Multiple Matrix Gaussian Graphs Estimation

Yunzhang Zhu[†] and Lexin Li[‡]

[†]Department of Statistics, Ohio State University

[‡]Division of Biostatistics, University of California at Berkeley

Abstract

Matrix-valued data, where the sampling unit is a matrix consisting of rows and columns of measurements, are emerging in numerous scientific and business applications. Matrix Gaussian graphical model is a useful tool to characterize the conditional dependence structure of rows and columns. In this article, we employ nonconvex penalization to tackle the estimation of multiple graphs from matrix-valued data under a matrix normal distribution. We propose a highly efficient nonconvex optimization algorithm that can scale up for graphs with hundreds of nodes. We establish the asymptotic properties of the estimator, which requires less stringent conditions and has a sharper probability error bound than existing results. We demonstrate the efficacy of our proposed method through both simulations and real functional magnetic resonance imaging analyses.

Key Words: Conditional independence; Gaussian graphical model; Matrix normal distribution; Nonconvex penalization; Resting-state functional magnetic resonance imaging; Sparsistency.

[1] Address for correspondence: Lexin Li, Division of Biostatistics, University of California at Berkeley, Berkeley, CA, 94720 U.S.A. Email: lexinli@berkeley.edu.

1 Introduction

Gaussian graphical model has been widely used to describe the conditional dependence relationship, which is encoded in a partial correlation matrix, among a set of interacting variables. There have been a large number of statistical methods proposed to estimate a sparse Gaussian graphical model (Meinshausen and Bühlmann, 2006; Yuan and Lin, 2007; Friedman et al., 2008; Ravikumar et al., 2011; Cai et al., 2011, among others). There are also extensions from estimation of a single graph to multiple graphs across groups (Guo et al., 2011; Danaher et al., 2014; Zhu et al., 2014; Lee and Liu, 2015; Cai et al., 2016). All those methods assume the vector of interacting variables follow a normal distribution. In recent years, matrix-valued data, where each sampling unit is a matrix consisting of rows and columns of measurements, are rapidly emerging. Accordingly, the matrix normal distribution is becoming increasingly popular in modeling such matrix-valued observations (Zhou, 2014). Under this distribution, there have been some recent development of sparse graphical model estimation that aims to characterize the dependence of rows and columns of matrix data (Yin and Li, 2012; Leng and Tang, 2012; Tsiligkaridis et al., 2013). In this article, we aim at estimation of multiple graphs for matrix data under a matrix normal distribution.

Our motivation is brain connectivity analysis based on resting-state functional magnetic resonance imaging (fMRI). Meanwhile, our proposal is equally applicable to many other network data analyses. Brain functional connectivity reveals the synchronization of brain systems through correlations in neurophysiological measures of brain activity. When measured during resting-state, it maps the intrinsic functional architecture of the brain (Varoquaux and Craddock, 2013). Brain connectivity analysis is now at the foreground of neuroscience research. Accumulated evidences have suggested that connectivity network alters with the presence of numerous neurological disorders, and such alternations hold useful insights of disease pathologies (Fox and Greicius, 2010). In a typical functional connectivity study, the fMRI data are collected for multiple subjects from the disease group and normal control. For each individual subject, the observed fMRI data takes the form of a region by time *matrix*, where the number of brain regions is usually in the order of 10^2 and the number of time points around 150 to 200. From this matrix, a region by region correlation matrix is estimated to describe the brain connectivity graph, one for each diagnostic group *separately*. In this graph, nodes represent brain regions, and links measure dependence between the brain regions, where partial correlation is a commonly used correlation measure (Fornito et al., 2013). Brain connectivity analysis is then turned into the problem of estimation of partial correlation matrices under Gaussian graphical models across multiple groups.

Our proposal integrates matrix normal distribution, multiple partial correlation matrices estimation, and nonconvex penalization. Such an integration distinguishes our proposal from the existing solutions. For the matrix-valued data, directly applying the existing graphical model estimation methods assuming a vector normal distribution, in effect, requires the columns of the matrix data to be independent, which is obviously not true. For instance, for fMRI, the columns correspond to time series of repeatedly measured brain activities and are highly correlated. Whitening may help reduce the between-column correlation. In Section 5, we compare and show that our method substantially outperforms two state-of-the-art vector normal based multi-graph estimation methods, Lee and Liu (2015) and Cai et al. (2016), both facilitated by whitening. Among the few solutions on graphical model estimation under a matrix normal distribution (Yin and Li, 2012; Leng and Tang, 2012; Zhou, 2014), none tackle estimation of multiple graphs across different populations, but instead only focus on a single graph. Our proposal is also different from two recent multi-graph estimation methods of Qiu et al. (2016) and Han et al. (2016), in both study goals and estimation approaches. Specifically, Qiu et al. (2016) aimed to estimate a graph at any given location, e.g., age, whereas Han et al. (2016) aimed to capture and summarize the commonality underlying a collection of individual graphs. By contrast, our goal is to simultaneously estimate multiple graphs, one from each of a given group of subjects. Besides, Qiu et al. (2016) proposed a two-step procedure, which first obtained a smoothed estimate of the sample covariance matrix through kernel smoothing, then plugged into the constrained ℓ_1 minimization method of Cai et al. (2011) for precision matrix estimation. Han et al. (2016) first obtained an estimate of all the individual graphs, using again Cai et al. (2011), then plugged into an objective function that minimizes the Hamming distance between the targeting median graph and the individual graphs. For our proposal, we employ a likelihood based loss function, plus a combination of a nonconvex sparsity penalty and a nonconvex group sparsity penalty to induce both sparsity and similarity across multiple partial correlation matrices. Our choice of loss function is to ensure both theoretical properties and positive-definiteness of the estimator. Meanwhile, our choice of penalty function is motivated by the belief that, the true graph is approximately sparse, and the difference of graphs across multiple groups is approximately sparse too. In other words, those graphs may exhibit different connectivity patterns, but are also encouraged to be similar to each other. Moreover, nonconvex penalization in high-dimensinoal models has often been shown to outperform its convex counterpart both in theory and in practice (Fan and Li, 2001a; Zhang, 2010; Shen et al., 2012). In the context of graphical model under a vector normal distribution, nonconvex penalization has been shown to deliver more precise and concise graph estimates (Fan et al., 2009; Shen et al., 2012).

The novelty of our proposal lies in both the computational and the theoretical contributions. Computationally, recognizing that nonconvex optimization is more challenging than convex optimization, we propose a highly efficient and scalable algorithm through a combination of two modern optimization techniques, the minorize-maximization algorithm (MM, Hunter and Lange, 2004), and the alternating direction method of multipliers (ADMM, Boyd et al., 2011). The proposed algorithm is fast, yielding a comparable computation time as its convex counterpart. It is also much faster than the competing methods of Lee and Liu (2015) and Cai et al. (2016). In addition, our method scales reasonably well and can work for graphs with the number of nodes ranging to hundreds. It is noteworthy that this range covers the typical size of a brain connectivity network in neuroimaging analysis.

Theoretically, we study the asymptotic properties of the proposed optimization problem and establish sharp theoretical results. We focus on two scenarios: imposing only the sparsity penalty, and imposing only the group sparsity penalty. Such an investigation would shed new insights on the connection and difference of the two types of penalties, and also facilitate a direct comparison with existing theoretical results.

Specifically, the first scenario corresponds to performing sparse graph estimation across multiple groups *separately*. In the context of single graph estimation under the vector normal distribution, theoretical analysis of correct identification of sparse structure has been investigated in Ravikumar et al. (2011); Fan et al. (2014); Loh and Wainwright (2014). Compared to Ravikumar et al. (2011), who employed a convex ℓ_1 penalty and thus required the irrepresentable condition, our sparsistency result does not require this rather stringent condition due to the use of the nonconvex penalty. Compared to Fan et al. (2014), we obtain a sharper probability error bound and an improved minimum signal strength condition. Moreover, we do not require a consistent initial estimator as Fan et al. (2014) did. Compared to Loh and Wainwright (2014), our result is directly comparable. But we develop a new proof technique that can easily generalize to multiple graphs. In the context of single graph estimation under the matrix normal distribution, Leng and Tang (2012) provided the estimation and sparseness pursuit guarantee; however, their results were established for some unknown local minimizer of their optimization function. By contrast, we obtain the theoretical properties for the actual local optimum computed by the optimization algorithm. In addition, Zhou (2014) studied estimation error, whereas we focus on the sparsity pattern reconstruction of the graphical dependency.

The second scenario corresponds to estimation of multiple graphs *jointly*. Both Danaher et al. (2014) and Zhu et al. (2014) studied multiple graph estimation with fusion type penalties. However, Danaher et al. (2014) did not provide any theoretical result on graph recovery, whereas Zhu et al. (2014) obtained sparsistency for the global, but not local, solution of their optimization function. Both Lee and Liu (2015) and Cai et al. (2016) provided the theoretical guarantee for multi-graph structure recovery, but none provided the positive-definiteness guarantee for the resulting estimator.

The rest of the article is organized as follows. Section 2 presents the model and the penalized objective function. Section 3 develops the optimization algorithm. Section 4 studies the asymptotic properties. Section 5 presents the simulations, and Section 6 the fMRI data analyses. Section 7 concludes the paper with a discussion. All technical proofs are relegated to an online Supplementary Appendix.

2 Model

2.1 Penalized optimization

Suppose the observed data, $\boldsymbol{X}_{ki}, i = 1, \ldots, n_k, k = 1, \ldots, K$, are from K heterogeneous populations, with n_k number of observations from the kth group. Each observation $\boldsymbol{X}_{ki}$ is a $p \times q$ matrix, with p denoting the spatial dimension and q the temporal dimension. We assume $\boldsymbol{X}_{ki}$ follows a matrix normal distribution, i.e.,

$$\boldsymbol{X}_{k1}, \ldots, \boldsymbol{X}_{kn_k} \overset{\text{i.i.d.}}{\sim} N(\boldsymbol{M}_k, \boldsymbol{\Sigma}_{kS} \otimes \boldsymbol{\Sigma}_{kT}), \quad k = 1, \cdots, K,$$

where $\boldsymbol{M}_k = \mathbb{E}[\boldsymbol{X}_{ki}]$, $\boldsymbol{\Sigma}_{kS} \in \mathbb{R}^{p\times p}$ and $\boldsymbol{\Sigma}_{kT} \in \mathbb{R}^{q\times q}$ denote the spatial and temporal covariance matrices, respectively, and $\otimes$ is the Kronecker product. This assumption of matrix normal distribution has been frequently adopted in numerous applications involving matrix-valued observations (Yin and Li, 2012; Leng and Tang, 2012). It is also scientifically plausible in the context of neuroimaging analysis. For instance, the standard neuroimaging processing software, such as SPM (Friston et al., 2007) and FSL (Smith et al., 2004), adopt a framework that assumes the data are normally distributed per voxel (location) with a noise factor and an autoregressive structure, which shares a similar spirit as the matrix normal formulation. We further discuss potential relaxation of this assumption in Section 7.

Our primary object of interest is the spatial partial correlation matrix,

$$\boldsymbol{\Omega}_k = \text{Diag}(\boldsymbol{\Sigma}_{kS})^{-1/2}\boldsymbol{\Sigma}_{kS}^{-1}\,\text{Diag}(\boldsymbol{\Sigma}_{kS})^{-1/2}, \quad k = 1, \cdots, K.$$

Under the normal distribution, a zero partial correlation coefficient implies the conditional independence of two nodes given all others in the graph. By contrast, the mean term $\boldsymbol{M}_k$ and the temporal correlation matrix are to be treated as nuisance parameters. This is mainly driven by our motivating application of brain connectivity analysis, where the primary interest is to estimate the connectivity pattern of spatial regions of the brain. Nevertheless we note that our proposed methodology is applicable to estimation of the temporal partial correlation matrix as well.

Under the matrix normal distribution, a natural solution seeks to minimize over

$(\boldsymbol{\Omega}_1, \ldots, \boldsymbol{\Omega}_K)$ the negative log likelihood function, aside from a constant,

$$\sum_{k=1}^{K} n_k \left\{ \mathrm{trace}(\boldsymbol{\Omega}_k \widehat{\boldsymbol{\Gamma}}_k) - \log\det(\boldsymbol{\Omega}_k) \right\}, \tag{1}$$

where $\widehat{\boldsymbol{\Gamma}}_k$ is a sample correlation matrix $\boldsymbol{\Gamma}_k$; for instance,

$$\widehat{\boldsymbol{\Gamma}}_k = \mathrm{DiagScale}\left\{ \sum_{i=1}^{n_k} (\boldsymbol{X}_{ki} - \bar{\boldsymbol{X}}_k)(\boldsymbol{X}_{ki} - \bar{\boldsymbol{X}}_k)^T \right\}, \quad k = 1, \cdots, K,$$

where $\mathrm{DiagScale}(\boldsymbol{C}) = \mathrm{Diag}(\boldsymbol{C})^{-1/2}\boldsymbol{C}\mathrm{Diag}(\boldsymbol{C})^{-1/2}$ for any square matrix $\boldsymbol{C}$. That is, we plug into (1) a set of consistent correlation estimators. The estimator $\widehat{\boldsymbol{\Gamma}}_k$ was studied by Zhou (2014), and its rate of convergence has been established in the high-dimensional regime, which would facilitate our subsequent asymptotic investigation. Directly solving (1), however, may encounter some challenges. First, the number of unknown parameters in $\{\boldsymbol{\Omega}_k\}_{k=1}^K$ may far exceed the sample size, causing inversion of $\widehat{\boldsymbol{\Gamma}}_k$ problematic. Second, we are generally interested in finding pairs of nodes that are conditionally independent given the others. However, minimizing (1) would not yield any exact zero estimates in $\{\boldsymbol{\Omega}_k\}_{k=1}^K$, rendering the interpretation difficult. Third, it is often desirable to encourage the estimated graphs to be similar across groups, under the belief that the differences of graphical structure would usually concentrate on some local areas of the nodes. For instance, in brain connectivity analysis, the brain region connections are usually sparse (Zhang et al., 2015), and the differences of brain connections across different populations usually localize in some subnetworks of the brain (Toussaint et al., 2014).

To address those challenges, we propose to estimate the K partial correlation matrices $\{\boldsymbol{\Omega}_k\}_{k=1}^K$ by solving the following penalized optimization,

$$\begin{aligned} \underset{\lambda_{\max}(\boldsymbol{\Omega}_k) \le R;\, k=1,\ldots,K}{\text{minimize}} \quad & \sum_{k=1}^{K} n_k \left\{ \mathrm{trace}(\boldsymbol{\Omega}_k \widehat{\boldsymbol{\Gamma}}_k) - \log\det(\boldsymbol{\Omega}_k) \right\} + \\ & \sum_{k=1}^{K} n_k \sum_{i \ne j} p_{\lambda_{1k}}\left(|\omega_{kij}|\right) + n_{\min} \sum_{i \ne j} p_{\lambda_2}\left(\sqrt{\omega_{1ij}^2 + \cdots \omega_{Kij}^2} \right) \end{aligned} \tag{2}$$

where $\lambda_{\max}(\boldsymbol{\Omega}_k)$ denotes the largest eigenvalue of $\boldsymbol{\Omega}_k$, $n_{\min} = \min_{1 \le k \le K} n_k$, $a, R > 0$, $\lambda_{1k}; k = 1, \cdots, K$, and λ_2 are the tuning parameters, and the penalty function $p_\lambda(\cdot) : \mathbb{R}^+ \to \mathbb{R}^+$ satisfies the following conditions:

(i) $p_\lambda(x)$ is nondecreasing and differentiable on $\mathbb{R}^+$ and $p_\lambda(0) = 0$;

(ii) $\lim_{x\to 0^+} p'_\lambda(x) = \lambda$;

(iii) $p_\lambda(x) + x^2/b$ is convex for some constant $b > 0$;

(iv) $p'_\lambda(x) = 0$ for $|x| > a\lambda$ for some constant $a \geq b/2$.

A few remarks are in order. First, the condition $a \geq b/2$ ensures the existence of $p_\lambda(x)$, and different choices of a, b correspond to different nonconvex penalties. For instance, $a > 2, b = 2/(a-1)$ leads to the penalty function of Fan and Li (2001a), and $a = b/2, b > 0$ to that of Zhang (2010). Other types of nonconvex penalty can also be used here, e.g., the truncated ℓ_1 penalty (Shen et al., 2012), or the ℓ_q penalty with $q < 1$. Second, our penalty function consists of two parts, a sparsity penalty that encourages sparsity in each individual partial correlation matrix, and a group sparsity penalty that encourages common sparsity patterns across different partial correlation matrices. Third, our penalty function is in general nonconvex, and using a nonconvex penalty is beneficial in several ways. It leads to nearly unbiased parameter estimation, is to facilitate cross-validation for parameter tuning, and can achieve a better sparsity pursuit guarantee under less stringent assumptions (Fan et al., 2009; Shen et al., 2012).

2.2 Parameter tuning

Parameter tuning is always challenging for high-dimensional models, and we propose the following cross-validation approach to tune the parameters in (2). Motivated by our theoretical analysis in Section 4, we let $\lambda_{11} = \lambda_1\sqrt{\frac{\log(p\vee q)}{n_1 q}}, \cdots, \lambda_{1K} = \lambda_1\sqrt{\frac{\log(p\vee q)}{n_K q}}$, where $p \vee q = \max(p, q)$, and let $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)^T$. We select $\boldsymbol{\lambda}$ by minimizing a prediction criterion using 5-fold cross-validation. That is, we divide the data set for each group into five parts $\mathcal{D}_1, \cdots, \mathcal{D}_5$. Under group k, define $\widehat{\boldsymbol{\Gamma}}_k^l$ and $\widehat{\boldsymbol{\Gamma}}_k^{-l}$ to be the sample correlation matrices calculated based on samples in $\mathcal{D}_l$ and $\{\mathcal{D}_1, \cdots, \mathcal{D}_5\} \setminus \mathcal{D}_l$, $l = 1, \cdots, 5$, respectively. Similarly, define $\widehat{\boldsymbol{\Omega}}_k^{-l}(\boldsymbol{\lambda})$ to be the partial correlation matrix calculated based on $\widehat{\boldsymbol{\Gamma}}_k^{-l}$, $l = 1, \cdots, 5$, under the tuning parameter $\boldsymbol{\lambda}$. Then we define the criterion as,

$$\mathrm{CV}(\boldsymbol{\lambda}) = \frac{1}{5K}\sum_{l=1}^{5}\sum_{k=1}^{K}\left[-\log\det\left\{\widehat{\boldsymbol{\Omega}}_k^{-l}(\boldsymbol{\lambda})\right\} + \mathrm{trace}\left\{\widehat{\boldsymbol{\Gamma}}_k^l\widehat{\boldsymbol{\Omega}}_k^{-l}(\boldsymbol{\lambda})\right\} - p\right].$$

The optimal tuning parameter for each data partition is selected as $\boldsymbol{\lambda}^\star = \arg\min_{\boldsymbol{\lambda}} \mathrm{CV}(\boldsymbol{\lambda})$, which is then used to obtain the final cross-validated estimator $(\widehat{\boldsymbol{\Omega}}_1, \cdots, \widehat{\boldsymbol{\Omega}}_K)$. Minimization of $\mathrm{CV}(\boldsymbol{\lambda})$ is carried out using a simple grid search over the domain of the tuning parameters. Following both the common practice in nonconvex penalization and our own theoretical analysis, we choose not to tune a and b in $p_\lambda(\cdot)$, but instead set $b = 2a$ and a equal to some constant divided by λ_1. We choose not to tune R either, since our method is not sensitive to the value of R as long as it is reasonably large. We also make some remarks comparing the cross-validation based tuning under a convex versus a nonconvex penalty. When comparing the goodness-of-fit of two selected models, it is essentially comparing the likelihood function evaluated at the constrained maximum likelihood estimator (MLE), i.e., the MLE over the selected support of the parameters. Since a convex penalty such as ℓ_1 does not yield a constrained MLE; rather, it shrinks the MLE to achieve an optimal bias-variance trade-off, the convex penalized estimator's cross-validation score is not suitable for model comparison. By contrast, a nonconvex penalized estimator is nearly identical to the constrained MLE given the selected support (Fan and Li, 2001a; Zhang, 2010; Shen et al., 2012). As such, a nonconvex penalty is better suited to cross-validation tuning for sparsity identification. In graphical model estimation with a convex penalty, cross-validation and the more traditional Bayesian information criterion have been shown to perform poorly (Liu et al., 2010). We further compare the two penalty functions numerically in Section 5.

