Joint Estimation of Multiple High-dimensional Precision Matrices *

T. Tony Cai, Hongzhe Li, Weidong Liu and Jichun Xie

Abstract

Motivated by the analysis of gene expression data measured in different tissues or disease states, we consider joint estimation of multiple precision matrices to effectively utilize the partially shared graphical structures of the corresponding graphs. The procedure is based on a weighted constrained ℓ_∞/ℓ_1 minimization approach, which can be effectively implemented by a second-order cone programming. Both theoretical and numerical properties of the procedure are investigated. It is shown that the proposed joint estimation procedure leads to a faster convergence rate than estimating the precision matrices individually under various losses. The supports of the precision matrices can also be recovered after an additional thresholding step. Under regularity conditions, the proposed procedure leads to an exact graph structure recovery with probability tending to 1. The method is illustrated through an analysis of an ovarian cancer gene expression data. The results indicate that the patients of the poor prognostic subtype lack some important links between the genes of the apoptosis pathway.

KEYWORDS: Constrained optimization; Convergence rate; Graph recovery; Precision matrices; Second-order cone programming; Sparsity

*Tony Cai is Dorothy Silberberg Professor of Statistics, Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104 (Email: tcai@wharton.upenn.edu). Hongzhe Li is Professor of Biostatistics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA 19104 (Email: hongzhe@upenn.edu). Weidong Liu is Professor, Department of Mathematics, Institute of Natural Sciences and MOE-LSC, Shanghai Jiao Tong University, Shanghai, China (Email: weidongl@sjtu.edu.cn). Jichun Xie is Assistant Professor, Department of Statistics, Fox School of Business, Temple University, Philadelphia, PA 19122 (Email: jichun@temple.edu). The research was supported in part by NSF FRG Grant DMS-0854973, NSF MRI Grant No. CNS-09-58854, NIH grants CA127334, GM097505. Weidong Liu's research was also supported by NSFC Grant No.11201298 and No.11322107, the Program for Professor of Special Appointment (Eastern Scholar) at Shanghai Institutions of Higher Learning, Shanghai Pujiang Program, Foundation for the Author of National Excellent Doctoral Dissertation of PR China and Program for New Century Excellent Talents in University.

1. INTRODUCTION

Gaussian graphical models provide a natural tool for modeling the conditional independence relationships among a set of random variables (Lauritzen, 1996; Whittaker, 1990). Such models have been applied to infer the relationships between genes at transcriptional level (Schäfer and Strimmer, 2005; Li and Gui, 2006; Li et al., 2013b), where the precision matrix, which is defined to be the inverse of the covariance matrix, of a multivariate normal distribution has an interpretation of conditional dependence. Compared with marginal dependence, conditional dependence can capture the direct "link" between two variables when other variables are conditioned on. Based on a precision matrix $\mathbf{\Omega} = (\omega_{ij})_{p\times p}$ of a p-dimensional random vector, we define its corresponding graphical structure by connecting variable i and variable j if and only if $\omega_{ij} \neq 0$. We define the support of $\mathbf{\Omega}$ by the set of nonzero entries, $\mathcal{S} = \{(i,j) : \ \omega_{ij} \neq 0\}$. If the maximum degree of $\mathbf{\Omega}$, $\max_i \sum_{j=1}^{p} I(\omega_{ij} \neq 0)$, is relaltively small, we call $\mathbf{\Omega}$ sparse. Since the expression variation of a gene can usually be explained by a small subset of other genes, the precision matrix for gene expression data is expected to be sparse.

Many methods for estimating the Gaussian graphical models in high-dimensional settings have been developed in recent years. Meinshausen and Bühlmann (2006) introduced a neighborhood selection approach to this problem by fitting an ℓ_1 penalized regression to each variable using the other variables as predictors. It was shown that this neighborhood selection procedure estimates consistently the set of non-zero elements of the precision matrix. Algorithms for exact maximization of the l_1-penalized log-likelihood have also been proposed. Yuan and Lin (2007), Banerjee et al. (2008) and Dahl et al. (2008) adapted an interior point optimization method to solve this problem. Based on the work of Banerjee et al. (2008) and a block-wise coordinate descent algorithm, Friedman et al. (2008) developed the graphical Lasso (GLASSO) for sparse precision matrix estimation, which is computationally efficient even when the dimension is greater than the sample size. Yuan (2010) developed a linear programming procedure for high dimensional precision matrix estimation and obtained oracle inequalities for the estimation error in terms of several matrix norms. Cai et al. (2011) developed a constrained ℓ_1 minimization approach (CLIME) to sparse precision matrix estimation.

