbol{\Delta})$ is given by

$$\hat{\boldsymbol{\Delta}} = \boldsymbol{Q}_x \Big[\boldsymbol{B} \circ [\boldsymbol{Q}_x^\top \big(\hat{\boldsymbol{\Sigma}}_x - \hat{\boldsymbol{\Sigma}}_y + \rho(\boldsymbol{W} - \boldsymbol{U})\big)\boldsymbol{Q}_y] \Big] \boldsymbol{Q}_y^\top \tag{11}$$

where the symbol $\circ$ denotes the Hadamard product and $\boldsymbol{B} \in \mathbb{R}^{mp \times mp}$ organizes the diagonal of $(\boldsymbol{D}_y \otimes \boldsymbol{D}_x + \rho \boldsymbol{I})^{-1}$ in a matrix with $\boldsymbol{B}_{jk} = 1/([\boldsymbol{D}_x]_{jj}[\boldsymbol{D}_y]_{kk} + \rho)$. Note that the eigendecomposition of $\hat{\boldsymbol{\Sigma}}_x$ and $\hat{\boldsymbol{\Sigma}}_y$ has to be done only once. Thus

$$\boldsymbol{\Delta}^{(i+1)} = \boldsymbol{Q}_x \Big[\boldsymbol{B} \circ [\boldsymbol{Q}_x^\top \big(\hat{\boldsymbol{\Sigma}}_x - \hat{\boldsymbol{\Sigma}}_y + \rho(\boldsymbol{W}^{(i)} - \boldsymbol{U}^{(i)})\big)\boldsymbol{Q}_y] \Big] \boldsymbol{Q}_y^\top \tag{12}$$

Update (b): Here we have the group lasso solution [35]–[37]

$$(\boldsymbol{W}^{(k\ell)})^{(i+1)} = \Big(1 - \frac{(\lambda/\rho)}{\|(\boldsymbol{\Delta}^{(i+1)} + \boldsymbol{U}^{(i)})^{(k\ell)}\|_F}\Big)_+ (\boldsymbol{\Delta}^{(i+1)} + \boldsymbol{U}^{(i)})^{(k\ell)} \tag{13}$$

where $(a)_+ = \max(0, a)$.

A pseudocode for the ADMM algorithm, MA-ADMM, used in this paper is given in Algorithm 1 where we use the stopping (convergence) criterion following [38, Sec. 3.3.1] and varying penalty parameter ρ following [38, Sec. 3.4.1]. The stopping criterion is based on primal and dual residuals being small where, in our case, at $(i+1)$st iteration, the primal residual is given by $\boldsymbol{\Delta}^{(i+1)} - \boldsymbol{W}^{(i+1)}$ and the dual residual by $\rho^{(i)}(\boldsymbol{W}^{(i+1)} - \boldsymbol{W}^{(i)})$. Convergence criterion is met when the norms of these residuals are below primary and dual tolerances τ_{pri} and τ_{dual}, respectively; see line 10 of Algorithm 1. In turn, τ_{pri} and τ_{dual} are chosen using an absolute and relative criterion as in line 10 of Algorithm 1 where τ_{abs} and τ_{rel} are user chosen absolute and relative tolerances, respectively. Line 10 of Algorithm 1 follows typical choices given in [38, Sec. 3.4.1]. For all numerical results presented later, we used $\rho_0 = 2$, $\mu = 10$, and $\tau_{abs} = \tau_{rel} = 10^{-4}$.

We will compare our approach with three other approaches in Sec. V-A. One of them is the single attribute (SA) based ADMM approach (see [8], [28]. A pseudocode of our implementation of this approach, SA-ADMM, is in Algorithm 2 which differs from in Algorithm 1 only in line 8 where we replace group lasso with elementwise lasso.

Algorithm 1 ADMM Algorithm MA-ADMM

Input: Data $\{\boldsymbol{x}(t)\}_{t=1}^{n_x}$, $\boldsymbol{x} \in \mathbb{R}^{mp}$, and $\{\boldsymbol{y}(t)\}_{t=1}^{n_y}$, $\boldsymbol{y} \in \mathbb{R}^{mp}$, regularization and penalty parameters λ and ρ_0, tolerances τ_{abs} and τ_{rel}, variable penalty factor μ, maximum number of iterations i_{max}.

Output: estimated $\hat{\boldsymbol{\Delta}}_{sym}$ and $\hat{\mathcal{E}}_\Delta$.

1: Calculate sample covariances $\hat{\boldsymbol{\Sigma}}_x = \frac{1}{n_x}\sum_{t=1}^{n_x} \boldsymbol{x}(t)\boldsymbol{x}^\top(t)$ and $\hat{\boldsymbol{\Sigma}}_y = \frac{1}{n_y}\sum_{t=1}^{n_y} \boldsymbol{y}(t)\boldsymbol{y}^\top(t)$.
2: Initialize: $\boldsymbol{\Delta}^{(0)} = \boldsymbol{U}^{(0)} = \boldsymbol{W}^{(0)} = \boldsymbol{0}$, where $\boldsymbol{\Delta}, \boldsymbol{U}, \boldsymbol{W} \in \mathbb{R}^{(mp)\times(mp)}$, $\rho^{(0)} = \rho_0$.
3: Eigendecompose $\hat{\boldsymbol{\Sigma}}_x$ and $\hat{\boldsymbol{\Sigma}}_y$ as $\hat{\boldsymbol{\Sigma}}_x = \boldsymbol{Q}_x \boldsymbol{D}_x \boldsymbol{Q}_x^\top$ and $\hat{\boldsymbol{\Sigma}}_y = \boldsymbol{Q}_y \boldsymbol{D}_y \boldsymbol{Q}_y^\top$.
4: converged = FALSE, $i = 0$
5: **while** converged = FALSE AND $i \le i_{max}$, **do**
6: Construct $\boldsymbol{B} \in \mathbb{R}^{mp \times mp}$ with $\boldsymbol{B}_{jk} = 1/([\boldsymbol{D}_x]_{jj}[\boldsymbol{D}_y]_{kk} + \rho^{(i)})$.
7: Set $\boldsymbol{\Delta}^{(i+1)} = \boldsymbol{Q}_x \Big[\boldsymbol{B} \circ [\boldsymbol{Q}_x^\top \big(\hat{\boldsymbol{\Sigma}}_x - \hat{\boldsymbol{\Sigma}}_y + \rho(\boldsymbol{W}^{(i)} - \boldsymbol{U}^{(i)})\big)\boldsymbol{Q}_y] \Big] \boldsymbol{Q}_y^\top$.
8: With $(a)_+ := \max(0, a)$, $\boldsymbol{A} = (\boldsymbol{\Delta}^{(i+1)} + \boldsymbol{U}^{(i)})^{(k\ell)}$ and $k, \ell \in [p]$, update $m \times m$ subblocks of $\boldsymbol{W}$ as

$$(\boldsymbol{W}^{(i+1)})^{(k\ell)} = \Big(1 - \frac{(\lambda/\rho)}{\|\boldsymbol{A}\|_F}\Big)_+ \boldsymbol{A}^{(k\ell)}.$$

9: Dual update $\boldsymbol{U}^{(i+1)} = \boldsymbol{U}^{(i)} + \big(\boldsymbol{\Delta}^{(i+1)} - \boldsymbol{W}^{(i+1)}\big)$.
10: Check convergence. Set tolerances

$$\tau_{pri} = mp\,\tau_{abs} + \tau_{rel}\,\max(\|\boldsymbol{\Delta}^{(i+1)}\|_F, \|\boldsymbol{W}^{(i+1)}\|_F)$$
$$\tau_{dual} = mp\,\tau_{abs} + \tau_{rel}\,\|\boldsymbol{U}^{(i+1)}\|_F/\rho^{(i)}.$$

Define $e_p = \|\boldsymbol{\Delta}^{(i+1)} - \boldsymbol{W}^{(i+1)}\|_F$ and $e_d = \rho^{(i)}\|\boldsymbol{W}^{(i+1)} - \boldsymbol{W}^{(i)}\|_F$. If $(e_p \le \tau_{pri})$ AND $(e_d \le \tau_{dual})$, set converged = TRUE .
11: Update penalty parameter ρ : If $e_p > \mu e_d$, set $\rho^{(i+1)} = 2\rho^{(i)}$, else if $e_d > \mu e_p$, set $\rho^{(i+1)} = \rho^{(i)}/2$, otherwise $\rho^{(i+1)} = \rho^{(i)}$. We also need to set $\boldsymbol{U}^{(i+1)} = \boldsymbol{U}^{(i+1)}/2$ for $e_p > \mu e_d$ and $\boldsymbol{U}^{(i+1)} = 2\boldsymbol{U}^{(i+1)}$ for $e_d > \mu e_p$.
12: $i \leftarrow i + 1$
13: **end while**
14: Set $\hat{\boldsymbol{\Delta}}_{sym} = \frac{1}{2}(\boldsymbol{W} + \boldsymbol{W}^\top)$. If $\|\hat{\boldsymbol{\Delta}}_{sym}^{(jk)}\|_F > 0$, assign edge $\{j,k\} \in \hat{\mathcal{E}}_\Delta$, else $\{j,k\} \notin \hat{\mathcal{E}}_\Delta$.

Algorithm 2 ADMM Algorithm SA-ADMM

Input: As in Algorithm 1.

Output: Estimated $\hat{\boldsymbol{\Delta}}_{sym}$ and $\hat{\mathcal{E}}_\Delta$.

1: Follow lines 1-7 of Algorithm 1
2: With $(a)_+ := \max(0, a)$, $\boldsymbol{A} = (\boldsymbol{\Delta}^{(i+1)} + \boldsymbol{U}^{(i)})^{(k\ell)}$ and $k, \ell \in [mp]$, update $[\boldsymbol{W}]_{k\ell} \in \mathbb{R}$ as

$$[\boldsymbol{W}^{(i+1)}]_{k\ell} = \Big(1 - \frac{(\lambda/\rho)}{|[\boldsymbol{A}]_{k\ell}|}\Big)_+ [\boldsymbol{A}]_{k\ell}.$$

3: Follow lines 9-14 of Algorithm 1

This article has been accepted for publication in IEEE Transactions on Signal Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TSP.2023.3343553

B. Proximal Gradient Descent Approach

It is a first-order method that is based on objective function values and gradient evaluations. A pseudocode of the PGD method of [11], MA-proximal, is in Algorithm 3, and that of [9], SA-proximal, is in Algorithm 4. Algorithm 4 differs from in Algorithm 3 only in line 8 where we replace group lasso with element-wise lasso. For all numerical results presented later, we used $\epsilon = 10^{-3}$ in line 7.

Algorithm 3 PGD Algorithm MA-PGD

Input: Data $\{\boldsymbol{x}(t)\}_{t=1}^{n_x}$, $\boldsymbol{x} \in \mathbb{R}^{mp}$, and $\{\boldsymbol{y}(t)\}_{t=1}^{n_y}$, $\boldsymbol{y} \in \mathbb{R}^{mp}$, tolerance ϵ, maximum number of iterations i_{max}
Output: Estimated $\hat{\boldsymbol{\Delta}}_{sym}$ and $\hat{\mathcal{E}}_{\Delta}$.

1: Calculate sample covariances $\hat{\boldsymbol{\Sigma}}_x = \frac{1}{n_x}\sum_{t=1}^{n_x} \boldsymbol{x}(t)\boldsymbol{x}^\top(t)$ and $\hat{\boldsymbol{\Sigma}}_y = \frac{1}{n_y}\sum_{t=1}^{n_y} \boldsymbol{y}(t)\boldsymbol{y}^\top(t)$.
2: Set $\eta = 1/\big(\phi_{max}(\hat{\boldsymbol{\Sigma}}_x)\,\phi_{max}(\hat{\boldsymbol{\Sigma}}_x)\big)$. Initialize: $\boldsymbol{\Delta}^{(0)} = \mathbf{0}$.
3: converged = FALSE, $i = 0$
4: **while** converged = FALSE AND $i \le i_{max}$, **do**
5: Set $\boldsymbol{A} = \boldsymbol{\Delta}^{(i)} - \eta\Big(\hat{\boldsymbol{\Sigma}}_x\boldsymbol{\Delta}^{(i)}\hat{\boldsymbol{\Sigma}}_y - (\hat{\boldsymbol{\Sigma}}_x - \hat{\boldsymbol{\Sigma}}_y)\Big)$.
6: For $k, \ell \in [p]$, update $m \times m$ subblocks as

$$(\boldsymbol{\Delta}^{(i+1)})^{(k\ell)} = \Big(1 - \frac{\lambda\eta}{\|\boldsymbol{A}\|_F}\Big)_+ \boldsymbol{A}^{(k\ell)}\,.$$

7: If $\frac{L_\lambda(\boldsymbol{\Delta}^{(i+1)},\hat{\boldsymbol{\Sigma}}_x,\hat{\boldsymbol{\Sigma}}_y) - L_\lambda(\boldsymbol{\Delta}^{(i)},\hat{\boldsymbol{\Sigma}}_x,\hat{\boldsymbol{\Sigma}}_y)}{L_\lambda(\boldsymbol{\Delta}^{(i)},\hat{\boldsymbol{\Sigma}}_x,\hat{\boldsymbol{\Sigma}}_y)} \le \epsilon$, set converged = TRUE .
8: $i \leftarrow i + 1$
9: **end while**
10: Set $\hat{\boldsymbol{\Delta}}_{sym} = \frac{1}{2}(\boldsymbol{\Delta} + \boldsymbol{\Delta}^\top)$. If $\|\hat{\boldsymbol{\Delta}}_{sym}^{(jk)}\|_F > 0$, assign edge $\{j,k\} \in \hat{\mathcal{E}}_{\Delta}$, else $\{j,k\} \notin \hat{\mathcal{E}}_{\Delta}$.

Algorithm 4 PGD Algorithm SA-PGD

Input: As in Algorithm 3
Output: Estimated $\hat{\boldsymbol{\Delta}}$ and $\hat{\mathcal{E}}_{\Delta}$

1: Follow lines 1-5 of Algorithm 3
2: For $k, \ell \in [mp]$, update

$$[\boldsymbol{\Delta}^{(i+1)}]_{k\ell} = \Big(1 - \frac{\lambda\eta}{\big|[\boldsymbol{A}]_{k\ell}\big|}\Big)_+ [\boldsymbol{A}]_{k\ell}\,.$$

3: Follow lines 7-10 of Algorithm 3

C. Computational Complexity

The computational complexity of ADMM and PGD methods has been discussed in [9] for SA differential graphs, and it is of the same order for MA graphs, because the difference lies only in lasso versus group lasso, i.e., element-wise soft-thresholding versus group-wise soft-thresholding. Noting that we have $mp \times mp$ precision matrices, by [9], the computational complexity of the ADMM approaches (our proposed and that of [8], [28]) is $\mathcal{O}((mp)^3)$ while that of the PGD methods of [9], [11] is either $\mathcal{O}((mp)^3)$ when as implemented in Algorithms 3 and 4, or $\mathcal{O}((n_x+n_y)(mp)^2)$ when an alternative implementation of the cost gradient in line 5 of Algorithms 3 and 4 is used (see [9, Sec. 2.2]). For $n_x + n_y \ge mp$, there is no advantage to this alternative approach.

D. Convergence of ADMM

The objective function (4), is strictly convex. It is also closed, proper and lower semi-continuous. Hence, for any fixed $\rho > 0$, the (two-block) ADMM algorithm is guaranteed to converge [38, Sec. 3.2], in the sense that we have primal residual convergence to 0, dual residual convergence to 0, and objective function convergence to the optimal value. For varying ρ, the convergence of ADMM has not been proven, but if we additionally impose $\rho^{(i)} = \rho^{(i_0)} > 0$ for $i \ge i_0$ for some i_0, we have convergence [38, Sec. 3.4.1].

E. Model Selection

Following the lasso penalty work of [8] (who invokes [12]), we use the following criterion for selection of group lasso penalty:

$$\begin{aligned} BIC(\lambda) =& (n_x + n_y)\,\|\hat{\boldsymbol{\Sigma}}_x\hat{\boldsymbol{\Delta}}\hat{\boldsymbol{\Sigma}}_y - (\hat{\boldsymbol{\Sigma}}_x - \hat{\boldsymbol{\Sigma}}_y)\|_F \\ &+ \ln(n_x + n_y)\,|\hat{\boldsymbol{\Delta}}|_0 \end{aligned} \tag{14}$$

where $|\boldsymbol{A}|_0$ denotes number of nonzero elements in $\boldsymbol{A}$ and $\hat{\boldsymbol{\Delta}}$ obeys (5). Choose λ to minimize $BIC(\lambda)$. Following [8] we term it BIC (Bayesian information criterion) even though the cost function used is not negative log-likelihood although $\ln(n_x + n_y)\,|\hat{\boldsymbol{\Delta}}|_0$ penalizes over-parametrization as in BIC. It is based on the fact that true $\boldsymbol{\Delta}^*$ satisfies $\boldsymbol{\Sigma}_x^*\boldsymbol{\Delta}^*\boldsymbol{\Sigma}_y^* - (\boldsymbol{\Sigma}_x^* - \boldsymbol{\Sigma}_y^*) = \mathbf{0}$. Since (14) is not scale invariant, we scale both $\hat{\boldsymbol{\Sigma}}_x$ and $\hat{\boldsymbol{\Sigma}}_y$ (and $\hat{\boldsymbol{\Delta}}$ commensurately) by $\bar{\boldsymbol{\Sigma}}^{-1}$ where $\bar{\boldsymbol{\Sigma}} = \text{diag}\{\hat{\boldsymbol{\Sigma}}_x\}$ is a diagonal matrix of diagonal elements of $\hat{\boldsymbol{\Sigma}}_x$.

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 [17]. Find the smallest λ, labeled λ_{sm} for which we get a no-edge model; then we set $\lambda_u = \lambda_{sm}/2$ and $\lambda_\ell = \lambda_u/10$.