3 Computation

Nonconvex optimization is in general more challenging than convex optimization. In this section, we develop a highly efficient and scalable optimization algorithm for nonconvex minimization of (2). The algorithm consists of two core components: the minorize-maximization algorithm that optimizes (2) through a sequence of convex relaxations (Hunter and Lange, 2004), and the alternating direction method of multipliers that optimizes each convex relaxation (Boyd et al., 2011). We first summarize our optimization procedure in Algorithm 1, then discuss each individual component in detail. We conclude this section with a discussion regarding the overall computational cost.

```
Initialize solution $\left(\widehat{\mathbf{\Omega}}_1^{(0)}, \cdots, \widehat{\mathbf{\Omega}}_K^{(0)}\right) = \left(\mathrm{Diag}(\widehat{\mathbf{\Gamma}}_1)^{-1}, \cdots, \mathrm{Diag}(\widehat{\mathbf{\Gamma}}_K)^{-1}\right)$.
Initialize weights $b_{kll'}^{(0)} = \lambda_1 \mathbb{I}(l \neq l'), c_{ll'}^{(0)} = \lambda_2 \mathbb{I}(l \neq l'), 1 \leq l, l' \leq p, k = 1, \cdots, K$.
Initialize MM iteration counter $t = 0$.                                    ▷ MM updates
repeat
    Initialize $\mathbf{\Theta}_k^{(0)} = \mathbf{0}, \widetilde{\mathbf{\Omega}}_k^{(0)} = \widehat{\mathbf{\Omega}}_k^{(t)}$, $\mathbf{\Delta}_k^{(0)} = \widehat{\mathbf{\Omega}}_k^{(t)}$, and ADMM iteration counter
     $m = 0$.
    repeat                                                                  ▷ ADMM updates
        Decompose $\rho\left(\mathbf{\Delta}_k^{(m)} - \mathbf{\Theta}_k^{(m)}\right) - \frac{n_k}{n_{\min}} \widehat{\mathbf{\Gamma}}_k = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^\top$.
        Define the diagonal matrix $\boldsymbol{Q}$ with
         $Q_{ii} = \arg\min_{0<x\leq R} x D_{ii} - \log(x) + \frac{cx^2}{2}; i = 1, \ldots, p$.
        Compute $\mathbf{\Omega}_k^{(m+1)} = \boldsymbol{U}\boldsymbol{Q}\boldsymbol{U}^T$.                        ▷ ADMM primal update
        For $1 \leq l, l' \leq p, k = 1, \cdots, K$, let
            $s_k = S_{b_{kll'}^{(t)}/\rho}\left(\tilde{\omega}_{kll'}^{(m+1)} + \theta_{kll'}^{(m)}\right)$,
            $\left(\delta_{1ll'}^{(m+1)}, \cdots, \delta_{Kll'}^{(m+1)}\right) = \left(1 - \frac{c_{ll'}^{(t)}}{\rho\|\boldsymbol{s}\|_2}\right)_+ (s_1, \cdots, s_K)$.
                                                                            ▷ ADMM primal update
        Set $\mathbf{\Theta}_k^{(m+1)} = \mathbf{\Theta}_k^{(m)} + \widetilde{\mathbf{\Omega}}_k^{(m+1)} - \mathbf{\Delta}_k^{(m+1)}$.                ▷ ADMM dual update
        Set $m = m + 1$.
    until ADMM stopping criteria is satisfied.
    Set $\mathbf{\Omega}_k^{(t+1)} = \mathbf{\Delta}_k^{(m+1)}, k = 1, \cdots, K$.
    Set $b_{kll'}^{(t+1)} = \frac{n_k}{n_{\min}} p'_{\lambda_1}(|\omega_{kll'}^{(t)}|), 1 \leq l, l' \leq p, k = 1, \cdots, K$.    ▷ Weights updating
    Set $c_{ll'}^{(t+1)} = p'_{\lambda_2}\left(\sqrt{\sum_{k=1}^K \left(\omega_{kll'}^{(t)}\right)^2}\right), 1 \leq l, l' \leq p$.       ▷ Weights updating
    Set $t = t + 1$.
until $\boldsymbol{B}_k^{(t+1)} = \boldsymbol{B}_k^{(t)}, k = 1, \cdots, K$ and $\boldsymbol{C}^{(t+1)} = \boldsymbol{C}^{(t)}$.
```

Algorithm 1: The MM algorithm and ADMM algorithm for solving (2).

3.1 Sequential convex relaxation through MM algorithm

The MM algorithm is commonly employed for solving nonconvex optimization approximately. Its key idea is to decompose the objective function into difference of two convex functions. In our setting, we linearize the nonconvex penalty based on the previous iterate $x^{(t)}$, i.e.,

$$p_\lambda(|x|) = p_\lambda\left(|x^{(t)}|\right) + p'_\lambda\left(|x^{(t)}|\right)(|x| - |x^{(t)}|),$$

to obtain a convex approximation at $x^{(t)}$. Accordingly, we solve the nonconvex optimization (2) by considering a sequence of convex relaxations until we get a stationary point. Specifically, based on $\left(\widehat{\mathbf{\Omega}}_1^{(t)}, \cdots, \widehat{\mathbf{\Omega}}_K^{(t)}\right)$ at step t, we minimize the following convex relaxation,

$$\sum_{k=1}^{K} \frac{n_k}{n_{\min}} \left\{ \operatorname{trace}(\mathbf{\Omega}_k \widehat{\mathbf{\Gamma}}_k) - \log\det(\mathbf{\Omega}_k) \right\} + \sum_{k=1}^{K} \sum_{l<l'} b_{kll'}^{(t)} |\omega_{kll'}| + \sum_{i<i'} c_{ll'}^{(t)} \sqrt{\sum_{k=1}^{K} \omega_{kll'}^2}, \qquad (3)$$

subject to $\lambda_{\max}(\mathbf{\Omega}_k) \leq R; k = 1, \ldots, K$, where

$$b_{kll'}^{(t)} = \frac{n_k}{n_{\min}} p'_{\lambda_1}(|\omega_{kll'}^{(t)}|), \quad c_{ll'}^{(t)} = p'_{\lambda_2} \left(\sqrt{\sum_{k=1}^{K} \left(\omega_{kll'}^{(t)}\right)^2} \right).$$

We then obtain the solution $\left(\widehat{\mathbf{\Omega}}_1^{(t+1)}, \cdots, \widehat{\mathbf{\Omega}}_K^{(t+1)}\right)$ at the $(t+1)$th step, and iterate over t until convergence.

3.2 Alternating direction method of multipliers

To solve each relaxation (3), we propose an ADMM algorithm. Specifically, we introduce K new variables $\mathbf{\Delta}_k = (\delta_{kll'})_{1 \leq l, l' \leq p}$, such that $\mathbf{\Delta}_k = \mathbf{\Omega}_k$, and K associated dual variables $\mathbf{\Theta}_k = (\theta_{kll'})_{1 \leq l, l' \leq p}$, $k = 1, \cdots, K$. The ADMM algorithm solves (3) through iteratively applying the following updating scheme, for $k = 1, \ldots, K$, and $1 \leq l < l' \leq p$,

$$\begin{aligned}
\mathbf{\Omega}_k^{(m+1)} &= \underset{\lambda_{\max}(\mathbf{\Omega}) \leq R}{\arg\min} \left\{ \frac{n_k}{n_{\min}} \left(\operatorname{trace}(\mathbf{\Omega}\widehat{\mathbf{\Gamma}}_k) - \log\det\mathbf{\Omega} \right) + \frac{\rho}{2} \left\| \mathbf{\Omega} - \mathbf{\Delta}_k^{(m)} + \mathbf{\Theta}_k^{(m)} \right\|_2^2 \right\}, && (4a)\\
\left(\delta_{kll'}^{(m+1)}\right)_{k=1}^{K} &= \underset{\boldsymbol{\delta} \in \mathbb{R}^K}{\arg\min} \left\{ \frac{\rho}{2} \sum_{k=1}^{K} \left(\delta_k - \omega_{kll'}^{(m+1)} - \theta_{kll'}^{(m)} \right)^2 + \sum_{k=1}^{K} b_{kll'}^{(t)} |\delta_k| + c_{ll'}^{(t)} \|\boldsymbol{\delta}\|_2 \right\}, && (4b)\\
\mathbf{\Theta}_k^{(m+1)} &= \mathbf{\Theta}_k^{(m)} + \mathbf{\Omega}_k^{(m+1)} - \mathbf{\Delta}_k^{(m+1)},
\end{aligned}$$

The first update (4a) can be carried out efficiently according to the next lemma.

Lemma 1. *Consider the following optimization problem,*

$$\underset{\mathbf{\Omega} \succeq 0,\, \lambda_{\max}(\mathbf{\Omega}) \leq R}{\text{minimize}} \operatorname{trace}(\mathbf{\Omega}\mathbf{\Delta}) - \log\det\mathbf{\Omega} + \frac{c}{2}\|\mathbf{\Omega}\|_F^2$$

Let $\mathbf{\Delta} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^\top$ be the eigen-decomposition of $\mathbf{\Delta}$. The solution to the above problem is given by

$$\mathbf{\Omega}^\star = \boldsymbol{U}\boldsymbol{Q}\boldsymbol{U}^\top,$$

where $\boldsymbol{Q}$ is a diagonal matrix with diagonal elements

$$Q_{ii} = \operatorname*{arg\,min}_{0<x\leq R} xD_{ii} - \log(x) + \frac{cx^2}{2}; \;\; i = 1, \dots, p.$$

The second update (4b) has an analytical solution according to the next lemma.

Lemma 2. *Consider the following generic minimization problem,*

$$\operatorname*{minimize}_{\boldsymbol{x}\in\mathbb{R}^K} \frac{1}{2}\sum_{k=1}^{K}(x_k - a_k)^2 + \sum_{k=1}^{K} b_k|x_k| + \nu\sqrt{\sum_{k=1}^{K} x_k^2}.$$

Its solution is given by

$$\boldsymbol{x}^{\star} = \left\{1 - \nu\left[\sum_{k=1}^{K}\left(S_{b_k}(a_k)\right)^2\right]^{-1/2}\right\}_{+} \left\{S_{b_1}(a_1), \cdots, S_{b_K}(a_K)\right\},$$

where $S_b(a) = \operatorname{Sign}(a)(|a| - b)_+$ is the soft-thresholding function.

The proofs of Lemma 1 and 2 are given in the Appendix.

3.3 Overall computational cost

We make a few remarks regarding the overall computation of our algorithm. First, the per-iteration computational complexity for carrying out the ADMM step in Algorithm 1 is $O(Kp^3)$. Such a cubic dependence on p is essentially inevitable if one is to obtain a positive-definite matrix estimate. If positive-definiteness is not required, there are some alternative loss functions such as the pseudo-likelihood loss, and faster algorithms are possible. We have chosen the likelihood loss partly because of the positive-definiteness requirement, and partly because it is more amenable to the theoretical analysis. Second, although nonconvex optimization is in general more challenging, our nonconvex algorithm achieves a comparable computation time as its convex counterpart, as we report in Section 5. This is due to the fast convergence of the step that tackles nonconvexity, i.e., the MM step of convex relaxations. Our numerical study shows that the MM step usually converges in only a few iterations. Consequently, the main computational cost of the algorithm is dominated by the convex optimization step of ADMM. Third, our optimization algorithm scales reasonably well, and can handle networks with the number of nodes up to a few hundreds. It is noteworthy that, in functional connectivity analysis, the typical size of a region-based brain network is in tens to a few hundreds. As such, our method is well suited for brain connectivity type applications. Finally, we comment that some of the steps in our algorithm can be parallelized to further speed up the computation.

4 Asymptotics

Our asymptotic analysis focuses on two scenarios. We first study the case when there is only the sparsity penalty, i.e., when $\lambda_2 = 0$. We then study the case when there is only the group sparsity penalty, i.e., when $\lambda_1 = 0$. Considering these cases provides new insights to the connection and difference of the two types of penalty functions. Meanwhile, it allows a direct comparison with existing theoretical results in Gaussian graphical models. We also note that we did not pursue the scenario where both λ_1 and λ_2 are non-zero, for two reasons. Although it is undoubtedly of interest to study the theoretical properties when both penalties are present, such a characterization would naturally require an explicit quantification of similarity between the true graphs. This kind of knowledge is almost surely unknown in reality, making the asymptotic result less relevant practically. Moreover, there are lack of tools to overcome some technical difficulties in analyzing the KKT conditions when both sparse and group sparse penalties are employed. There is no existing work of this type even for the vector normal case. We defer this pursuit as potential future research.

That being said, we also clarify on our theoretical contributions. For the separate graph estimation scenario with $\lambda_2 = 0$, we provide a new sparsistency result that achieves a sharper error bound, requires less stringent conditions, and holds for the actual local optimum of the estimation algorithm. For the multi-graph joint estimation scenario with $\lambda_1 = 0$, we establish the sparsistency for the actual local instead of global minimizer, and guarantee both multi-graph structure recovery, symmetry and positive-definiteness. Moreover, we develop a new proof technique that permits a direct generalization from the single graph case to the multi-graph case. This proof technique is new in the literature, and is potentially useful for theoretical analysis of other models as well.

4.1 Sparsity penalty only with $\lambda_2 = 0$

First we consider the case where we impose the sparsity penalty only and set $\lambda_2 = 0$ in (2). Let $A_k^0 = \{(i,j) : \omega_{kij}^0 \neq 0\}$ denote the support of the true partial correlation matrix $\mathbf{\Omega}_k^0 = (\omega_{kij}^0)$, $i, j = 1, \ldots, p, k = 1, \cdots, K$. We define the oracle estimator $\widehat{\mathbf{\Omega}}_{k,A_k^0}$ as

$$\widehat{\mathbf{\Omega}}_{k,A_k^0} \quad = \quad \underset{(i,j):\omega_{kij}^0=0,(i,j)\notin A_k^0}{\arg\min} \left\{ \operatorname{trace}\big(\widehat{\mathbf{\Gamma}}_k \mathbf{\Omega}\big) - \log\det\big(\mathbf{\Omega}\big) \right\}, \tag{5}$$

which is essentially the MLE over $\{A_k^0\}_{k=1}^K$. Moreover, let $n_{\min} = \min_{1\le k\le K} n_k$ and $n_{\max} = \max_{1\le k\le K} n_k$. We impose the following assumptions.

(A1) Let $\mathbf{\Gamma}_k^0 = \operatorname{Diag}(\mathbf{\Sigma}_{kS})^{-1/2}\mathbf{\Sigma}_{kS}\operatorname{Diag}(\mathbf{\Sigma}_{kS})^{-1/2}$ denote the true correlation matrix. Assume that, for all $k = 1, \cdots, K$,

$$c_0^{-1} < \lambda_{min}(\mathbf{\Gamma}_k^0) \le \lambda_{max}(\mathbf{\Gamma}_k^0) < c_0 \text{ and } c_0^{-1} < \lambda_{min}(\mathbf{\Sigma}_{kT}^0) \le \lambda_{max}(\mathbf{\Sigma}_{kT}^0) < c_0,$$

holds for some positive real number c_0.

(A2) Let $c_1 = \max_k \|\mathbf{\Gamma}_k^0\|_{\infty,\infty}, c_2 = \max_k \|\boldsymbol{I}_k\|_{\infty,\infty}$, where $\boldsymbol{I}_k = \frac{1}{2}[\mathbf{\Omega}_k^0]^{-1} \otimes_s [\mathbf{\Omega}_k^0]^{-1}$ is the Fisher information matrix in group k, and $\|\boldsymbol{A}\|_{\infty,\infty} = \max_{1\le j\le p}\sum_{k=1}^n |A_{jk}|$ is the ℓ_∞/ℓ_∞-operator norm of matrix $\boldsymbol{A}$. Let $s_0 = \max_{1\le k\le K}\max_{1\le j\le p}\sum_{i=1}^p \mathbb{I}((i,j) \in A_k^0)$, where $\mathbb{I}$ is the indicator function. Assume that

$$2c_1c_2c_3\left(1 + 2c_1^2c_2\right)s_0\sqrt{\frac{\log(p \vee q)}{n_k q}} \le 1,\ k = 1, \cdots, K\,,$$

where c_3 is some absolute constant.

Assumption (A1) is a commonly imposed condition when analyzing the theoretical properties of many types of precision matrix estimators; see, for example, Fan et al. (2009); Cai et al. (2016). Assumption (A2) restricts the scaling of the graph sparsity level measured by s_0 as a function of sample size n and graph size p. Similar scaling has been used in Fan et al. (2009); Loh and Wainwright (2014). It is also noteworthy that the quantities c_0, c_1, c_2, and s_0 can grow with the sample size, the spatial dimension, and the temporal dimension. Under these assumptions, we have the following result.

Theorem 1. *Under Assumptions (A1) and (A2), and the condition that,*

$$\min_{(i,j)\in A_k^0} |\omega_{kij}^0| > \left\{2c_2c_3 + (1+c_1^2c_2)c_3\left(c_0 + 2c_1^{-1}\right)^2\right\}\sqrt{\frac{\log(p\vee q)}{n_kq}}, \tag{6}$$

for $k = 1, \cdots, K$, there exist λ_1 and a such that the oracle estimator $\widehat{\mathbf{\Omega}}_{k,A_k^0}$ is the unique minimizer of problem (2) *when $R = \sqrt{2a}$, $b = 2a$, and $\lambda_2 = 0$, with probability at least $1 - \frac{6K}{(p\vee q)^2}$, as $n, p \to \infty$.*

This theorem shows that the oracle estimator is the unique minimizer of (2) under $\lambda_2 = 0$. That is, when the maximum node degree s_0 does not grow too fast as (n, p) goes to infinity, for some choice of the tuning parameters, solving (2) could reconstruct the true structure of the K graphs with probability tending to one. This result holds when the minimum signal satisfies the condition (6). If we further assume $c_i, i = 1, 2, 3$, are all constants, then the minimum signal condition (6) roughly requires that

$$\min_{(i,j)\in A_k^0} |\omega_{kij}^0| \geq O\left(\sqrt{\frac{\log(p\vee q)}{n_kq}}\right), \; k = 1, \cdots, K. \tag{7}$$

Comparing (7) to the minimum signal strength condition in Fan et al. (2014), their condition is suboptimal in terms of dependence on column/row sparsity s_0, in that it requires $\min_{(i,j)\in A^u} |\omega_{ij}^0| > O\left(s_0^2\sqrt{\frac{\log p}{n}}\right)$. By contrast, we only require $\min_{(i,j)\in A^u} |\omega_{ij}^0| > O\left(\sqrt{\frac{\log p}{n}}\right)$. Our result is comparable to that of Loh and Wainwright (2014). However, their proof used a primal-dual witness technique, whereas our proof proceeds in two steps, by first establishing the rate of convergence for the oracle estimator, and then proving that the oracle estimator is the unique local minimum. An advantage of our two-step proof is that it is straightforward to generalize to the multiple partial correlation matrices case when a group sparsity penalty is further imposed. Finally, unlike Leng and Tang (2012) that established the oracle property for some unknown local minimizer of their objective function, we obtain the result for our actual local minimizer.

4.2 Group sparsity penalty only with $\lambda_1 = 0$

Next we consider the case where we impose the group sparsity penalty only and set $\lambda_1 = 0$ in (2). For this case, it is impossible to recover the oracle estimator unless

$A_1^0 = \cdots = A_K^0$, since the graph estimators obtained by using only the group sparsity penalty would be identical across all groups. On the other hand, it is still feasible to recover the oracle estimator over $A^u = \cup_{k=1}^K A_k^0$. Specifically, we define the oracle estimator $\widehat{\mathbf{\Omega}}_{1,A^u}, \cdots, \widehat{\mathbf{\Omega}}_{K,A^u}$ as

$$\widehat{\mathbf{\Omega}}_{k,A^u} \;=\; \underset{(i,j):\omega^0_{kij}=0,(i,j)\notin A^u}{\arg\min} \left\{ \operatorname{trace}(\widehat{\mathbf{\Gamma}}_k \mathbf{\Omega}) - \log\det(\mathbf{\Omega}) \right\},$$

which is essentially the MLE over the joint set A^u. We also modify the assumption (A2) slightly and introduce the next assumption.

(A3) Let $\tilde{s}_0 = \max_{1\le j\le p} \sum_{i=1}^p \mathbb{I}((i,j)\in A^u)$. Assume that

$$2c_1c_2c_3\left(1+2c_1^2c_2\right)\tilde{s}_0\sqrt{\frac{\log(p\vee q)}{n_k q}} \le 1,\; k=1,\cdots,K.$$

Assumption (A3) is directly comparable to (A2). In (A3), $\tilde{s}_0$ is the sparsity level of the joint of all K graphs, whereas s_0 in (A2) is the maximum sparsity level of all graphs. Easily $s_0 \le \tilde{s}_0$; and when the sparisity pattern differs significantly across different groups, $\tilde{s}_0$ can be much larger than s_0. In this sense, the group sparsity penalty is most effective when the sparsity patterns are similar across different groups. Under (A1) and (A3), we have the following result.