These methods have focused on estimating a single precision matrix or a single Gaussian graph-

ical model. However, in many applications it is advantageous to jointly estimate multiple precision matrices and their corresponding graphical strucutures, especially when the graphical structures share some common edges. A good motivating example is that gene expression data are often measured over different tissues or in different populations, and it is expected that the underlying Gaussian graphs share many common links, which may reflect the common regulatory relationships among a set of genes across different tissues or in different populations. However, we also expect certain tissue-specific or population-specific links among the genes. This raises an important statistical problem of jointly estimating multiple precision matrices. Guo et al. (2011) proposed a method that jointly estimates several graphical models (JEMGM) corresponding to the different groups. The method aims to preserve the common structure, while allowing for differences between the groups. This is achieved through a hierarchical penalty that targets the removal of common zeros in the precision matrices across groups. Danaher et al. (2013) proposed the joint graphical Lasso (FGL and GGL), which borrows strength across the groups in order to estimate multiple graphical models that share certain characteristics, such as the locations or weights of nonzero edges. Their approach is based upon maximizing a penalized log likelihood, where generalized fused Lasso or group Lasso penalty is used. In both papers, the authors show that their joint estimators achieve the same asymptotic convergence rate as the individual estimators.

In this paper, we propose a weighted constrained ℓ_∞/ℓ_1 minimization estimation method to jointly estimate K sparse precision matrices (MPE). We aim to minimize the maximum of the K matrix ℓ_1 norms under a constraint that encourages group-wise sparsity. We show that the joint estimation procedure leads to a faster convergence rate than estimating the precision matrices individually under the entry-wise ℓ_∞ norm loss. An additional thresholding step on the estimators with a careful chosen threshold leads to a more accurate recovery of the graphical structure of the precision matrices. After thresholding, the resulting estimator has a faster rate of convergence than estimators obtained from individual samples under the matrix ℓ_1 norm. We also show that when the multiple precision matrices have common graphical structures, our procedure leads to the exact recovery of the graph structure with probability tending to 1.

Different from Guo et al. (2011) and Danaher et al. (2013), our method does not require independence assumptions among the random variables across different groups. In genetic applications,

the expression levels of the same subject across difference tissues are often correlated. Therefore, the independence assumption sometimes fails to hold in real applications. Furthermore, we demonstrate from the theoretical perspective the importance of joint estimation when multiple precision matrices share common graphical structures, since the joint estimators achieve faster convergence rates compared to the individual estimators.

The rest of the paper is organized as follows. Section 2 presents the estimation method and the optimization algorithm. Theoretical properties of the estimation procedure and accuracy of the graph structure recovery are studied in Section 3. Section 4 investigates the numerical performance of the method through a simulation study. The proposed procedure is compared with other alternative approaches. The method is also illustrated via an analysis of human heart gene expression data in Section 5. A brief discussion is given in Section 6 and technical proofs are presented in the Appendix.

2. METHODOLOGY

We begin by introducing the basic notation and definitions used in this paper. For a vector $a = (a_1, \ldots, a_p)^T \in \mathbb{R}^p$, define $|a|_1 = \sum_{j=1}^p |a_j|$ and $|a|_2 = (\sum_{j=1}^p a_j^2)^{1/2}$. For a matrix $A = (a_{ij}) \in \mathbb{R}^{p\times q}$, the elementwise ℓ_r norm is given by $|A|_r = (\sum_{i,j} |a_{ij}|^r)^{1/r}$ and the matrix 1-norm by the maximum absolute column sum, $\|A\|_{l_1} = \max_{1\le j\le q} \sum_{i=1}^p |a_{ij}|$. The spectral norm of A is denoted as $\|A\|_2$. Let $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ be the largest and smallest eigenvalues of A respectively. For two sequences of real numbers $\{a_n\}$ and $\{b_n\}$, write $a_n = O(b_n)$ if there exists a constant C such that $|a_n| \le C|b_n|$ holds for all sufficiently large n, write $a_n = o(b_n)$ if $\lim_{n\to\infty} a_n/b_n = 0$.

2.1 The Joint Estimation Method

We introduce a joint estimation method for simultaneously estimating K precision matrices that compeletely or partially share common support. The method is related to the constrained ℓ_1 minimization approach for high dimensional regression and high dimensional precision matrix estimation which has been demonstrated to be effective for recovering sparse vector (Donoho et al., 2006; Candés and Tao, 2007) and a single sparse precision matrix (Cai et al., 2011).