IV. Theoretical Analysis

Here we analyze the properties of $\hat{\boldsymbol{\Delta}}$. Theorem 1 follows the approach(es) of [8], [16], [28], [29], [31] while Theorem 2 follows the general framework of [30]. The general method of [31] used in [8], [16], [28], [29] requires an irrepresentability condition (see (20)) which is not needed by the method of [30]. On the other hand, our Theorem 2 does not have a result like Theorem 1(ii), the oracle property, or support recovery Theorem 1(iv).

First some notation. Define the true differential edgeset

$$S = \mathcal{E}_{\Delta^*} = \{\{k,\ell\} : \|\boldsymbol{\Delta}^{*(k\ell)}\|_F \ne 0\}\,, \quad s = |S|\,. \tag{15}$$

Define

$$\boldsymbol{\Gamma}^* = \boldsymbol{\Sigma}_y^* \boxtimes \boldsymbol{\Sigma}_x^*\,, \quad \hat{\boldsymbol{\Gamma}} = \hat{\boldsymbol{\Sigma}}_y \boxtimes \hat{\boldsymbol{\Sigma}}_x\,. \tag{16}$$

Also, recall the operator $\boldsymbol{\mathcal{C}}(\cdot)$ defined in Sec. I-C. In the rest of this section, we allow p, s and λ to be a functions of sample size n, denoted as p_n, s_n and λ_n, respectively. Define

$$M = \max\{\|\boldsymbol{\mathcal{C}}(\boldsymbol{\Sigma}_x^*)\|_\infty\,, \|\boldsymbol{\mathcal{C}}(\boldsymbol{\Sigma}_y^*)\|_\infty\}\,, \tag{17}$$

$$M_\Sigma = \max\{\|\boldsymbol{\mathcal{C}}(\boldsymbol{\Sigma}_x^*)\|_{1,\infty}\,, \|\boldsymbol{\mathcal{C}}(\boldsymbol{\Sigma}_y^*)\|_{1,\infty}\}\,, \tag{18}$$

$$\kappa_\Gamma = \|\boldsymbol{\mathcal{C}}(\Gamma_{S,S}^*)^{-1}\|_{1,\infty}\,, \tag{19}$$

$$\alpha = 1 - \max_{e \in S^c} \|\boldsymbol{\mathcal{C}}(\boldsymbol{\Gamma}_{e,S}^*(\boldsymbol{\Gamma}_{S,S}^*)^{-1})\|_1\,, \tag{20}$$

$$\bar{\sigma}_{xy} = \max\{\max_i [\boldsymbol{\Sigma}_x^*]_{ii}, \max_i [\boldsymbol{\Sigma}_y^*]_{ii}\} \tag{21}$$

$$C_0 = 40\,m\,\bar{\sigma}_{xy}\sqrt{2\big(\tau + \ln(4m^2)/\ln(p_n)\big)} \tag{22}$$

where S and $\mathbf{\Gamma}^*$ have been defined in (15) and (16). In (20), we require $0 < \alpha < 1$, and the expression

$$\max_{e \in S^c} \|\mathcal{C}(\mathbf{\Gamma}^*_{e,S}(\mathbf{\Gamma}^*_{S,S})^{-1})\|_1 \leq 1 - \alpha$$

for some $\alpha \in (0,1)$ is called the *irrepresentability condition*. Similar conditions are also used in [8], [16], [28], [29], [31].

Let $\hat{\mathbf{\Delta}}$ be as in (5).

Theorem 1. For the system model of Sec. II, under (15) and the irrepresentability condition (20) for some $\alpha \in (0,1)$, if

$$\lambda_n = \max\Big\{\frac{8}{\alpha}, \frac{3}{\alpha \bar{C}_\alpha}\, s_n \kappa_\Gamma M C_{M\kappa}\Big\} C_0 \sqrt{\frac{\ln(p_n)}{n}} \tag{23}$$

$$n = \min(n_x, n_y) > \max\Big\{\frac{1}{\min\{M^2, 1\}}, 81 M^2 s_n^2 \kappa_\Gamma^2, \frac{9 s_n^2}{(\alpha \bar{C}_\alpha)^2}(\kappa_\Gamma M C_{M\kappa})^2\Big\} C_0^2 \ln(p_n) \tag{24}$$

where $\bar{C}_\alpha = \frac{1-\alpha}{2(2M+1)-2\alpha M}$ and $C_{M\kappa} = 1.5\big(1 + \kappa_\Gamma \min\{s_n M^2, M_\Sigma^2\}\big)$, then with probability $> 1 - 2/p_n^{\tau-2}$, for any $\tau > 2$, we have

(i) $\|\mathcal{C}(\hat{\mathbf{\Delta}} - \mathbf{\Delta}^*)\|_\infty \leq (C_{b1} + C_{b2}) C_0 \sqrt{\frac{\ln(p_n)}{n}}$

where $C_{b1} = 3\kappa_\Gamma \max\Big\{\frac{8}{\alpha}, \frac{3}{\alpha \bar{C}_\alpha}\, s_n \kappa_\Gamma M C_{M\kappa}\Big\}$

$C_{b2} = 9 s_n \kappa_\Gamma^2 M^2$.

(ii) $\hat{\mathbf{\Delta}}_{S^c} = \mathbf{0}$.

(iii) $\|\mathcal{C}(\hat{\mathbf{\Delta}} - \mathbf{\Delta}^*)\|_F \leq \sqrt{s_n}\, \|\mathcal{C}(\hat{\mathbf{\Delta}} - \mathbf{\Delta}^*)\|_\infty$.

(iv) Additionally, if $\min_{(k,\ell) \in S} \|(\mathbf{\Delta}^*)^{(k\ell)}\|_F \geq 2(C_{b1} + C_{b2}) C_0 \sqrt{\frac{\ln(p_n)}{n}}$, then $P(\mathcal{G}_{\hat{\Delta}} = \mathcal{G}_{\Delta^*}) > 1 - 2/p_n^{\tau-2}$ (support recovery) •

The proof of Theorem 1 is given in Appendix A.

Now we present Theorem 2 that follows the general framework of [30]. Let $\hat{\mathbf{\Delta}}$ be as in (5).

Theorem 2. For the system model of Sec. II, under (15), if

$$\lambda_n \geq (4 + 6 M s_n C_1) C_0 \sqrt{\frac{\ln(p_n)}{n}} \tag{25}$$

$$n = \min(n_x, n_y) > \max\Big\{\frac{1}{M^2}, \Big(\frac{96 M s_n}{\phi^*_{min}}\Big)^2\Big\} C_0^2 \ln(p_n) \tag{26}$$

where $C_1 = \max_{\{k,\ell\} \in V \times V} \|(\mathbf{\Delta}^*)^{(k\ell)}\|_F$ and $\phi^*_{min} = \phi_{min}(\mathbf{\Sigma}^*_x)\phi_{min}(\mathbf{\Sigma}^*_y)$, then with probability $> 1 - 2/p_n^{\tau-2}$, for any $\tau > 2$, we have

$$\|\hat{\mathbf{\Delta}} - \mathbf{\Delta}^*\|_F \leq \frac{12\sqrt{s_n}}{\phi^*_{min}}(4 + 6 M s_n C_1) C_0 \sqrt{\frac{\ln(p_n)}{n}} \quad \bullet \tag{27}$$

The proof of Theorem 2 is given in Appendix B. Note that $\|\mathcal{C}(\hat{\mathbf{\Delta}} - \mathbf{\Delta}^*)\|_F = \|\hat{\mathbf{\Delta}} - \mathbf{\Delta}^*\|_F$ when comparing Theorems 1 and 2.

Remark 1: *Convergence Rate*. If M, M_Σ and κ_Γ stay bounded with increasing sample size n, we have $\|\mathcal{C}(\hat{\mathbf{\Delta}} - \mathbf{\Delta}^*)\|_F = \mathcal{O}_P(s_n^{1.5}\sqrt{\ln(p_n)/n})$. Therefore, for $\|\mathcal{C}(\hat{\mathbf{\Delta}} - \mathbf{\Delta}^*)\|_F \to 0$ as $n \to \infty$, we must have $s_n^{1.5}\sqrt{\ln(p_n)/n} \to 0$. The SA results in [8] need $s_n^{3.5}\sqrt{\ln(p_n)/n} \to 0$ when we take into account the dependence of various constants on s_n in [8]. Notice that M_Σ constraints covariances $\mathbf{\Sigma}^*_x$ and $\mathbf{\Sigma}^*_y$ which can be dense even if $\mathbf{\Omega}^*_x$ and $\mathbf{\Omega}^*_y$ are sparse (they need not be sparse for differential estimation), making them possibly unbounded with increasing sample size n. In this case we use $\min\{s_n M^2, M_\Sigma^2\} = s_n M^2$ in $C_{M\kappa}$ and C_{b1}, with M bounded, leading to $\|\mathcal{C}(\hat{\mathbf{\Delta}} - \mathbf{\Delta}^*)\|_F = \mathcal{O}_P(s_n^{2.5}\sqrt{\ln(p_n)/n})$. On the other hand, in Theorem 2, we always have $\|\mathcal{C}(\hat{\mathbf{\Delta}} - \mathbf{\Delta}^*)\|_F = \mathcal{O}_P(s_n^{1.5}\sqrt{\ln(p_n)/n})$. □

Remark 2: Our results assume Gaussian data. Theorems 1 and 2 continue to hold for more general sub-Gaussian distributions (Gaussian distribution is a sub-Gaussian distribution) except that the zeros in the precision matrix (see (1)), or in the difference of two precision matrices, no longer signify conditional independence (or change in conditional independence) of the random vectors associated with the respective nodes; they only imply zero partial correlation. We state Lemma 1 in Appendix A for sub-Gaussian distributions, following [31, Lemma 1]. Lemma 2 is then specialized to Gaussian distributions by setting the sub-Gaussian parameter $\sigma_{sg} = 1$. If $\sigma_{sg} \neq 1$, then the only changes required in Theorems 1 and 2 is a scaling of C_0 in (22) where one needs to replace the factor of 40 with $8(1 + 4\sigma_{sg}^2)$. □

Remark 3: Theorems 1 and 2 assume a constant number of attributes m, with only p, s and λ allowed to be functions of sample size n, and our Remark 1 reflects this fact. In terms of m, the bounds in Theorem 1(i) and Theorem 2 are $\mathcal{O}(m^3)$, which follows from $M = \mathcal{O}(m)$ and $C_{M\kappa} = \mathcal{O}(m^2)$ in Theorem 1, and $M = \mathcal{O}(m)$, $C_1 = \mathcal{O}(m)$ and $C_0 = \mathcal{O}(m)$ in Theorem 2. Therefore, for large m, one would need much higher number of samples n, and it is not clear how to circumvent this bottleneck. A different class of models based on matrix-valued graphical modeling [40]–[42] is a potential solution. In a matrix graph set-up, one has matrix-valued observations $\boldsymbol{Z}$, which in our context would require attributes (components of $\boldsymbol{z}_i$) to be arranged along rows and nodes i along columns, with the covariance of $\text{vec}(\boldsymbol{Z})$ having a Kronecker-product structure. This structure drastically reduces the number of unknowns from $\mathcal{O}((mp)^2)$ to $\mathcal{O}(m^2+p^2)$. Prior reported work ([40]–[42]) is on matrix graph estimation, with no reported work on differential matrix graphs. □

V. Numerical Examples

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

A. Synthetic Data: Erdös-Rènyi and Barabási-Albert Graphs

We consider two types of graphs: Erdös-Rènyi (ER) graph and Barabási-Albert (BA) graph [43], [44]. The BA graphs are an example of scale-free graphs with power law degree distribution [43]. In the ER graph, $p = 100$ nodes are connected to each other with probability $p_{er} = 0.5$ and there are $m = 3$ 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 [44] (MATLAB

This article has been accepted for publication in IEEE Transactions on Signal Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TSP.2023.3343553

function BAmodel.m from https://github.com/ShanLu1984/Scale-Free-Network-Generation-and-Comparison). In the upper triangular $\mathbf{\Omega}_x$, we set $[\mathbf{\Omega}_x^{(jk)}]_{st} = 0.5^{|s-t|}$ for $j = k = 1, \cdots, p$, $s, t = 1, \cdots, m$. For $j \neq k$, if the two nodes are not connected in the graph (ER or BA), we have $\mathbf{\Omega}^{(jk)} = \mathbf{0}$, and if nodes j and k are connected, then $[\mathbf{\Omega}^{(jk)}]_{st}$ is uniformly distributed over $[-0.4, -0.1] \cup [0.1, 0.4]$. Then add lower triangular elements to make $\mathbf{\Omega}_x$ a symmetric matrix. To generate $\mathbf{\Omega}_y$, we follow [8] and first generate a differential graph with $\mathbf{\Delta} \in \mathbb{R}^{(mp)\times(mp)}$ as an ER graph (regardless of whether $\mathbf{\Omega}_x$ is based on ER or BA model), with connection probability $p_{er} = 0.05$ (sparse): if nodes j and k are connected in the $\mathbf{\Omega}_x$ model, then each of m^2 elements of $\mathbf{\Delta}^{(jk)}$ is independently set to ± 0.9 with equal probabilities. Then $\mathbf{\Omega}_y = \mathbf{\Omega}_x + \mathbf{\Delta}$. Finally add $\gamma \boldsymbol{I}$ to $\mathbf{\Omega}_y$ and to $\mathbf{\Omega}_x$ and pick γ so that $\mathbf{\Omega}_y$ and $\mathbf{\Omega}_x$ are both positive definite. With $\mathbf{\Phi}_x \mathbf{\Phi}_x^\top = \mathbf{\Omega}_x^{-1}$, we generate $\boldsymbol{x} = \mathbf{\Phi}\boldsymbol{w}$ with $\boldsymbol{w} \in \mathbb{R}^{mp}$ as zero-mean Gaussian, with identity covariance, and similarly for $\boldsymbol{y}$. We generate $n = n_x = n_y$ i.i.d. observations for $\boldsymbol{x}$ and $\boldsymbol{y}$, with $m = 3$, $p = 100$, $n \in \{100, 200, 300, 400, 800, 1200, 1600\}$.

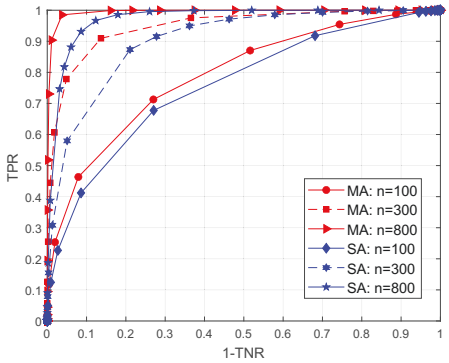


Fig. 1: ROC curves for ER graph based on ADMM approaches. TPR=true positive rate, TNR=true negative rate

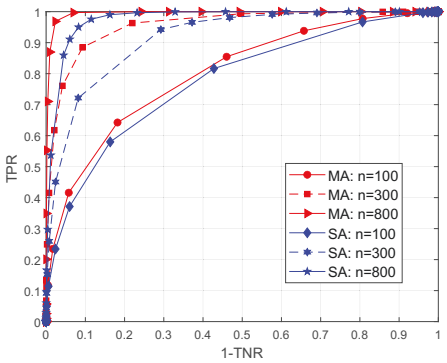


Fig. 2: ROC curves for BA graph based on ADMM approaches. TPR=true positive rate, TNR=true negative rate

Simulation results based on 100 runs are shown in Figs. 1-4. By changing the penalty parameter λ and determining the resulting edges, we calculated the true positive rate (TPR) and false positive rate 1-TNR (where TNR is the true negative rate) over 100 runs. The receiver operating characteristic (ROC) for ER graphs is shown in Fig. 1 for our MA-ADMM approach (labeled "MA") as well as for a SA-ADMM approach (labeled "SA"), based on [28], where we first estimate an mp-node differential graph, and then use $\|\hat{\mathbf{\Delta}}^{(k\ell)}\|_F \neq 0 \Leftrightarrow \{\{k, \ell\} \in \mathcal{E}_\Delta$. It is seen from Fig. 1 that our MA-ADMM approach outperforms the SA-ADMM approach (that uses the same cost but element-wise lasso penalty instead of group-lasso penalty). Fig. 2 is the counterpart of Fig. 1 for BA graphs., and comments made regarding Fig. 1 apply here too.

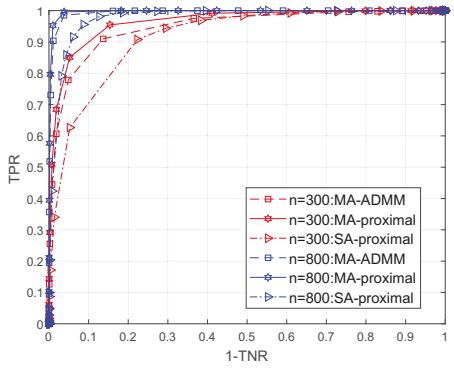


Fig. 3: ROC curves for ER graph based on ADMM as well as proximal approaches.

In Fig. 3 we compare ROC curves of our MA-ADMM approach with that for the MA-proximal and SA-proximal approaches of [11] and [9], respectively, for $n = 300$ and 800. It is seen that the MA-proximal approach outperforms our MA-ADMM approach, while both significantly outperform the SA-proximal approach (and the SA-ADMM approach whose ROC curves are in Fig. 1). In Table I, for the ER graph ($n = 300, 800$, $p = 100$, $m = 3$) we compare the four approaches (MA-ADMM, SA-ADMM, MA-proximal, SA-proximal) in terms of the F_1 score, execution time (based on tic-toc functions in MATLAB), TPR and 1-TNR, for fixed penalty parameter λ selected from a grid of values (the same as for computing the ROC curves) to maximize the F_1 score averaged over 100 runs. All algorithms were run on a Window 10 Pro operating system with processor Intel(R) Core(TM) i7-10700 CPU @2.90 GHz with 32 GB RAM, using MATLAB R2023a. Notice that while the MA-proximal approach outperforms our MA-ADMM approach, it also takes more than twice the computation time for the MA-ADMM approach. Similarly, the SA-proximal approach takes more than twice the computation time for the SA-ADMM approach. The latter observation is consistent with the findings of [9].