Theorem 2. *Under Assumptions (A1) and (A3), and the condition that*

$$\begin{aligned}\min_{(i,j)\in A^u}\sqrt{\sum_{k=1}^K\left(\omega^0_{kij}\right)^2} &> 2c_2c_3\sqrt{\frac{K\log(p\vee q)}{n_{\min}q}}\\ &+(1+c_1^2c_2)c_3\left(c_0+2c_1^{-1}\right)^2\sqrt{\frac{n_{\max}}{n_{\min}}}\sqrt{\frac{K\log(p\vee q)}{n_{\min}q}},\end{aligned} \tag{8}$$

for $k=1,\ldots,K$, there exist λ_2 and a such that the oracle estimator $\widehat{\mathbf{\Omega}}_{k,A^u}, k=1,\cdots,K$ is the unique minimizer of (2) *when $R=\sqrt{2a}$, $b=2a$, and $\lambda_1=0$, with probability at least $1-\frac{6K}{(p\vee q)^2}$, as $n,p\to\infty$.*

This theorem says that, if the size of the union of supports A_k^0 is not too large, the oracle estimator is the unique local optimum of (2) under $\lambda_1 = 0$, and can recover the true graph structure with probability tending to one. Again, if we treat $c_i, i = 1, 2, 3$, as constants, then the condition (8) becomes

$$\min_{(i,j)\in A^u} \sqrt{\sum_{k=1}^{K} \left(\omega_{kij}^0\right)^2} > O\left(\sqrt{\frac{n_{\max}}{n_{\min}}}\sqrt{\frac{K\log(p\vee q)}{n_{\min}q}}\right). \tag{9}$$

Comparing the two minimum signal strength conditions (7) and (9) reveals some useful insights about the two penalties. It is noted that neither condition is stronger nor weaker than the other. When the sample sizes $n_1, \cdots, n_K$ are well balanced, and the sparsity patterns are similar across all groups, adding a sparsity group penalty is to facilitate the graph recovery. This can be seen by inspecting the extreme case where $n_1 = \cdots = n_K = \tilde{n}$ and the sparsity patterns are identical. In this case, the condition for using the group sparsity penalty reduces to $\min_{(i,j)\in A^u} \sqrt{\frac{\sum_{k=1}^K \left(\omega_{kij}^0\right)^2}{K}} \geq O\left(\sqrt{\frac{\log(p\vee q)}{\tilde{n}q}}\right)$, which is clearly less stringent than the condition (7) required for using the sparsity penalty, because $\min_{(i,j)\in A^u} \sqrt{\frac{\sum_{k=1}^K \left(\omega_{kij}^0\right)^2}{K}} \geq \min_k \min_{(i,j)\in A_k^0} |\omega_{kij}^0|$. On the other hand, if the sample sizes are highly unbalanced, or the sparsity patterns are markedly different across groups, then using the sparsity penalty would require a less stringent condition. Comparing to some existing vector-based multi-graph analysis, our result is for the actual local minimizer, rather than the global minimizer as in Zhu et al. (2014). Moreover, we guarantee both multi-graph structure recovery and ensure positive-definiteness of the estimator, while Lee and Liu (2015); Cai et al. (2016) can not guarantee the latter.

5 Simulations

5.1 Setup

We study the finite-sample performance of our method through simulations. To evaluate the accuracy of sparsity identification, we employ the average false positive (FP) and average false negative (FN) rates, defined as,

$$\begin{aligned}
\mathrm{FP} &= \frac{1}{K}\sum_{k=1}^{K} \frac{\sum_{1\leq l<l'\leq p} \mathbb{I}(\omega_{ll'k} = 0, \hat{\omega}_{ll'k} \neq 0)}{\sum_{1\leq l<l'\leq p} \mathbb{I}(\omega_{ll'k} = 0)} \Big\{1 - \mathbb{I}\big(\boldsymbol{\Omega}_{k,-ll} \neq \mathbf{0}\big)\Big\}, \\
\mathrm{FN} &= \frac{1}{K}\sum_{k=1}^{K} \frac{\sum_{1\leq l<l'\leq p} \mathbb{I}(\omega_{ll'k} \neq 0, \hat{\omega}_{ll'k} = 0)}{\sum_{1\leq l<l'\leq p} \mathbb{I}(\omega_{ll'k} \neq 0)} \mathbb{I}\big(\boldsymbol{\Omega}_{k,-ll} \neq \mathbf{0}\big),
\end{aligned}$$

where $\mathbf{\Omega}_{k,-ll}$ are the off-diagonal elements of $\mathbf{\Omega}_k$. To evaluate the accuracy of parameter estimation, we employ the entropy loss (EL) and quadratic loss (QL), defined as,

$$\begin{aligned}
\mathrm{EL}_k &= \mathrm{trace}\big(\mathbf{\Omega}_k^{-1}\widehat{\mathbf{\Omega}}_k\big) - \log\det\big(\mathbf{\Omega}_k^{-1}\widehat{\mathbf{\Omega}}_k\big) - p, \\
\mathrm{QL}_k &= \mathrm{trace}\left\{\big(\mathbf{\Omega}_k^{-1}\widehat{\mathbf{\Omega}}_k - \boldsymbol{I}\big)^2\right\}, \quad k = 1, \cdots, K.
\end{aligned}$$

We generate the data from a matrix normal distribution. We consider three spatial dependence structures: a chain graph, a hub graph, and a random graph, as shown in Figure 1. We fix the temporal dependence structure as an order-one autoregressive model. We focus on the two-group graph estimation, i.e., with $K = 2$, although our method can be equally applied to more than two groups. We first generate a graph following one of the three structures in Figure 1 for one group, then construct the graph for the other group by randomly adding a few edges to the first graph. We vary the number of per-group subjects $n_k = \{10, 20\}$, the spatial dimension $p = \{100, 200\}$, and the temporal dimension $q = \{50, 100\}$. In the interest of space, we report the results when $n_k = 10$ in the online Supplementary Appendix.

Figure 1 about here

5.2 Comparison

We compare our method with some competing alternative solutions. The first is a matrix Gaussian multi-graph estimation method using the convex penalty, i.e., a combination of the ℓ_1 and the group ℓ_1 penalty. The second category are two state-of-the-art vector Gaussian multi-graph estimation methods, Lee and Liu (2015) and Cai et al. (2016). Both estimate multiple graphs that share a common structure, and both utilize the convex penalty. Since both methods have been designed for the vector-valued rather than the matrix-valued data, we first apply whitening to reduce the temporal correlations among the columns of the matrix data, then apply Lee and Liu (2015) and Cai et al. (2016). All parameter tunings are done via 5-fold cross-validation. Tables 1 to 3 summarize the results based on 100 data replications for the three spatial graph structures in Figure 1. In summary, our proposed method clearly outperforms the alternative solutions in terms of both sparsity identification and graph estimation accuracy.

Compared with the convex counterpart, our proposed nonconvex method achieves a smaller false positive, as well as a smaller estimation error. For instance, under the chain graph and $n_k = 20, p = q = 100$, the average entropy loss for the first graph for our method is 0.093, with the standard deviation SD = 0.015, and that for the convex method is 4.030, with SD = 2.160. Meanwhile, the average false positive rate for our method is 0.003, with SD = 0.005, and that for the convex method is 0.059, with SD = 0.058. Similar numerical advantages of the nonconvex solution are consistently observed for different graph structures, sample sizes, and spatial and temporal dimensions. These results, to some extent, also reflect the advantage of a nonconvex penalty compared to a convex one when the parameter tuning is done via cross-validation.

Compared with the method of Lee and Liu (2015), again, our proposal performs much better in both sparsity identification and graph estimation accuracy. In particular, the method of Lee and Liu (2015) yields a much higher false positive rate than our approach, while the false negative rates of the two are comparable. Besides, the estimation error of our method is 3 to 10 times smaller than that of Lee and Liu (2015). Compared with the method of Cai et al. (2016), our proposal performs about the same in terms of sparsity identification, but substantially improves in graph estimation. Actually, the graph estimation error of Cai et al. (2016) is the worst among all solutions, and in some situations, its estimation error is 1000 times higher than that of our proposed method. Since both Lee and Liu (2015) and Cai et al. (2016) relied on some convex penalties, these results partially reflect the conflict between selection consistency and estimation accuracy that is not uncommon when employing a convex penalty (Shen et al., 2012). Moreover, it shows the advantage of directly working with the matrix data, rather than working with the vector-valued data after whitening.

Tables 1-3 about here

5.3 Computation

We also examine in detail the computational cost of our proposed solution. All computations were done on a single core, Xeon E5-2690 v3 at 2.6GHz and 128G memory. We first report and compare the running time of various methods for the simulation examples in Section 5.2. The last column of Tables 1 to 3 records the average running time, in seconds, rounded up to integers. It is seen that our proposed method is slower than its convex counterpart, but only slightly, and the two running times are comparable. For instance, for the chain graph with $n_k = 20, p = 200, q = 100$, the average running time for our method is 474 seconds, and that for the convex solution is 398 seconds. This is due to that the MM step of convex relaxation of our nonconvex objective function usually converges in only a few iterations. Consequently, the main computational cost of our algorithm is dominated by the convex optimization step of ADMM. On the other hand, we have observed a 4 to 50 fold slowdown in running time for the other two competing methods of Lee and Liu (2015) and Cai et al. (2016). For the aforementioned setup, the average time for Lee and Liu (2015) is 9,308 seconds, and for Cai et al. (2016) is 7,875 seconds. This is partly because those two alternatives use the interior point method in optimization, which slows down significantly when the graph size increases. As a further illustration, we also report the computational time when the number of network nodes gradually increases from $p = 25$ to $p = 500$ in the online Supplementary Appendix. Our method is found to be comparable to the convex solution in terms of running time, but is much faster than Lee and Liu (2015) and Cai et al. (2016), especially when the graph dimension p is large.

6 Data analysis

6.1 Autism spectrum disorder study

Autism spectrum disorder (ASD) is an increasingly prevalent neurodevelopmental disorder, and its estimated prevalence was 1 in every 68 American children according to the Centers for Disease Control and Prevention in 2014. It is characterized by symptoms such as social difficulties, communication deficits, stereotyped behaviors and cognitive delays (Rudie et al., 2013). We analyzed a resting-state fMRI dataset from the Autism Brain Imaging Data Exchange (ABIDE) study (Di Martino et al., 2014). The imaging was performed on Siemens magneto trio scanners, with the scan parameters: voxel size

$= 3 \times 3 \times 4$mm, slice thickness = 4mm, number of slices = 34, repetition time = 3s, and echo time = 28ms. During imaging acquisition, all subjects were asked to lie still, stay awake, and keep eyes open under a white background with a black central fixation cross. After removing the images with poor quality or substantial missing values, we focused on a dataset of 795 subjects, among whom 362 have ASD, and 433 are normal controls. See Table 4 for the basic demographic information of the study subjects. All fMRI scans have been preprocessed through a standard pipeline, including slice timing correction, motion correction, denoising by regressing out motion parameters and white matter and cerebrospinal fluid time courses, spatial smoothing, band-pass filtering, and registration. Each brain image was then parcellated into 116 regions of interest using the Anatomical Automatic Labeling (AAL) atlas (Tzourio-Mazoyer et al., 2002). The time series of voxels within the same region were then averaged, resulting in a spatial-temporal data matrix for each individual subject, with the spatial dimension $p = 116$ and the temporal dimension $q = 146$. We also comment that, other than simple averaging, there are alternative approaches to summarize the voxel data within each region (Kang et al., 2016). Our proposed method is equally applicable to the data with a different summary.

Of scientific interest is to understand how brain functional connectivity differs between the ASD subjects and normal controls. We applied our nonconvex penalized multi-graph estimation method to this data, and tuned the parameters using 5-fold cross-validation. A quick examination of the quantile-quantile plot (not shown here) suggested the normality holds approximately for this data. Figure 2 reports the results. To facilitate the graphical presentation, we plot only the top 2% of the identified links for the autism and normal control groups. The dashed links are the ones found common in both groups, while the solid links are unique to each group. Our findings are in general consistent with the ASD literature. For instance, we have observed decreased connectivity between the two hemispheres, as shown by the solid links in the control group between the left and right half of the graph (Vissers et al., 2012). We also found some brain regions with different connectivity patterns between the two groups of subjects, such as inferior frontal gyrus and fusiform gyrus, which have been noted in previous studies too (Rudie et al., 2013; Di Martino et al., 2014; Tyszka et al., 2014).

Table 4 about here

Figure 2 about here

6.2 Attention deficit hyperactivity disorder study

Attention deficit hyperactivity disorder (ADHD) is one of the most commonly diagnosed child-onset neurodevelopmental disorders and has an estimated childhood prevalence of 5 to 10% worldwide (Pelham et al., 2007). Symptoms include difficulty in staying focused and paying attention, difficulty in controlling behavior, and over-activity. These symptoms may persist into adolescence and adulthood, resulting in a lifelong impairment (Biederman et al., 2000). We analyzed a resting-state fMRI dataset from the ADHD-200 Global Competition. The fMRI images were acquired on Siemens allegra 3T scanners at New York University, with the scan parameters: voxel size = $3 \times 3 \times 4$mm, slice thickness = 4mm, number of slices = 33, repetition time = 2s, and echo time = 15ms. During acquisition, all subject were asked to stay awake and not to think about anything under a black screen. For each subject, one or two fMRI scans were acquired, and for each scan, a quality control assessment (pass or questionable) was given by the data curators. We only used the scans that pass the quality control. If both scans of a subject passed the quality control, we arbitrarily chose the first scan. If neither scan passed the quality control, we removed that subject from further analysis. This results in 187 subjects, among whom 96 are combined ADHD subjects and 91 are typically developing controls. See Table 4 for the demographic information. All the scans have been preprocessed using the Athena pipeline, including slice timing correction, motion correction, denoising, spatial smoothing, band-pass filtering, and registration. Each image was then parcellated using the AAL atlas, and the resulting data is a spatial-temporal matrix, with the spatial dimension $p = 116$ and the temporal dimension $q = 172$.

Our study goal is to estimate and compare the functional connectivity network between the ADHD and control groups. We applied our method, tuned by 5-fold cross-validation. The quantile-quantile plot suggested the data are approximately normal. Figure 3 shows the results of the top 2% of the identified links for the two groups. We found a number of brain regions that exhibit different connectivity patterns between the ADHD and control groups, including frontal gyrus, cingulate gyrus, cerebellum and cerebellar vermis. Such finds are generally in agreement with the ADHD literature. Specifically, the prefrontal cortex is responsible for many higher-order mental functions, including those that regulate attention and behavior, and it is commonly believed that ADHD is associated with alterations in the prefrontal cortex (Arnsten and Li, 2005). The cingulate gyrus is associated with cognitive process, and there are evidences of anterior cingulate dysfunctions in ADHD patients (Bush et al., 2005). The cerebellum is responsible for motor control and cognitive functions such as attention and language, and dysfunction in the cerebellum and anomaly in the cerebellar vermis in ADHD patients have been reported (Toplak et al., 2006; Goetz et al., 2014).

Figure 3 about here

7 Discussion

In this article, we have proposed a nonconvex penalized method to simultaneously estimate multiple graphs from matrix-valued data. We have developed an efficient optimization algorithm, and established some sharp theoretical results. Numerical analysis has demonstrated clear advantages of our method compared to some alternative solutions.

We have advocated a nonconvex penalty, since it produces a nearly unbiased estimator, is better suited for cross-validation tuning, and can achieve a better theoretical guarantee under less stringent assumptions. Meanwhile, we recognize its potential limitations. In terms of prediction and estimation accuracy, a nonconvex penalty tends to work better when the signal in the data is sparse and has a relatively large magnitude. On the other hand, a convex penalty tends to perform better if the signal is not sparse and if there are many small signals. This phenomenon has been constantly observed in the context of high-dimensional linear model selection and graph estimation (Fan et al., 2009; Zhang, 2010; Shen et al., 2012). We also clarify that, the proposed penalized optimization formulation in (2) is in general a nonconvex problem. However, (2) can be convex for the special cases when $\lambda_1 = 0$ or $\lambda_2 = 0$ under some choice of the parameters, e.g., $R = \sqrt{2a}, b = 2a$. The convexity allows us to establish the desired theoretical properties for those special cases. This is a strategy commonly used in high-dimensional theoretical analysis. For instance, for variable selection, typically, it is only shown that there exist some tuning parameter values at which the solution is selection consistent or attains the oracle property (Fan and Li, 2001b; Zhang, 2010; Shen et al., 2013).

A key assumption for our proposal is the matrix normal distribution. Such an assumption is widely used, and is scientifically plausible in the context of brain connectivity analysis. On the other hand, we recognize that this assumption may not always hold. There are two possible ways to relax this assumption. The first is to consider a different loss function; for instance, the D-trace loss function (Zhang and Zou, 2012), or the pseudo-likelihood loss function (Lee and Hastie, 2015). The penalty function developed in our solution can be coupled with those alternative loss functions. We have chosen the likelihood based loss function, because it is more amenable to the theoretical analysis thanks to the strong convexity property of the negative log-likehood loss function, and because it yields a positive-definite estimator for the precision matrix. The second type of relaxation comes from recent development that extend a vector Gaussian graphical model to a semiparametric model (Liu et al., 2012), or a fully nonparametric model (Lee et al., 2016). Parallel extension of those methods to matrix-valued data is warranted for future research.

We have primarily focused on *graph estimation* in this article, which is a different problem than *graph inference*, even though both can produce, in effect, a sparse representation of the graph structure. We recognize that graph-based inference is a very challenging problem, and is currently an active area of research that receives increasing attention (Janková and van de Geer, 2015; Xia and Li, 2017). An alternative solution is Bayesian graph estimation (Peterson et al., 2015; Zhu et al., 2016), which could automatically produce a valid inference for all the parameters, provided that the prior is appropriately specified. However, a major challenge for the class of Bayesian solutions is the computation and the scalability to large graphs. The sampling method used in Bayesian analysis is computationally much more expensive than the optimization in the frequentist solution. For the Gaussian graphical model, each Markov chain Monte Carlo

(MCMC) iteration requires $O(p^3)$ operations, and the number of MCMC steps required for mixing is usually much larger than the number of ADMM steps in our optimization. Scalable Bayesian graph estimation is an important future direction.

8 Acknowledgement

Zhu's research was supported in part by NSF grants DMS-1721445 and DMS-1712580. Li's research was supported in part by NSF grant DMS-1613137 and NIH grant AG034570. The authors thank the editor, the associate editor and three referees for their valuable comments and suggestions.

References

Arnsten, A. F. and Li, B.-M. (2005). Neurobiology of executive functions: Catecholamine influences on prefrontal cortical functions. *Biological Psychiatry*, 57(11):1377 – 1384.

Biederman, J., Mick, E., and Faraone, S. V. (2000). Age-dependent decline of symptoms of attention deficit hyperactivity disorder: Impact of remission definition and symptom type. *American Journal of Psychiatry*, 157(5):816–818.

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122.

Bush, G., Valera, E. M., and Seidman, L. J. (2005). Functional neuroimaging of attention-deficit/hyperactivity disorder: A review and suggested future directions. *Biological Psychiatry*, 57(11):1273 – 1284.

Cai, T. T., Li, H., Liu, W., and Xie, J. (2016). Joint estimation of multiple high-dimensional precision matrices. *Statistica Sinica*, 26:445–464.

Cai, T. T., Liu, W., and Luo, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *J. Amer. Statist. Assoc.*, 106(494):594–607.

Danaher, P., Wang, P., and Witten, D. M. (2014). The joint graphical lasso for inverse covariance estimation across multiple classes. *J. R. Statist. Soc. B*, 76(2):373–397.

Di Martino, A., Yan, C.-G., Li, Q., Denio, E., Castellanos, F. X., Alaerts, K., Anderson, J. S., Assaf, M., Bookheimer, S. Y., Dapretto, M., Deen, B., Delmonte, S., Dinstein, I., Ertl-Wagner, B., Fair, D. A., Gallagher, L., Kennedy, D. P., Keown, C. L., Keysers, C., Lainhart, J. E., Lord, C., Luna, B., Menon, V., Minshew, N. J., Monk, C. S., Mueller, S., Muller, R.-A., Nebel, M. B., Nigg, J. T., O'Hearn, K., Pelphrey, K. A., Peltier, S. J., Rudie, J. D., Sunaert, S., Thioux, M., Tyszka, J. M., Uddin, L. Q., Verhoeven, J. S., Wenderoth, N., Wiggins, J. L., Mostofsky, S. H., and Milham, M. P.

(2014). The autism brain imaging data exchange: Towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular Psychiatry*, 19(6):659–667.

Fan, J., Feng, Y., and Wu, Y. (2009). Network exploration via the adaptive lasso and SCAD penalties. *Ann. Appl. Stat.*, 3(2):521–541.

Fan, J. and Li, R. (2001a). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.*, 96(456):1348–1360.

Fan, J. and Li, R. (2001b). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360.

Fan, J., Xue, L., and Zou, H. (2014). Strong oracle optimality of folded concave penalized estimation. *Annals of statistics*, 42(3):819.

Fornito, A., Zalesky, A., and Breakspear, M. (2013). Graph analysis of the human connectome: Promise, progress, and pitfalls. *NeuroImage*, 80:426–444.

Fox, M. D. and Greicius, M. (2010). Clinical applications of resting state functional connectivity. *Frontiers in Systems Neuroscience*, 4(19).

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441.

Friston, K., Ashburner, J., Kiebel, S., Nichols, T., and Penny, W., editors (2007). *Statistical Parametric Mapping: the Analysis of Functional Brain Images*. Academic Press, London.

Goetz, M., Vesela, M., and Ptacek, R. (2014). Notes on the role of the cerebellum in adhd. *Austin J Psychiatry Behav Sci*, 1(3):1013.

Guo, J., Levina, E., Michailidis, G., and Zhu, J. (2011). Joint estimation of multiple graphical models. *Biometrika*, 98(1):1–15.

Han, F., Han, X., Liu, H., and Caffo, B. (2016). Sparse median graphs estimation in a high-dimensional semiparametric model. *The Annals of Applied Statistics*, 10(3):1397–1426.

Hunter, D. R. and Lange, K. (2004). A tutorial on mm algorithms. *The American Statistician*, 58(1):30–37.

Janková, J. and van de Geer, S. (2015). Honest confidence regions and optimality in high-dimensional precision matrix estimation. *arXiv preprint arXiv:1507.02061*.

Kang, J., Bowman, F. D., Mayberg, H., and Liu, H. (2016). A depression network of functionally connected regions discovered via multi-attribute canonical correlation graphs. *NeuroImage*, 141:431–441.