For $1 \le k \le K$, let $\boldsymbol{X}^{(k)} \sim N(\boldsymbol{\mu}^{(k)}, \boldsymbol{\Sigma}^{(k)})$ be a p-dimensional random vector for the kth group. The precision matrix of $\boldsymbol{X}^{(k)}$, denoted by $\boldsymbol{\Omega}^{(k)} = (\omega_{ij}^{(k)})$, is the inverse of the covariance matrix

$\mathbf{\Sigma}_k$. Assume that $\boldsymbol{X}^{(k)}$'s are independent of each other. Suppose there are n_k identically and independently distributed random samples from $\mathbf{X}^{(k)}$: $\{\boldsymbol{X}_j^{(k)}, 1 \le j \le n_k\}$. The sample covariance matrix for the kth group is

$$\hat{\mathbf{\Sigma}}^{(k)} = \frac{1}{n_k - 1} \sum_{j=1}^{n_k} (\boldsymbol{X}_j^{(k)} - \bar{\boldsymbol{X}}^{(k)})(\boldsymbol{X}_j^{(k)} - \bar{\boldsymbol{X}}^{(k)})^{\mathrm{T}},$$

where $\bar{\boldsymbol{X}}^{(k)} = \sum_{j=1}^{n_k} \boldsymbol{X}_j^{(k)}/n_k$. We shall denote $n = n_1 + \cdots + n_K$.

Our goal is to simultaneously estimate the precision matrices $\mathbf{\Omega}^{(k)}$ for $1 \le k \le K$ based on the observed samples from each of the K groups. We propose a weighted constrained ℓ_∞/ℓ_1 minimization method which utilizes the potential shared support among the K groups. However, the graphical structures of the K matrices do not need to be identical. Specifically, we estimate $\mathbf{\Omega}^{(k)} = (\omega_{ij}^{(k)})$ for $k = 1, \cdots, K$ by the following constrained optimization,

$$\begin{aligned} &\min_{\mathbf{\Omega}_1^{(k)} \in \mathbb{R}^{p\times p},\ 1\le k\le K} \Big(\max_{1\le k\le K} |\mathbf{\Omega}_1^{(k)}|_1 \Big), \\ &\text{subject to } \max_{i,j} \Big\{ \sum_{k=1}^{K} w_k |(\hat{\mathbf{\Sigma}}^{(k)} \mathbf{\Omega}_1^{(k)} - \mathbf{I})_{ij}|^2 \Big\}^{1/2} \le \lambda_n, \end{aligned} \tag{1}$$

where $w_k = n_k/n$ is the weight for the kth group, and $\lambda_n = C(\log p/n)^{1/2}$ is a tuning parameter. The ℓ_∞/ℓ_1 objective function is used to encourage the sparsity of all K precision matrices. The constraint is imposed on the maximum of the element-wise group ℓ_2 norm to encourage the groups to share a common sparsity pattern.

Denote by $\hat{\mathbf{\Omega}}_1^{(k)}$ $(1 \le k \le K)$ the solution to (1). Then $\hat{\mathbf{\Omega}}_1^{(k)}$ are not necessarily symmetric in general. Our final estimator $\hat{\mathbf{\Omega}}^{(k)} = (\hat{\omega}_{ij}^{(k)})$ of $\mathbf{\Omega}^{(k)}$ is obtained by symmetrizing $\hat{\mathbf{\Omega}}_1^{(k)}$. This is done by comparing the pair of the non-diagonal entries at symmetric positions $\hat{\omega}_{1ij}^{(k)}$ and $\hat{\omega}_{1ji}^{(k)}$ and by assigning the one with a smaller magnitude at both entries. That is,

$$\hat{\omega}_{ij}^{(k)} = \hat{\omega}_{ji}^{(k)} := \hat{\omega}_{1ij}^{(k)} I(|\hat{\omega}_{1ij}^{(k)}| \le |\hat{\omega}_{1ji}^{(k)}|) + \hat{\omega}_{1ji}^{(k)} I(|\hat{\omega}_{1ij}^{(k)}| > |\hat{\omega}_{1ji}^{(k)}|).$$

It is worthwhile to point out that the symmetrizing procedure is not ad-hoc. The procedure assures the final estimator $\hat{\mathbf{\Omega}}^{(k)}$ to obtain the same entry-wise ℓ_∞ estimation error as $\hat{\mathbf{\Omega}}_1^{(k)}$. The details are discussed in Section 3.