In Table I, we also show results for $n = 3000$ and 6000 for MA-ADMM and MA-proximal approaches in order to provide further empirical validation of the theoretical results stated in Theorem 1. Theorem 1 states that for sufficiently large sample size n, one can recover the differential graph structure exactly w.h.p. It is seen that the F_1 score approaches one with increasing n, implying graph support recovery, as claimed in Theorem 1(iv). Note that Theorem 1(iv) holds w.h.p., not with probability one, implying a nonzero probability of possibly inexact graph recovery, yielding an $F1$-score less than 1 for $n = 3000, 6000$. The sample sizes of $n = 300, 800$ are not large enough to yield an F_1-score close to 1. There is a lower bound (24) on n for Theorem 1(iv) to hold w.h.p. The sample sizes of $n = 300, 800$ are apparently less than bound (which is not easily computable since it needs α and κ_Γ).

This article has been accepted for publication in IEEE Transactions on Signal Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TSP.2023.3343553

Approach	F_1 score ($\pm\sigma$)	timing (s) ($\pm\sigma$)	TPR ($\pm\sigma$)	1-TNR ($\pm\sigma$)
		n = 300		
MA-ADMM	0.6152 $\pm$0.0705	2.5044 $\pm$0.2939	0.6067 $\pm$0.1230	0.0184 $\pm$0.0080
MA-proximal	0.6686 $\pm$0.0639	6.6931 $\pm$0.3743	0.6845 $\pm$0.1253	0.0184 $\pm$0.0085
SA-ADMM	0.4549 $\pm$0.0332	0.1739 $\pm$0.0173	0.5795 $\pm$0.1132	0.0506 $\pm$0.0186
SA-proximal	0.4772 $\pm$0.0328	0.3517 $\pm$0.0195	0.6263 $\pm$0.1157	0.0524 $\pm$0.0193
		n = 800		
MA-ADMM	0.8537 $\pm$0.0491	2.1911 $\pm$0.0639	0.9037 $\pm$0.0703	0.0111 $\pm$0.0041
MA-proximal	0.8898 $\pm$0.0408	4.8277 $\pm$0.2647	0.9526 $\pm$0.0537	0.0009 $\pm$0.0041
SA-ADMM	0.6336 $\pm$0.0055	0.1510 $\pm$0.0081	0.7468 $\pm$0.1017	0.0316 $\pm$0.0090
SA-proximal	0.6612 $\pm$0.0401	0.2533 $\pm$0.0143	0.7917 $\pm$0.0931	0.0314 $\pm$0.0090
		n = 3000		
MA-ADMM	0.9795 $\pm$0.0152	2.1436 $\pm$0.0624	0.9928 $\pm$0.0018	0.0012 $\pm$0.0041
MA-proximal	0.9914 $\pm$0.0106	3.8312 $\pm$0.3948	0.9964 $\pm$0.0121	0.0007 $\pm$0.0007
		n = 6000		
MA-ADMM	0.9914 $\pm$0.0087	1.9459 $\pm$0.0544	0.9997 $\pm$0.0013	0.0008 $\pm$0.0009
MA-proximal	0.9979 $\pm$0.0037	3.517 $\pm$0.1829	0.9998 $\pm$0.0011	0.0002 $\pm$0.0004

TABLE I: *Comparisons among various approaches: Erdös-Rènyi graph,* $n = 300, 800, 3000, 6000$, $p = 100$, $m = 3$. *Tuning parameter* λ *picked to yield the highest* F_1 *score. Results based on 100 runs.*

In Fig. 4 we show the results based on 100 runs for our approach when BIC parameter selection method (Sec. III-E) is applied in conjunction with the MA-ADMM approach. Here we show the TPR, 1-TNR and F_1 score values along with the $\pm\sigma$ error bars. The proposed approach works well both in terms of F_1 score and TPR vs 1-TNR.

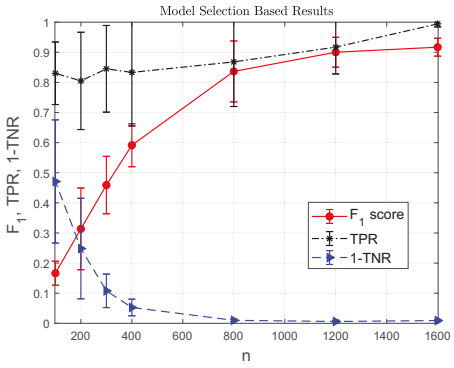


Fig. 4: BIC based results for ER graph: F_1-scores, TPR and 1-TNR

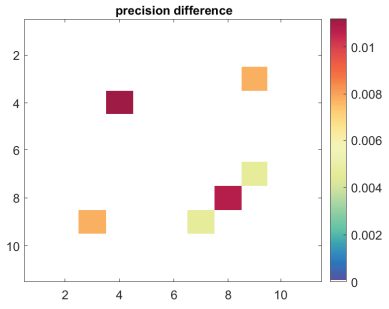
B. Real Data: Beijing air-quality dataset [45]

Here we consider Beijing air-quality dataset [45], [46], downloaded from https://archive.ics.uci.edu/ml/datasets/Beijing+Multi-Site+Air-Quality+Data. This data set includes hourly air pollutants data from 12 nationally-controlled air-quality monitoring sites in the Beijing area from the Beijing Municipal Environmental Monitoring Center, and meteorological data in each air-quality site are matched with the nearest weather station from the China Meteorological Administration. The time period is from March 1st, 2013 to February 28th, 2017.The six air pollutants are $PM_{2.5}$, PM_{10}, SO_2, NO_2, CO, and O_3, and the meteorological data is comprised of five features: temperature, atmospheric pressure, dew point, wind speed, and rain; we did not use wind direction. Thus we have eleven features. We used data from 8 sites: 4 suburban/rural sites – Changping, Huairou, Shunyi, Dingling, and 4 urban area stations – Aotizhongxin, Dongsi, Guanyuan, Gucheng [46, Fig. 1]. The data are averaged over 24 hour period to yield daily averages. We used one year of daily data resulting in $n_x = n_y = 365$ days. The stations are used as attributes, with $m = 8$ for comparison between years 2013-14 and 2014-15, and $m = 4$ for comparison between suburban/rural sites and urban sites using 2013-14 year data.

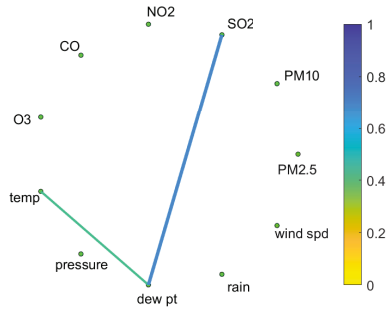
We pre-process the data as follows. Given ith feature data $\boldsymbol{z}_i(t) \in \mathbb{R}^m$, we transform it to $\bar{\boldsymbol{z}}_i(t) = \ln(\boldsymbol{z}_i(t)/\boldsymbol{z}_i(t-1))$ and then detrend it (i.e., remove the best straight-line fit using the MATLAB function detrend). Finally, we scale the detrended scalar sequence to have a mean-square value of one over n_x or n_y samples. The logarithmic transformation and detrending of each feature sequence makes the sequence closer to (univariate) stationary and Gaussian, while scaling "balances" the possible wide variations in the scale of various feature measurements. All temperatures were converted from Celsius to Kelvin to avoid negative numbers, and if a value of a feature is zero (e.g., wind speed), we added a small positive number to it, so that the logarithmic transformation is well-defined.

Fig. 5 shows the estimated differential graphs when comparing daily-averaged data from 2013-14 (x-data) to that from 2014-15 (y-data), with air-quality and meteorological variables as $p = 11$ features measured at 8 monitoring sites (m=8). The objective is to visualize and explore differential conditional dependency relationships among the 11 variables, comparing one year to another, to investigate if pollution reduction measures have had any impact. Our intuition is that one does not expect such rapid changes within a short period of one year (see also [45]), therefore, our method should confirm our intuition. Figs. 5(a)-(b) show estimated $\|\hat{\boldsymbol{\Delta}}^{(k\ell)}\|_F$ for various edges $\{k, \ell\}$, where it is unscaled in Fig. 5(a) but scaled in Fig. 5(b) so that the largest $\|\hat{\boldsymbol{\Delta}}^{(k\ell)}\|_F$ (including $k = \ell$) is normalized to one. It is seen that differential graph weights are essentially zero (very sparse), implying that there are no year-to-year changes in the conditional dependency relationships among the 11 variables. This observation conforms to the findings of [45], [46]: no significant year-to-year changes. We also estimated the MA graphs for each year separately as $\|\hat{\boldsymbol{\Omega}}_x^{(k\ell)}\|_F$ and $\|\hat{\boldsymbol{\Omega}}_y^{(k\ell)}\|_F$ (shown in Figs. 5(c)-(d)), using the approach of [17], and based on the individual estimates, we computed the differential graph $\|\hat{\boldsymbol{\Omega}}_y^{(k\ell)} - \hat{\boldsymbol{\Omega}}_x^{(k\ell)}\|_F$ (shown in Figs. 5(e)-(f)). It is seen the separate-estimation approach

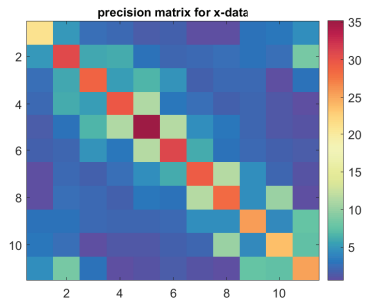
This article has been accepted for publication in IEEE Transactions on Signal Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TSP.2023.3343553



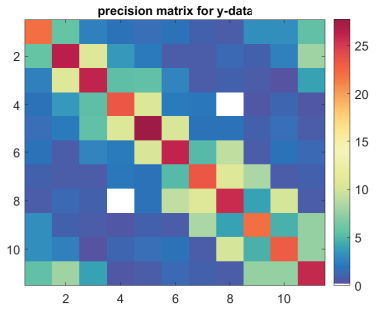
(a) Proposed approach: $\|\hat{\mathbf{\Delta}}^{(k\ell)}\|_F$, $k, \ell = 1, 2, \cdots, 11$



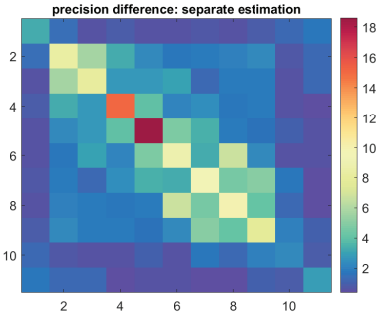
(b) Proposed approach: edges $\big\{\{k, \ell\} : \|\hat{\mathbf{\Delta}}^{(k\ell)}\|_F > 0\big\}$



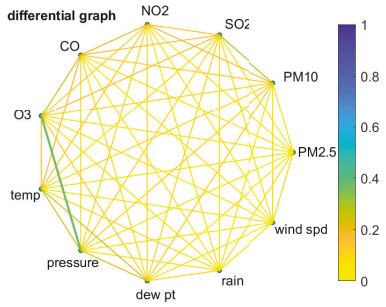
(c) $\|\hat{\mathbf{\Omega}}_x^{(k\ell)}\|_F$, $k, \ell = 1, 2, \cdots, 11$ [17]



(d) $\|\hat{\mathbf{\Omega}}_y^{(k\ell)}\|_F$, $k, \ell = 1, 2, \cdots, 11$ [17]

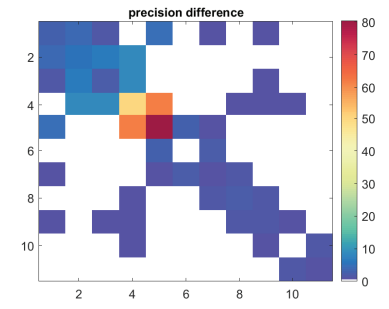


(e) $\|\hat{\mathbf{\Omega}}_y^{(k\ell)} - \hat{\mathbf{\Omega}}_x^{(k\ell)}\|_F$ ([17])

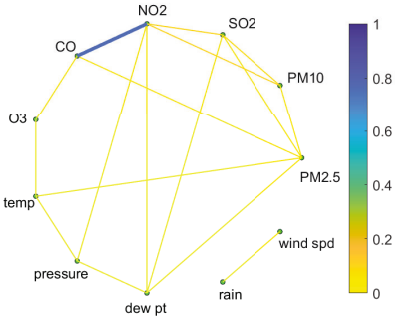


(f) Estimated edges $\big\{\{k, \ell\} : \|\hat{\mathbf{\Omega}}_y^{(k\ell)} - \hat{\mathbf{\Omega}}_x^{(k\ell)}\|_F > 0\big\}$

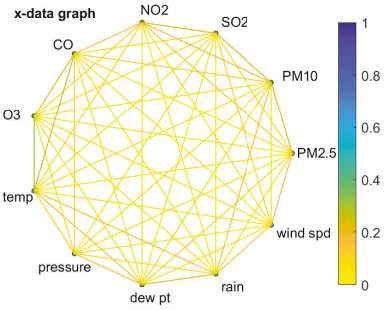
Fig. 5: Differential graphs comparing Beijing air-quality datasets [45] for years 2013-14 and 2014-15: 8 monitoring stations and 11 features ($m = 8$, $p = 11$, $n_x = n_y = 365$). The features are numbered 1-11 beginning PM2.5 ($\text{PM}_{2.5}$) and moving counter-clockwise in Fig. 5(b).



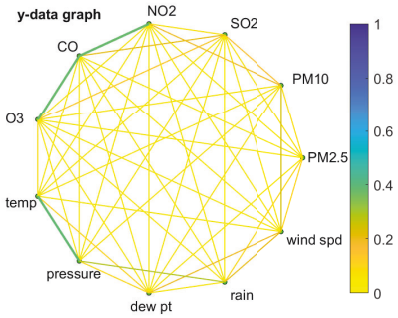
(a) Proposed approach: $\|\hat{\mathbf{\Delta}}^{(k\ell)}\|_F$, $k, \ell = 1, 2, \cdots, 11$



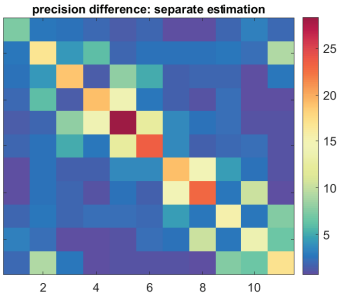
(b) Proposed approach: edges $\big\{\{k, \ell\} : \|\hat{\mathbf{\Delta}}^{(k\ell)}\|_F > 0\big\}$



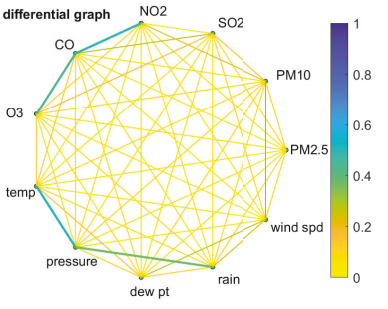
(c) Estimated edges $\big\{\{k, \ell\} : \|\hat{\mathbf{\Omega}}_x^{(k\ell)}\|_F > 0\big\}$ [17]



(d) Estimated edges $\big\{\{k, \ell\} : \|\hat{\mathbf{\Omega}}_y^{(k\ell)}\|_F > 0\big\}$ [17]



(e) $\|\hat{\mathbf{\Omega}}_y^{(k\ell)} - \hat{\mathbf{\Omega}}_x^{(k\ell)}\|_F$ ([17])



(f) Estimated edges $\big\{\{k, \ell\} : \|\hat{\mathbf{\Omega}}_y^{(k\ell)} - \hat{\mathbf{\Omega}}_x^{(k\ell)}\|_F > 0\big\}$

Fig. 6: Differential graphs comparing Beijing air-quality datasets [45] acquired from two sets of monitoring stations, 4 stations per set, year 2013-14: 4 monitoring stations and 11 features ($m = 4$, $p = 11$, $n_x = n_y = 365$).

does not yield a sparse differential graph, even though the two individual graphs in Figs. 5(c)-(d) are not all that different.

Fig. 6 shows the estimated differential graphs when comparing daily-averaged data over the period 2013-14, from four suburban/rural sites (x-data) to that from four urban sites (y-data), with air-quality and meteorological variables as $p = 11$ features measured at two sets of 4 monitoring sites (m=4). The objective again is to visualize and explore differential conditional dependency relationships among the 11 variables, but in this case comparing one subregion to another. There are significant differences in meteorological conditions and pollutant sources, levels and mutual interactions, among suburban and urban areas [45], [46]. The suburban areas (located toward north) are less polluted than the urban areas (located toward south) [45], [46]. Automobile exhaust is the main cause of NO_2 which is likely to undergo a chemical reaction with Ozone O_3, thereby, lowering its concentration [46]. Cold, dry air from the north reduces both dew point and $PM_{2.5}$ particle concentration in suburban areas while southerly wind brings warmer and more humid air from the more polluted south that elevates the $PM_{2.5}$ concentration [45]. The urban stations neighbor the south of Beijing which is heavily installed with iron, steel and cement industries in Hebei province [45]. Figs. 6(a)-(b) show estimated $\|\hat{\boldsymbol{\Delta}}^{(k\ell)}\|_F$ for various edges $\{k, \ell\}$, where it is unscaled in Fig. 6(a) but scaled in Fig. 6(b) so that the largest $\|\hat{\boldsymbol{\Delta}}^{(k\ell)}\|_F$ (including $k = \ell$) is normalized to one. It is seen that quite a few of the differential graph weights are significantly non-zero in Fig. 6(a), unlike that in Fig. 5(a), implying significant differences in the conditional dependency relationships among the 11 variables for suburban and urban areas. This observation conforms to the findings of [45], [46]. The comments made regarding Figs. 5(c)-(f) apply as well to Figs. 6(c)-(f).

VI. Conclusions

A group lasso penalized D-trace loss function approach for differential graph learning from multi-attribute data was presented. An ADMM algorithm was presented to optimize the convex objective function. Theoretical analysis establishing consistency of the estimator in high-dimensional settings was performed. We tested the proposed approach on synthetic as well as real data. In the synthetic data example, the multi-attribute approach is shown to outperform a single-attribute approach in correctly detecting the differential graph edges with ROC as the performance metric.