Lee, J. D. and Hastie, T. J. (2015). Learning the structure of mixed graphical models. *Journal of Computational and Graphical Statistics*, 24(1):230–253.

Lee, K.-Y., Li, B., and Zhao, H. (2016). On additive partial correlation operator and nonparametric estimation of graphical models. *Biometrika*, in press.

Lee, W. and Liu, Y. (2015). Joint estimation of multiple precision matrices with common structures. *Journal of Machine Learning Research*, 16:1035–1062.

Leng, C. and Tang, C. Y. (2012). Sparse matrix graphical models. *J. Amer. Statist. Assoc.*, 107(499):1187–1200.

Liu, H., Han, F., Yuan, M., Lafferty, J., and Wasserman, L. (2012). High-dimensional semiparametric Gaussian copula graphical models. *Ann. Statist.*, 40(4):2293–2326.

Liu, H., Roeder, K., and Wasserman, L. (2010). Stability approach to regularization selection (stars) for high dimensional graphical models. In Lafferty, J. D., Williams, C. K. I., Shawe-Taylor, J., Zemel, R. S., and Culotta, A., editors, *Advances in Neural Information Processing Systems 23*, pages 1432–1440. Curran Associates, Inc.

Loh, P.-L. and Wainwright, M. J. (2014). Support recovery without incoherence: A case for nonconvex regularization. *arXiv preprint arXiv:1412.5632*.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, pages 1436–1462.

Pelham, W. E., Foster, E. M., and Robb, J. A. (2007). The economic impact of attention-deficit/hyperactivity disorder in children and adolescents. *Ambulatory Pediatrics*, 7(1, Supplement):121 – 131. Measuring Outcomes in Attention Deficit Hyperactivity Disorder.

Peterson, C., Stingo, F. C., and Vannucci, M. (2015). Bayesian inference of multiple gaussian graphical models. *Journal of the American Statistical Association*, 110(509):159–174.

Qiu, H., Han, F., Liu, H., and Caffo, B. (2016). Joint estimation of multiple graphical models from high dimensional time series. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(2):487–504.

Ravikumar, P., Wainwright, M. J., Raskutti, G., and Yu, B. (2011). High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electron. J. Stat.*, 5:935–980.

Rudie, J., Brown, J., Beck-Pancer, D., Hernandez, L., Dennis, E., Thompson, P., Bookheimer, S., and Dapretto, M. (2013). Altered functional and structural brain network organization in autism. *NeuroImage: Clinical*, 2:79 – 94.

Shen, X., Pan, W., and Zhu, Y. (2012). Likelihood-based selection and sharp parameter estimation. *J. Amer. Statist. Assoc.*, 107(497):223–232.

Shen, X., Pan, W., Zhu, Y., and Zhou, H. (2013). On constrained and regularized high-dimensional regression. *Annals of the Institute of Statistical Mathematics*, 65(5):807–832.

Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, T. E., Johansen-Berg, H., Bannister, P. R., Luca, M. D., Drobnjak, I., Flitney, D. E., Niazy, R. K., Saunders, J., Vickers, J., Zhang, Y., Stefano, N. D., Brady, J. M., and Matthews, P. M. (2004). Advances in functional and structural {MR} image analysis and implementation as {FSL}. *NeuroImage*, 23, Supplement 1:S208 – S219. Mathematics in Brain Imaging.

Toplak, M. E., Dockstader, C., and Tannock, R. (2006). Temporal information processing in adhd: Findings to date and new methods. *Journal of Neuroscience Methods*, 151(1):15 – 29. Towards a Neuroscience of Attention-Deficit/Hyperactivity Disorder (ADHD).

Toussaint, P.-J., Maiz, S., Coynel, D., Doyon, J., Messé, A., de Souza, L. C., Sarazin, M., Perlbarg, V., Habert, M.-O., and Benali, H. (2014). Characteristics of the default mode functional connectivity in normal ageing and alzheimer's disease using resting state fmri with a combined approach of entropy-based and graph theoretical measurements. *NeuroImage*, 101:778 – 786.

Tsiligkaridis, T., Hero, III, A. O., and Zhou, S. (2013). On convergence of Kronecker graphical lasso algorithms. *IEEE Trans. Signal Process.*, 61(7):1743–1755.

Tyszka, J. M., Kennedy, D. P., Paul, L. K., and Adolphs, R. (2014). Largely typical patterns of resting-state functional connectivity in high-functioning adults with autism. *Cerebral Cortex*, 24(7):1894–1905.

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., and Joliot, M. (2002). Automated anatomical labeling of activations in {SPM} using a macroscopic anatomical parcellation of the {MNI} {MRI} single-subject brain. *NeuroImage*, 15(1):273 – 289.

Varoquaux, G. and Craddock, R. C. (2013). Learning and comparing functional connectomes across subjects. *NeuroImage*, 80:405 – 415. Mapping the Connectome.

Vissers, M. E., X Cohen, M., and Geurts, H. M. (2012). Review. *Neuroscience and Biobehavioral Reviews*, 36(1):604–625.

Xia, Y. and Li, L. (2017). Hypothesis testing of matrix graph model with application to brain connectivity analysis. *Biometrics*, in press.

Yin, J. and Li, H. (2012). Model selection and estimation in the matrix normal graphical model. *Journal of Multivariate Analysis*, 107:119 – 140.

Yuan, M. and Lin, Y. (2007). Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1):19–35.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.*, 38(2):894–942.

Zhang, T., Wu, J., Li, F., Caffo, B., and Boatman-Reich, D. (2015). A dynamic directional model for effective brain connectivity using electrocorticographic (ECoG) time series. *J. Amer. Statist. Assoc.*, 110(509):93–106.

Zhang, T. and Zou, H. (2012). Sparse precision matrix estimation via lasso penalized d-trace loss. *Biometrika*, 99(1):1–18.

Zhou, S. (2014). Gemini: graph estimation with matrix variate normal instances. *Ann. Statist.*, 42(2):532–562.

Zhu, H., Strawn, N., and Dunson, D. B. (2016). Bayesian graphical models for multivariate functional data. *Journal of Machine Learning Research*, 17(204):1–27.

Zhu, Y., Shen, X., and Pan, W. (2014). Structural Pursuit Over Multiple Undirected Graphs. *J. Amer. Statist. Assoc.*, 109(508):1683–1696.

Table 1: Chain graph. Reported are the average and standard deviation (in parenthesis) of the accuracy criteria based on 100 data replications. Also reported is the average running time (in seconds). Evaluation criteria include the false positive rate (FP), the false negative rate (FN), the entropy loss (EL_k), and the quadratic loss (QL_k). We compare the proposed nonconvex based multi-graph estimation method (denoted as Nonconvex) with its convex counterpart (denoted as Convex), the method of Lee and Liu (2015) (denoted as Lee & Liu), and the method of Cai et al. (2016) (denoted as Cai et al.).

n_k	p	q	Method	FP	FN	EL_1	EL_2	QL_1	QL_2	Time
20	100	100	Nonconvex	0.003 (0.005)	0.000 (0.000)	0.093 (0.015)	0.105 (0.015)	0.230 (0.036)	0.265 (0.038)	116
			Convex	0.059 (0.058)	0.000 (0.000)	4.030 (2.160)	3.520 (1.700)	7.850 (4.410)	7.080 (3.640)	85
			Lee & Liu	0.413 (0.058)	0.000 (0.000)	1.475 (0.043)	0.960 (0.059)	3.991 (0.121)	2.562 (0.181)	1050
			Cai et al.	0.005 (0.005)	6e-04 (0.002)	14.40 (1.100)	16.90 (2.000)	48.00 (5.000)	55.50 (9.300)	631
		50	Nonconvex	0.001 (0.003)	0.000 (0.000)	0.194 (0.028)	0.216 (0.027)	0.484 (0.072)	0.547 (0.069)	159
			Convex	0.053 (0.045)	0.000 (0.000)	6.500 (3.350)	5.460 (2.440)	13.10 (7.410)	11.70 (5.800)	102
			Lee & Liu	0.382 (0.008)	0.000 (0.000)	1.835 (0.070)	1.253 (0.070)	5.063 (0.218)	3.486 (0.216)	1096
			Cai et al.	7e-04 (7e-04)	0.000 (0.000)	7.200 (0.750)	9.300 (1.500)	22.10 (2.300)	28.30 (4.700)	608
	200	100	Nonconvex	0.000 (0.001)	0.000 (0.000)	0.192 (0.023)	0.196 (0.018)	0.475 (0.056)	0.475 (0.044)	474
			Convex	0.032 (0.032)	0.000 (0.000)	11.40 (5.020)	9.110 (3.780)	22.70 (10.40)	18.70 (8.130)	398
			Lee & Liu	0.166 (0.002)	0.000 (0.000)	2.800 (0.059)	1.600 (0.044)	7.200 (0.170)	4.000 (0.120)	9308
			Cai et al.	1e-04 (2e-04)	0.000 (0.000)	13.50 (1.200)	19.10 (2.400)	41.20 (3.800)	54.20 (7.800)	7875
		50	Nonconvex	0.000 (0.000)	0.000 (0.000)	0.390 (0.039)	0.384 (0.034)	0.972 (0.099)	0.933 (0.085)	730
			Convex	0.023 (0.027)	0.000 (0.000)	16.70 (6.510)	13.10 (4.550)	34.10 (14.50)	27.20 (10.400)	672
			Lee & Liu	0.212 (0.031)	0.000 (0.000)	3.400 (0.088)	1.900 (0.180)	8.900 (0.220)	5.000 (0.520)	11770
			Cai et al.	0.001 (0.001)	0.000 (0.000)	8.200 (0.680)	10.50 (1.800)	23.70 (1.700)	29.00 (4.300)	7281

Table 2: Hub graph. The setup is the same as Table 1.

n_k	p	q	Method	FP	FN	EL_1	EL_2	QL_1	QL_2	Time
20	100	100	Nonconvex	0.006 (0.006)	0.000 (0.000)	0.086 (0.013)	0.111 (0.018)	0.408 (0.070)	0.459 (0.077)	226
			Convex	0.199 (0.049)	0.000 (0.000)	1.290 (0.894)	1.360 (0.861)	4.31 (2.76)	4.180 (2.580)	197
			Lee & Liu	0.467 (0.024)	0.000 (0.000)	1.100 (0.033)	1.000 (0.037)	3.500 (0.180)	3.90 (0.260)	1072
			Cai et al.	5e-04 (0.001)	0.000 (0.000)	20.90 (3.300)	22.40 (3.200)	577.5 (169.4)	574.5 (155.4)	603
		50	Nonconvex	0.012 (0.007)	0.001 (0.003)	0.183 (0.025)	0.309 (0.049)	0.875 (0.149)	1.130 (0.162)	272
			Convex	0.193 (0.055)	0.001 (0.003)	2.450 (1.330)	2.530 (1.170)	8.010 (4.370)	7.340 (3.620)	257
			Lee & Liu	0.399 (0.016)	0.000 (0.000)	1.500 (0.060)	1.600 (0.073)	6.000 (0.460)	6.500 (0.570)	1108
			Cai et al.	0.005 (0.003)	9e-05 (6e-04)	17.00 (2.300)	18.40 (2.400)	412.4 (94.70)	419.5 (94.20)	611
	200	100	Nonconvex	0.001 (0.001)	0.000 (0.000)	0.171 (0.020)	0.198 (0.024)	0.818 (0.106)	0.797 (0.097)	915
			Convex	0.099 (0.028)	0.000 (0.000)	2.870 (1.680)	2.630 (1.390)	10.40 (4.950)	8.280 (3.880)	857
			Lee & Liu	0.247 (0.021)	0.000 (0.000)	2.100 (0.054)	1.900 (0.065)	6.600 (0.250)	7.200 (0.430)	9073
			Cai et al.	0.002 (0.001)	0.000 (0.000)	48.80 (15.10)	47.50 (13.40)	1615 (835.8)	1418 (680.7)	7533
		50	Nonconvex	0.001 (0.000)	0.002 (0.003)	0.354 (0.034)	0.470 (0.073)	1.710 (0.214)	1.740 (0.249)	1295
			Convex	0.098 (0.029)	0.000 (0.001)	7.480 (3.840)	6.280 (2.890)	25.20 (13.90)	18.90 (9.770)	1089
			Lee & Liu	0.222 (0.011)	1e-04 (5e-04)	3.300 (0.089)	2.700 (0.079)	11.70 (0.560)	11.80 (0.660)	11136
			Cai et al.	0.006 (0.001)	5e-05 (3e-04)	30.90 (3.500)	30.60 (3.100)	721.7 (141.4)	649.1 (112.8)	7297

Table 3: Random graph. The setup is the same as Table 1.

n_k	p	q	Method	FP	FN	EL$_1$	EL$_2$	QL$_1$	QL$_2$	Time
20	100	100	Nonconvex	0.005 (0.005)	0.000 (0.000)	0.121 (0.016)	0.146 (0.016)	0.328 (0.044)	0.457 (0.054)	195
			Convex	0.143 (0.045)	0.001 (0.002)	2.230 (1.160)	3.010 (1.430)	4.850 (2.730)	6.540 (3.510)	178
			Lee & Liu	0.471 (0.030)	0.000 (0.000)	0.950 (0.034)	1.400 (0.043)	2.600 (0.120)	3.700 (0.150)	1107
			Cai et al.	0.006 (0.006)	0.030 (0.008)	10.50 (1.000)	6.900 (0.660)	40.80 (5.700)	23.90 (3.700)	605
		50	Nonconvex	0.029 (0.008)	0.001 (0.002)	0.276 (0.034)	0.409 (0.094)	0.736 (0.094)	1.200 (0.293)	176
			Convex	0.304 (0.061)	0.001 (0.002)	1.690 (0.854)	2.050 (0.883)	3.720 (1.990)	4.510 (2.190)	153
			Lee & Liu	0.390 (0.011)	3e-04 (8e-04)	1.300 (0.048)	1.800 (0.068)	3.800 (0.160)	5.200 (0.280)	1145
			Cai et al.	0.005 (0.005)	0.026 (0.009)	8.100 (0.800)	6.100 (0.600)	29.20 (3.900)	22.00 (3.600)	580
	200	100	Nonconvex	0.022 (0.001)	0 (0.001)	0.284 (0.021)	0.459 (0.077)	0.776 (0.057)	1.830 (0.355)	1206
			Convex	0.333 (0.016)	0.001 (0.001)	1.610 (0.073)	2.510 (0.075)	3.78 (0.18)	6.120 (0.193)	1082
			Lee & Liu	0.204 (0.033)	0.002 (0.001)	3.000 (0.057)	4.000 (0.089)	8.000 (0.180)	10.90 (0.310)	9600
			Cai et al.	0.007 (0.010)	0.079 (0.006)	14.60 (0.800)	13.10 (0.780)	60.00 (6.000)	96.20 (11.60)	8310
		50	Nonconvex	0.020 (0.003)	0.020 (0.006)	0.700 (0.058)	1.810 (0.169)	1.850 (0.157)	6.480 (0.638)	1303
			Convex	0.299 (0.033)	0.009 (0.002)	3.400 (0.645)	5.030 (0.873)	7.900 (1.470)	11.90 (2.090)	1071
			Lee & Liu	0.334 (0.024)	0.006 (0.003)	3.800 (0.088)	6.000 (0.140)	10.70 (0.290)	22.70 (1.400)	11904
			Cai et al.	0.004 (0.004)	0.077 (0.008)	11.10 (0.880)	13.20 (0.980)	38.50 (3.700)	98.60 (14.10)	7959

Table 4: Demographic information of the ASD dataset and the ADHD dataset.

	ASD study		ADHD study	
Group	Case	Control	Case	Control
Sample size	362	433	96	91
Age (mean ± sd)	16.72 ± 8.253	16.27 ± 6.893	11.38 ± 2.757	12.38 ± 3.112
Male/female	341/48	348/85	73/23	44/47

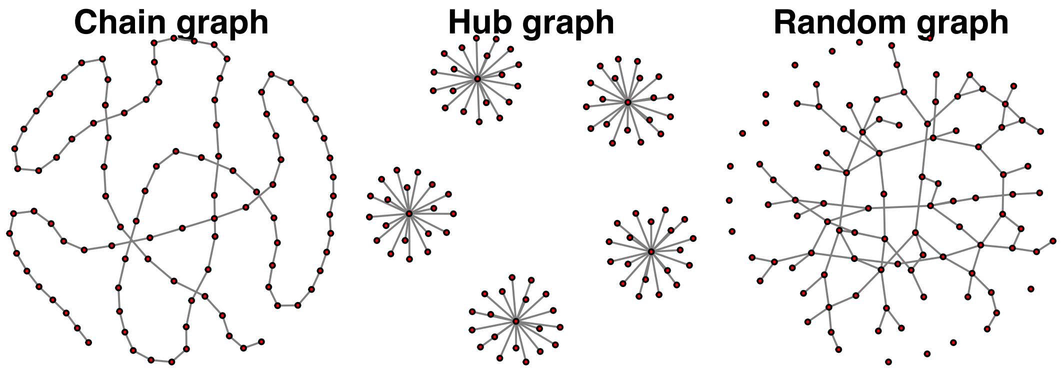
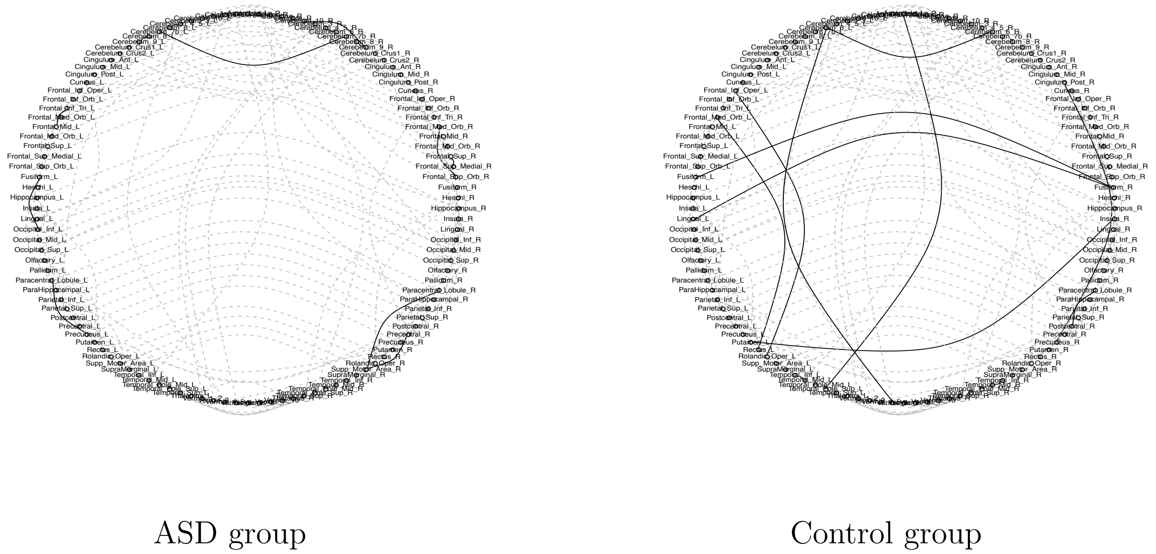


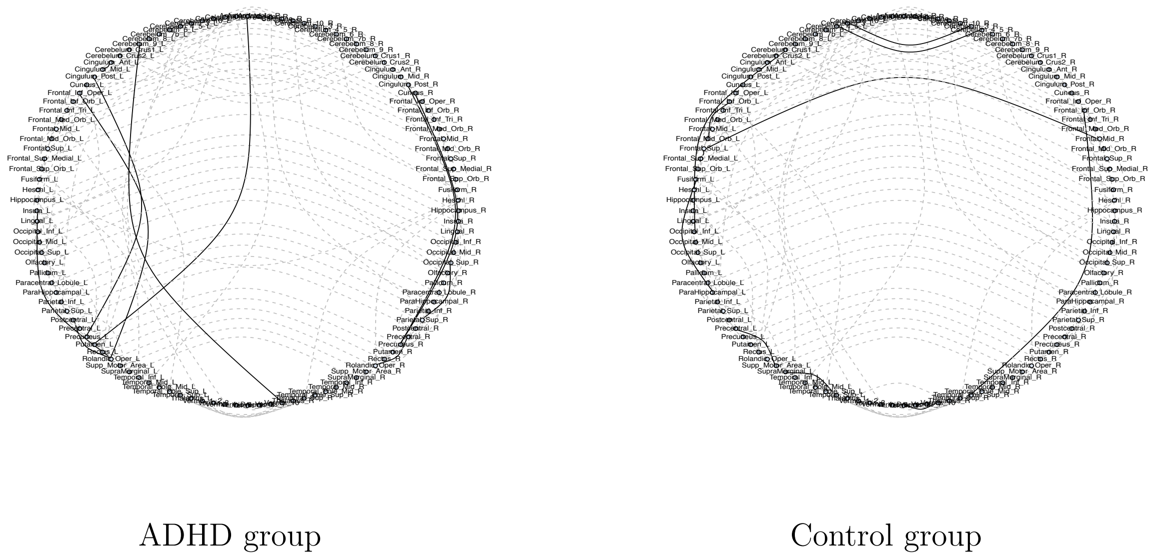
Figure 1: Three types of graphs used in our simulation studies



ASD group

Control group

Figure 2: Estimated connectivity networks for the ABIDE data. The left panel is for the ASD group, and the right panel for the normal control. Shown are the top 2% links, where the dashed links are the ones found common in both groups, and the solid links are unique to each group.