2.2 Computational algorithm

The convex optimization problem (1) involves estimating K $p \times p$ precision matrices. To reduce the computation complexity, it can be further decomposed into p sub-problems that involve estimating K $p \times 1$ sparse vectors:

$$\min_{\boldsymbol{\beta}_j^{(k)} \in \mathbb{R}^p,\ 1 \le k \le K} \Big(\max_{1 \le k \le K} |\beta_j^{(k)}|_1 \Big),$$
$$\text{subject to } \max_i \Big\{ \sum_{k=1}^{K} w_k |(\hat{\boldsymbol{\Sigma}}^{(k)} \beta_j^{(k)} - e_j)_i|^2 \Big\}^{1/2} \le \lambda_n \tag{2}$$

for $1 \le j \le p$, where $e_j \in \mathbb{R}^p$ is the unit vector with the j-th element being 1 and other elements being 0. The following lemma shows that solving (2) is equivalent to solving (1).

Lemma 1 *Suppose* $\hat{\boldsymbol{\Omega}}_1^{(k)}$ *is the solution to (1) and* $\hat{B}^{(k)} := (\hat{\beta}_1^{(k)}, \ldots, \hat{\beta}_p^{(k)})$*, where* $\hat{\beta}_j^{(k)}$ *is the solution to (2). Then* $\hat{\boldsymbol{\Omega}}_1^{(k)} = \hat{\mathbf{B}}^{(k)}$ *for* $1 \le k \le K$*.*

Problem (2) can be solve by a second-order cone programming. There are existing packages that can be used to solve (2), such as the SDTP3 and the SeDuMi package in Matlab, and the CLSOCP package in R. CLSOCP uses a one-step smoothing Newton method of Liang et al. (2009). This algorithm has good precision but works relatively slowly for high dimensional problem. SeDuMi and SDTP3 adopted the primal-dual infeasible-interior point algorithm (Newsterov and Todd, 1998). The most time-consuming part of the algorithm is to solve the Schur complement equation, which involves Cholesky factorization. The sparsity and the size of the Schur complement matrix are two factors that affect the efficiency. SDTP3 is able to divide a high dimensional optimization problem into sparse blocks and uses the sparse solver for Cholesky factorizations. It is therefore faster than SeDuMi in solving (2). In this paper, we used the SDTP3 package. For a problem with $p = 200$, $n_k = 150$ and $K = 3$, it takes a dual-core 2.7 GHz Intel Core i7 laptop approximately 11 minutes to solve (1).

2.3 Tuning Parameter Selection

The tuning parameter λ_n in (1) and (2) determines the sparsity of the estimators, where a larger λ_n leads to sparser solutions. But such solutions are often biased. To prevent the over-fitting and reduce the bias, we calculate a BIC score using a re-estimated precision matrix based on the selected coefficients. The procedure can be summarized as the following:

1. For a given λ, calculate the estimator $\hat{\mathbf{\Omega}}^{(k)}$. Based on the support of $\hat{\mathbf{\Omega}}^{(k)}$, we use least squares and neighborhood selection to re-fit the precision matrix estimator $\hat{\mathbf{\Omega}}_2^{(k)}$.

2. Define $\mathcal{S}_j^{(k)} = \{i : \hat{\omega}_{ij}^{(k)} \neq 0, i \neq j\}$, which is the set of non-zero none-diagonal elements of the jth column of $\hat{\mathbf{\Omega}}^{(k)}$.

3. If $\text{Card}(\mathcal{S}_j^{(k)}) \geq n_k$, let the jth column of $\hat{\mathbf{\Omega}}^{(k)}$ equal to the jth column of $\hat{\mathbf{\Omega}}_{\cdot j}^{(k)}$, *i.e.* $\hat{\mathbf{\Omega}}_{2,\cdot j}^{(k)} = \hat{\mathbf{\Omega}}_{\cdot j}^{(k)}$. If $\text{Card}(\mathcal{S}_j^{(k)}) < n_k$, fit the regression model

$$X_j^{(k)} = \sum_{i \in \mathcal{S}_j^{(k)}} \beta_{ij}^{(k)} X_i^{(k)} + \epsilon_j^{(k)}. \tag{3}$$

   It is easy to show that if $\mathcal{S}_j^{(k)}$ equal to the true support $\mathcal{S}_{0,j}^{(k)} = \{l : \omega_{0,ij}^{(k)} \neq 0, i \neq j\}$, $\beta_{lj}^{(k)} = -\omega_{0,ij}^{(k)}/\omega_{0,jj}^{(k)}$ and $\mathsf{Var}(\epsilon_j^{(k)}) = 1/\omega_{0,jj}^{(k)}$. Thus, after fitting Model (3), we let $\hat{\omega}_{2,jj}^{(k)} = 1/\mathsf{Var}(\hat{\epsilon}_j^{(k)})$, and $\hat{\omega}_{2,ij}^{(k)} = -\hat{\beta}_{ij}^{(k)} \hat{\omega}_{2,jj}^{(k)}$.