Appendix A
Technical Lemmas and Proof of Theorem 1

In this Appendix, we provide a proof of Theorem 1. A necessary and sufficient condition for minimization of convex $L_\lambda(\boldsymbol{\Delta}, \hat{\boldsymbol{\Sigma}}_x, \hat{\boldsymbol{\Sigma}}_y)$ given by (4) w.r.t. $\boldsymbol{\Delta} \in \mathbb{R}^{mp\times mp}$ is that $\hat{\boldsymbol{\Delta}}$ minimizes (4) iff the zero matrix belongs to the sub-differential of $L_\lambda(\boldsymbol{\Delta}, \hat{\boldsymbol{\Sigma}}_x, \hat{\boldsymbol{\Sigma}}_y)$. That is,

$$\begin{aligned}\mathbf{0} =& \frac{\partial L(\boldsymbol{\Delta}, \hat{\boldsymbol{\Sigma}}_x, \hat{\boldsymbol{\Sigma}}_y)}{\partial \boldsymbol{\Delta}} + \lambda \boldsymbol{Z}(\boldsymbol{\Delta})\Big|_{\boldsymbol{\Delta}=\hat{\boldsymbol{\Delta}}} \\ =& \hat{\boldsymbol{\Sigma}}_x \hat{\boldsymbol{\Delta}} \hat{\boldsymbol{\Sigma}}_y - (\hat{\boldsymbol{\Sigma}}_x - \hat{\boldsymbol{\Sigma}}_y) + \lambda \boldsymbol{Z}(\hat{\boldsymbol{\Delta}}) \end{aligned} \tag{28}$$

where $\boldsymbol{Z}(\boldsymbol{\Delta}) \in \partial \sum_{k,\ell=1}^{p} \|\boldsymbol{\Delta}^{(k\ell)}\|_F \in \mathbb{R}^{mp\times mp}$, the sub-differential of group lasso penalty term, is given by [36], [37]

$$(\boldsymbol{Z}(\boldsymbol{\Delta}))^{(k\ell)} = \begin{cases} \frac{\boldsymbol{\Delta}^{(k\ell)}}{\|\boldsymbol{\Delta}^{(k\ell)}\|_F} & \text{if } \|\boldsymbol{\Delta}^{(k\ell)}\|_F \neq 0 \\ \boldsymbol{V} \in \mathbb{R}^{m\times m},\ \|\boldsymbol{V}\|_F \leq 1, & \text{if } \|\boldsymbol{\Delta}^{(k\ell)}\|_F = 0 \end{cases}. \tag{29}$$

In terms of $m \times m$ submatrices of $\boldsymbol{\Delta}$, $\hat{\boldsymbol{\Sigma}}_x$, $\hat{\boldsymbol{\Sigma}}_y$ and $\boldsymbol{Z}(\boldsymbol{\Delta})$ corresponding to various graph edges, using $\text{bvec}(\boldsymbol{ADB}) = (\boldsymbol{B}^\top \boxtimes \boldsymbol{A})\text{bvec}(\boldsymbol{D})$ [33, Lemma 1], we may rewrite (28) as

$$(\hat{\boldsymbol{\Sigma}}_y \boxtimes \hat{\boldsymbol{\Sigma}}_x)\text{bvec}(\hat{\boldsymbol{\Delta}}) - \text{bvec}(\hat{\boldsymbol{\Sigma}}_x - \hat{\boldsymbol{\Sigma}}_y) + \lambda\, \text{bvec}(\boldsymbol{Z}(\hat{\boldsymbol{\Delta}})) = \mathbf{0} \tag{30}$$

Then (30) can be rewritten as

$$\begin{aligned}&\begin{bmatrix} \hat{\boldsymbol{\Gamma}}_{S,S} & \hat{\boldsymbol{\Gamma}}_{S,S^c} \\ \hat{\boldsymbol{\Gamma}}_{S^c,S} & \hat{\boldsymbol{\Gamma}}_{S^c,S^c} \end{bmatrix} \begin{bmatrix} \text{bvec}(\hat{\boldsymbol{\Delta}}_S) \\ \text{bvec}(\hat{\boldsymbol{\Delta}}_{S^c}) \end{bmatrix} - \begin{bmatrix} \text{bvec}((\hat{\boldsymbol{\Sigma}}_x - \hat{\boldsymbol{\Sigma}}_y)_S) \\ \text{bvec}((\hat{\boldsymbol{\Sigma}}_x - \hat{\boldsymbol{\Sigma}}_y)_{S^c}) \end{bmatrix} \\ &\quad + \lambda \begin{bmatrix} \text{bvec}(\boldsymbol{Z}(\hat{\boldsymbol{\Delta}}_S)) \\ \text{bvec}(\boldsymbol{Z}(\hat{\boldsymbol{\Delta}}_{S^c})) \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}. \end{aligned} \tag{31}$$

The general approach of [31] (followed in [8], [16], [28], [29]) is to first solve the hypothetical constrained optimization problem with known edgeset S

$$\tilde{\boldsymbol{\Delta}} = \arg \min_{\boldsymbol{\Delta}:\boldsymbol{\Delta}_{S^c}=\mathbf{0}} L_\lambda(\boldsymbol{\Delta}, \hat{\boldsymbol{\Sigma}}_x, \hat{\boldsymbol{\Sigma}}_y) \tag{32}$$

where S^c is the complement of S. Since, by construction, $\tilde{\boldsymbol{\Delta}}_{S^c} = \mathbf{0}$, in this case (31) reduces to

$$\hat{\boldsymbol{\Gamma}}_{S,S}\text{bvec}(\tilde{\boldsymbol{\Delta}}_S) - \text{bvec}((\hat{\boldsymbol{\Sigma}}_x - \hat{\boldsymbol{\Sigma}}_y)_S) + \lambda\, \text{bvec}(\boldsymbol{Z}(\tilde{\boldsymbol{\Delta}}_S)) = \mathbf{0}\,. \tag{33}$$

In the approach of [31], one investigates conditions under which the solution $\hat{\boldsymbol{\Delta}}$ to (4) is the same as the solution $\tilde{\boldsymbol{\Delta}}$ to (32). This is done by showing that $\hat{\boldsymbol{\Delta}}$ satisfies (31). The choice $\hat{\boldsymbol{\Delta}} = \tilde{\boldsymbol{\Delta}}$ implies that $\hat{\boldsymbol{\Delta}}_{S^c} = \mathbf{0}$ and (33) is true with $\tilde{\boldsymbol{\Delta}}$ replaced with $\hat{\boldsymbol{\Delta}}$. In order to satisfy (31), it remains to show that for any edge $e \in S^c$, we have strict feasibility

$$\|\hat{\boldsymbol{\Gamma}}_{e,S}\text{bvec}(\tilde{\boldsymbol{\Delta}}_S) - \text{bvec}((\hat{\boldsymbol{\Sigma}}_x - \hat{\boldsymbol{\Sigma}}_y)_e)\|_2 < \lambda\,, \tag{34}$$

where for $\boldsymbol{a} \in \mathbb{R}^q$, $\|\boldsymbol{a}\|_2 = \sqrt{\boldsymbol{a}^\top \boldsymbol{a}}$. This requires a set of sufficient conditions stated in Theorem 1.

Lemma 1 follows from [31, Lemma 1]. It is stated for more general sub-Gaussian distributions as in [31, Lemma 1], but will be used later for Gaussian distributions, a subset of sub-Gaussian distributions.

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

$$\tilde{C}_0 = 8(1 + 4\sigma_{sg}^2)m\sigma_{max}\sqrt{2\big(\tau + \ln(4m^2)/\ln(p_n)\big)}. \tag{35}$$

Then

$$P\Big(\|\boldsymbol{\mathcal{C}}(\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}^*)\|_\infty > \tilde{C}_0\sqrt{\ln(p_n)/n}\Big) \leq 1/p_n^{\tau-2} \tag{36}$$

for any $\tau > 2$ and $n > 2m^2(\ln(4m^2) + \tau\ln(p_n))$. •

Proof. By [31, Lemma 1], with $b = 8(1 + 4\sigma_{sg}^2)$, we have

$$P\Big(|[\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}^*]_{ij}| > \delta\Big) \leq 4\,\exp(-c_* n\delta^2) \tag{37}$$

for any $\delta \in (0, b\,\sigma_{max})$ where $c_*^{-1} = 2b^2\sigma_{max}^2$. For any edge $\{k,\ell\}$ of the MA graph, with m^2 edges $\{i,j\}$ of the corresponding SA graph associated with $\{k,\ell\}$, using the union bound, we have

$$\begin{aligned} &P\Big(|[\boldsymbol{\mathcal{C}}(\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}^*)]_{kl}| > \delta\Big) \\ &\le P\Big(\max_{\{i,j\}\in\{k,\ell\}} ([\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}^*]_{ij})^2 > \frac{\delta^2}{m^2}\Big) \\ &\le m^2 P\Big(|[\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}^*]_{ij}| > \frac{\delta}{m}\Big) = 4m^2 \exp\big(-c_* n \frac{\delta^2}{m^2}\big)\,. \end{aligned} \tag{38}$$

Applying the union bound once more over all p_n^2 entries

$$P\Big(\|\boldsymbol{\mathcal{C}}(\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}^*)\|_\infty > \delta\Big) \le 4(mp_n)^2 \exp\big(-c_* n \frac{\delta^2}{m^2}\big) =: P_{tb}\,. \tag{39}$$

Choose $\delta = \tilde{C}_0\sqrt{\ln(p_n)/n} = bm\sigma_{max}\sqrt{2\ln(4p_n^\tau m^2)/n}$. Then we have

$$P_{tb} = 4(mp_n)^2 \exp\big(\ln(4p_n^\tau m^2)^{-1}\big) = 1/p_n^{\tau-2} \tag{40}$$

provided $\delta \in (0, b\sigma_{max})$. Therefore, we need to have $\tilde{C}_0\sqrt{\ln(p_n)/n} < b\sigma_{max}$ requiring $n > 2m^2(\ln(4m^2) + \tau\ln(p_n))$. This completes the proof. □

Using the union bound, Lemma 1 and Gaussian assumption, we have Lemma 2.

Lemma 2. Let $\hat{\boldsymbol{\Sigma}}_x$ and $\hat{\boldsymbol{\Sigma}}_y$ be as in (2), $\bar{\sigma}_{xy}$ as in (21), C_0 as in (22) and assume data are Gaussian. Define $n = \min(n_x, n_y)$ and

$$\mathcal{A} = \max\big\{\|\boldsymbol{\mathcal{C}}(\hat{\boldsymbol{\Sigma}}_x - \boldsymbol{\Sigma}_x^*)\|_\infty, \|\boldsymbol{\mathcal{C}}(\hat{\boldsymbol{\Sigma}}_y - \boldsymbol{\Sigma}_y^*)\|_\infty\big\}\,. \tag{41}$$

Then for any $\tau > 2$ and $n > 2m^2\ln(4m^2p_n^\tau)$,

$$P\Big(\mathcal{A} > C_0\sqrt{\ln(p_n)/n}\Big) \le 2/p_n^{\tau-2} \quad \bullet \tag{42}$$

Proof. For Gaussian distribution, the sub-Gaussian parameter σ_{sg} of Lemma 1 equals 1. Then $8(1+4\sigma_{sg}^2) = 40$. Let $C_{0x} = 40m(\max_i \Sigma_{x,ii}^*)\sqrt{2\big(\tau + \ln(4m^2)/\ln(p_n)\big)}$ and $C_{0y} = 40m(\max_i \Sigma_{y,ii}^*)\sqrt{2\big(\tau + \ln(4m^2)/\ln(p_n)\big)}$ (where $\Sigma_{x,ii}^* = [\boldsymbol{\Sigma}_x^*)]_{ii}$, etc.). Then using Lemma 1 and union bound,

$$\begin{aligned} &P\Big(\mathcal{A} > C_0\sqrt{\ln(p_n)/n}\Big) \\ &\le P\Big(\|\boldsymbol{\mathcal{C}}(\hat{\boldsymbol{\Sigma}}_x - \boldsymbol{\Sigma}_x^*)\|_\infty > C_0\sqrt{\ln(p_n)/n}\Big) \\ &\quad + P\Big(\|\boldsymbol{\mathcal{C}}(\hat{\boldsymbol{\Sigma}}_y - \boldsymbol{\Sigma}_y^*)\|_\infty > C_0\sqrt{\ln(p_n)/n}\Big) \\ &\le 2/p_n^{\tau-2} \end{aligned} \tag{43}$$

since $C_0 \ge C_{0x}$ and $C_0 \ge C_{0y}$. □

Recall (16)-(20) and define

$$\boldsymbol{\Delta}_x = \hat{\boldsymbol{\Sigma}}_x - \boldsymbol{\Sigma}_x^*\,,\ \boldsymbol{\Delta}_y = \hat{\boldsymbol{\Sigma}}_y - \boldsymbol{\Sigma}_y^*\,,\ \boldsymbol{\Delta}_\Gamma = \hat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}^*\,, \tag{44}$$

$$\boldsymbol{\Delta}_\Sigma = \boldsymbol{\Delta}_x - \boldsymbol{\Delta}_y\,,\ \epsilon_x = \|\boldsymbol{\mathcal{C}}(\boldsymbol{\Delta}_x)\|_\infty\,, \tag{45}$$

$$\epsilon_y = \|\boldsymbol{\mathcal{C}}(\boldsymbol{\Delta}_y)\|_\infty\,,\ \epsilon > \max\{\epsilon_x, \epsilon_y\}\,. \tag{46}$$

Lemma 3. Assume that

$$\kappa_\Gamma < \frac{1}{3s_n(\epsilon^2 + 2M\epsilon)}\,. \tag{47}$$

Let ($\boldsymbol{\Gamma}_{S,S}^{-*}$ denotes $(\boldsymbol{\Gamma}_{S,S}^*)^{-1}$)

$$R(\boldsymbol{\Delta}_\Gamma) = \hat{\boldsymbol{\Gamma}}_{S,S}^{-1} - \boldsymbol{\Gamma}_{S,S}^{-*} + \boldsymbol{\Gamma}_{S,S}^{-*}(\boldsymbol{\Delta}_\Gamma)_{S,S}\boldsymbol{\Gamma}_{S,S}^{-*}\,. \tag{48}$$

Then we have

$$\|\boldsymbol{\mathcal{C}}(R(\boldsymbol{\Delta}_\Gamma))\|_\infty \le \frac{3}{2}\kappa_\Gamma^3 s_n(\epsilon^2 + 2M\epsilon)^2\,, \tag{49}$$

$$\|\boldsymbol{\mathcal{C}}(R(\boldsymbol{\Delta}_\Gamma))\|_{1,\infty} \le \frac{3}{2}\kappa_\Gamma^3 s_n^2(\epsilon^2 + 2M\epsilon)^2\,, \tag{50}$$

$$\begin{aligned} &\|\boldsymbol{\mathcal{C}}(\hat{\boldsymbol{\Gamma}}_{S,S}^{-1} - \boldsymbol{\Gamma}_{S,S}^{-*})\|_\infty \\ &\quad \le \kappa_\Gamma^2(\epsilon^2 + 2M\epsilon)\big(1 + 1.5s_n\kappa_\Gamma(\epsilon^2 + 2M\epsilon)\big)\,, \end{aligned} \tag{51}$$

$$\|\boldsymbol{\mathcal{C}}(\hat{\boldsymbol{\Gamma}}_{S,S}^{-1} - \boldsymbol{\Gamma}_{S,S}^{-*})\|_{1,\infty} \le s_n\|\boldsymbol{\mathcal{C}}(\hat{\boldsymbol{\Gamma}}_{S,S}^{-1} - \boldsymbol{\Gamma}_{S,S}^{-*})\|_\infty \quad \bullet \tag{52}$$

Proof. We have

$$\begin{aligned} \boldsymbol{\Delta}_\Gamma &= \hat{\boldsymbol{\Sigma}}_y \boxtimes \hat{\boldsymbol{\Sigma}}_x - \boldsymbol{\Sigma}_y^* \boxtimes \boldsymbol{\Sigma}_x^* \\ &= \boldsymbol{\Delta}_y \boxtimes \boldsymbol{\Delta}_x + \boldsymbol{\Sigma}_y^* \boxtimes \boldsymbol{\Delta}_x + \boldsymbol{\Delta}_y \boxtimes \boldsymbol{\Sigma}_x^*. \end{aligned} \tag{53}$$

By [16, Lemma 14],

$$\|\boldsymbol{\mathcal{C}}(\boldsymbol{AB})\|_{1,\infty} \le \|\boldsymbol{\mathcal{C}}(\boldsymbol{A})\|_{1,\infty}\|\boldsymbol{\mathcal{C}}(\boldsymbol{B})\|_{1,\infty} \tag{54}$$

and by [16, Lemma 15],

$$\|\boldsymbol{\mathcal{C}}(\boldsymbol{AB})\|_\infty \le \|\boldsymbol{\mathcal{C}}(\boldsymbol{A})\|_\infty\|\boldsymbol{\mathcal{C}}(\boldsymbol{B}^\top)\|_{1,\infty}\,. \tag{55}$$

Since $\|\boldsymbol{A}\otimes\boldsymbol{B}\|_F = \|\boldsymbol{A}\|_F\,\|\boldsymbol{B}\|_F$ and $\boldsymbol{A}\boxtimes\boldsymbol{B} = [\boldsymbol{A}_{ij}\boxtimes\boldsymbol{B}]_{ij} = [[\boldsymbol{A}_{ij}\otimes\boldsymbol{B}_{k\ell}]_{k\ell}]_{ij}$, we have

$$\|\boldsymbol{\mathcal{C}}(\boldsymbol{A}\boxtimes\boldsymbol{B})\|_\infty \le \|\boldsymbol{\mathcal{C}}(\boldsymbol{A})\|_\infty\|\boldsymbol{\mathcal{C}}(\boldsymbol{B})\|_\infty\,. \tag{56}$$