ADHD group

Control group

Figure 3: Estimated connectivity networks for the ADHD data. The left panel is for the ADHD group, and the right panel for the normal control. Shown are the top 2% links, where the dashed links are the ones found common in both groups, and the solid links are unique to each group.

Supplementary Appendix for "Multiple Matrix Gaussian Graphs Estimation"

Yunzhang Zhu and Lexin Li

This appendix collects the proofs of the two lemmas in the ADMM step of the optimization algorithm in Section 3, a number of technical lemmas, the proofs of Theorems 1 and 2 in Section 4, and some additional simulation results.

A Proof of Lemmas 1 and 2 of the ADMM optimization

Proof of Lemma 1. It suffices to show that $\mathbf{\Omega}^\star$ satisfies the optimality condition,

$$\text{trace}\left[\{\mathbf{\Delta} - (\mathbf{\Omega}^\star)^{-1} + c\mathbf{\Omega}^\star\}(\mathbf{\Omega} - \mathbf{\Omega}^\star)\right] \geq 0 \text{ for any } \lambda_{\max}(\mathbf{\Omega}) \leq R \text{ and } \mathbf{\Omega} \succeq 0\,.$$

Substituting $\mathbf{\Omega}^\star = \boldsymbol{U}\boldsymbol{Q}\boldsymbol{U}^\top$ and $\mathbf{\Delta} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^\top$, it suffices to show that

$$\text{trace}\left\{(\boldsymbol{D} - \boldsymbol{Q}^{-1} + c\boldsymbol{Q})(\boldsymbol{U}^\top\mathbf{\Omega}\boldsymbol{U} - \boldsymbol{Q})\right\} \geq 0 \text{ for any } \lambda_{\max}(\mathbf{\Omega}) \leq R \text{ and } \mathbf{\Omega} \succeq 0\,.$$

Let $\widetilde{\mathbf{\Omega}} = \boldsymbol{U}^\top\mathbf{\Omega}\boldsymbol{U}$. Since $\lambda_{\max}(\mathbf{\Omega}) = \lambda_{\max}(\widetilde{\mathbf{\Omega}})$, it is then equivalent to show

$$(D_{ii} - Q_{ii}^{-1} + cQ_{ii})(\widetilde{\Omega}_{ii} - Q_{ii}) \geq 0 \text{ for any } \lambda_{\max}(\widetilde{\mathbf{\Omega}}) \leq R \text{ and } \widetilde{\mathbf{\Omega}} \succeq 0\,.$$

This holds true, because for any $0 < x \leq R$,

$$(D_{ii} - Q_{ii}^{-1} + cQ_{ii})(x - Q_{ii}) \geq 0,$$

and $0 < \widetilde{\Omega}_{ii} \leq \lambda_{\max}(\widetilde{\mathbf{\Omega}}) \leq R$ for any $1 \leq i \leq p$. This completes the proof. □

Proof of Lemma 2. Note that the subgradient of $\sum_{k=1}^K b_k|x_k| + \nu\sqrt{\sum_{k=1}^K x_k^2}$ is of the structure $\{\boldsymbol{t} + \boldsymbol{s} : \boldsymbol{t}, \boldsymbol{s} \in \mathbb{R}^k, |t_k| \leq b_k; k = 1, \ldots, K, \text{ and } \|\boldsymbol{s}\|_2 \leq \nu\}$. Therefore, if $\|\left(S_{b_1}(a_1), \cdots, S_{b_K}(a_K)\right)\|_2 \leq \nu$, we have that

$$\begin{aligned}\boldsymbol{a} &= \{b_1\text{Sign}(a_1), \ldots, b_K\text{Sign}(b_K)\} + \{S_{b_1}(a_1), \cdots, S_{b_K}(a_K)\} \\ &\in \{\boldsymbol{t} + \boldsymbol{s} : \boldsymbol{t}, \boldsymbol{s} \in \mathbb{R}^k, |t_k| \leq b_k; k = 1, \ldots, K, \text{ and } \|\boldsymbol{s}\|_2 \leq \nu\},\end{aligned}$$

which implies that $\boldsymbol{x}^\star = 0$ when $\|\left\{S_{b_1}(a_1), \cdots, S_{b_K}(a_K)\right\}\|_2 \leq \nu$. Moreover, when $\|\left\{S_{b_1}(a_1), \cdots, S_{b_K}(a_K)\right\}\|_2 > \nu$, the optimality condition becomes

$$x_k = \begin{cases} a_k - \text{Sign}(x_k)b_k - \nu\frac{x_k}{\|\boldsymbol{x}\|_2}, & \text{if } |a_k| > b_k, \\ 0, & \text{if } |a_k| \leq b_k, \end{cases}$$

which implies that

$$x_k^\star = \left[1 - \frac{\nu}{\sqrt{\sum_{k=1}^K \{S_{b_k}(a_k)\}^2}}\right] S_{b_k}(a_k);\ k = 1, \ldots, K,$$

where we have used the fact that $x_k^\star$, if nonzero, must have the same sign as a_k. This completes the proof. □

B Technical Lemmas in preparation for Theorems 1 and 2

We present a number of technical lemmas that would facilitate the proofs of Theorems 1 and 2. We first introduce some notations. For a function $f(x)$, let $\partial f(x)$ denote the subgradient of $f(\cdot)$ at x. For a symmetric matrix $\boldsymbol{C} \in \mathbb{R}^{p\times p}$, let $\boldsymbol{C}_{ij}$ or $(\boldsymbol{C})_{ij}$ denote its (i,j)th entry, and let $\mathrm{vec}(\boldsymbol{C}) = \left\{\sqrt{1+\mathbb{I}(i\neq j)}\boldsymbol{C}_{ij}\right\}_{i\leq j} \in \mathbb{R}^{\frac{p(p+1)}{2}}$ denote its scaled vectorization (Alizadeh et al., 1998), where $\mathbb{I}(\cdot)$ is the indicator function. Let $\mathrm{vec}_B(\boldsymbol{C}) = \left\{\sqrt{1+\mathbb{I}(i\neq j)}\boldsymbol{C}_{ij}\right\}_{(i,j) \text{ or } (j,i)\in B}$ denote a sub-vector of $\mathrm{vec}(\boldsymbol{C})$ excluding components with indices not in the index set B. Let $\boldsymbol{C}_B$ denote a matrix of the same dimension of $\boldsymbol{C}$, such that $(\boldsymbol{C}_B)_{ij} = \boldsymbol{C}_{ij}$ when $(i,j)\in B$, and $(\boldsymbol{C}_B)_{ij} = 0$ when $(i,j)\notin B$. Define the symmetric Kronecker product $\boldsymbol{C}\otimes_s \boldsymbol{C} \in \mathbb{R}^{\frac{p(p+1)}{2}\times\frac{p(p+1)}{2}}$ as $(\boldsymbol{C}\otimes_s\boldsymbol{C})\,\mathrm{vec}(\boldsymbol{\Delta}) = \mathrm{vec}\left(\boldsymbol{C}\boldsymbol{\Delta}\boldsymbol{C}\right)$ for any symmetric matrix $\boldsymbol{\Delta}$ (Alizadeh et al., 1998).

Lemma B.1. *For any positive-definite symmetric matrix $\boldsymbol{C}\succ 0$,*

$$\nabla\left(\log\det\boldsymbol{C}\right) = -\,\mathrm{vec}(\boldsymbol{C}^{-1}), \tag{S1}$$

$$\nabla^2\left(\log\det\boldsymbol{C}\right) = \boldsymbol{C}^{-1}\otimes_s\boldsymbol{C}^{-1}. \tag{S2}$$

Moreover, for a positive-definite symmetric matrix $\boldsymbol{\Gamma}^0$ and its inverse $\boldsymbol{\Omega}^0 = (\boldsymbol{\Gamma}^0)^{-1}$, define $\boldsymbol{I} = \nabla^2\left(-\frac{1}{2}\log\det\boldsymbol{\Omega}^0\right)$. Then,

$$\mathrm{vec}(\boldsymbol{C})^\top\boldsymbol{I}\,\mathrm{vec}(\boldsymbol{C}) = \frac{1}{2}\mathrm{trace}\left(\boldsymbol{\Gamma}^0\boldsymbol{C}\boldsymbol{\Gamma}^0\boldsymbol{C}\right). \tag{S3}$$

Proof of Lemma B.1: We first show that, for any symmetric matrices $\boldsymbol{C}_1$ and $\boldsymbol{C}_2$,

$$\mathrm{vec}(\boldsymbol{C}_1)^\top\mathrm{vec}(\boldsymbol{C}_2) = \sum_{i\leq j}(1+\mathbb{I}(i\neq j))(\boldsymbol{C}_1)_{ij}(\boldsymbol{C}_2)_{ij} = \sum_{i,j}(\boldsymbol{C}_1)_{ij}(\boldsymbol{C}_2)_{ij} = \mathrm{trace}(\boldsymbol{C}_1\boldsymbol{C}_2).$$

This implies (S1). Next it follows from the Taylor's expansion of the $\log\det$ function that

$$\begin{aligned}
&\log\det(\boldsymbol{C}+\boldsymbol{\Delta}) - \log\det(\boldsymbol{C})\\
=\ &\mathrm{trace}(\boldsymbol{C}^{-1}\boldsymbol{\Delta}) - \frac{1}{2}\mathrm{trace}\left((\boldsymbol{C}^{-1}\boldsymbol{\Delta})^2\right) + o(\|\boldsymbol{C}^{-1/2}\boldsymbol{\Delta}\boldsymbol{C}^{-1/2}\|_F^2)\\
=\ &\mathrm{vec}(\boldsymbol{C}^{-1})^\top\mathrm{vec}(\boldsymbol{\Delta}) - \frac{1}{2}\mathrm{vec}(\boldsymbol{\Delta})^\top\mathrm{vec}(\boldsymbol{C}^{-1}\boldsymbol{\Delta}\boldsymbol{C}^{-1}) + o(\|\boldsymbol{C}^{-1/2}\boldsymbol{\Delta}\boldsymbol{C}^{-1/2}\|_F^2)\\
=\ &\mathrm{vec}(\boldsymbol{C}^{-1})^\top\mathrm{vec}(\boldsymbol{\Delta}) - \frac{1}{2}\mathrm{vec}(\boldsymbol{\Delta})^\top\left(\boldsymbol{C}^{-1}\otimes_s\boldsymbol{C}^{-1}\right)\mathrm{vec}(\boldsymbol{\Delta}) + o(\|\boldsymbol{C}^{-1/2}\boldsymbol{\Delta}\boldsymbol{C}^{-1/2}\|_F^2).
\end{aligned}$$

This implies (S2) holds true.

Given (S2), $\boldsymbol{I} = \frac{1}{2}\boldsymbol{\Gamma}^0 \otimes_s \boldsymbol{\Gamma}^0$. Then,

$$\operatorname{vec}(\boldsymbol{C})^\top \boldsymbol{I} \operatorname{vec}(\boldsymbol{C}) = \frac{1}{2} \operatorname{vec}(\boldsymbol{C})^\top \operatorname{vec}(\boldsymbol{\Gamma}^0 \boldsymbol{C} \boldsymbol{\Gamma}^0) = \frac{1}{2} \operatorname{trace}(\boldsymbol{C} \boldsymbol{\Gamma}^0 \boldsymbol{C} \boldsymbol{\Gamma}^0),$$

which proves (S3). This completes the proof. □

The next lemma gives the optimality condition for a general convex optimization problem. Its proof can be found in a standard convex analysis textbook, e.g., Hiriart-Urruty and Lemaréchal (2012), and as such is omitted.

Lemma B.2. *Let $f(x)$ be a subdifferentiable convex function and $\mathcal{X}$ be a closed convex set. Then $x^\star \in \mathcal{X}$ is a solution of the convex optimization problem:* $\operatorname{minimize}_{x \in X} f(x)$*, if and only if there exists a subdifferentiable $g \in \partial f(x^\star)$ such that $g^\top (x - x^\star) \geq 0$ for all $x \in \mathcal{X}$. Sufficiently, if there exists an $x^\star \in \mathcal{X}$ such that $0 \in \partial f(x^\star)$, then $x^\star$ is a solution to this optimization problem.*

For a $p \times p$ symmetric positive-definite matrix $\boldsymbol{\Gamma}^0$, let $\boldsymbol{\Omega}^0 = \left[\boldsymbol{\Gamma}^0\right]^{-1} = (\omega_{ij})_{1 \leq i,j \leq p}$. Let A^0 denote the support of $\boldsymbol{\Omega}^0$, and $s_0 = \max_i \sum_{j=1}^p \omega_{ij}^0 \neq 0$ is the maximum number of nonzeros across rows/columns of $\boldsymbol{\Omega}^0$. Let $\gamma_1 = \|\boldsymbol{\Gamma}^0\|_{\infty,\infty}$ and $\gamma_2 = \left\|\boldsymbol{I}_{A^0,A^0}^{-1}\right\|_{\infty,\infty}$, where $\boldsymbol{I} = \nabla^2 \left(-\frac{1}{2} \log \det \boldsymbol{\Omega}^0\right)$, $\boldsymbol{I}_{A^0,A^0}$ is the $|A^0| \times |A^0|$ submatrix of $\boldsymbol{I}$ that extracts the corresponding entries whose indices belong to A^0, and $|A^0|$ is the size of A^0. Furthermore, let $\widehat{\boldsymbol{\Gamma}}$ be an estimator of $\boldsymbol{\Gamma}^0$. Consider the following optimization problem,

$$\underset{\boldsymbol{\Omega} \succeq 0,\, \omega_{ij}=0, (i,j) \notin A^0}{\operatorname{minimize}} \operatorname{trace}(\boldsymbol{\Omega} \widehat{\boldsymbol{\Gamma}}) - \log \det(\boldsymbol{\Omega}). \tag{S4}$$

We call its solution an oracle estimator with respect to A^0. The next two lemmas first give the sufficient optimality condition for this oracle estimator, then establish its existence, uniqueness, and its consistency property.

Lemma B.3. *(Optimality condition for the oracle estimator) If an estimator $\widehat{\boldsymbol{\Omega}}_{A^0} = (\hat{\omega}_{ij})_{1 \leq i,j \leq p}$ satisfies that,*

$$\widehat{\boldsymbol{\Omega}}_{A^0} \succeq 0, \hat{\omega}_{ij} = 0 \textit{ for } (i,j) \notin A^0, \textit{ and } \operatorname{vec}_{A^0}\left(\widehat{\boldsymbol{\Gamma}} - \widehat{\boldsymbol{\Omega}}_{A^0}^{-1}\right) = \mathbf{0},$$

then it must be a solution to the optimization problem (S4)*.*

Proof of Lemma B.3: By focusing on $\operatorname{vec}_{A^0}(\boldsymbol{\Omega})$, (S4) becomes a convex optimization problem with the constraint $\boldsymbol{\Omega} \succeq 0$. Applying Lemma B.2 to this problem, we obtain a sufficient optimality condition,

$$\boldsymbol{\Omega} \succeq 0, \omega_{ij} = 0 \text{ for } (i,j) \notin A^0, \text{ and } \operatorname{vec}_{A^0}\left\{\nabla \log \det(\boldsymbol{\Omega})\right\} = \mathbf{0}.$$

Moreover, together with $\nabla \log \det(\boldsymbol{\Omega}) = -\operatorname{vec}(\boldsymbol{\Omega}^{-1})$ from Lemma B.1-(S1), we have that $\operatorname{vec}_{A^0}\left\{\nabla \log \det(\boldsymbol{\Omega})\right\} = \operatorname{vec}_{A^0}\left\{\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Omega}^{-1}\right\}$. This completes the proof. □

Lemma B.4. *(Existence, uniqueness and consistency of the oracle estimator) On the event that*

$$\left\{ \|\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}^0\|_\infty \leq \frac{1}{2\gamma_1\gamma_2 s_0(1+2\gamma_1^2\gamma_2)} \right\} \tag{S5}$$

the oracle estimator, $\widehat{\boldsymbol{\Omega}}_{A^0} = \arg\min_{\boldsymbol{\Omega}\succeq 0,\, \omega_{ij}=0,(i,j)\notin A^0} \operatorname{trace}(\boldsymbol{\Omega}\widehat{\boldsymbol{\Sigma}}) - \log\det(\boldsymbol{\Omega})$*, exists and is unique, and we have that*

$$\|\widehat{\boldsymbol{\Omega}}_{A^0} - \boldsymbol{\Omega}^0\|_\infty \;\; \leq \;\; 2\gamma_2 \|\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}^0\|_\infty. \tag{S6}$$

Proof of Lemma B.4: We outline the major steps of our proof. First we show that the oracle estimator, if exists, must be unique. Then we construct a mapping from an L_∞-band to itself and use the Brouwer fixed point theorem to prove the existence of an oracle estimator based on the fixed point of this mapping. Finally, the L_∞ bound (S6) for the oracle estimator follows from the the fact that the fixed point must belong to the L_∞-band.

First, we show that, if the oracle estimator exists, then it must be unique. Toward that end, we note that the objective function of the optimization problem (S4) is strongly convex in $\boldsymbol{\Omega}$, because $\nabla^2 l(\boldsymbol{\Omega}) = \boldsymbol{\Omega}^{-1} \otimes_s \boldsymbol{\Omega}^{-1} \succ 0$. Moreover, the constraint set at which $\boldsymbol{\Omega}_{(A^0)^c} = 0$ is convex. These facts, together with Theorem 27.1(e) of Rockafellar (1997), implies that the oracle estimator is unique if it exists.

Next, we prove the existence of the oracle estimator, as well as (S6) on the event (S5). For any $\boldsymbol{\Omega}$ with $\omega_{ij} = 0$ for $(i,j) \notin A^0$, let $\boldsymbol{\delta} = \operatorname{vec}_{A^0}(\boldsymbol{\Omega} - \boldsymbol{\Omega}^0)$. The construction of an oracle estimator relies on the Brouwer fixed point theorem. Toward that end, we construct a mapping $T_{A^0}(\cdot)$ and a L_∞-band $\boldsymbol{B}_r = \left\{\boldsymbol{\delta} \in \mathbb{R}^{|A^0|} : \|\boldsymbol{\delta}\|_\infty \leq r\right\}$ with a suitable $r > 0$, such that: (i) $\boldsymbol{\delta}$ is the fixed point of $T_{A^0}(\cdot)$ if and only if $\boldsymbol{\Omega} = \boldsymbol{\Omega}^0 + \boldsymbol{\Delta}$ satisfies the score equation $\operatorname{vec}_{A^0}\left(\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Omega}^{-1}\right) = \mathbf{0}$ with $\omega_{ij} = 0$ for $(i,j) \notin A^0$; (ii) $T_{A^0}(\cdot)$ maps $\boldsymbol{B}_r$ into itself; and (iii) $\boldsymbol{\Omega}^0 + \boldsymbol{\Delta} \succeq 0$ for any $\|\boldsymbol{\Delta}\|_\infty \leq r$. Suppose such a mapping can be constructed. Then by the Brouwer fixed point theorem, $T_{A^0}(\boldsymbol{\delta})$ must have a fixed point $\boldsymbol{\delta}^\star$, and it satisfies that $\|\boldsymbol{\delta}^\star\|_\infty \leq r$. Let $\boldsymbol{\Delta}^\star$ be a matrix such that $\boldsymbol{\Delta}^\star_{ij} = 0$ for any $(i,j) \notin A^0$ and $\operatorname{vec}(\boldsymbol{\Delta}^\star) = \boldsymbol{\delta}^\star$. Then, $\boldsymbol{\Omega}^\star = \boldsymbol{\Omega}^0 + \boldsymbol{\Delta}^\star$ must satisfies the score equation and is positive-definite. Hence, by Lemma B.3, $\boldsymbol{\Omega}^\star$ is an oracle estimator, and the size of the L_∞-band gives the rate of convergence of this oracle estimator in that $\|\boldsymbol{\Omega}^\star - \boldsymbol{\Omega}^0\|_\infty = \|\boldsymbol{\Delta}^\star\|_\infty \leq r$.

There remains to be shown that we can indeed construct such a mapping that satifies (i)–(iii) with $r \leq 2\gamma_2\|\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}^0\|_\infty$. The construction of the mapping $T_{A^0}(\cdot)$ is motivated by the score equation of the oracle estimator $\widehat{\boldsymbol{\Omega}}_{A^0}$. In particular, note that $\widehat{\boldsymbol{\Omega}}_{A^0}$ satisfies a score equation in terms of $p \times p$ matrix $\boldsymbol{\Omega}_{A^0}$: $\widehat{\boldsymbol{\Gamma}}_{A^0} - \{(\boldsymbol{\Omega}_{A^0})^{-1}\}_{A^0} = \mathbf{0}$, or equivalently, $\widehat{\boldsymbol{\Delta}} = \widehat{\boldsymbol{\Omega}}_{A^0} - \boldsymbol{\Omega}^0$ satisfies an equation in $\boldsymbol{\Delta}$,

$$\operatorname{vec}_{A^0}\left(\boldsymbol{\Gamma}^0 + \boldsymbol{\Lambda} - (\boldsymbol{\Delta} + \boldsymbol{\Omega}^0)^{-1}\right) \;\; = \;\; \mathbf{0}, \tag{S7}$$

where $\boldsymbol{\Lambda} = \widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}^0$. Then we define a mapping $T_{A^0}(\cdot) : \mathbb{R}^{|A^0|} \to \mathbb{R}^{|A^0|}$ as

$$T_{A^0}(\boldsymbol{\delta}) \;\; = \;\; \boldsymbol{\delta} - 2\boldsymbol{I}^{-1}_{A^0,A^0} \operatorname{vec}_{A^0}\left\{\boldsymbol{\Gamma}^0 + \boldsymbol{\Lambda} - (\boldsymbol{\Delta} + \boldsymbol{\Omega}^0)^{-1}\right\}, \tag{S8}$$

where $\boldsymbol{\delta} = \mathrm{vec}_{A^0}\big(\boldsymbol{\Delta}\big) \in \mathbb{R}^{|A^0|}$. Then $\boldsymbol{\delta}$ is the fixed point of $T_{A^0}(\cdot)$ if and only if $\boldsymbol{\Delta}$ satisfies the score equation (S7). This proves (i).