4. Repeat Step 3 for $j = 1, \ldots, p$ and $k = 1, \ldots, K$. The resulting matrices $\hat{\mathbf{\Omega}}_2^{(k)}$, $k = 1, \ldots, K$ are not symmetric. We symmetrize $\hat{\mathbf{\Omega}}_2^{(k)}$ by the same procedure as we do to $\hat{\mathbf{\Omega}}_1^{(k)}$:

$$\hat{\omega}_{3,ji}^{(k)} = \hat{\omega}_{3,ij}^{(k)} := \hat{\omega}_{2,ij}^{(k)} I(|\hat{\omega}_{2,ij}^{(k)}| \leq |\hat{\omega}_{2,ji}^{(k)}|) + \hat{\omega}_{2,ji}^{(k)} I(|\hat{\omega}_{2,ij}^{(k)}| > |\hat{\omega}_{2,ji}^{(k)}|).$$

We use $\hat{\mathbf{\Omega}}_3^{(k)} = (\hat{\omega}_{3,ij}^{(k)})$, $k = 1, \ldots, K$ as the estimators corresponding to the tuning parameter λ. Compared with the original estimator $\hat{\mathbf{\Omega}}^{(k)}$, the re-fitted estimator improved the tuning parameter selection in the simulations.

The optimal tuning parameter can be selected by Bayesian information criterion (BIC),

$$\text{BIC}(\lambda) = \sum_{k=1}^{K} \left\{ n_k \text{tr}\left(\hat{\mathbf{\Sigma}}^{(k)} \hat{\mathbf{\Omega}}_3^{(k)}\right) - n_k \log(\det \hat{\mathbf{\Omega}}_3^{(k)}) + \log(n_k) s_k \right\}, \tag{4}$$

where $s_k = \text{Card}\{(i,j) : \hat{\omega}_{i,j} \neq 0, 1 \leq i < j \leq p\}$. We obtain the solution to our method over a wide range of tuning parameters and choose $\hat{\lambda}_n$ that minimizes $\text{BIC}(\lambda)$.

3. THEORETICAL PROPERTIES

3.1 Estimation Error Bound

We investigate the properties of the proposed estimator by considering the convergence rates of $\hat{\mathbf{\Omega}}^{(k)} - \mathbf{\Omega}^{(k)}$, including estimation error bounds and graph structure recovery. We assume the following conditions:

(C1). Suppose there exists some constant $a > 0$, such that

$$\log p = o\left(\frac{n}{K^{2a}(\log n)^2}\right), \quad \text{and} \quad \max(K, K^{4-a}\log K) = o(\log p).$$

(C2). Let $\max_{1\le k\le K}\{\lambda_{\max}(\mathbf{\Omega}^{(k)})/\lambda_{\min}(\mathbf{\Omega}^{(k)})\} \le M_0$ for some bounded constant $M_0 > 0$.

(C3). Suppose that $n_1 \asymp n_2 \asymp \cdots \asymp n_K$, where $n = \sum_{k=1}^K n_k$ and $w_k = n_k/n$.

Let $M_n = \max_{1\le k\le K}\max_j \sum_{i=1}^p |\omega_{ij}^{(k)}| = \max_{1\le k\le K} \|\mathbf{\Omega}^{(k)}\|_{\ell_1}$ be the maximum matrix ℓ_1 norms of the K matrices. The following theorem establishes the convergence rate of the precision matrix estimates under the element-wise ℓ_∞ norm.

Theorem 1 *Let $\lambda_n = C_0(\log p/n)^{1/2}$ for some constant $C_0 > \sqrt{2M_0+2}$. Suppose that (C1)-(C3) hold. We have*

$$\max_{i,j}\Big\{\sum_{k=1}^K w_k|(\hat{\mathbf{\Omega}}^{(k)} - \mathbf{\Omega}^{(k)})_{i,j}|^2\Big\}^{1/2} \le C_1 M_n\left(\frac{\log K\cdot\log p}{n}\right)^{1/2} \tag{5}$$

with a high probability converging to 1 and $C_1 = 2C_0$.

Remark 1: The value of C_0 depends on M_0. In practice, M_0 is often unknown. However, we can can use tuning parameter selection method, such as BIC in (4), to choose λ_n. The details are discussed in Section 2.3.

Remark 2: Theorem 1 (and Theorem 2 and 3) does not require the true precision matrices $\mathbf{\Omega}^{(k)}$ to have identical graphical structures. Both the values and locations of non-zero entries can differ across $\mathbf{\Omega}^{(k)}$, $k = 1, \ldots, K$.