From (17), (46), (53) and (56),

$$\|\boldsymbol{\mathcal{C}}(\boldsymbol{\Delta}_\Gamma)\|_\infty \le \epsilon_x\epsilon_y + M\epsilon_x + M\epsilon_y < \epsilon^2 + 2M\epsilon \tag{57}$$

and since $|S| = s_n$,

$$\begin{aligned} \|\boldsymbol{\mathcal{C}}((\boldsymbol{\Delta}_\Gamma)_{S,S})\|_{1,\infty} &\le s_n\|\boldsymbol{\mathcal{C}}((\boldsymbol{\Delta}_\Gamma)_{S,S})\|_\infty \le s_n\|\boldsymbol{\mathcal{C}}(\boldsymbol{\Delta}_\Gamma)\|_\infty \\ &< s_n(\epsilon^2 + 2M\epsilon)\,. \end{aligned} \tag{58}$$

By assumption (47),

$$\begin{aligned} \kappa_\Gamma\|\boldsymbol{\mathcal{C}}((\boldsymbol{\Delta}_\Gamma)_{S,S})\|_{1,\infty} &= \|\boldsymbol{\mathcal{C}}((\Gamma_{S,S}^*)^{-1})\|_{1,\infty}\|\boldsymbol{\mathcal{C}}((\boldsymbol{\Delta}_\Gamma)_{S,S})\|_{1,\infty} \\ &< \frac{1}{3}\,. \end{aligned} \tag{59}$$

By (59) we can invoke [31, Lemma 5] to have

$$R(\boldsymbol{\Delta}_\Gamma) = \boldsymbol{\Gamma}_{S,S}^{-*}(\boldsymbol{\Delta}_\Gamma)_{S,S}\boldsymbol{\Gamma}_{S,S}^{-*}(\boldsymbol{\Delta}_\Gamma)_{S,S}\boldsymbol{J}\boldsymbol{\Gamma}_{S,S}^{-*} \tag{60}$$

where $\boldsymbol{J} = \sum_{k=0}^\infty(-1)^k\big(\boldsymbol{\Gamma}_{S,S}^{-*}(\boldsymbol{\Delta}_\Gamma)_{S,S}\big)^k$. Using (54), (55) and (60), we have

$$\begin{aligned} \|\boldsymbol{\mathcal{C}}(R(\boldsymbol{\Delta}_\Gamma))\|_\infty &\le \|\boldsymbol{\mathcal{C}}(\boldsymbol{\Gamma}_{S,S}^{-*}(\boldsymbol{\Delta}_\Gamma)_{S,S})\|_\infty \\ &\quad \times \|\boldsymbol{\mathcal{C}}(\boldsymbol{\Gamma}_{S,S}^{-*}(\boldsymbol{\Delta}_\Gamma)_{S,S}J\boldsymbol{\Gamma}_{S,S}^{-*})^\top\|_{1,\infty} \\ &\le \|\boldsymbol{\mathcal{C}}(\boldsymbol{\Gamma}_{S,S}^{-*})\|_{1,\infty}^3\|\boldsymbol{\mathcal{C}}((\boldsymbol{\Delta}_\Gamma)_{S,S})\|_\infty\|\boldsymbol{\mathcal{C}}((\boldsymbol{\Delta}_\Gamma)_{S,S})\|_{1,\infty} \\ &\quad \times \|\boldsymbol{\mathcal{C}}(\boldsymbol{J}^\top)\|_{1,\infty}\,. \end{aligned} \tag{61}$$

Now using (59),

$$\begin{aligned} \|\boldsymbol{\mathcal{C}}(\boldsymbol{J}^\top)\|_{1,\infty} &\le \sum_{k=0}^\infty\|\boldsymbol{\mathcal{C}}(\boldsymbol{\Gamma}_{S,S}^{-*})\|_{1,\infty}^k\|\boldsymbol{\mathcal{C}}((\boldsymbol{\Delta}_\Gamma)_{S,S})\|_{1,\infty}^k \\ &= \frac{1}{1 - \|\boldsymbol{\mathcal{C}}(\boldsymbol{\Gamma}_{S,S}^{-*})\|_{1,\infty}\|\boldsymbol{\mathcal{C}}((\boldsymbol{\Delta}_\Gamma)_{S,S})\|_{1,\infty}} \\ &= \frac{1}{1 - \kappa_\Gamma\|\boldsymbol{\mathcal{C}}((\boldsymbol{\Delta}_\Gamma)_{S,S})\|_{1,\infty}} \overset{(59)}{<} \frac{1}{1-(1/3)} = \frac{3}{2}\,. \end{aligned} \tag{62}$$

Using (57), (59), (61) and (62), we have

$$\begin{aligned}\|\boldsymbol{\mathcal{C}}(R(\boldsymbol{\Delta}_\Gamma))\|_\infty \;&\le\; \frac{3}{2}\kappa_\Gamma^3 s_n \|\boldsymbol{\mathcal{C}}((\boldsymbol{\Delta}_\Gamma)_{S,S})\|_\infty^2 \\ &< \frac{3}{2}\kappa_\Gamma^3 s_n(\epsilon^2+2M\epsilon)^2\,. \end{aligned} \tag{63}$$

This proves (49), from which (50) immediately follows.

Using (19), (48), (54), (55) and (57) we have

$$\begin{aligned}&\|\boldsymbol{\mathcal{C}}(\hat{\boldsymbol{\Gamma}}_{S,S}^{-1}-\boldsymbol{\Gamma}_{S,S}^{-*})\|_\infty \le \|\boldsymbol{\mathcal{C}}(R(\boldsymbol{\Delta}_\Gamma))\|_\infty \\ &\quad + \|\boldsymbol{\mathcal{C}}(\boldsymbol{\Gamma}_{S,S}^{-*}(\boldsymbol{\Delta}_\Gamma)_{S,S}\boldsymbol{\Gamma}_{S,S}^{-*})\|_\infty \\ &\le \|\boldsymbol{\mathcal{C}}(R(\boldsymbol{\Delta}_\Gamma))\|_\infty + \|\boldsymbol{\mathcal{C}}(\boldsymbol{\Gamma}_{S,S}^{-*})\|_{1,\infty}^2\|\boldsymbol{\mathcal{C}}((\boldsymbol{\Delta}_\Gamma)_{S,S})\|_\infty \\ &< \kappa_\Gamma^2(\epsilon^2+2M\epsilon)\big(1+1.5s_n\kappa_\Gamma(\epsilon^2+2M\epsilon)\big)\,. \end{aligned} \tag{64}$$

This proves (51). The claim (52) follows by noting that $|S| = s_n$. This completes the proof. □

Lemma 4. Assume (47) and the following conditions:

$$0<\alpha<1 \text{ where } \alpha \text{ is as in } (20)\,, \tag{65}$$

$$\epsilon < \min\left\{M, \frac{\alpha\lambda_n}{2(2-\alpha)}\right\}, \tag{66}$$

$$\alpha C_\alpha \min\{\lambda_n, 1\} \ge 3s_n\epsilon M\kappa_\Gamma B_s \tag{67}$$

where

$$C_\alpha = \frac{\alpha\lambda_n + 2\epsilon\alpha - 4\epsilon}{2M\alpha\lambda_n + \alpha\lambda_n + 2\epsilon\alpha}\,, \tag{68}$$

$$\begin{aligned}B_s = \Big[1+\kappa_\Gamma\Big(3s_n\epsilon M + \min\{s_n M^2, M_\Sigma^2\}\Big) \\ \times\big(4.5s_n\epsilon M\kappa_\Gamma+1\big)\Big]\,. \end{aligned} \tag{69}$$

Then we have

(i) $\text{bvec}(\hat{\boldsymbol{\Delta}}_{S^c}) = \mathbf{0}$.

(ii) $\|\boldsymbol{\mathcal{C}}(\hat{\boldsymbol{\Delta}}-\boldsymbol{\Delta}^*)\|_\infty \le 2\lambda_n\kappa_\Gamma + 3s_n\epsilon M\kappa_\Gamma^2\big(4.5s_n\epsilon M\kappa_\Gamma + 1\big)\big(2M+2\lambda_n\big)$ •

Proof. To establish part (i), we need to show that (34) is true. Let d denote the left-side of (34). It follow from (33) that

$$\text{bvec}(\tilde{\boldsymbol{\Delta}}_S) = \hat{\boldsymbol{\Gamma}}_{S,S}^{-1}\Big(\text{bvec}((\hat{\boldsymbol{\Sigma}}_x-\hat{\boldsymbol{\Sigma}}_y)_S) - \lambda\,\text{bvec}(\boldsymbol{Z}(\tilde{\boldsymbol{\Delta}}_S))\Big)\,. \tag{70}$$

Substitute (70) in the left-side of (34) to yield

$$\begin{aligned}d =& \|\hat{\boldsymbol{\Gamma}}_{e,S}\big[\hat{\boldsymbol{\Gamma}}_{S,S}^{-1}\big(\text{bvec}((\hat{\boldsymbol{\Sigma}}_x-\hat{\boldsymbol{\Sigma}}_y)_S) - \lambda\,\text{bvec}(\boldsymbol{Z}(\tilde{\boldsymbol{\Delta}}_S))\big)\big] \\ &- \text{bvec}((\hat{\boldsymbol{\Sigma}}_x-\hat{\boldsymbol{\Sigma}}_y)_e)\|_2\,. \end{aligned} \tag{71}$$

At the true values we have

$$\mathbf{0} = \frac{\partial L_\lambda(\boldsymbol{\Delta},\boldsymbol{\Sigma}_x^*,\boldsymbol{\Sigma}_y^*)}{\partial\boldsymbol{\Delta}}\Big|_{\boldsymbol{\Delta}=\boldsymbol{\Delta}^*} = \boldsymbol{\Sigma}_x^*\boldsymbol{\Delta}^*\boldsymbol{\Sigma}_y^* - (\boldsymbol{\Sigma}_x^*-\boldsymbol{\Sigma}_y^*)$$

implying

$$\boldsymbol{\Gamma}^*\text{bvec}(\boldsymbol{\Delta}^*) - \text{bvec}(\boldsymbol{\Sigma}_x^*-\boldsymbol{\Sigma}_y^*) = \mathbf{0}\,, \tag{72}$$

which, noting that $(\boldsymbol{\Delta}^*)_{S^c} = \mathbf{0}$, can be rewritten as (cf. (31))

$$\boldsymbol{\Gamma}_{S,S}^*\text{bvec}(\boldsymbol{\Delta}_S^*) = \text{bvec}(\boldsymbol{\Sigma}_x^*)_S - \text{bvec}(\boldsymbol{\Sigma}_y^*)_S\,, \tag{73}$$

$$\boldsymbol{\Gamma}_{e,S}^*\text{bvec}(\boldsymbol{\Delta}_S^*) = \text{bvec}(\boldsymbol{\Sigma}_x^*)_e - \text{bvec}(\boldsymbol{\Sigma}_y^*)_e\,. \tag{74}$$

Therefore, $(\boldsymbol{A}^{-*} = (\boldsymbol{A}^*)^{-1})$,

$$\begin{aligned}&\boldsymbol{\Gamma}_{e,S}^*\boldsymbol{\Gamma}_{S,S}^{-*}\big(\text{bvec}(\boldsymbol{\Sigma}_x^*)_S - \text{bvec}(\boldsymbol{\Sigma}_y^*)_S\big) \\ &\qquad - \text{bvec}(\boldsymbol{\Sigma}_x^*)_e + \text{bvec}(\boldsymbol{\Sigma}_y^*)_e = \mathbf{0}\,. \end{aligned} \tag{75}$$

Recalling (44) and using (75) in (71),

$$\begin{aligned}d =& \|\hat{\boldsymbol{\Gamma}}_{e,S}\hat{\boldsymbol{\Gamma}}_{S,S}^{-1}\text{bvec}((\boldsymbol{\Delta}_\Sigma)_S) \\ &+ \big(\hat{\boldsymbol{\Gamma}}_{e,S}\hat{\boldsymbol{\Gamma}}_{S,S}^{-1} - \boldsymbol{\Gamma}_{e,S}^*\boldsymbol{\Gamma}_{S,S}^{-*}\big)\big(\text{bvec}(\boldsymbol{\Sigma}_x^*)_S - \text{bvec}(\boldsymbol{\Sigma}_y^*)_S\big) \\ &- \lambda\,\hat{\boldsymbol{\Gamma}}_{e,S}\hat{\boldsymbol{\Gamma}}_{S,S}^{-1}\text{bvec}(\boldsymbol{Z}(\tilde{\boldsymbol{\Delta}}_S)) - \text{bvec}((\boldsymbol{\Delta}_\Sigma)_e)\|_2\,. \end{aligned} \tag{76}$$

We now bound various terms in (76). Note that $\hat{\boldsymbol{\Gamma}}_{e,S} \in \mathbb{R}^{m^2\times(m^2 s_n)}$, $\hat{\boldsymbol{\Gamma}}_{S,S}^{-1} \in \mathbb{R}^{(m^2 s_n)\times(m^2 s_n)}$, and $\text{bvec}((\boldsymbol{\Delta}_\Sigma)_S) \in \mathbb{R}^{m^2 s_n}$ where $\boldsymbol{\Delta}_\Sigma \in \mathbb{R}^{(mp_n)\times(mp_n)}$. Consider $\boldsymbol{A}_{e,S} \in \mathbb{R}^{m^2\times(m^2 s_n)}$. Then

$$\|\boldsymbol{A}_{e,S}\text{bvec}((\boldsymbol{\Delta}_\Sigma)_S)\|_2 = \|\sum_{f\in S}\boldsymbol{A}_{e,f}\text{vec}((\boldsymbol{\Delta}_\Sigma)_f)\|_2 \tag{77}$$

where edge $f \in S$, $\boldsymbol{A}_{e,f} \in \mathbb{R}^{m^2\times m^2}$ and $(\boldsymbol{\Delta}_\Sigma)_f \in \mathbb{R}^{m\times m}$. By the triangle inequality

$$\|\boldsymbol{A}_{e,S}\text{bvec}((\boldsymbol{\Delta}_\Sigma)_S)\|_2 \le \sum_{f\in S}\|\boldsymbol{A}_{e,f}\text{vec}((\boldsymbol{\Delta}_\Sigma)_f)\|_2\,. \tag{78}$$

With $\boldsymbol{B}_{i.}$ denoting the ith row of matrix $\boldsymbol{B}$ and using Cauchy-Schwarz inequality, we have

$$\begin{aligned}\|\boldsymbol{A}_{e,f}\text{vec}((\boldsymbol{\Delta}_\Sigma)_f)\|_2 &= \Big(\sum_{i=1}^{m^2}\big([\boldsymbol{A}_{e,f}]_{i.}\text{vec}((\boldsymbol{\Delta}_\Sigma)_f)\big)^2\Big)^{1/2} \\ &\le \Big(\sum_{i=1}^{m^2}\|[\boldsymbol{A}_{e,f}]_{i.}\|_2^2\,\|\text{vec}((\boldsymbol{\Delta}_\Sigma)_f)\|_2^2\Big)^{1/2} \\ &= \|\text{vec}((\boldsymbol{\Delta}_\Sigma)_f)\|_2\Big(\sum_{i=1}^{m^2}\|[\boldsymbol{A}_{e,f}]_{i.}\|_2^2\Big)^{1/2} \\ &= \|\text{vec}((\boldsymbol{\Delta}_\Sigma)_f)\|_2\,\|\boldsymbol{A}_{e,f}\|_F\,. \end{aligned} \tag{79}$$

Therefore, using (78) and (79),

$$\begin{aligned}\|\boldsymbol{A}_{e,S}\text{bvec}((\boldsymbol{\Delta}_\Sigma)_S)\|_2 &\le \Big(\sum_{f\in S}\|\boldsymbol{A}_{e,f}\|_F\Big)\max_{g\in S}\|\text{vec}((\boldsymbol{\Delta}_\Sigma)_g)\|_2 \\ &= \|\boldsymbol{\mathcal{C}}(\boldsymbol{A}_{e,S})\|_1\,\|\boldsymbol{\mathcal{C}}(\boldsymbol{\Delta}_\Sigma)\|_\infty\,. \end{aligned} \tag{80}$$

Using (80), we have

$$\begin{aligned}&\|\hat{\boldsymbol{\Gamma}}_{e,S}\hat{\boldsymbol{\Gamma}}_{S,S}^{-1}\text{bvec}((\boldsymbol{\Delta}_\Sigma)_S)\|_2 \\ &\quad\le \|\boldsymbol{\mathcal{C}}(\hat{\boldsymbol{\Gamma}}_{e,S}\hat{\boldsymbol{\Gamma}}_{S,S}^{-1})\|_1\,\|\boldsymbol{\mathcal{C}}(\boldsymbol{\Delta}_\Sigma)\|_\infty\,, \end{aligned} \tag{81}$$

$$\begin{aligned}&\|\big(\hat{\boldsymbol{\Gamma}}_{e,S}\hat{\boldsymbol{\Gamma}}_{S,S}^{-1} - \boldsymbol{\Gamma}_{e,S}^*\boldsymbol{\Gamma}_{S,S}^{-*}\big)\big(\text{bvec}(\boldsymbol{\Sigma}_x^*)_S - \text{bvec}(\boldsymbol{\Sigma}_y^*)_S\big)\|_2 \\ &\quad\le \|\boldsymbol{\mathcal{C}}(\hat{\boldsymbol{\Gamma}}_{e,S}\hat{\boldsymbol{\Gamma}}_{S,S}^{-1} - \boldsymbol{\Gamma}_{e,S}^*\boldsymbol{\Gamma}_{S,S}^{-*})\|_1\,\|\boldsymbol{\mathcal{C}}(\boldsymbol{\Sigma}_x^*-\boldsymbol{\Sigma}_y^*)\|_\infty\,, \end{aligned} \tag{82}$$

$$\begin{aligned}&\|\hat{\boldsymbol{\Gamma}}_{e,S}\hat{\boldsymbol{\Gamma}}_{S,S}^{-1}\text{bvec}(\boldsymbol{Z}(\tilde{\boldsymbol{\Delta}}_S))\|_2 \\ &\quad\le \|\boldsymbol{\mathcal{C}}(\hat{\boldsymbol{\Gamma}}_{e,S}\hat{\boldsymbol{\Gamma}}_{S,S}^{-1})\|_1\,\|\boldsymbol{\mathcal{C}}(\boldsymbol{Z}(\tilde{\boldsymbol{\Delta}}_S))\|_\infty\,, \end{aligned} \tag{83}$$

$$\|\text{bvec}((\boldsymbol{\Delta}_\Sigma)_e)\|_2 \le \|\boldsymbol{\mathcal{C}}(\boldsymbol{\Delta}_\Sigma)\|_\infty\,. \tag{84}$$