Next we show that there exists an L_∞-band $\boldsymbol{B}_r = \big\{\boldsymbol{\delta} \in \mathbb{R}^{|A^0|} : \|\boldsymbol{\delta}\|_\infty \le r\big\}$, with a suitable $r > 0$, such that $T_{A^0}(\cdot)$ maps $\boldsymbol{B}_r$ into itself. For any $\boldsymbol{\Delta}$, suppose that $\boldsymbol{\delta} = \mathrm{vec}_{A^0}\big(\boldsymbol{\Delta}\big) \in \boldsymbol{B}_r$, where r is to be determined later. We expand $(\boldsymbol{\Delta} + \boldsymbol{\Omega}^0)^{-1}$ at $\boldsymbol{\Gamma}^0$ via the matrix perturbation formula as follows,

$$\begin{aligned} &-(\boldsymbol{\Delta} + \boldsymbol{\Omega}^0)^{-1} + \boldsymbol{\Gamma}^0 = -\boldsymbol{\Gamma}^0\left(\boldsymbol{I}_{p\times p} + \boldsymbol{\Delta}\boldsymbol{\Gamma}^0\right)^{-1} + \boldsymbol{\Gamma}^0 \\ &= -\boldsymbol{\Gamma}^0 \textstyle\sum_{k=0}^\infty \left(- \boldsymbol{\Delta}\boldsymbol{\Gamma}^0\right)^k + \boldsymbol{\Gamma}^0 = \boldsymbol{\Gamma}^0\boldsymbol{\Delta}\boldsymbol{\Gamma}^0 + R(\boldsymbol{\Delta}), \end{aligned} \tag{S9}$$

where $R(\boldsymbol{\Delta}) = -\boldsymbol{\Gamma}^0 \sum_{k=2}^\infty \left(-\boldsymbol{\Delta}\boldsymbol{\Gamma}^0\right)^k$. Validity of this expansion, that is, the convergence of the above infinite series, is to be justified later by appropriately choosing a suitable radius r for $\boldsymbol{\Delta}$. Combining (S9) with (S8) leads to

$$\begin{aligned} T_{A^0}(\boldsymbol{\delta}) &= \boldsymbol{\delta} - 2\boldsymbol{I}^{-1}_{A^0,A^0}\mathrm{vec}_{A^0}\left\{-(\boldsymbol{\Delta} + \boldsymbol{\Omega}^0)^{-1} + \boldsymbol{\Gamma}^0 + \boldsymbol{\Lambda}\right\} \\ &= \boldsymbol{\delta} - 2\boldsymbol{I}^{-1}_{A^0,A^0}\mathrm{vec}_{A^0}\left(\boldsymbol{\Gamma}^0\boldsymbol{\Delta}\boldsymbol{\Gamma}^0\right) - 2\boldsymbol{I}^{-1}_{A^0,A^0}\mathrm{vec}_{A^0}\left\{R(\boldsymbol{\Delta}) + \boldsymbol{\Lambda}\right\}. \end{aligned} \tag{S10}$$

Moreover, by the definition of $\boldsymbol{I}$ and Lemma B.1-(S3), we have that $\boldsymbol{I}\,\mathrm{vec}(\boldsymbol{\Delta}) = \frac{1}{2}\,\mathrm{vec}(\boldsymbol{\Gamma}^0\boldsymbol{\Delta}\boldsymbol{\Gamma}^0)$, which further implies that

$$\boldsymbol{I}_{A^0,A^0}\mathrm{vec}_{A^0}(\boldsymbol{\Delta}) = \frac{1}{2}\,\mathrm{vec}_{A^0}\big(\boldsymbol{\Gamma}^0\boldsymbol{\Delta}\boldsymbol{\Gamma}^0\big).$$

This, together with (S10), implies that $T_{A^0}(\boldsymbol{\delta}) = -\boldsymbol{I}^{-1}_{A^0,A^0}\mathrm{vec}_{A^0}\left\{R(\boldsymbol{\Delta}) + \boldsymbol{\Lambda}\right\}$. Therefore,

$$\|T_{A^0}(\boldsymbol{\delta})\|_\infty \;\le\; 2\gamma_2\left\|\{R(\boldsymbol{\Delta})\}_{A^0}\right\|_\infty + 2\gamma_2\|\boldsymbol{\Lambda}\|_\infty.$$

To bound $\|R(\boldsymbol{\Delta})\|_\infty$, note that, for any matrix $\boldsymbol{C}$,

$$\|\boldsymbol{\Gamma}^0\boldsymbol{\Delta}\boldsymbol{\Gamma}^0\|_\infty \le \gamma_1^2\|\boldsymbol{\Delta}\|_\infty, \quad \|\boldsymbol{\Gamma}^0\boldsymbol{\Delta}\boldsymbol{C}\|_\infty \;\le\; \gamma_1\|\boldsymbol{\Delta}\boldsymbol{C}\|_\infty \le \gamma_1 s_0\|\boldsymbol{\Delta}\|_\infty\|\boldsymbol{C}\|_\infty.$$

Then $\|\boldsymbol{\Gamma}^0\big(\boldsymbol{\Delta}\boldsymbol{\Gamma}^0\big)^k\|_\infty \le \big(\gamma_1 s_0\|\boldsymbol{\Delta}\|_\infty\big)^{k-1}\gamma_1^2\|\boldsymbol{\Delta}\|_\infty$ for any integer $k \ge 1$. Henceforth,

$$\|R(\boldsymbol{\Delta})\|_\infty \le \sum_{k=2}^\infty \|\boldsymbol{\Gamma}^0\big(\boldsymbol{\Delta}\boldsymbol{\Gamma}^0\big)^k\|_\infty \le \gamma_1^3 s_0\|\boldsymbol{\Delta}\|_\infty^2 \sum_{k=1}^\infty \big(\gamma_1 s_0\|\boldsymbol{\Delta}\|_\infty\big)^{k-1} = \frac{\gamma_1^3 s_0\|\boldsymbol{\Delta}\|_\infty^2}{1 - \gamma_1 s_0\|\boldsymbol{\Delta}\|_\infty},$$

where the requirement that $\gamma_1 s_0\|\boldsymbol{\Delta}\|_\infty < 1$ would ensure the validity of the expansion in (S9), and is to be verified later by an appropriate choice of r. Now

$$\|T_{A^0}(\boldsymbol{\delta})\|_\infty \;\le\; \gamma_2\big\|R(\boldsymbol{\Delta})\big\|_\infty + \gamma_2\|\boldsymbol{\Lambda}\|_\infty \le \frac{\gamma_2\gamma_1^3 s_0\|\boldsymbol{\delta}\|_\infty^2}{1 - \gamma_1 s_0\|\boldsymbol{\delta}\|_\infty} + \gamma_2\|\boldsymbol{\Lambda}\|_\infty.$$

Henceforth, to ensure $\|T_{A^0}(\boldsymbol{\delta})\|_\infty \le r$, we require that $\frac{\gamma_2\gamma_1^3 s_0 r^2}{1-\gamma_1 s_0 r} + \gamma_2\|\boldsymbol{\Lambda}\|_\infty \le r$, or $(\gamma_1^3\gamma_2 s_0 + \gamma_1 s_0)r^2 - (1 + \gamma_1\gamma_2 s_0\|\boldsymbol{\Lambda}\|_\infty)r + \gamma_2\|\boldsymbol{\Lambda}\|_\infty \le 0$. A solution to this quadratic inequality exists if $J = (1 - \gamma_1\gamma_2 s_0\|\boldsymbol{\Lambda}\|_\infty)^2 - 4\gamma_1^3\gamma_2^2 s_0\|\boldsymbol{\Lambda}\|_\infty \ge 0$. A sufficient condition for this is

$$\Big(1 + 2\gamma_1^2\gamma_2 + \sqrt{(1 + 2\gamma_1^2\gamma_2)^2 - 1}\Big)\gamma_1\gamma_2 s_0\|\boldsymbol{\Lambda}\|_\infty \le 1, \tag{S11}$$

which is ensured by (S5).

The choice of r is important as it determines the convergence rate of the oracle estimator in the ℓ_∞ norm. We choose the smallest possible $r > 0$ satisfying the quadratic inequality,

$$r = \frac{1 + \gamma_1\gamma_2 s_0\|\mathbf{\Lambda}\|_\infty - \sqrt{I}}{2\gamma_1 s_0(1+\gamma_1^2\gamma_2)} = \frac{2\gamma_2\|\mathbf{\Lambda}\|_\infty}{1 + \gamma_1\gamma_2 s_0\|\mathbf{\Lambda}\|_\infty + \sqrt{I}}. \tag{S12}$$

This choice of r ensures that the expansion in (S9) is valid, because $\gamma_1 s_0\|\mathbf{\Delta}\|_\infty \leq \gamma_1 s_0 r \leq 2\gamma_1\gamma_2 s_0\|\mathbf{\Lambda}\|_\infty < 1$ by (S11). Moreover, the mapping $T_{A^0}(\cdot)$ constructed as above maps $\boldsymbol{B}_r$ into itself with r specified in (S12). This proves (ii).

It is also ensured that $\mathbf{\Omega}^0 + \mathbf{\Delta} \succeq 0$ for any $\|\mathbf{\Delta}\|_\infty \leq r$, because $\mathbf{\Omega}^0 + \mathbf{\Delta}$ is invertible, with its inverse equal to $\mathbf{\Gamma}^0 - \mathbf{\Gamma}^0\mathbf{\Delta}\mathbf{\Gamma}^0 - R(\mathbf{\Delta})$, for any $\|\mathbf{\Delta}\|_\infty \leq r$. This proves (iii).

Finally, an application of the fixed point theorem yields that there exists $\boldsymbol{\delta}^\star$ such that $\mathbf{\Omega}^\star = \mathbf{\Omega}^0 + \mathbf{\Delta}^\star$ satisfies the score equation, is positive-definite, and

$$\|\mathbf{\Omega}^\star - \mathbf{\Omega}^0\|_\infty = \|\mathbf{\Delta}^\star\|_\infty \leq r \leq 2\gamma_2\|\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma}^0\|_\infty.$$

This completes the proof. □

We make a few remarks regarding to these two lemmas. First, a similar convergence rate has been obtained in Zhou et al. (2011). The key difference between our result and Zhou et al. (2011) is the norm used. We focus on the convergence rate in terms of the L_∞ norm, whereas Zhou et al. (2011) used the Frobenius norm. The L_∞ norm result is more refined, in that one can easily bound the Frobenius norm by the size of the support times the L_∞ norm bound. Second, the main technique in the proof of Lemma B.4 is inspired by the fixed point argument as used in Loh and Wainwright (2014). It is different from the approach of Zhou et al. (2011).

Recall our proposed penalized estimation formulation in Equation (2) of the paper,

$$\begin{aligned} \underset{\lambda_{\max}(\mathbf{\Omega}_k)\leq R;\, k=1,\ldots,K}{\text{minimize}} \; & \sum_{k=1}^K n_k\left\{\text{trace}(\mathbf{\Omega}_k\widehat{\mathbf{\Gamma}}_k) - \log\det(\mathbf{\Omega}_k)\right\} + \\ & \sum_{k=1}^K n_k \sum_{i\neq j} p_{\lambda_{1k}}\left(|\omega_{kij}|\right) + n_{\min}\sum_{i\neq j} p_{\lambda_2}\left(\sqrt{\omega_{1ij}^2 + \cdots \omega_{Kij}^2}\right). \end{aligned}$$

Denote the objective function in this problem as $\mathcal{L}(\mathbf{\Omega}_1, \ldots, \mathbf{\Omega}_K)$. The next two lemmas give the sufficient optimality condition for this optimization problem when $\lambda_1 = \lambda_{11} = \ldots = \lambda_{1K}$ or λ_2 is equal to zero.

Lemma B.5. *If there exists $A_k \subseteq \{(i,j) : 1 \leq i \neq j \leq p\}$ such that $\widehat{\mathbf{\Omega}}_k$ and A_k satisfy the following optimality conditions:*

$$\begin{aligned} &\hat{\omega}_{kij} \neq 0 \textit{ and } \left(\widehat{\mathbf{\Omega}}_k^{-1}\right)_{ij} - \hat{\gamma}_{kij} + p'_\lambda(|\hat{\omega}_{ij}|) = 0 \textit{ for any } (i,j) \in A_k, \\ &\hat{\omega}_{kij} = 0 \textit{ and } \left|\left(\widehat{\mathbf{\Omega}}_k^{-1}\right)_{ij} - \hat{\gamma}_{kij}\right| \leq \lambda \textit{ for any } (i,j) \notin A_k, \\ &\left(\widehat{\mathbf{\Omega}}_k^{-1}\right)_{ii} = \hat{\gamma}_{kii}, i = 1, \ldots, p,\, \widehat{\mathbf{\Omega}}_k \succeq 0, \textit{ and } \lambda_{\max}(\widehat{\mathbf{\Omega}}_k) \leq \sqrt{2a}. \end{aligned} \tag{S13}$$

Then $\widehat{\boldsymbol{\Omega}}_k$ must be a solution of the optimization problem (2) of the paper when $\lambda_2 = 0$ and $R = \sqrt{2a}$, $k = 1, \ldots, K$.

Proof of Lemma B.5: We first show that the objective function $\mathcal{L}(\boldsymbol{\Omega}_1, \ldots, \boldsymbol{\Omega}_K)$ in (2) is convex when $\lambda_2 = 0$ and $R = \sqrt{2a}$. Applying B.1-(S2), we obtain that

$$\begin{aligned} \nabla^2 n_k \left\{ \mathrm{trace}(\boldsymbol{\Omega}_k \widehat{\boldsymbol{\Gamma}}_k) - \log\det(\boldsymbol{\Omega}_k) \right\} = n_k \boldsymbol{\Omega}_k^{-1} \otimes_s \boldsymbol{\Omega}_k^{-1} \quad &\succeq \quad \frac{n_{\min}}{\lambda^2_{\max}(\boldsymbol{\Omega}_k)} \boldsymbol{I}_{\frac{p(p+1)}{2} \times \frac{p(p+1)}{2}} \\ &\succeq \quad \frac{n_{\min}}{2a} \boldsymbol{I}_{\frac{p(p+1)}{2} \times \frac{p(p+1)}{2}} . \end{aligned}$$

This result, combined with the fact that $p_\lambda(x) + x^2/2a$ is convex, implies that $\mathcal{L}(\boldsymbol{\Omega}_1, \ldots, \boldsymbol{\Omega}_K)$ is convex when $\lambda_2 = 0$ and $R = \sqrt{2a}$.

Next, in view of Lemma B.2, we only need to show that (S13) constitutes conditions for $\widehat{\boldsymbol{\Omega}}_k$ to be in the constraint set, and the subgradient of the objective function contains $\mathbf{0}$ at $\widehat{\boldsymbol{\Omega}}_k$, which together would imply that $\widehat{\boldsymbol{\Omega}}_k$ must be the solution of the optimization problem (2). Toward that end, we first note that the last condition in (S13) is only the constraint of the optimization problem. We then show that the first two conditions are equivalent to saying that the subgradient of the objective function contains $\mathbf{0}$. This is true because the subderivative of the objective function with respect to ω_{kij} at $\widehat{\boldsymbol{\Omega}}_k$ is

$$\left(\widehat{\boldsymbol{\Omega}}_k^{-1} \right)_{ij} - \hat{\gamma}_{kij} + p'_\lambda(|\hat{\omega}_{kij}|),$$

when $\hat{\omega}_{kij} \neq 0$; and is the interval

$$\left(\left(\widehat{\boldsymbol{\Omega}}_k^{-1} \right)_{ij} - \hat{\gamma}_{kij} - \lambda \, , \, \left(\widehat{\boldsymbol{\Omega}}_k^{-1} \right)_{ij} - \hat{\gamma}_{kij} + \lambda \right) ,$$

when $\hat{\omega}_{kij} = 0$. Here we have used the fact that $\partial p_{\lambda_1}(0) = [-\lambda_1, \lambda_1]$. This completes the proof. □

Lemma B.6. *If there exists $A^u \subseteq \{(i,j) : 1 \leq i \neq j \leq p\}$ such that $\widehat{\boldsymbol{\Omega}}_k$ and A^u satisfy the following optimality conditions:*

$$\begin{aligned} &\sqrt{\textstyle\sum_{k=1}^K \hat{\omega}^2_{kij}} \neq 0 \textit{ for any } (i,j) \in A^u \textit{ and } \sqrt{\textstyle\sum_{k=1}^K \hat{\omega}^2_{kij}} = 0 \textit{ for any } (i,j) \notin A^u \\ &n_k \left\{ \left(\widehat{\boldsymbol{\Omega}}_k^{-1} \right)_{ij} - \hat{\gamma}_{kij} \right\} + n_{\min} \frac{\partial p_{\lambda_2}\left(\sqrt{\sum_{k=1}^K \hat{\omega}^2_{kij}} \right)}{\partial \omega_{kij}} = 0 \textit{ for } (i,j) \in A^u, k = 1, \ldots, K \, , \\ &\sqrt{\textstyle\sum_{k=1}^K n_k^2 \left\{ \left(\widehat{\boldsymbol{\Omega}}_k^{-1} \right)_{ij} - \hat{\gamma}_{kij} \right\}^2} \leq \lambda_2 n_{\min} \textit{ for } (i,j) \notin A^u, k = 1, \ldots, K \, , \\ &\left(\widehat{\boldsymbol{\Omega}}_k^{-1} \right)_{ii} = \hat{\gamma}_{kii}, \widehat{\boldsymbol{\Omega}}_k \succeq 0, \textit{ and } \lambda_{\max}(\widehat{\boldsymbol{\Omega}}_k) \leq \sqrt{2a} \textit{ for } k = 1, \ldots, K \, , \end{aligned} \tag{S14}$$

Then $\widehat{\boldsymbol{\Omega}}_k$ must be a solution of the optimization problem (2) of the paper when $\lambda_1 = 0$ and $R = \sqrt{2a}$, $k = 1, \ldots, K$.

Proof of Lemma B.6: The proof again makes use of Lemma B.2 and is similar to that of Lemma B.5. First, we show that the objective function $\mathcal{L}(\boldsymbol{\Omega}_1, \ldots, \boldsymbol{\Omega}_K)$ in (2) is convex when $\lambda_1 = 0$ and $R = \sqrt{2a}$. We apply Lemma B.1 and obtain that

$$\nabla^2 \left\{ \operatorname{trace}(\boldsymbol{\Omega}\widehat{\boldsymbol{\Gamma}}_k) - \log\det(\boldsymbol{\Omega}) \right\} = \boldsymbol{\Omega}^{-1} \otimes_s \boldsymbol{\Omega}^{-1} \succeq \frac{1}{\lambda^2_{\max}(\boldsymbol{\Omega})} \boldsymbol{I}_{\frac{p(p+1)}{2} \times \frac{p(p+1)}{2}} \succeq \frac{1}{2a} \boldsymbol{I}_{\frac{p(p+1)}{2} \times \frac{p(p+1)}{2}}.$$

This result, combined with the fact that $p_\lambda(x) + \frac{x^2}{2a}$ is convex, implies that $\mathcal{L}(\boldsymbol{\Omega}_1, \ldots, \boldsymbol{\Omega}_K)$ is convex when $\lambda_1 = 0$ and $R = \sqrt{2a}$.