Remark 3: It is not necessary to assume the independence between the groups $\boldsymbol{X}^{(k)}$. Let $Y_{ij}^{(k)} = (\boldsymbol{X}^{(k)}\boldsymbol{X}^{(k)\prime}\mathbf{\Omega}^{(k)})_{ij} - e_{ij}$ and $Y_{ij} = (Y_{ij}^{(1)}, \ldots, Y_{ij}^{(K)})$. Let $\lambda_{\max,ij}$ be the largest eigenvalue of $\mathsf{Cov}(Y_{ij})$. If we replace the condition (C2) by $\max_{ij}\lambda_{\max,ij} \le M$, then Theorem 1 still hold; so as Theorem 2 and Theorem 3.

By Theorem 1, the average rate under the element-wise ℓ_∞ norm of the K estimators is of the order of $(\log K/K)^{1/2}M_n(\log p/n_1)^{1/2}$. Here the number of groups K can grow with n and p. Suppose the matrix ℓ_1 norm of the K-matrices are of the same order. Cai et al. (2012) showed that the minimax rate for estimating the precision matrices separately is $CM_n(\log p/n_1)^{1/2}$, which leads to the following proposition.

Proposition 1 *Let $\hat{\mathcal{U}}$ be the set of estimators $(\tilde{\mathbf{\Omega}}^{(1)},\ldots,\tilde{\mathbf{\Omega}}^{(K)})$, where $\tilde{\mathbf{\Omega}}^{(k)}$ only depends on the k-th sample $\{\boldsymbol{X}_j^{(k)};1\le j\le n_k\}$. Then*

$$\min_{(\tilde{\mathbf{\Omega}}^{(1)},\ldots,\tilde{\mathbf{\Omega}}^{(K)})\in\hat{\mathcal{U}}}\max_{(\mathbf{\Omega}^{(1)},\ldots,\mathbf{\Omega}^{(K)})\in\mathcal{U}} pr\left[\max_{i,j}\Big\{\sum_{k=1}^{K}|(\tilde{\mathbf{\Omega}}^{(k)}-\mathbf{\Omega}^{(k)})_{i,j}|^2\Big\}^{1/2} \ge CM_n\left(\frac{K\cdot\log p}{n}\right)^{1/2}\right]\ge\alpha>0,$$

for some $\alpha>0$ and sufficiently large $n_1,\ldots,n_K$.

Clearly, joint estimation of precision matrices leads to a faster convergence rate under the entry-wise ℓ_∞ norm than estimating the precision matrices individually, especially when K is large. An additional thresholding step on the estimators with a careful chosen threshold leads to more accurate recovery of the precision matrices. Define the thresholded estimator $\breve{\mathbf{\Omega}}^{(k)}=(\breve{\omega}_{ij}^{(k)})$ as follows:

$$\breve{\omega}_{ij}^{(k)}=\hat{\omega}_{ij}^{(k)}I\left\{\left(\sum_{k=1}^{K}w_k(\hat{\omega}_{ij}^{(k)})^2\right)^{1/2}>C_1M_n\left(\frac{\log K\cdot\log p}{n}\right)^{1/2}\right\}.$$

Here, C_1 is the same constant as in (5).

Joint estimation can also lead to a faster rate under the matrix ℓ_1 norm under certain sparsity assumption on the precision matrices. Let $\mathcal{S}_j^{(k)}=\{(i,j):\ \omega_{ij}^{(k)}\neq 0,\ i<j\}$ and $\mathcal{S}_j=\cup_{k=1}^{k}S_j^{(k)}$. Let $s_0(p)=\max_{1\le j\le p}\mathrm{Card}(\mathcal{S}_j)$ is the union sparcity. The next theorem shows the convergence rate under the matrix ℓ_1 norm.

Theorem 2 *Suppose that (C1)-(C3) hold. Then*

$$\max_j\sum_{i=1}^{p}\Big\{\sum_{k=1}^{K}w_k(\breve{\mathbf{\Omega}}^{(k)}-\mathbf{\Omega}^{(k)})_{ij}^2\Big\}^{1/2}\le C_1M_ns_0(p)\left(\frac{\log K\cdot\log p}{n}\right)^{1/2} \tag{6}$$

with a high probability converging to 1 and C_1 is as same as in (5).