By (29), (44) and (46)

$$\|\boldsymbol{\mathcal{C}}(\boldsymbol{\Sigma}_x^*-\boldsymbol{\Sigma}_y^*)\|_\infty \le \|\boldsymbol{\mathcal{C}}(\boldsymbol{\Sigma}_x^*)\|_\infty + \|\boldsymbol{\mathcal{C}}(\boldsymbol{\Sigma}_y^*)\|_\infty \le 2M\,, \tag{85}$$

$$\|\boldsymbol{\mathcal{C}}(\boldsymbol{Z}(\tilde{\boldsymbol{\Delta}}_S))\|_\infty \le 1\,, \tag{86}$$

$$\|\boldsymbol{\mathcal{C}}(\boldsymbol{\Delta}_\Sigma)\|_\infty \le \|\boldsymbol{\mathcal{C}}(\boldsymbol{\Delta}_x)\|_\infty + \|\boldsymbol{\mathcal{C}}(\boldsymbol{\Delta}_y)\|_\infty < 2\epsilon\,. \tag{87}$$

Using (76) and (81)-87),

$$\begin{aligned}d \;<\; &2\epsilon\|\boldsymbol{\mathcal{C}}(\hat{\boldsymbol{\Gamma}}_{e,S}\hat{\boldsymbol{\Gamma}}_{S,S}^{-1})\|_1 + 2M\|\boldsymbol{\mathcal{C}}(\hat{\boldsymbol{\Gamma}}_{e,S}\hat{\boldsymbol{\Gamma}}_{S,S}^{-1} - \boldsymbol{\Gamma}_{e,S}^*\boldsymbol{\Gamma}_{S,S}^{-*})\|_1 \\ &+ \lambda\,\|\boldsymbol{\mathcal{C}}(\hat{\boldsymbol{\Gamma}}_{e,S}\hat{\boldsymbol{\Gamma}}_{S,S}^{-1})\|_1 + 2\epsilon\,. \end{aligned} \tag{88}$$

Therefore, $d < \lambda$ for any edge $e \in S^c$ if

$$U_{b1} := \max_{e\in S^c} 2M\|\boldsymbol{\mathcal{C}}(\hat{\boldsymbol{\Gamma}}_{e,S}\hat{\boldsymbol{\Gamma}}_{S,S}^{-1} - \boldsymbol{\Gamma}^*_{e,S}\boldsymbol{\Gamma}^{-*}_{S,S})\|_1 + 2\epsilon(1 + \|\boldsymbol{\mathcal{C}}(\hat{\boldsymbol{\Gamma}}_{e,S}\hat{\boldsymbol{\Gamma}}_{S,S}^{-1})\|_1) \le \alpha\lambda_n(1 - C_\alpha)\,, \quad (89)$$

$$U_{b2} := \max_{e\in S^c} \|\boldsymbol{\mathcal{C}}(\hat{\boldsymbol{\Gamma}}_{e,S}\hat{\boldsymbol{\Gamma}}_{S,S}^{-1})\|_1 \le 1 - (1 - C_\alpha)\alpha\,. \quad (90)$$

It remains to show that (89) and (90) are true under the assumptions of Lemma 4. Since

$$\hat{\boldsymbol{\Gamma}}_{e,S}\hat{\boldsymbol{\Gamma}}_{S,S}^{-1} - \boldsymbol{\Gamma}^*_{e,S}\boldsymbol{\Gamma}^{-*}_{S,S} = (\hat{\boldsymbol{\Gamma}}_{e,S} - \boldsymbol{\Gamma}^*_{e,S})\boldsymbol{\Gamma}^{-*}_{S,S} + \boldsymbol{\Gamma}^*_{e,S}(\hat{\boldsymbol{\Gamma}}_{S,S}^{-1} - \boldsymbol{\Gamma}^{-*}_{S,S}) + (\hat{\boldsymbol{\Gamma}}_{e,S} - \boldsymbol{\Gamma}^*_{e,S})(\hat{\boldsymbol{\Gamma}}_{S,S}^{-1} - \boldsymbol{\Gamma}^{-*}_{S,S})\,, \quad (91)$$

we have

$$\begin{aligned}\|\boldsymbol{\mathcal{C}}(\hat{\boldsymbol{\Gamma}}_{e,S}\hat{\boldsymbol{\Gamma}}_{S,S}^{-1} - \boldsymbol{\Gamma}^*_{e,S}\boldsymbol{\Gamma}^{-*}_{S,S})\|_\infty &\le \|\boldsymbol{\mathcal{C}}(\hat{\boldsymbol{\Gamma}}_{e,S} - \boldsymbol{\Gamma}^*_{e,S})\|_\infty\|\boldsymbol{\mathcal{C}}(\boldsymbol{\Gamma}^{-*}_{S,S})\|_{1,\infty} \\ &+ \|\boldsymbol{\mathcal{C}}(\boldsymbol{\Gamma}^*_{e,S})\|_\infty\|\boldsymbol{\mathcal{C}}(\hat{\boldsymbol{\Gamma}}_{S,S}^{-1} - \boldsymbol{\Gamma}^{-*}_{S,S})\|_{1,\infty} \\ &+ \|\boldsymbol{\mathcal{C}}(\hat{\boldsymbol{\Gamma}}_{e,S} - \boldsymbol{\Gamma}^*_{e,S})\|_\infty\|\boldsymbol{\mathcal{C}}(\hat{\boldsymbol{\Gamma}}_{S,S}^{-1} - \boldsymbol{\Gamma}^{-*}_{S,S})\|_{1,\infty}\,. \quad (92)\end{aligned}$$

With edge $e = \{i,k\} \in S^c$ and edge $f = \{j,\ell\} \in S$, consider

$$\begin{aligned}\hat{\boldsymbol{\Gamma}}_{e,f} - \boldsymbol{\Gamma}^*_{e,f} &= \hat{\boldsymbol{\Gamma}}_{ik,j\ell} - \boldsymbol{\Gamma}^*_{ik,j\ell} \\ &= \hat{\boldsymbol{\Sigma}}_y^{(ij)} \otimes \hat{\boldsymbol{\Sigma}}_x^{(k\ell)} - \boldsymbol{\Sigma}_y^{*(ij)} \otimes \boldsymbol{\Sigma}_x^{*(k\ell)} \\ &= \boldsymbol{\Delta}_y^{(ij)} \otimes \boldsymbol{\Delta}_x^{(k\ell)} + \boldsymbol{\Sigma}_y^{*(ij)} \otimes \boldsymbol{\Delta}_x^{(k\ell)} + \boldsymbol{\Delta}_y^{(ij)} \otimes \boldsymbol{\Sigma}_x^{*(k\ell)}. \quad (93)\end{aligned}$$

It then follows that

$$\begin{aligned}|\boldsymbol{\mathcal{C}}(\hat{\boldsymbol{\Gamma}}_{e,f} - \boldsymbol{\Gamma}^*_{e,f})| &\le \|\boldsymbol{\Delta}_y^{(ij)}\|_F\,\|\boldsymbol{\Delta}_x^{(k\ell)}\|_F \\ &+ \|\boldsymbol{\Sigma}_y^{*(ij)}\|_F\,\|\boldsymbol{\Delta}_x^{(k\ell)}\|_F + \|\boldsymbol{\Delta}_y^{(ij)}\|_F\,\|\boldsymbol{\Sigma}_x^{*(k\ell)}\|_F \\ &\le \epsilon_y\epsilon_x + M\epsilon_x + M\epsilon_y < \epsilon^2 + 2M\epsilon. \quad (94)\end{aligned}$$

Hence

$$\|\boldsymbol{\mathcal{C}}(\hat{\boldsymbol{\Gamma}}_{e,S} - \boldsymbol{\Gamma}^*_{e,S})\|_\infty < \epsilon^2 + 2M\epsilon\,, \quad (95)$$

$$\|\boldsymbol{\mathcal{C}}(\hat{\boldsymbol{\Gamma}}_{e,S} - \boldsymbol{\Gamma}^*_{e,S})\|_1 < s_n(\epsilon^2 + 2M\epsilon). \quad (96)$$

Since $\boldsymbol{\Gamma}^*_{e,f} = \boldsymbol{\Sigma}_y^{*(ij)} \otimes \boldsymbol{\Sigma}_x^{*(k\ell)}$, we have $|\boldsymbol{\Gamma}^*_{e,f}| \le M^2$ and

$$\|\boldsymbol{\mathcal{C}}(\boldsymbol{\Gamma}^*_{e,S})\|_\infty \le M^2, \quad \|\boldsymbol{\mathcal{C}}(\boldsymbol{\Gamma}^*_{e,S})\|_1 \le s_n M^2\,. \quad (97)$$

Alternatively, with $e = \{i,k\} \in S^c$ and $f = \{j,\ell\} \in S$,

$$\begin{aligned}\|\boldsymbol{\mathcal{C}}(\boldsymbol{\Gamma}^*_{e,S})\|_1 &= \sum_{f\in S} |\boldsymbol{\mathcal{C}}(\boldsymbol{\Sigma}_y^{*(ij)} \otimes \boldsymbol{\Sigma}_x^{*(k\ell)})| \\ &\le \sum_{\{j,\ell\}\in S} \|\boldsymbol{\Sigma}_y^{*(ij)}\|_F\,\|\boldsymbol{\Sigma}_x^{*(k\ell)}\|_F \\ &\le (\sum_{j=1}^p \|\boldsymbol{\Sigma}_y^{*(ij)}\|_F)(\sum_{\ell=1}^p \|\boldsymbol{\Sigma}_x^{*(k\ell)}\|_F) \\ &\le \|\boldsymbol{\mathcal{C}}(\boldsymbol{\Sigma}_y^*)\|_{1,\infty}\,\|\boldsymbol{\mathcal{C}}(\boldsymbol{\Sigma}_x^*)\|_{1,\infty} \le M_\Sigma^2\,. \quad (98)\end{aligned}$$

From (91)-(98) and Lemma 3, we have

$$\begin{aligned}&\|\boldsymbol{\mathcal{C}}(\hat{\boldsymbol{\Gamma}}_{e,S}\hat{\boldsymbol{\Gamma}}_{S,S}^{-1} - \boldsymbol{\Gamma}^*_{e,S}\boldsymbol{\Gamma}^{-*}_{S,S})\|_1 \\ &\le \|\boldsymbol{\mathcal{C}}(\hat{\boldsymbol{\Gamma}}_{e,S} - \boldsymbol{\Gamma}^*_{e,S})\|_1\|\boldsymbol{\mathcal{C}}(\boldsymbol{\Gamma}^{-*}_{S,S})\|_{1,\infty} \\ &\quad + \|\boldsymbol{\mathcal{C}}(\boldsymbol{\Gamma}^*_{e,S})\|_1\|\boldsymbol{\mathcal{C}}(\hat{\boldsymbol{\Gamma}}_{S,S}^{-1} - \boldsymbol{\Gamma}^{-*}_{S,S})\|_{1,\infty} \\ &\quad + \|\boldsymbol{\mathcal{C}}(\hat{\boldsymbol{\Gamma}}_{e,S} - \boldsymbol{\Gamma}^*_{e,S})\|_1\|\boldsymbol{\mathcal{C}}(\hat{\boldsymbol{\Gamma}}_{S,S}^{-1} - \boldsymbol{\Gamma}^{-*}_{S,S})\|_{1,\infty} \\ &\le s_n(\epsilon^2 + 2M\epsilon)\kappa_\Gamma + \big[\min\{s_nM^2, M_\Sigma^2\} + s_n(\epsilon^2 + 2M\epsilon)\big] \\ &\qquad \times \big[s_n(\epsilon^2 + 2M\epsilon)\kappa_\Gamma^2\big]\big(1 + 1.5s_n(\epsilon^2 + 2M\epsilon)\kappa_\Gamma\big) \\ &\overset{\epsilon<M}{\le} 3s_n\epsilon M\kappa_\Gamma B_s \le \alpha C_\alpha \min\{\lambda_n, 1\}\,, \quad (99)\end{aligned}$$

where B_s is as in (69) and we used $\epsilon < M$ to infer $\epsilon^2 + 2M\epsilon < 3M\epsilon$. Using the triangle inequality $|a| - |b| \le |a - b| \le |a| + |b|$, we have $\|\boldsymbol{\mathcal{C}}(\hat{\boldsymbol{\Gamma}}_{e,S}\hat{\boldsymbol{\Gamma}}_{S,S}^{-1} - \boldsymbol{\Gamma}^*_{e,S}\boldsymbol{\Gamma}^{-*}_{S,S})\|_1 \ge \|\boldsymbol{\mathcal{C}}(\hat{\boldsymbol{\Gamma}}_{e,S}\hat{\boldsymbol{\Gamma}}_{S,S}^{-1})\|_1 - \|\boldsymbol{\mathcal{C}}(\boldsymbol{\Gamma}^*_{e,S}\boldsymbol{\Gamma}^{-*}_{S,S})\|_1$, which, using (65) and (99), leads to

$$\begin{aligned}\|\boldsymbol{\mathcal{C}}(\hat{\boldsymbol{\Gamma}}_{e,S}\hat{\boldsymbol{\Gamma}}_{S,S}^{-1})\|_1 &\le \|\boldsymbol{\mathcal{C}}(\boldsymbol{\Gamma}^*_{e,S}\boldsymbol{\Gamma}^{-*}_{S,S})\|_1 + \alpha C_\alpha \min\{\lambda_n, 1\} \\ &\le 1 - \alpha + \alpha C_\alpha \min\{\lambda_n, 1\} \le 1 - (1 - C_\alpha)\alpha\,. \quad (100)\end{aligned}$$

This establishes (90). To show (89), using (99)-(100),

$$\begin{aligned}U_{b1} &\le 2M\alpha C_\alpha \min\{\lambda_n, 1\} + 2\epsilon(1 + 1 - (1 - C_\alpha)\alpha) \\ &\le 2M\alpha C_\alpha\lambda_n + 2\epsilon(2 - (1 - C_\alpha)\alpha) \overset{(68)}{=} \alpha\lambda_n(1 - C_\alpha)\,. \quad (101)\end{aligned}$$

This proves (89), and thus, part (i) of Lemma 4.

We now turn to the proof of Lemma 4(ii). Since $\hat{\boldsymbol{\Delta}} = \tilde{\boldsymbol{\Delta}}$, for any edge $\{k,\ell\} \in S$, we have

$$\begin{aligned}\|(\hat{\boldsymbol{\Delta}} - \boldsymbol{\Delta}^*)^{(k\ell)}\|_F &= \|(\tilde{\boldsymbol{\Delta}} - \boldsymbol{\Delta}^*)^{(k\ell)}\|_F \\ &= \|\mathrm{vec}(\tilde{\boldsymbol{\Delta}}^{(k\ell)}) - \mathrm{vec}((\boldsymbol{\Delta}^*)^{(k\ell)})\|_2\,. \quad (102)\end{aligned}$$

Using (33) and (73)

$$\begin{aligned}\mathrm{bvec}((\tilde{\boldsymbol{\Delta}} - \boldsymbol{\Delta}^*)_S) &= \hat{\boldsymbol{\Gamma}}_{S,S}^{-1}\mathrm{bvec}((\boldsymbol{\Delta}_\Sigma)_S) + (\hat{\boldsymbol{\Gamma}}_{S,S}^{-1} - \boldsymbol{\Gamma}^{-*}_{S,S}) \\ &\quad \times \mathrm{bvec}((\boldsymbol{\Sigma}_x^* - \boldsymbol{\Sigma}_y^*)_S) - \lambda_n\hat{\boldsymbol{\Gamma}}_{S,S}^{-1}\,\mathrm{bvec}(\boldsymbol{Z}(\tilde{\boldsymbol{\Delta}}_S))\,. \quad (103)\end{aligned}$$

Since $\hat{\boldsymbol{\Gamma}}_{S,S}^{-1} = \hat{\boldsymbol{\Gamma}}_{S,S}^{-1} - \boldsymbol{\Gamma}^{-*}_{S,S} + \boldsymbol{\Gamma}^{-*}_{S,S}$,

$$\|\boldsymbol{\mathcal{C}}(\hat{\boldsymbol{\Gamma}}_{S,S}^{-1})\|_{1,\infty} \le \|\boldsymbol{\mathcal{C}}(\hat{\boldsymbol{\Gamma}}_{S,S}^{-1} - \boldsymbol{\Gamma}^{-*}_{S,S})\|_{1,\infty} + \|\boldsymbol{\mathcal{C}}(\boldsymbol{\Gamma}^{-*}_{S,S})\|_{1,\infty}\,. \quad (104)$$