Next, in view of Lemma B.2 again, we only need to show that (S14) constitutes conditions for $\widehat{\boldsymbol{\Omega}}_k$ to be in the constraint set, and the subgradient of the objective function contains $\mathbf{0}$ at $\widehat{\boldsymbol{\Omega}}_k$. Toward that end, we note that the last condition in (S14) is the constraint set. We then show that the first three conditions ensure that $\mathbf{0}$ is contained in the subgradient of the objective function. This is true because the subderivative of the objective function with respect to ω_{kij} at $\widehat{\boldsymbol{\Omega}}_k$ is

$$n_k \left\{ \left(\widehat{\boldsymbol{\Omega}}_k^{-1} \right)_{ij} - \hat{\gamma}_{kij} \right\} + n_{\min} \frac{\partial p_{\lambda_2} \left(\sqrt{\sum_{k=1}^K \hat{\omega}_{kij}^2} \right)}{\partial \omega_{kij}}$$

when $\sqrt{\sum_{k=1}^K \hat{\omega}_{kij}^2} \neq 0$; and is the interval

$$\left[\sqrt{\sum_{k=1}^K n_k^2 \left\{ \left(\widehat{\boldsymbol{\Omega}}_k^{-1} \right)_{ij} - \hat{\gamma}_{kij} \right\}^2} - \lambda_2 n_{\min} \, , \, \sqrt{\sum_{k=1}^K n_k^2 \left\{ \left(\widehat{\boldsymbol{\Omega}}_k^{-1} \right)_{ij} - \hat{\gamma}_{kij} \right\}^2} + \lambda_2 n_{\min} \right]$$

when $\sqrt{\sum_{k=1}^K \hat{\omega}_{kij}^2} = 0$. Here we have used the fact that the subgradient of $p_{\lambda_2}(\| \cdot \|_2)$ at $\mathbf{0}$ is $\partial p_{\lambda_2}(\|\boldsymbol{x}\|_2)\big|_{\boldsymbol{x}=\mathbf{0}} = \{\boldsymbol{x} : \|\boldsymbol{x}\|_2 \leq \lambda_2\}$. This completes the proof. □

Lemmas B.5 and B.6 provide sets of sufficient optimality conditions for the solution of problem (2) in the paper, which are to be used later in the proofs of Theorem 1 and 2. Finally, we have the following result for $\widehat{\boldsymbol{\Gamma}}_k$ in our optimization problem (2).

Lemma B.7. *Under Assumption (A1), we have that*

$$\mathbb{P}\left(\|\widehat{\boldsymbol{\Gamma}}_k - \boldsymbol{\Gamma}_k^0\|_\infty \geq c_3 \sqrt{\frac{\log(p \vee q)}{nq}} \right) \leq 1 - \frac{2}{(p \vee q)^2}, \quad k = 1, \cdots, K,$$

where c_3 is some constant.

Proof of Lemma B.7: By Theorem 4.1 of Zhou (2014), we have that

$$\mathbb{P}\left(\|\widehat{\boldsymbol{\Gamma}}_k - \boldsymbol{\Gamma}_k^0\|_\infty \geq c_3' \frac{\sqrt{q}\|\Sigma_{kT}^0\|_F}{\operatorname{trace}(\Sigma_{kT}^0)} \sqrt{\frac{\log(p \vee q)}{nq}} \right) \leq 1 - \frac{2}{(p \vee q)^2}, \quad k = 1, \cdots, K,$$

for some constant c_3'. Then the conclusion of this lemma follows immediately, by setting $c_3 = c_3' c_0^2$, and the inequality that $\frac{\sqrt{q}\|\Sigma_{kT}^0\|_F}{\operatorname{trace}(\Sigma_{kT}^0)} \leq \kappa(\Sigma_{kT}^0) \leq c_0^2$. □

C Proof of Theorem 1

We first outline the main steps of our proof. Our aim is to prove that, there exist λ_1 and a such that the oracle estimator $\widehat{\mathbf{\Omega}}_{k,A_k^0}$ is indeed a minimizer of the optimization problem (2) of the paper when $\lambda_2 = 0$. By Lemma B.5, we only need to show that the oracle estimator satisfies (S13). We plan to verify this in three steps.

In the first step, we plan to show that, when

$$\lambda_1 \leq a^{-1}\left\{\min_{(i,j)\in A_k^0}|\omega_{kij}^0| - 2c_2c_3\sqrt{\frac{\log(p\vee q)}{n_kq}}\right\}, \tag{S15}$$

with probability at least $1-\frac{2K}{(p\vee q)^2}$, we have $\left(\widehat{\mathbf{\Omega}}_{k,A_k^0}^{-1}\right)_{ij} - \hat{\gamma}_{kij} + \frac{\partial p_{\lambda_1}(|\hat{\omega}_{kij}|)}{\partial \omega_{kij}} = 0$ holds for any $(i,j)\in A_k^0, k=1,\ldots,K$. Sufficiently, we plan to show that

$$|\hat{\omega}_{kij}| \geq a\lambda_1 \text{ and } \left(\widehat{\mathbf{\Omega}}_{k,A_k^0}^{-1}\right)_{ij} - \hat{\gamma}_{kij} = 0 \text{ for any } (i,j)\in A_k^0 \text{ and } k=1,\ldots,K. \tag{S16}$$

In the second step, we plan to show that, when

$$\lambda_1 \geq 2(1+c_1^2c_2)c_3\sqrt{\frac{\log(p\vee q)}{n_kq}}, \tag{S17}$$

with probability at least $1-\frac{2K}{(p\vee q)^2}$, we have

$$\left|\left(\widehat{\mathbf{\Omega}}_{k,A_k^0}^{-1}\right)_{ij} - \hat{\gamma}_{kij}\right| \leq \lambda_1 \text{ for any } (i,j)\notin A_k^0 \text{ and } k=1,\ldots,K\,. \tag{S18}$$

Finally, in the third step, we plan to show that, when a satisfies that

$$a > \frac{1}{2}\left\{c_0 + 4c_2c_3s_0\sqrt{\frac{\log(p\vee q)}{n_kq}}\right\}^2, \tag{S19}$$

the oracle estimator satisfies that, with probability at least $1-\frac{2K}{(p\vee q)^2}$,

$$\lambda_{\max}(\widehat{\mathbf{\Omega}}_{k,A_k^0}) \leq \sqrt{2a} \text{ for } k=1,\ldots,K. \tag{S20}$$

For step 1, we prove (S16). Note that $\widehat{\mathbf{\Omega}}_{k,A_k^0}$ is the oracle estimator over $A_k^0, k = 1,\ldots,K$. By applying Lemma B.3 with $A^0 = A_k^0$, we have that $\left(\widehat{\mathbf{\Omega}}_{k,A_k^0}^{-1}\right)_{ij} - \hat{\gamma}_{kij} = 0$ for any $(i,j)\in A_k^0$, $k=1,\ldots,K$. This proves the second part in (S16). For the first part in (S16), by the triangular inequality, Lemma B.4, and Lemma B.7, with probability at least $1-\frac{2}{(p\vee q)^2}$, we have that, for each $k=1,\ldots,K$,

$$\begin{aligned}\min_{(i,j)\in A_k^0}|\hat{\omega}_{kij}| &\geq \min_{(i,j)\in A_k^0}|\omega_{kij}^0| - \max_{(i,j)\in A_k^0}|\hat{\omega}_{kij}-\omega_{kij}^0| \geq \min_{(i,j)\in A_k^0}|\omega_{kij}^0| - 2c_2\|\widehat{\mathbf{\Gamma}}_k - \mathbf{\Gamma}_k^0\|_\infty \\ &\geq a\lambda_1 + 2c_2c_3\sqrt{\frac{\log(p\vee q)}{n_kq}} - 2c_2\|\widehat{\mathbf{\Gamma}}_k - \mathbf{\Gamma}_k^0\|_\infty \geq a\lambda_1,\end{aligned}$$

where the second to last inequality uses (S15). Henceforth, with probability at least $1 - \frac{2K}{(p\vee q)^2}$, (S16) holds for all $k = 1, \dots, K$.

For step 2, we prove (S18). Let $\widehat{\boldsymbol{\Delta}}_k = \widehat{\boldsymbol{\Omega}}_{k,A_k^0} - \boldsymbol{\Omega}_k^0$, and $\boldsymbol{\Lambda}_k = \widehat{\boldsymbol{\Gamma}}_k - \boldsymbol{\Gamma}_k^0$. Note that under Assumption (A2),

$$\begin{aligned}
\big\|\widehat{\boldsymbol{\Omega}}_{k,A_k^0}^{-1} - \widehat{\boldsymbol{\Gamma}}_k\big\|_\infty &= \big\|(\widehat{\boldsymbol{\Delta}}_k + \boldsymbol{\Omega}_k^0)^{-1} - \boldsymbol{\Gamma}_k^0 - \boldsymbol{\Lambda}_k\big\|_\infty \le \big\|(\widehat{\boldsymbol{\Delta}}_k + \boldsymbol{\Omega}_k^0)^{-1} - \boldsymbol{\Gamma}_k^0\big\|_\infty + \|\boldsymbol{\Lambda}_k\|_\infty \\
&= \big\|\sum_{j=1}^{\infty} \boldsymbol{\Gamma}_k^0(-\widehat{\boldsymbol{\Delta}}_k \boldsymbol{\Gamma}_k^0)^j\big\|_\infty + \|\boldsymbol{\Lambda}_k\|_\infty \le c_1^2 \|\widehat{\boldsymbol{\Delta}}_k\|_\infty + \big\|R(\widehat{\boldsymbol{\Delta}})\big\|_\infty + \|\boldsymbol{\Lambda}\|_\infty \\
&\le 2c_1^2 c_2 \|\boldsymbol{\Lambda}_k\|_\infty + \|\boldsymbol{\Lambda}_k\|_\infty + \|\boldsymbol{\Lambda}_k\|_\infty = 2(1 + c_1^2 c_2)\|\boldsymbol{\Lambda}_k\|_\infty,
\end{aligned}$$

where the second to last inequality uses the fact that $\|\widehat{\boldsymbol{\Delta}}_k\|_\infty \le \|\boldsymbol{\Lambda}_k\|$ and Lemma B.4. This, together with (S17), implies that, for each $k = 1, \dots, K$,

$$\begin{aligned}
\mathbb{P}\left(\big\|\widehat{\boldsymbol{\Omega}}_{k,A_k^0}^{-1} - \widehat{\boldsymbol{\Gamma}}_k\big\|_\infty \le \lambda_1\right) &\ge \mathbb{P}\left(2(1 + c_1^2 c_2)\|\boldsymbol{\Lambda}_k\|_\infty \le \lambda_1\right) \\
&\ge \mathbb{P}\left(\|\boldsymbol{\Lambda}_k\|_\infty \le 2c_3\sqrt{\frac{\log(p\vee q)}{n_k q}}\right) \ge 1 - \frac{2}{(p\vee q)^2}.
\end{aligned}$$

Henthforth, with probability at least $1 - \frac{2K}{(p\vee q)^2}$, (S18) holds for all $k = 1, \dots, K$.

For step 3, we prove (S20); i.e., we show that $\widehat{\boldsymbol{\Omega}}_{k,A_k^0}$ are the interior points of the constraints $\lambda_{\max}(\boldsymbol{\Omega}_k) \le \sqrt{2a}; k = 1, \cdots, K$, with probability at least $1 - \frac{2K}{(p\vee q)^2}$. We first note that,

$$\lambda_{\max}(\widehat{\boldsymbol{\Omega}}_{k,A_k^0}) = \lambda_{\max}(\boldsymbol{\Omega}_{k,A_k^0}^0 + \widehat{\boldsymbol{\Delta}}_k) \le \lambda_{\max}(\boldsymbol{\Omega}_{k,A_k^0}^0) + \lambda_{\max}(\widehat{\boldsymbol{\Delta}}_k) \le c_0 + \lambda_{\max}(\widehat{\boldsymbol{\Delta}}_k). \quad \text{(S21)}$$

To bound $\lambda_{\max}(\widehat{\boldsymbol{\Delta}}_k)$, we note that

$$\begin{aligned}
\lambda_{\max}(\widehat{\boldsymbol{\Delta}}_k) &= \sup_{\|\boldsymbol{u}\|_2=1} \boldsymbol{u}^T \widehat{\boldsymbol{\Delta}} \boldsymbol{u} = \sup_{\|\boldsymbol{u}\|_2=1} \sum_{(i,j)\in A_k^0} u_i u_j \widehat{\Delta}_{kij} \le \|\widehat{\boldsymbol{\Delta}}_k\|_\infty \sup_{\|\boldsymbol{u}\|_2=1} \sum_{(i,j)\in A^0} |u_i u_j| \\
&\le \|\widehat{\boldsymbol{\Delta}}_k\|_\infty \sup_{\|\boldsymbol{u}\|_2=1} \sqrt{\Big(\sum_{(i,j)\in A^0} u_i^2\Big)\Big(\sum_{(i,j)\in A^0} u_j^2\Big)} \\
&\le \|\widehat{\boldsymbol{\Delta}}_k\|_\infty \sup_{\|\boldsymbol{u}\|_2=1} \sqrt{s_0\|\boldsymbol{u}\|_2^2 s_0 \|\boldsymbol{u}\|_2^2} = s_0 \|\widehat{\boldsymbol{\Delta}}_k\|_\infty.
\end{aligned}$$

Then, applying Lemmas B.4 and B.7, we obtain that $\lambda_{\max}(\widehat{\boldsymbol{\Delta}}_k) \le 2c_2 s_0 \max_k \|\boldsymbol{\Lambda}_k\|_\infty \le 4c_2 c_3 s_0 \sqrt{\frac{\log(p\vee q)}{n_k q}}$, with probability at least $1 - \frac{2}{(p\vee q)^2}$. This result, in combination with (S19) and (S21), implies that $\lambda_{\max}(\widehat{\boldsymbol{\Omega}}_{k,A_k^0}) \le c_0 + 4c_2 c_3 s_0 \sqrt{\frac{\log(p\vee q)}{n_k q}} \le \sqrt{2a}$. That is, (S20) holds with probability at least $1 - \frac{2K}{(p\vee q)^2}$.

Combining steps 1 to 3, if the tuning parameters λ_1, R, and a satisfy the condition

$$\begin{aligned}
&2(1 + c_1^2 c_2) c_3 \sqrt{\tfrac{\log(p\vee q)}{n_k q}} \le \lambda_1 \le a^{-1}\left\{\min_{(i,j)\in A^u} |\omega_{kij}^0| - 2c_2 c_3 \sqrt{\tfrac{\log(p\vee q)}{n_k q}}\right\} \\
&R = \sqrt{2a} \quad \text{and} \quad a > \frac{\left(c_0 + 4c_2 c_3 s_0 \sqrt{\frac{\log(p\vee q)}{n_k q}}\right)^2}{2},
\end{aligned} \quad \text{(S22)}$$

then with probability at least $1-\frac{6K}{(p\vee q)^2}$, the optimization problem (2) is convex and the oracle estimator is the unique minimizer of (2). The existence of λ_1 and a that satisfies (S22) is ensured by

$$\min_{(i,j)\in A_k^0}|\omega_{kij}^0| \geq \left\{c_0+4c_2c_3s_0\sqrt{\frac{\log(p\vee q)}{n_kq}}\right\}^2(1+c_1^2c_2)c_3\sqrt{\frac{\log(p\vee q)}{n_kq}}+2c_2c_3\sqrt{\frac{\log(p\vee q)}{n_kq}},$$

which is true due to Assumption (A2) and the minimum signal condition (6) of Theorem 1. This completes the proof. □

D Proof of Theorem 2

The proof of this theorem follows a similar structure as that of Thoerem 1. Again we first outline the main steps of the proof. Our aim is to prove that, there exist λ_2 and a such that the oracle estimator $\widehat{\mathbf{\Omega}}_{k,A^u}$ is indeed a minimizer of the optimization problem (2) of the paper when $\lambda_1=0$. By Lemma B.6, we only need to show that the oracle estimator satisfies (S14). We plan to verify it in three steps.

In the first step, we plan to show that, when

$$\lambda_2 \leq a^{-1}\left(\min_{(i,j)\in A^u}\sqrt{\sum_{k=1}^K\left(\omega_{kij}^0\right)^2}-2c_2c_3\sqrt{\frac{K\log(p\vee q)}{n_{\min}q}}\right) \tag{S23}$$

with probability at least $1-\frac{2K}{(p\vee q)^2}$, we have

$$n_k\left(\left[\widehat{\mathbf{\Omega}}_{k,A^u}^{-1}\right]_{ij}-\hat{\gamma}_{kij}\right)+n_{\min}\frac{\partial p_{\lambda_2}\left(\sqrt{\sum_{k=1}^K\hat{\omega}_{kij}^2}\right)}{\partial\omega_{kij}}=0$$

for any $(i,j)\in A^u$ and $k=1,\ldots,K$. Sufficiently, we plan to show that,

$$\sqrt{\sum_{k=1}^K\hat{\omega}_{kij}^2}\geq a\lambda_2 \text{ and } \left[\widehat{\mathbf{\Omega}}_{k,A^u}^{-1}\right]_{ij}=\hat{\gamma}_{kij} \text{ for any } (i,j)\in A^u \text{ and } k=1,\ldots,K. \tag{S24}$$

In the second step, we plan to show that, when λ_2 satisfies

$$\frac{n_{\min}\lambda_2}{n_k\sqrt{K}}\geq 2(1+c_1^2c_2)c_3\sqrt{\frac{\log(p\vee q)}{n_kq}} \text{ for any } k=1,\cdots,K\,, \tag{S25}$$

with probability at least $1-\frac{2K}{(p\vee q)^2}$, we have that

$$\sqrt{\sum_{k=1}^K n_k^2\left\{\left(\widehat{\mathbf{\Omega}}_{k,A^u}^{-1}\right)_{ij}-\hat{\gamma}_{kij}\right\}^2}\leq\lambda_2 n_{\min} \text{ for any } (i,j)\notin A^u \tag{S26}$$

Sufficiently, we plan to show that

$$\left|\left(\widehat{\mathbf{\Omega}}_{k,A^u}^{-1}\right)_{ij} - \hat{\gamma}_{kij}\right| \leq \frac{n_{\min}\lambda_2}{n_k\sqrt{K}}; \text{ for } (i,j) \notin A^u \text{ and } k = 1,\ldots,K. \tag{S27}$$

Finally, in the third step, we plan to show that, when a satisfies that

$$a \geq \frac{1}{2}\left(c_0 + 4c_2c_3\tilde{s}_0\sqrt{\frac{\log(p\vee q)}{n_{\min}q}}\right)^2 \tag{S28}$$

the oracle estimator satisfies that, with probability at least $1 - \frac{2K}{(p\vee q)^2}$,

$$\lambda_{\max}(\widehat{\mathbf{\Omega}}_{k,A^u}) \leq \sqrt{2a} \text{ for } k = 1,\ldots,K. \tag{S29}$$

For step 1, we prove (S24). Note that $\widehat{\mathbf{\Omega}}_{k,A^u}$ is the oracle estimator over A^u. By applying Lemma B.3 with $A^0 = A^u$, we have that $\left(\widehat{\mathbf{\Omega}}_{k,A^u}^{-1}\right)_{ij} - \hat{\gamma}_{kij} = 0$ for any $(i,j) \in A^u$ and $k = 1,\ldots,K$. Moreover, by triangular inequality, the condition (S23), and Lemma B.7, with probability at least $1 - \frac{2}{(p\vee q)^2}$, we have that, for each $k = 1,\ldots,K$,

$$\begin{aligned}
\min_{(i,j)\in A^u}\sqrt{\sum_{k=1}^K \hat{\omega}_{kij}^2} &\geq \min_{(i,j)\in A^u}\sqrt{\sum_{k=1}^K \left(\omega_{kij}^0\right)^2} - \max_{(i,j)\in A^u}\sqrt{\sum_{k=1}^K(\hat{\omega}_{kij} - \omega_{kij}^0)^2} \\
&\geq a\lambda_2 + 2c_2c_3\sqrt{\frac{K\log(p\vee q)}{n_{\min}q}} - \sqrt{K}\max_{1\leq k\leq K}\|\widehat{\mathbf{\Omega}}_{k,A^u} - \mathbf{\Omega}_k^0\|_\infty \\
&\geq a\lambda_2 + 2c_2c_3\sqrt{\frac{K\log(p\vee q)}{n_{\min}q}} - 2c_2\sqrt{K}\max_{1\leq k\leq K}\|\widehat{\mathbf{\Gamma}}_k - \mathbf{\Gamma}_k^0\|_\infty \geq a\lambda_2
\end{aligned}$$

Henceforth, with probability at least $1 - \frac{2K}{(p\vee q)^2}$, (S24) holds.

For step 2, we prove (S26) and (S27). Let $\widehat{\mathbf{\Delta}}_k = \widehat{\mathbf{\Omega}}_{k,A^u} - \mathbf{\Omega}_k^0$, and $\mathbf{\Lambda}_k = \widehat{\mathbf{\Gamma}}_k - \mathbf{\Gamma}_k^0$. Note that under Assumption (A2),

$$\begin{aligned}
\left\|\widehat{\mathbf{\Omega}}_{k,A^u}^{-1} - \widehat{\mathbf{\Gamma}}_k\right\|_\infty &= \left\|(\widehat{\mathbf{\Delta}}_k + \mathbf{\Omega}_k^0)^{-1} - \mathbf{\Gamma}_k^0 - \mathbf{\Lambda}_k\right\|_\infty \leq \left\|(\widehat{\mathbf{\Delta}}_k + \mathbf{\Omega}_k^0)^{-1} - \mathbf{\Gamma}_k^0\right\|_\infty + \|\mathbf{\Lambda}_k\|_\infty \\
&= \left\|\sum_{j=1}^\infty \mathbf{\Gamma}_k^0(-\widehat{\mathbf{\Delta}}_k\mathbf{\Gamma}_k^0)^j\right\|_\infty + \|\mathbf{\Lambda}_k\|_\infty \leq c_1^2\|\widehat{\mathbf{\Delta}}_k\|_\infty + \left\|R(\widehat{\mathbf{\Delta}})\right\|_\infty + \|\mathbf{\Lambda}\|_\infty \\
&\leq 2c_1^2c_2\|\mathbf{\Lambda}_k\|_\infty + \|\mathbf{\Lambda}_k\|_\infty + \|\mathbf{\Lambda}_k\|_\infty = 2(1 + c_1^2c_2)\|\mathbf{\Lambda}_k\|_\infty,
\end{aligned}$$

where the second to last inequality uses the fact that $\|\widehat{\mathbf{\Delta}}_k\|_\infty \leq \|\mathbf{\Lambda}_k\|$ and Lemma B.4. This, together with (S25), implies that, for each $k = 1,\ldots,K$,

$$\begin{aligned}
\mathbb{P}\left(\left\|\widehat{\mathbf{\Omega}}_{k,A^u}^{-1} - \widehat{\mathbf{\Gamma}}_k\right\|_\infty \leq \frac{n_{\min}\lambda_2}{n_k\sqrt{K}}\right) &\geq \mathbb{P}\left(2(1 + c_1^2c_2)\|\mathbf{\Lambda}_k\|_\infty \leq \frac{n_{\min}\lambda_2}{n_k\sqrt{K}}\right) \\
&\geq \mathbb{P}\left(\|\mathbf{\Lambda}_k\|_\infty \leq 2c_3\sqrt{\frac{\log(p\vee q)}{n_kq}}\right) \geq 1 - \frac{2}{(p\vee q)^2}.
\end{aligned}$$

Henceforth, with probability at least $1-\frac{2K}{(p\vee q)^2}$, (S26) and (S27) hold.