The convergence rates of $\breve{\mathbf{\Omega}}^{(k)}$ depend on the union sparsity level $s_0(p)$. When the precision matrices share the same graphical structure, $s_0(p)=\max_{1\le j\le p}\mathrm{Card}(\mathcal{S}_j^{(k)})$, for all $k=1,\ldots,K$. If there are more shared elements in the supports of the precision matrices, the union sparsity $s_0(p)$ becomes smaller, which leads to smaller estimation error.

Let $\hat{a}_{ij}=\sqrt{\sum_{k=1}^{K}w_k(\breve{\mathbf{\Omega}}^{(k)}-\mathbf{\Omega}^{(k)})_{ij}^2}$. Then the matrix $\hat{\mathbf{A}}=(\hat{a}_{ij})_{p\times p}$ measures the overall errors between the entries of $\breve{\mathbf{\Omega}}^{(k)}$ and $\mathbf{\Omega}^{(k)}$ for $k=1,...,K$. Theorem 2 leads to the following corollary.

Corollary 1 *With a high probability converging to 1, for some constant $\epsilon > 0$,*

$$\|\hat{\mathbf{A}}\|_2 \leq \|\hat{\mathbf{A}}\|_1 \leq C_1 M_n s_0(p) \left(\frac{\log K \cdot \log p}{n} \right)^{1/2},$$

where C_1 is as same as in (5).

3.2 Graphical Structure Recovery

For graphical structure recovery the analysis becomes very complicated when the corresponding graphical structures of the precision matrices are different across the K groups. We shall focus on the case that the K precision matrices have a common support. Let $\mathcal{S}_k = \{(i,j) : \omega_{ij}^{(k)} \neq 0\}$ be the support for the kth precision matrix. Assuming a common support, $\mathcal{S} = \mathcal{S}_1 = \cdots = \mathcal{S}_K$, then by Theorem 1, we estimate $\mathcal{S}$ by

$$\hat{\mathcal{S}} = \left[(i,j) : \left\{ \sum_{k=1}^{K} w_k \left(\hat{\omega}_{ij}^{(k)} \right)^2 \right\}^{1/2} > C_1 M_n \left(\frac{\log K \cdot \log p}{n} \right)^{1/2} \right],$$

where C_1 is a constant given in Theorem 1. Let

$$\theta_n = \min_{(i,j) \in \mathcal{S}} \left\{ \sum_{k=1}^{K} w_k \left(\omega_{ij}^{(k)} \right)^2 \right\}^{1/2}.$$

We have the following theorem on the support recovery.

Theorem 3 *Suppose that the conditions in Theorem 1 hold. Assume that*

$$\theta_n > 2C_1 M_n \left(\frac{\log K \cdot \log p}{n} \right)^{1/2}. \tag{7}$$

We have $\hat{\mathcal{S}} = \mathcal{S}$ with a high probability converging to 1 and some constant $\epsilon > 0$.

The lower bound condition (7) is necessary for graphical structure recovery. When the precision matrices are the same across all K groups, condition (7) is weaker than the necessary condition (8) for single graphical structure recovery using only the data for the kth group $\boldsymbol{X}_j^{(k)} (1 \leq j \leq n_k)$:

$$\min_{(i,j) \in S^{(1)}} |\omega_{ij}^{(k)}| \geq 2C M_n \left(\frac{\log p}{n_k} \right)^{1/2}. \tag{8}$$

4. SIMULATION STUDIES

4.1 Data generation

We present in this section simulation results to evaluate the numerical performance of the proposed method and other methods, including single precision matrix estimation procedures proposed by Friedman et al. (2008) and Cai et al. (2011) and joint estimation method proposed by Guo et al. (2011) and Danaher et al. (2013). The single precision matrix estimation methods are applied to each group, and therefore ignore the common structures between the groups. In all numerical studies, we set $p = 200$, $K = 3$ and $n_k = 80, 120, 150$, respectively for $k = 1, 2, 3$. The simulated observations in each group have identically and independently distributed multivariate normal distribution $\mathrm{N}\{0, (\mathbf{\Omega}^{(k)})^{-1}\}$, where $\mathbf{\Omega}^{(k)}$ is the precision matrix for the kth group. For each model, 100 replications are performed.

We consider four different types of graph structures, Barabási and Albert graph (Barabási and Albert, 1999), Erdös and Rényi graph (Erdös and Rényi, 1960) , random geometric graph (Penrose, 2003), and the undirected graph corresponding to the Watts-Strogatz network (Watts and Strogatz, 1998). For each graph structure, we consider three different ratios of the numbers of individual-specific edges to the number of common edges, $\rho = 0,\ 1/4,\ 1$.