By (103), for any edge $f = \{k,\ell\} \in S$, we have

$$\begin{aligned}\|\mathrm{vec}((\tilde{\boldsymbol{\Delta}} - \boldsymbol{\Delta}^*)^{(k\ell)})\|_2 &\le \|(\hat{\boldsymbol{\Gamma}}_{S,S}^{-1} - \boldsymbol{\Gamma}^{-*}_{S,S})_{f,S} \\ &\quad \times \mathrm{bvec}\big((\boldsymbol{\Delta}_\Sigma)_S + (\boldsymbol{\Sigma}_x^* - \boldsymbol{\Sigma}_y^*)_S - \lambda_n\boldsymbol{Z}(\tilde{\boldsymbol{\Delta}}_S)\big)\|_2 \\ &\quad + \|(\boldsymbol{\Gamma}^{-*}_{S,S})_{f,S}\,\mathrm{bvec}\Big((\boldsymbol{\Delta}_\Sigma)_S - \lambda_n\boldsymbol{Z}(\tilde{\boldsymbol{\Delta}}_S)\Big)\|_2 \\ &\le \|\boldsymbol{\mathcal{C}}(\hat{\boldsymbol{\Gamma}}_{S,S}^{-1} - \boldsymbol{\Gamma}^{-*}_{S,S})\|_{1,\infty}\Big(\|\boldsymbol{\mathcal{C}}(\boldsymbol{\Delta}_\Sigma)\|_\infty + \|\boldsymbol{\mathcal{C}}(\boldsymbol{\Sigma}_x^* - \boldsymbol{\Sigma}_y^*)\|_\infty \\ &\quad + \lambda_n\Big) + \|\boldsymbol{\mathcal{C}}(\boldsymbol{\Gamma}^{-*}_{S,S})\|_{1,\infty}\big(\|\boldsymbol{\mathcal{C}}(\boldsymbol{\Delta}_\Sigma)\|_\infty + \lambda_n\big) \\ &\le s_n\kappa_\Gamma^2(\epsilon^2 + 2M\epsilon)\big(1 + 1.5s_n(\epsilon^2 + 2M\epsilon)\kappa_\Gamma\big) \\ &\quad \times (2\epsilon + 2M + \lambda_n) + \kappa_\Gamma(2\epsilon + \lambda_n) =: U_{b3}\,. \quad (105)\end{aligned}$$

By (66), for $0 < \alpha < 1$, we have $2\epsilon < \alpha\lambda_n/(2-\alpha) < \alpha\lambda_n < \lambda_n$. Therefore, $\kappa_\Gamma(2\epsilon + \lambda_n) < 2\kappa_\Gamma\lambda_n$ and $2\epsilon + 2M + \lambda_n < 2M + 2\lambda_n$. Since $\epsilon < M$ by (66), we also have $\epsilon^2 + 2M\epsilon < 3M\epsilon$. Using these relations and (105), it follows that

$$U_{b3} \le 3s_n\epsilon M\kappa_\Gamma^2\big(1 + 4.5s_n\epsilon M\kappa_\Gamma\big)\big(2M + 2\lambda_n\big) + 2\lambda_n\kappa_\Gamma\,.$$

Finally,

$$\|\boldsymbol{\mathcal{C}}(\hat{\boldsymbol{\Delta}} - \boldsymbol{\Delta}^*)\|_\infty = \max_{f=\{k,\ell\}\in S} \|\mathrm{vec}((\tilde{\boldsymbol{\Delta}} - \boldsymbol{\Delta}^*)^{(k\ell)})\|_2\,,$$

proving the desired result. □

Proof of Theorem 1. Here we first show that under the sufficient conditions of Theorem 1, the assumptions of Lemmas 2-4 holds true. We pick $\epsilon = C_0\sqrt{ln(p_n)/n}$, implying, by Lemma 2, that $\|\boldsymbol{\mathcal{C}}(\hat{\boldsymbol{\Sigma}}_x - \boldsymbol{\Sigma}_x^*)\|_\infty \le \epsilon$ and $\|\boldsymbol{\mathcal{C}}(\hat{\boldsymbol{\Sigma}}_y - \boldsymbol{\Sigma}_y^*)\|_\infty \le \epsilon$. with probability $\ge 1 - 2/p_n^{\tau-2}$, $\tau > 2$. We first show that with this choice of ϵ, condition (66) of Lemma 4 holds true. By the

choice of λ_n, we have $\lambda_n \geq 8\epsilon/\alpha$. Clearly, for $0 < \alpha < 1$, $(\alpha/8) < \alpha/(4(2-\alpha))$. Therefore

$$\epsilon \leq \frac{\alpha\lambda_n}{8} < \frac{\alpha\lambda_n}{4(2-\alpha)} < \frac{\alpha\lambda_n}{2(2-\alpha)}\,. \tag{106}$$

By a choice of n in (24), we have $n > C_0^2 \ln(p_n)/M^2$. Hence, $\epsilon = C_0\sqrt{ln(p_n)/n} < M$. Thus, (66) of Lemma 4 holds true. Next we show that condition (47) of Lemma 3 holds. By a choice of n in (24), we have $n > 81C_0^2 \ln(p_n) M^2 s_n^2 \kappa_\Gamma^2$. Therefore,

$$\kappa_\Gamma < \frac{1}{C_0}\sqrt{\frac{n}{\ln(p_n)}} \times \frac{1}{9s_nM} = \frac{1}{9s_nM\epsilon} < \frac{1}{3s_n(\epsilon^2 + 2M\epsilon)} \tag{107}$$

since $\epsilon < M$. This proves (47) of Lemma 3 holds.

Now we show that (67) of Lemma 4 holds. Since $\epsilon < \alpha\lambda_n/(4(2-\alpha))$, by (68), we have

$$\begin{aligned} C_\alpha &= \frac{\alpha\lambda_n + 2\epsilon\alpha - 4\epsilon}{2M\alpha\lambda_n + \alpha\lambda_n + 2\epsilon\alpha} > \frac{\alpha\lambda_n - 4\epsilon}{2M\alpha\lambda_n + \alpha\lambda_n + 2\epsilon\alpha} \\ &> \frac{\alpha\lambda_n - \alpha\lambda_n/(2-\alpha)}{\alpha\lambda_n(2M+1) + 2\epsilon\alpha} = \frac{1-\alpha}{(2-\alpha)\big[2M + 1 + \frac{2\epsilon\alpha}{\lambda_n\alpha}\big]} \\ &> \frac{1-\alpha}{(2-\alpha)(2M+1)+\alpha} = \bar{C}_\alpha \end{aligned} \tag{108}$$

where in the last inequality above, we used $\epsilon < \alpha\lambda_n/(2(2-\alpha))$ from (106). Consider the right-side $3s_n\epsilon M\kappa_\Gamma B_s$ of (67). From (107), $\kappa_\Gamma < 1/(9s_nM\epsilon)$. Therefore,

$$\begin{aligned} B_s <& 1 + \kappa_\Gamma\big(\min\{s_nM^2, M_\Sigma^2\} + \frac{1}{3\kappa_\Gamma}\big)\big(\frac{1}{2}+1\big) \\ =& 1.5 + 1.5\kappa_\Gamma \min\{s_nM^2, M_\Sigma^2\} = C_{M\kappa}\,. \end{aligned} \tag{109}$$

By (23), $\lambda_n \geq 4.5\epsilon(\alpha\bar{C}_\alpha)^{-1}s_nM\kappa_\Gamma(1 + \kappa_\Gamma\min\{s_nM^2, M_\Sigma^2\})$. Hence, using (108), we have

$$3s_n\epsilon M\kappa_\Gamma B_s < \alpha\bar{C}_\alpha\lambda_n < \alpha C_\alpha\lambda_n\,, \tag{110}$$

proving part of (67) for our choice of λ_n. To show that we also have $3s_n\epsilon M\kappa_\Gamma B_s < \alpha\bar{C}_\alpha$, consider the choice of n in (24) given by

$$n > C_0^2 \ln(p_n)\,\frac{9s_n^2}{(\alpha\bar{C}_\alpha)^2}(\kappa_\Gamma M C_{M\kappa})^2\,. \tag{111}$$

Then

$$\epsilon = C_0\sqrt{\frac{\ln(p_n)}{n}} < \frac{\alpha\bar{C}_\alpha}{3s_nM\kappa_\Gamma C_{M\kappa}}\,, \tag{112}$$

and from (110),

$$3s_n\epsilon M\kappa_\Gamma B_s < \alpha\bar{C}_\alpha < \alpha C_\alpha\,. \tag{113}$$

Thus, all assumptions of Lemma 4 hold true.

Therefore, Lemma 4(i) applies, proving Theorem 1(ii). By (107), $9s_n\epsilon M\kappa_\Gamma < 1$. Using this fact in Lemma 4(ii),

$$\begin{aligned} \|\boldsymbol{\mathcal{C}}(\hat{\boldsymbol{\Delta}} - \boldsymbol{\Delta}^*)\|_\infty &\leq 2\lambda_n\kappa_\Gamma + 9s_n\epsilon M\kappa_\Gamma^2\big(M + \lambda_n\big) \\ &\leq 3\lambda_n\kappa_\Gamma + 9s_n\epsilon M^2\kappa_\Gamma^2\,. \end{aligned} \tag{114}$$

Since, by (23), $3\lambda_n\kappa_\Gamma = C_0\sqrt{\ln(p_n)/n}\,C_{b1}$ and we picked $\epsilon = C_0\sqrt{\ln(p_n)/n}$, we have

$$\|\boldsymbol{\mathcal{C}}(\hat{\boldsymbol{\Delta}} - \boldsymbol{\Delta}^*)\|_\infty \leq (C_{b1} + C_{b2})C_0\sqrt{\ln(p_n)/n}\,,$$

proving Theorem 1(i). To prove part (iii), since $\hat{\boldsymbol{\Delta}}_{S^c} = \tilde{\boldsymbol{\Delta}}_{S^c} = \boldsymbol{\Delta}^*_{S^c} = \mathbf{0}$, we have

$$\begin{aligned} \|\boldsymbol{\mathcal{C}}(\hat{\boldsymbol{\Delta}} - \boldsymbol{\Delta}^*)\|_F &= \Big(\sum_{\{k,\ell\}\in S} \|\hat{\boldsymbol{\Delta}}^{(k\ell)} - (\boldsymbol{\Delta}^*)^{(k\ell)}\|_F^2\Big)^{1/2} \\ &\leq \|\boldsymbol{\mathcal{C}}(\hat{\boldsymbol{\Delta}} - \boldsymbol{\Delta}^*)\|_\infty\sqrt{s_n}\,. \end{aligned} \tag{115}$$

Finally, to establish part (iv), note that parts (i)-(iii) hold with probability $> 1 - 2/p_n^{\tau-2}$ (with high probability (w.h.p.)). Recall that $\mathcal{G}_\Delta = (V, \mathcal{E}_\Delta)$ denotes the MA differential graph with edgeset $\mathcal{E}_\Delta = \{\{k,\ell\} : \|\boldsymbol{\Delta}^{(k\ell)}\|_F > 0\}$. Let $\mathcal{G}_{\Delta^*}$ and $\mathcal{G}_{\hat{\Delta}}$ denoted true and estimated graphs based on $\boldsymbol{\Delta}^*$ and $\hat{\boldsymbol{\Delta}}$, respectively. If $\min_{(k,\ell)\in S}\|(\boldsymbol{\Delta}^*)^{(k\ell)}\|_F \geq 2\|\boldsymbol{\mathcal{C}}(\hat{\boldsymbol{\Delta}} - \boldsymbol{\Delta}^*)\|_\infty$, then $\boldsymbol{\mathcal{C}}(\hat{\boldsymbol{\Delta}} - \boldsymbol{\Delta}^*)\|_\infty = \boldsymbol{\mathcal{C}}((\hat{\boldsymbol{\Delta}} - \boldsymbol{\Delta}^*)_S)\|_\infty \leq (1/2)\min_{(k,\ell)\in S}\|(\boldsymbol{\Delta}^*)^{(k\ell)}\|_F$, therefore, $\min_{(k,\ell)\in S}\|(\hat{\boldsymbol{\Delta}}_S)^{(k\ell)}\|_F \geq (1/2)\min_{(k,\ell)\in S}\|(\boldsymbol{\Delta}^*)^{(k\ell)}\|_F > 0$, while $\hat{\boldsymbol{\Delta}}_{S^c} = \mathbf{0}$ w.h.p. □

Appendix B
Technical Lemmas and Proof of Theorem 2

In order to invoke [30], we first vectorize (3), using $\boldsymbol{\theta} = \text{bvec}(\boldsymbol{\Delta}) \in \mathbb{R}^{m^2p^2}$, as (cf. (30))

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2}\boldsymbol{\theta}^\top(\hat{\boldsymbol{\Sigma}}_y \boxtimes \hat{\boldsymbol{\Sigma}}_x)\boldsymbol{\theta} - \boldsymbol{\theta}^\top\text{bvec}(\hat{\boldsymbol{\Sigma}}_x - \hat{\boldsymbol{\Sigma}}_y) \tag{116}$$

where previous $L(\boldsymbol{\Delta}, \hat{\boldsymbol{\Sigma}}_x, \hat{\boldsymbol{\Sigma}}_y)$ is now $\mathcal{L}(\boldsymbol{\theta})$. To include sparse-group penalty, recall that the submatrix $\boldsymbol{\Delta}^{(k\ell)}$ of $\boldsymbol{\Delta}$ corresponds to the edge $\{k,\ell\}$ of the MA graph. We denote its vectorized version as $\boldsymbol{\theta}_{Gt} \in \mathbb{R}^{m^2}$ (subscript G for grouped variables [30]) with index $t = 1, 2, \cdots, p^2$. Then $\boldsymbol{\theta}_{Gt} = \text{vec}(\boldsymbol{\Delta}^{(k\ell)})$ where $t = (k-1)p + \ell$, $\ell = t \mod p$, and $k = \lfloor t/p \rfloor + 1$. Using this notation, the penalty $\lambda\sum_{k,\ell=1}^p \|\boldsymbol{\Delta}^{(k\ell)}\|_F = \lambda\sum_{t=1}^{p^2}\|\boldsymbol{\theta}_{Gt}\|_2$. In the notation of [30], the regularization penalty

$$\mathcal{R}(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_{\bar{\mathcal{G}},2} := \sum_{t=1}^{p^2}\|\boldsymbol{\theta}_{Gt}\|_2 \tag{117}$$

where the index set $\{1, 2, \cdots, (mp)^2\}$ is partitioned into a set of $N_G = p^2$ disjoint groups $\bar{\mathcal{G}} = \{G_1, G_2, \cdots, G_{p^2}\}$. As shown in [30, Sec. 2.2], $\mathcal{R}(\boldsymbol{\theta})$ is a norm. The counterpart to penalized $L_\lambda(\boldsymbol{\Delta}, \hat{\boldsymbol{\Sigma}}_x, \hat{\boldsymbol{\Sigma}}_y)$ of (4) is (we denote λ by λ_n, as in Appendix A)

$$\mathcal{L}_\lambda(\boldsymbol{\theta}) = \mathcal{L}(\boldsymbol{\theta}) + \lambda_n\mathcal{R}(\boldsymbol{\theta})\,. \tag{118}$$

As discussed in [30, Sec. 2.2], w.r.t. the usual Euclidean inner product $\langle \boldsymbol{u}, \boldsymbol{v}\rangle = \boldsymbol{u}^\top\boldsymbol{v}$ for $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^{m^2p^2}$ and given any subset $S_{\bar{\mathcal{G}}} \subseteq \{1, 2, \cdots, N_G\}$ of group indices, define the subspace

$$\mathcal{M} = \{\boldsymbol{\theta} \in \mathbb{R}^{m^2p^2} \,|\, \boldsymbol{\theta}_{Gt} = \mathbf{0} \text{ for all } t \notin S_{\bar{\mathcal{G}}}\} \tag{119}$$

and its orthogonal complement

$$\mathcal{M}^\perp = \{\boldsymbol{\theta} \in \mathbb{R}^{m^2p^2} \,|\, \boldsymbol{\theta}_{Gt} = \mathbf{0} \text{ for all } t \in S_{\bar{\mathcal{G}}}\}\,. \tag{120}$$

The chosen $\mathcal{R}(\boldsymbol{\theta})$ is decomposable w.r.t. $(\mathcal{M}, \mathcal{M}^\perp)$ since $\mathcal{R}(\boldsymbol{\theta}^{(1)} + \boldsymbol{\theta}^{(2)}) = \mathcal{R}(\boldsymbol{\theta}^{(1)}) + \mathcal{R}(\boldsymbol{\theta}^{(2)})$ for any $\boldsymbol{\theta}^{(1)} \in \mathcal{M}$ and $\boldsymbol{\theta}^{(2)} \in \mathcal{M}^\perp$ [30, Sec. 2.2, Example 2].

In order to invoke [30], we need the dual norm $\mathcal{R}^*$ of regularizer $\mathcal{R}$ w.r.t. the inner product $\langle \boldsymbol{u}, \boldsymbol{v} \rangle = \boldsymbol{u}^\top \boldsymbol{v}$. It is given by [30, Eqns. (14)-(15)]

$$\mathcal{R}^*(\boldsymbol{v}) = \sup_{\mathcal{R}(\boldsymbol{u}) \leq 1} \langle \boldsymbol{u}, \boldsymbol{v} \rangle = \max_{t=1,2,\cdots p^2} \|\boldsymbol{u}_{Gt}\|_2 \,. \tag{121}$$

We also need the subspace compatibility index [30], defined as

$$\Psi(\mathcal{M}) = \sup_{\boldsymbol{u} \in \mathcal{M} \setminus \{0\}} \frac{\mathcal{R}(\boldsymbol{u})}{\|\boldsymbol{u}\|_2} \,. \tag{122}$$

For group lasso penalty, $\Psi(\mathcal{M}) = \sqrt{s_n}$ [30, Sec. 9.2 (Supplementary)], where $s_n = |S_{\bar{\mathcal{G}}}|$ = number of edges in the true MA differential graph. We need to establish a restricted strong convexity condition [30] on $\mathcal{L}(\boldsymbol{\theta})$. With $\boldsymbol{\theta}^* = \text{bvec}(\boldsymbol{\Delta}^*)$ denoting the true value, and $\boldsymbol{\theta} = \boldsymbol{\theta}^* + \tilde{\boldsymbol{\theta}}$, define

$$\delta\mathcal{L}(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) = \mathcal{L}(\boldsymbol{\theta}^* + \tilde{\boldsymbol{\theta}}) - \mathcal{L}(\boldsymbol{\theta}^*) - \langle \nabla\mathcal{L}(\boldsymbol{\theta}^*), \tilde{\boldsymbol{\theta}} \rangle \tag{123}$$

where the gradient $\nabla\mathcal{L}(\boldsymbol{\theta}^*)$ at $\boldsymbol{\theta} = \boldsymbol{\theta}^*$ is

$$\nabla\mathcal{L}(\boldsymbol{\theta}^*) = (\hat{\boldsymbol{\Sigma}}_y \boxtimes \hat{\boldsymbol{\Sigma}}_x)\boldsymbol{\theta}^* - \text{bvec}(\hat{\boldsymbol{\Sigma}}_x - \hat{\boldsymbol{\Sigma}}_y) \,. \tag{124}$$

Hence (123) simplifies to

$$\delta\mathcal{L}(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) = \tilde{\boldsymbol{\theta}}^\top (\hat{\boldsymbol{\Sigma}}_y \boxtimes \hat{\boldsymbol{\Sigma}}_x)\tilde{\boldsymbol{\theta}} = \tilde{\boldsymbol{\theta}}^\top \hat{\boldsymbol{\Gamma}} \tilde{\boldsymbol{\theta}} \,, \tag{125}$$

which may be rewritten as

$$\delta\mathcal{L}(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) = \tilde{\boldsymbol{\theta}}^\top \boldsymbol{\Gamma}^* \tilde{\boldsymbol{\theta}} + \tilde{\boldsymbol{\theta}}^\top (\hat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}^*)\tilde{\boldsymbol{\theta}} \,. \tag{126}$$

By the sparsity assumption, $\boldsymbol{\theta}^* = \boldsymbol{\theta}^*_{\mathcal{M}}$, hence, $\boldsymbol{\theta}^*_{\mathcal{M}^\perp} = \mathbf{0}$, where $\boldsymbol{\theta}_{\mathcal{M}}$ and $\boldsymbol{\theta}_{\mathcal{M}^\perp}$ denote projection of $\boldsymbol{\theta}$ on subspaces $\mathcal{M}$ and $\mathcal{M}^\perp$, respectively.