For step 3, we prove (S29); i.e., we show that $\widehat{\boldsymbol{\Omega}}_{k,A^u}$ are the interior points of the constraints $\lambda_{\max}(\boldsymbol{\Omega}_k)\le\sqrt{2a}; k=1,\cdots,K$, with probability at least $1-\frac{2K}{(p\vee q)^2}$. We first note that,

$$\lambda_{\max}(\widehat{\boldsymbol{\Omega}}_{k,A^u})=\lambda_{\max}(\boldsymbol{\Omega}^0_{k,A^u}+\widehat{\boldsymbol{\Delta}}_k)\le\lambda_{\max}(\boldsymbol{\Omega}^0_{k,A^u})+\lambda_{\max}(\widehat{\boldsymbol{\Delta}}_k)\le c_0+\lambda_{\max}(\widehat{\boldsymbol{\Delta}}_k). \quad \text{(S30)}$$

To bound $\lambda_{\max}(\widehat{\boldsymbol{\Delta}}_k)$, we note that

$$\begin{aligned}\lambda_{\max}(\widehat{\boldsymbol{\Delta}}_k) &= \sup_{\|\boldsymbol{u}\|_2=1}\boldsymbol{u}^T\widehat{\boldsymbol{\Delta}}\boldsymbol{u}=\sup_{\|\boldsymbol{u}\|_2=1}\sum_{(i,j)\in A_k^0}u_iu_j\widehat{\Delta}_{kij}\le\|\widehat{\boldsymbol{\Delta}}_k\|_\infty\sup_{\|\boldsymbol{u}\|_2=1}\sum_{(i,j)\in A^0}|u_iu_j|\\ &\le \|\widehat{\boldsymbol{\Delta}}_k\|_\infty\sup_{\|\boldsymbol{u}\|_2=1}\sqrt{\Big(\sum_{(i,j)\in A^0}u_i^2\Big)\Big(\sum_{(i,j)\in A^0}u_j^2\Big)}\\ \tilde{s}_0 &\le \|\widehat{\boldsymbol{\Delta}}_k\|_\infty\sup_{\|\boldsymbol{u}\|_2=1}\sqrt{\|\boldsymbol{u}\|_2^2\tilde{s}_0\|\boldsymbol{u}\|_2^2}=\tilde{s}_0\|\widehat{\boldsymbol{\Delta}}_k\|_\infty.\end{aligned}$$

Then, applying Lemmas B.7 and B.4, we obtain that $\lambda_{\max}(\widehat{\boldsymbol{\Delta}}_k)\le 2c_2\tilde{s}_0\max_k\|\boldsymbol{\Lambda}_k\|_\infty\le 4c_2c_3\tilde{s}_0\sqrt{\frac{\log(p\vee q)}{n_kq}}$, with probability at least $1-\frac{2}{(p\vee q)^2}$. This result, in combination with (S19) and (S30), implies that $\lambda_{\max}(\widehat{\boldsymbol{\Omega}}_{k,A^u})\le c_0+4c_2c_3\tilde{s}_0\sqrt{\frac{\log(p\vee q)}{n_kq}}\le\sqrt{2a}$. That is, (S29) holds with probability at least $1-\frac{2K}{(p\vee q)^2}$.

Combining steps 1 to 3, if the tuning parameters λ_2, R, and a satisfy the condition

$$\begin{aligned}&\tfrac{2(1+c_1^2c_2)c_3\sqrt{n_{\max}}}{\sqrt{n_{\min}}}\sqrt{\tfrac{K\log(p\vee q)}{n_{\min}q}}\le\lambda_2\le\frac{\min\limits_{(i,j)\in A^u}\sqrt{\sum_{k=1}^K\left(\omega^0_{kij}\right)^2}-2c_2c_3\sqrt{\frac{K\log(p\vee q)}{n_{\min}q}}}{a},\\ &R=\sqrt{2a}\quad\text{and}\quad a\ge\tfrac{1}{2}\left(c_0+4c_2c_3\tilde{s}_0\sqrt{\tfrac{\log(p\vee q)}{n_{\min}q}}\right)^2,\end{aligned} \quad \text{(S31)}$$

then with probability at least $1-\frac{6K}{(p\vee q)^2}$, the optimization problem (2) is convex and the oracle estimator is the unique minimizer of (2). The existence of λ_1 and a that satisfies (S22) is ensured by

$$\begin{aligned}\min_{(i,j)\in A^u}\sqrt{\sum_{k=1}^K\left(\omega^0_{kij}\right)^2} &\ge 2c_2c_3\sqrt{\frac{K\log(p\vee q)}{n_{\min}q}}+\\ &\left\{c_0+4c_2c_3\tilde{s}_0\sqrt{\frac{\log(p\vee q)}{n_{\min}q}}\right\}^2\frac{(1+c_1^2c_2)c_3\sqrt{n_{\max}}}{\sqrt{n_{\min}}}\sqrt{\frac{K\log(p\vee q)}{n_{\min}q}},\end{aligned}$$

which is true due to Assumption (A3) and the minimum signal condition (8) of Theorem 2. This completes the proof. □

E Additional simulations

We report here the simulation results in Section 6 of the paper when the sample size $n_k = 10$. Tables S1-S3 correspond to Tables 1-3 of the paper.

We also graphically report the F_1 score, which is a composite measure of the accuracy of sparsity identification. It is defined as

$$F_1 = \left\{ \text{true positive rate}^{-1} + (1 - \text{false positive rate})^{-1} \right\}^{-1}.$$

In Figure S1, we report the F_1 scores when $n_k = 20$, $p = 200$, and $q = 100$.

Finally, as a further illustration, we report in Figure S2 the computational time, in seconds, when the number of network nodes gradually increases from $p = 25$ to $p = 500$, with the sample size fixed at $n_k = 10$, and the temporal dimension $q = 50$. It is seen that, for all three graph structures, our method is comparable to the convex solution in terms of running time, but is much faster than Lee and Liu (2015) and Cai et al. (2016), especially when the graph dimension p is large. This plot also reflects, to some extent, the scalability of our approach to networks with the number of nodes up to a few hundreds.

References

Alizadeh, F., Haeberly, J. A., and Overton, M. L. (1998). Primal-dual interior-point methods for semidefinite programming: convergence rates, stability and numerical results. *SIAM Journal on Optimization*, 8(3):746–768.

Cai, T. T., Li, H., Liu, W., and Xie, J. (2016). Joint estimation of multiple high-dimensional precision matrices. *Statistica Sinica*, 26:445–464.

Hiriart-Urruty, J.-B. and Lemaréchal, C. (2012). *Fundamentals of convex analysis*. Springer Science & Business Media.

Lee, W. and Liu, Y. (2015). Joint estimation of multiple precision matrices with common structures. *Journal of Machine Learning Research*, 16:1035–1062.

Loh, P.-L. and Wainwright, M. J. (2014). Support recovery without incoherence: A case for nonconvex regularization. *arXiv preprint arXiv:1412.5632*.

Rockafellar, R. T. (1997). Convex analysis. princeton landmarks in mathematics.

Zhou, S. (2014). Gemini: graph estimation with matrix variate normal instances. *Ann. Statist.*, 42(2):532–562.

Zhou, S., Rütimann, P., Xu, M., and Bühlmann, P. (2011). High-dimensional covariance estimation based on gaussian graphical models. *Journal of Machine Learning Research*, 12(Oct):2975–3026.

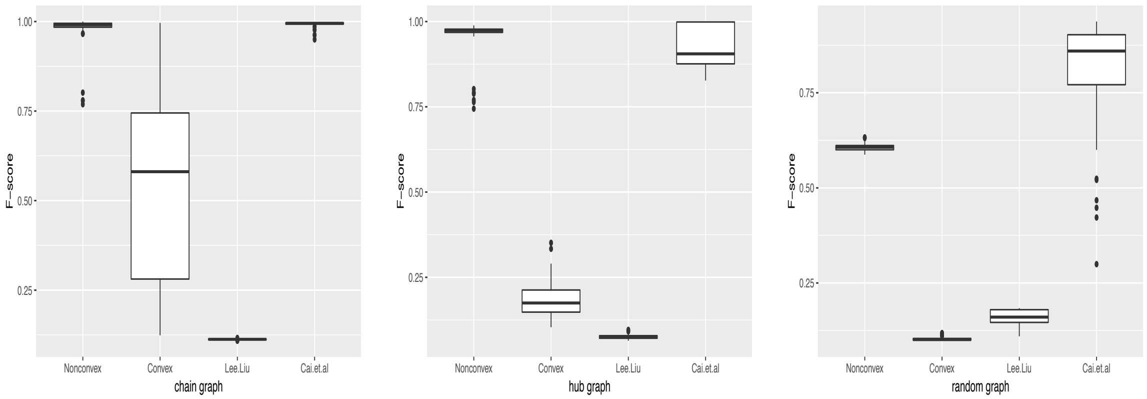


Figure S1: The F_1 score comparison of our method, the convex counterpart, Lee and Liu (2015), and Cai et al. (2016), with $n_k = 20$, $p = 200$, and $q = 100$.

Table S1: Chain graph. Reported are the average and standard deviation (in parenthesis) of the accuracy criteria based on 100 data replications. Also reported is the average running time (in seconds). Evaluation criteria include the false positive rate (FP), the false negative rate (FN), the entropy loss (EL_k), and the quadratic loss (QL_k). We compare the proposed nonconvex based multi-graph estimation method (denoted as Nonconvex) with its convex counterpart (denoted as Convex), the method of Lee and Liu (2015) (denoted as Lee & Liu), and the method of Cai et al. (2016) (denoted as Cai et al.).

n_k	p	q	Method	FP	FN	EL_1	EL_2	QL_1	QL_2	Time
10	100	100	Nonconvex	0.002 (0.004)	0.000 (0.000)	0.201 (0.033)	0.221 (0.028)	0.499 (0.083)	0.561 (0.071)	161
			Convex	0.053 (0.049)	0.000 (0.000)	7.59 (3.33)	6.460 (2.510)	15.50 (7.460)	13.80 (5.990)	149
			Lee & Liu	0.379 (0.009)	0.000 (0.000)	1.801 (0.051)	1.068 (0.049)	4.761 (0.159)	2.838 (0.146)	1159
			Cai et al.	7e-04 (7e-04)	0.000 (0.000)	7.200 (0.780)	9.500 (1.500)	22.20 (2.400)	28.70 (4.600)	634
		50	Nonconvex	0.001 (0.001)	0.000 (0.001)	0.411 (0.060)	0.502 (0.098)	1.04 (0.158)	1.290 (0.253)	326
			Convex	0.035 (0.015)	0.000 (0.001)	6.950 (2.280)	6.080 (1.560)	13.60 (5.040)	12.50 (3.700)	287
			Lee & Liu	0.270 (0.007)	0.000 (0.000)	2.849 (0.123)	1.597 (0.087)	7.597 (0.345)	4.499 (0.283)	1381
			Cai et al.	0.002 (0.002)	0.000 (0.000)	4.400 (0.430)	5.300 (0.390)	13.00 (1.100)	15.50 (1.300)	615
	200	100	Nonconvex	0.000 (0.000)	0.000 (0.000)	0.406 (0.043)	0.415 (0.039)	1.010 (0.107)	0.997 (0.095)	805
			Convex	0.019 (0.022)	0.000 (0.000)	15.00 (6.630)	10.80 (3.950)	30.10 (14.80)	22.3 (9.000)	720
			Lee & Liu	0.213 (0.035)	0.000 (0.000)	3.400 (0.096)	1.900 (0.170)	8.900 (0.250)	5.000 (0.480)	11769
			Cai et al.	6e-04 (0.001)	0.000 (0.000)	8.200 (0.930)	10.10 (1.500)	23.70 (2.100)	28.00 (3.600)	7790
		50	Nonconvex	0.000 (0.000)	0.000 (0.000)	0.817 (0.083)	0.855 (0.090)	2.060 (0.216)	2.100 (0.230)	2579
			Convex	0.015 (0.008)	0.000 (0.001)	16.00 (5.320)	13.00 (3.530)	31.40 (11.80)	26.40 (8.060)	2307
			Lee & Liu	0.215 (0.010)	0.000 (0.000)	5.200 (0.140)	3.100 (0.120)	14.000 (0.410)	8.400 (0.390)	24491
			Cai et al.	0.004 (9e-04)	0.000 (0.000)	5.900 (0.240)	5.000 (0.210)	16.00 (0.720)	13.70 (0.670)	6797

Table S2: Hub graph. The setup is the same as Table S1.

n_k	p	q	Method	FP	FN	EL$_1$	EL$_2$	QL$_1$	QL$_2$	Time
10	100	100	Nonconvex	0.008 (0.007)	0.002 (0.003)	0.190 (0.029)	0.320 (0.058)	0.920 (0.175)	1.120 (0.189)	302
			Convex	0.172 (0.054)	0.011 (0.1)	3.410 (1.880)	3.030 (1.490)	11.20 (7.210)	9.020 (5.350)	241
			Lee & Liu	0.388 (0.014)	0.000 (0.000)	1.600 (0.061)	1.300 (0.063)	5.700 (0.390)	5.400 (0.520)	1096
			Cai et al.	0.004 (0.003)	0.000 (0.000)	16.70 (2.000)	18.30 (2.000)	404.8 (81.20)	416.1 (79.10)	600
		50	Nonconvex	0.002 (0.001)	0.029 (0.011)	0.383 (0.067)	0.856 (0.129)	1.910 (0.431)	3.080 (0.493)	483
			Convex	0.063 (0.032)	0.014 (0.010)	5.540 (1.540)	5.400 (1.240)	18.60 (5.620)	16.60 (4.410)	278
			Lee & Liu	0.311 (0.020)	0.002 (0.003)	2.400 (0.130)	2.600 (0.130)	10.20 (1.000)	11.40 (1.200)	1292
			Cai et al.	0.008 (0.002)	0.005 (0.004)	9.200 (1.200)	10.30 (1.100)	138.6 (32.40)	146.8 (29.80)	610
	200	100	Nonconvex	0.001 (0.001)	0.003 (0.003)	0.377 (0.036)	0.550 (0.071)	1.810 (0.217)	1.900 (0.221)	1359
			Convex	0.09 (0.043)	0.001 (0.002)	8.690 (3.700)	6.540 (2.370)	29.30 (13.70)	19.10 (7.940)	1109
			Lee & Liu	0.217 (0.011)	2e-04 (6e-04)	3.000 (0.089)	3.000 (0.090)	11.40 (0.660)	12.60 (0.760)	11261
			Cai et al.	0.006 (0.001)	0.000 (0.000)	30.10 (3.300)	30.10 (3.000)	686.1 (127.7)	630.3 (108.5)	7528
		50	Nonconvex	0.000 (0.000)	0.016 (0.004)	0.789 (0.100)	1.210 (0.113)	3.930 (0.598)	4.710 (0.523)	3415
			Convex	0.025 (0.011)	0.008 (0.004)	11.60 (2.540)	10.80 (2.000)	39.30 (9.140)	33.70 (7.080)	2973
			Lee & Liu	0.160 (0.003)	0.004 (0.001)	5.200 (0.050)	4.600 (0.050)	21.60 (0.500)	21.80 (0.500)	17690
			Cai et al.	0.006 (5e-04)	0.002 (0.002)	9.800 (0.700)	9.900 (0.720)	83.20 (11.90)	81.90 (11.70)	6987

Table S3: Random graph. The setup is the same as Table S1.

n_k	p	q	Method	FP	FN	EL$_1$	EL$_2$	QL$_1$	QL$_2$	Time
10	100	100	Nonconvex	0.01 (0.007)	0 (0.001)	0.281 (0.037)	0.352 (0.051)	0.744 (0.098)	1.250 (0.214)	198
			Convex	0.151 (0.048)	0.001 (0.002)	2.860 (1.070)	3.840 (1.490)	6.250 (2.630)	9.060 (4.240)	144
			Lee & Liu	0.386 (0.010)	3e-04 (9e-04)	1.200 (0.036)	1.700 (0.051)	3.200 (0.120)	4.600 (0.170)	1172
			Cai et al.	0.005 (0.005)	0.024 (0.008)	7.900 (0.730)	6.000 (0.450)	28.10 (3.300)	21.10 (2.500)	618
		50	Nonconvex	0.006 (0.005)	0.022 (0.008)	0.692 (0.104)	1.310 (0.137)	1.870 (0.268)	4.100 (0.473)	335
			Convex	0.111 (0.045)	0.016 (0.012)	5.720 (1.610)	5.980 (1.470)	12.70 (4.100)	13.60 (4.020)	294
			Lee & Liu	0.304 (0.023)	0.005 (0.004)	2.000 (0.077)	2.200 (0.096)	5.400 (0.250)	6.200 (0.430)	1400
			Cai et al.	0.006 (0.003)	0.027 (0.006)	5.500 (0.450)	5.300 (0.370)	18.90 (2.000)	18.90 (1.900)	622
	200	100	Nonconvex	0.021 (0.006)	0.032 (0.008)	0.706 (0.067)	2.100 (0.184)	1.920 (0.185)	6.180 (0.507)	1386
			Convex	0.272 (0.061)	0.012 (0.004)	4.320 (1.620)	5.300 (1.410)	9.900 (3.800)	12.00 (3.570)	1070
			Lee & Liu	0.322 (0.020)	0.004 (0.002)	4.300 (0.099)	5.600 (0.120)	12.30 (0.360)	20.90 (1.100)	11977
			Cai et al.	0.005 (0.005)	0.076 (0.009)	11.10 (0.930)	13.00 (1.200)	38.60 (3.600)	94.00 (14.90)	8735
		50	Nonconvex	0.005 (0.004)	0.092 (0.011)	2.310 (0.239)	4.710 (0.246)	5.890 (0.614)	13.60 (0.921)	3589
			Convex	0.111 (0.029)	0.045 (0.006)	7.400 (1.510)	9.420 (1.580)	16.40 (3.810)	20.90 (4.340)	2769
			Lee & Liu	0.257 (0.014)	0.007 (0.003)	5.400 (0.150)	9.500 (0.290)	15.60 (0.530)	54.40 (5.100)	23671
			Cai et al.	0.008 (0.003)	0.064 (0.008)	8.000 (0.590)	12.50 (0.880)	25.00 (1.800)	91.00 (15.50)	7137

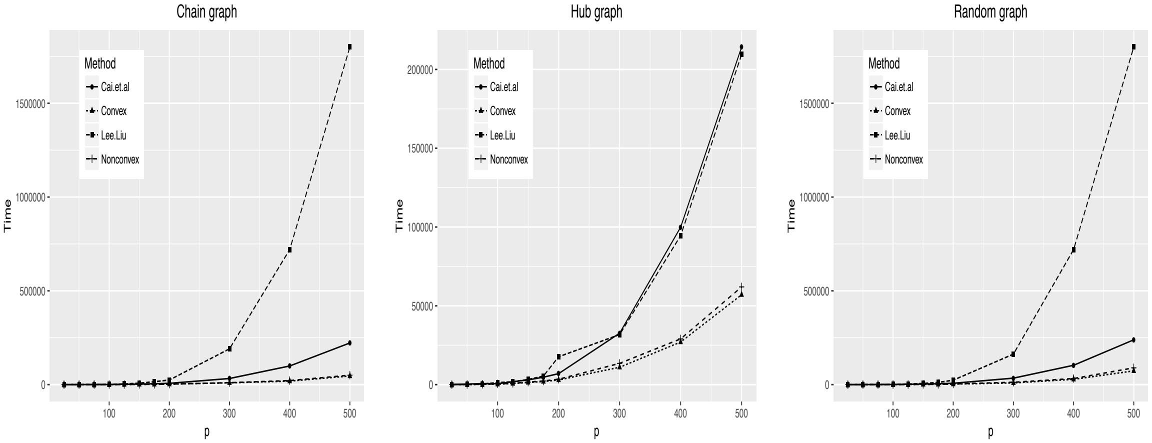


Figure S2: The running time comparison of our method, the convex counterpart, Lee and Liu (2015), and Cai et al. (2016). The number of network nodes gradually increases from $p = 25$ to $p = 500$, with $n_k = 10$ and $q = 50$.