Similar to (5), suppose

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \left\{ \mathcal{L}(\boldsymbol{\theta}) + \lambda_n \mathcal{R}(\boldsymbol{\theta}) \right\}, \tag{127}$$

and we consider (123) and (125) with $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}^* + \tilde{\boldsymbol{\theta}}$. Then

$$\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^* = \hat{\boldsymbol{\theta}}_{\mathcal{M}} - \boldsymbol{\theta}^* + \hat{\boldsymbol{\theta}}_{\mathcal{M}^\perp} = \tilde{\boldsymbol{\theta}}_{\mathcal{M}} + \tilde{\boldsymbol{\theta}}_{\mathcal{M}^\perp} \,. \tag{128}$$

By [30, Lemma 1],

$$\mathcal{R}(\tilde{\boldsymbol{\theta}}_{\mathcal{M}^\perp}) \leq 3\mathcal{R}(\tilde{\boldsymbol{\theta}}_{\mathcal{M}}) + 4\mathcal{R}(\boldsymbol{\theta}^*_{\mathcal{M}^\perp}) \,, \tag{129}$$

if

$$\lambda_n \geq 2\mathcal{R}^*(\nabla\mathcal{L}(\boldsymbol{\theta}^*)) \,. \tag{130}$$

Since in our case $\boldsymbol{\theta}^*_{\mathcal{M}^\perp} = \mathbf{0}$, we have $\mathcal{R}(\boldsymbol{\theta}^*_{\mathcal{M}^\perp}) = 0$.

Lemma 5. Under (15) and using the notation of Appendix A,

$$\mathcal{R}^*(\nabla\mathcal{L}(\boldsymbol{\theta}^*)) \leq (\epsilon^2 + 2M\epsilon) s_n \max_{t=1,\cdots,p^2} \|\boldsymbol{\theta}^*_{Gt}\|_2 + 2\epsilon \quad \bullet$$

Proof. Using (44), (72) and (124), we have

$$\nabla\mathcal{L}(\boldsymbol{\theta}^*) = \boldsymbol{\Delta}_\Gamma \boldsymbol{\theta}^* + \text{bvec}(\boldsymbol{\Delta}_y) - \text{bvec}(\boldsymbol{\Delta}_x) \,. \tag{131}$$

Expressing it group-wise, with groups t and t_1 corresponding to edges $\{j,k\}$ and $\{\ell,q\}$, respectively,

$$(\nabla\mathcal{L}(\boldsymbol{\theta}^*))_{Gt_1} = \sum_{t=1}^{p^2} (\boldsymbol{\Delta}_\Gamma)_{Gt_1,Gt} \boldsymbol{\theta}^*_{Gt} + \text{bvec}(\boldsymbol{\Delta}_y)_{Gt_1} - \text{bvec}(\boldsymbol{\Delta}_x)_{Gt_1} \,. \tag{132}$$

Therefore, by the Cauchy-Schwartz inequality, and using (45), (46) and (57), we have

$$\begin{aligned} \|(\nabla\mathcal{L}(\boldsymbol{\theta}^*))_{Gt_1}\|_2 &\leq \sum_{t=1}^{p^2} \|(\boldsymbol{\Delta}_\Gamma)_{Gt_1,Gt}\|_F \, \|\boldsymbol{\theta}^*_{Gt}\|_2 \\ &\quad + \|\text{bvec}(\boldsymbol{\Delta}_y)_{Gt_1}\|_2 + \|\text{bvec}(\boldsymbol{\Delta}_x)_{Gt_1}\|_2 \\ &\leq \|\boldsymbol{\mathcal{C}}(\boldsymbol{\Delta}_\Gamma)\|_\infty \sum_{t=1}^{p^2} \|\boldsymbol{\theta}^*_{Gt}\|_2 + \|\boldsymbol{\Delta}_y^{(\ell q)}\|_\infty + \|\boldsymbol{\Delta}_y^{(\ell q)}\|_\infty \\ &\leq (\epsilon^2 + 2M\epsilon) s_n \max_{t=1,\cdots,p^2} \|\boldsymbol{\theta}^*_{Gt}\|_2 + \epsilon + \epsilon \,. \end{aligned} \tag{133}$$

By (121) and (133) we have the desired result. □

Lemma 6. Under (15) and the notation of Appendix A,

$$\delta\mathcal{L}(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) \geq \kappa_{\mathcal{L}} \|\tilde{\boldsymbol{\theta}}\|_2^2 \tag{134}$$

where $\kappa_{\mathcal{L}} = \frac{1}{2}\phi^*_{min} - 8s_n(\epsilon^2 + 2M\epsilon)$. •

Proof. We have

$$\tilde{\boldsymbol{\theta}}^\top (\hat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}^*)\tilde{\boldsymbol{\theta}} = \sum_{t_1=1}^{p^2} \sum_{t_2=1}^{p^2} \tilde{\boldsymbol{\theta}}^\top_{Gt_1} (\boldsymbol{\Delta}_\Gamma)_{Gt_1,Gt_2} \tilde{\boldsymbol{\theta}}_{Gt_2} \,. \tag{135}$$

Therefore,

$$\begin{aligned} |\tilde{\boldsymbol{\theta}}^\top (\hat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}^*)\tilde{\boldsymbol{\theta}}| &\leq \sum_{t_1=1}^{p^2} \sum_{t_2=1}^{p^2} |\tilde{\boldsymbol{\theta}}^\top_{Gt_1} (\boldsymbol{\Delta}_\Gamma)_{Gt_1,Gt_2} \tilde{\boldsymbol{\theta}}_{Gt_2}| \\ &\leq \sum_{t_1=1}^{p^2} \sum_{t_2=1}^{p^2} \|\tilde{\boldsymbol{\theta}}_{Gt_1}\|_2 \, \|(\boldsymbol{\Delta}_\Gamma)_{Gt_1,Gt_2}\|_F \, \|\tilde{\boldsymbol{\theta}}_{Gt_2}\|_2 \\ &\leq \|\boldsymbol{\mathcal{C}}(\boldsymbol{\Delta}_\Gamma)\|_\infty \sum_{t_1=1}^{p^2} \sum_{t_2=1}^{p^2} \|\tilde{\boldsymbol{\theta}}_{Gt_1}\|_2 \, \|\tilde{\boldsymbol{\theta}}_{Gt_2}\|_2 \\ &\leq (\epsilon^2 + 2M\epsilon) \, \|\tilde{\boldsymbol{\theta}}\|^2_{\bar{\mathcal{G}},2} \,, \end{aligned} \tag{136}$$

where we used (117). We have

$$\begin{aligned} \|\tilde{\boldsymbol{\theta}}\|^2_{\bar{\mathcal{G}},2} &= \|\tilde{\boldsymbol{\theta}}_{\mathcal{M}} + \tilde{\boldsymbol{\theta}}_{\mathcal{M}^\perp}\|^2_{\bar{\mathcal{G}},2} = (\|\tilde{\boldsymbol{\theta}}_{\mathcal{M}}\|_{\bar{\mathcal{G}},2} + \|\tilde{\boldsymbol{\theta}}_{\mathcal{M}^\perp}\|_{\bar{\mathcal{G}},2})^2 \\ &\overset{(129)}{\leq} 16 \, \|\tilde{\boldsymbol{\theta}}_{\mathcal{M}}\|^2_{\bar{\mathcal{G}},2} \overset{(122)}{\leq} 16 s_n \|\tilde{\boldsymbol{\theta}}_{\mathcal{M}}\|_2^2 \leq 16 s_n \|\tilde{\boldsymbol{\theta}}\|_2^2 \,. \end{aligned} \tag{137}$$

Noting that $\tilde{\boldsymbol{\theta}}^\top \boldsymbol{\Gamma}^*)\tilde{\boldsymbol{\theta}} \geq \phi^*_{min} \|\tilde{\boldsymbol{\theta}}\|_2^2$ and using (126), (136) and (137), we have

$$\delta\mathcal{L}(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) \geq \big(\frac{1}{2}\phi^*_{min} - 8s_n(\epsilon^2 + 2M\epsilon)\big) \|\tilde{\boldsymbol{\theta}}\|_2^2 = \kappa_{\mathcal{L}} \|\tilde{\boldsymbol{\theta}}\|_2^2 \,,$$

proving the desired result. □

Proof of Theorem 2. First choose ϵ to make $\kappa_{\mathcal{L}} > 0$ in Lemma 6. For instance, suppose we take $8s_n(\epsilon^2 + 2M\epsilon) \leq \phi^*_{min}/4$. Then $\kappa_{\mathcal{L}} > \phi^*_{min}/4$. Now pick

$$\epsilon = C_0 \sqrt{ln(p_n)/n} \leq \min\left\{M, \frac{\phi^*_{min}}{96 s_n M}\right\}, \tag{138}$$

leading to $8s_n(\epsilon^2 + 2M\epsilon) \leq 24 s_n M\epsilon \leq \phi^*_{min}/4$. These upper bounds can be ensured by picking appropriate lower bounds to sample size n and invoking Lemma 2. The choice of n specified in (26) satisfies (138) with probability $> 1 - 2/p_n^{\tau-2}$. Using $\epsilon = C_0\sqrt{ln(p_n)/n} \leq M$, the lower bound on λ_n given

in (25) satisfies (130) with $\mathcal{R}^*(\nabla\mathcal{L}(\boldsymbol{\theta}^*))$ as in Lemma 5. By [30, Theorem 1], $\hat{\boldsymbol{\theta}}$ given by (127) satisfies

$$\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2 \leq \frac{3\lambda_n}{\kappa_{\mathcal{L}}}\Psi(\mathcal{M})\,. \tag{139}$$

The left-side of (139) equals $\|\hat{\boldsymbol{\Delta}} - \boldsymbol{\Delta}^*\|_F$ while the right-side of (139) equals right-side of (27) using $\Psi(\mathcal{M}) = \sqrt{s_n}$, $\kappa_{\mathcal{L}} > \phi^*_{min}/4$, and noting that $\max_{t=1,\cdots,p^2} \|\boldsymbol{\theta}^*_{Gt}\|_2 = \max_{\{k,\ell\}\in V\times V} \|(\boldsymbol{\Delta}^*)^{(k\ell)}\|_F$. This proves Theorem 2. □

REFERENCES

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

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

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

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

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

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

[7] Y. Wu, T. Li, X. Liu and L.l Chen, "Differential network inference via the fused D-trace loss with cross variables," *Electronic J. Statistics*, vol. 14, pp. 1269-1301, 2020.

[8] H. Yuan, R. Xi, C. Chen and M. Deng, "Differential network analysis via lasso penalized D-trace loss," *Biometrika*, vol. 104, pp. 755-770, 2017.

[9] Z. Tang, Z. Yu and C. Wang, "A fast iteraive algorithm for high-dimensional differential network," *Computational Statistics*, vol. 35, pp. 95-109, 2020.

[10] B. Zhao, Y.S. Wang and M. Kolar, "Direct estimation of differential functional graphical models," in *Proc. 33rd Conf. Neural Information Processing Systems (NeurIPS 2019)*, Vancouver, Canada, 2019.

[11] B. Zhao, Y.S. Wang and M. Kolar, "FuDGE: A method to estimate a functional differential graph in a high-dimensional setting," *J. Machine Learning Research*, vol. 23, pp. 1-82, 2022.

[12] S.D. Zhao, T.T. Cai and H. Li, "Direct estimation of differential networks," *Biometrika*, vol. 101, pp. 253-268, June 2014.

[13] E. Belilovsky, G. Varoquaux and M.B. Blaschko, "Hypothesis testing for differences in Gaussian graphical models: Applications to brain connectivity," *Advances in Neural Information Processing Systems (NIPS 2016)*, vol. 29, Dec. 2016.

[14] J. Chiquet, G. Rigaill and M. Sundquist, "A multiattribute Gaussian graphical model for inferring multiscale regulatory networks: an application in breast cancer." In: Sanguinetti G., Huynh-Thu V. (eds), *Gene Regulatory Networks. Methods in Molecular Biology*, vol 1883. Humana Press, New York, NY, 2019.

[15] M. Kolar, H. Liu and E.P. Xing, "Markov network estimation from multi-attribute data," in *Proc. 30th Intern. Conf. Machine Learning (ICML)*, Atlanta, GA, 2013.

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

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

[18] G. Marjanovic and V. Solo, "Vector l_0 sparse conditional independence graphs," in *Proc. IEEE ICASSP 2018*, pp. 2731-2735, 2018.

[19] Z. Yue, P. Sundaram and V. Solo, "Fast block-sparse estimation for vector networks," in *Proc. IEEE ICASSP 2020*, pp. 5505-5509, 2020.

[20] P. Sundaram, M. Luessi, M. Bianciardi, S. Stufflebeam, M. Hämäläinen and V. Solo, "Individual resting-state brain networks enabled by massive multivariate conditional mutual information," *IEEE Trans. Med. Imaging*, vol. 39, pp. 1957-1966, 2020.

[21] Z. Yue and V. Solo, "Comparing vector networks via frequency domain persistent homology," in *Proc. IEEE CDC*, pp. 126-131, Dec. 2021.

[22] K. Tsai, O. Koyejo, and M. Kolar, "Joint Gaussian graphical model estimation: A survey," *WIREs Computational Statistics*, vol. 14, no. 6, pp. e1582, Nov/Dec. 2022.

[23] B. Zhang, H. Li, R.B. Riggins, M. Zhan, J. Xuan, Z. Zhang, E.P. Hoffman, R. Clarke and Y. Wang, "Differential dependency network analysis to identify condition-specific topological changes in biological networks," *Bioinformatics*, vol. 25, no. 4, pp. 526-532, 2009.

[24] P. Xu and Q. Gu, "Semiparametric differential graph models," in *Proc. 30th Conference on Neural Information Processing Systems (NIPS 2016)*,Barcelona, Spain, 2016.

[25] S. Liu, J.A. Quinn, M.U. Gutmann, T. Suzuki and M. Sugiyama, "Direct learning of sparse changes in Markov networks by density ratio estimation," *Neural Computation*, 26(6):1169-1197, 2014.

[26] S. Liu, T. Suzuki, R. Relator, J. Sese, M. Sugiyama, and K. Fukumizu, "Support consistency of direct sparse-change learning in Markov networks" *Annals Statistics*, vol. 45, no. 3, pp. 959-990, 2017.

[27] S. Liu, K. Fukumizu and T. Suzuki, "Learning sparse structural changes in high-dimensional Markov network: A review on methodologies and theories," *Behaviormetrika*, vol. 44, pp. 265-286, 2017.

[28] B. Jiang, X. Wang and C. Leng, "A direct approach for sparse quadratic discriminant analysis," *J. Machine Learning Research*, vol. 19, pp. 1-37, 2018.

[29] T. Zhang and H. Zou, "Sparse precision matrix estimation via lasso penalized D-trace loss," *Biometrika*, vol. 101, pp. 103-120, 2014.

[30] S.N. Negahban, P. Ravikumar, M.J. Wainwright and B. Yu, "A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers," *Statistical Science*, vol. 27, No. 4, pp. 538-557, 2012.

[31] P. Ravikumar, M.J. Wainwright, G. Raskutti and B. Yu, "High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence," *Electronic J. Statistics*, vol. 5, pp. 935-980, 2011.

[32] J.K. Tugnait, "Estimation of high-dimensional differential graphs from multi-attribute data," in *Proc. 2023 IEEE Intern. Conf. Acoustics, Speech & Signal Processing (ICASSP 2023)*, pp. 1-5, Rhodes Island, Greece, June 4-9, 2023.

[33] D.S. Tracy and K.G. Jinadasa, "Partitioned Kronecker products of matrices and applications," *Canadian J. Statistics*, vol. 17, pp. 107-120, March 1989.

[34] S. Liu, "Matrix results on Khatri-Rao and Tracy-Singh products," *Linear Algebra & Its Applications*, vol. 289, pp. 267-277, 1999.

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

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

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

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

[39] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM Journal on Imaging Sciences*, vol. 2, no. 1, pp. 183-202, 2009.

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

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

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

[43] A-L. Barabási and R. Albert, Réka, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509-512, Oct. 1999.

[44] S. Lu, J. Kang, W. Gong and D. Towsley, "Complex network comparison using random walks," in *WWW '14 Companion: Proc. 23rd Intern. Conf. World Wide Web*, pp. 727-730, Seoul, Korea, April 2014.

[45] S. Zhang, B. Guo, A. Dong, J. He, Z. Xu and S.X. Chen, "Cautionary tales on air-quality improvement in Beijing," *Proc. Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 473, p. 20170457, 2017.

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