We first generate the common graph structure. For Barabási and Albert Model, a new vertex is added to the existing graph each time and the new vertex is connected to an existing old vertex with a probability proportional to the degree of the existing vertices plus one. For Erdös and Rényi graph, the common structure contains p vertices and each pair of vertices are connected with probability 0.05. For geometric random graph, p points are dropped on a unit square. Two vertex will be connected with an undirected edge if and only if their corresponding points are closer to each other than a radius of 0.05. For Watt-Strogatz network, first a ring lattice of p vertex is created. One vertice is connected with its neighbors within order distance of 15. Then the edges of the lattice are rewired uniformly randomly with probability 0.01. The resulting network has "small-world" property, which is shared by many protein networks (Vendrascolo et al., 2002; Greene and Higman, 2003). It is possible that the network contains loops or multiple edges, which are removed afterwards to create an undirected graph.

After we generate the common graph structure, we add individual edges with the individual to

common edges ratio of $\rho = 0,\ 1/4$ and 1. The first case ($\rho = 0$) represents the scenario where the precision matrices in different groups share exactly same support, but the values of the entries could be different. The second and the third cases ($\rho = 1/4, 1$) imply that among all the edges within each group, $1/(1+\rho)$ of the edges are shared by all groups, the remaining edges are group-specific. Let M be the number of the shared edges. For each individual graph, we randomly choose $\lfloor \rho M \rfloor$ pairs of new edges. After the support of the matrices are determined, the values of the non-zero entries are generated independently from the uniform distribution in $[-1, -0.5] \cup [0.5, 1]$. The diagonal values are assigned with a constant so that each matrix has the condition number equal to p.

4.2 Simulation results

Each method is evaluated for a range of tuning parameters under each model. The optimal tuning parameter is chosen by Bayesian information criterion (4). Several measures are used to compare the performance of these estimators. The estimation error is evaluated in terms of average matrix L_1 norm, L_2 norm (spectral norm) and Frobenius norm, which are defined as follows:

$$L_1 = \frac{1}{K}\sum_{k=1}^{K} \|\hat{\mathbf{\Omega}}^{(k)} - \mathbf{\Omega}_0^{(k)}\|_1,$$
$$L_2 = \frac{1}{K}\sum_{k=1}^{K} \|\hat{\mathbf{\Omega}}^{(k)} - \mathbf{\Omega}_0^{(k)}\|_2,$$
$$L_{\mathrm{F}} = \frac{1}{K}\sum_{k=1}^{K} \|\hat{\mathbf{\Omega}}^{(k)} - \mathbf{\Omega}_0^{(k)}\|_{\mathrm{F}}.$$

The graph structure recovery results are evaluated by average sensitivity (SEN), specificity (SPE) and Matthews correlation coefficient (MCC). Suppose a true precision matrix $\mathbf{\Omega}_0 = (\omega_{0,ij})$ has the support set $\mathcal{S}_0 = \{(i,j) : \omega_{0,ij} \neq 0 \text{ and } i \neq j\}$ and its estimator $\hat{\mathbf{\Omega}}$ has the support set $\hat{\mathcal{S}}$. Then the measures with respect to $\mathbf{\Omega}_0$ and $\hat{\mathbf{\Omega}}$ are defined as follows:

$$\mathrm{SPE} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}, \quad \mathrm{SEN} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}},$$
$$\mathrm{MCC} = \frac{\mathrm{TP}\times\mathrm{TN} - \mathrm{FP}\times\mathrm{FN}}{\{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{FP}+\mathrm{FN})\}^{1/2}}.$$

Here, TP, TN, FP, FN are the numbers of true positives, true negatives, false positives and false

enegatives, which are defined as

$$\text{TP} = \#\{(i,j) : (i,j) \in \mathcal{S}_0 \cap \hat{\mathcal{S}}\}, \quad \text{TN} = \#\{(i,j) : (i,j) \in \mathcal{S}_0^C \cap \hat{\mathcal{S}}^C\}$$
$$\text{FP} = \#\{(i,j) : (i,j) \in \mathcal{S}_0^C \cap \hat{\mathcal{S}}\}, \quad \text{FN} = \#\{(i,j) : (i,j) \in \mathcal{S}_0 \cap \hat{\mathcal{S}}^C\}.$$

We compare $\hat{\mathbf{\Omega}}^{(k)}$ and $\mathbf{\Omega}_0^{(k)}$ and report the average sensitivities (SEN), specificities (SPE) and Matthews correlation coefficient (MCC) among K groups.

The comparisons of the results for the four graphical models are shown in Tables 1, 2, 3 and 4. It shows that when $\rho = 0$, *i.e.*, the true graph structures are the same across all three groups, joint estimation methods perform much better than the separate estimation methods. As ρ increases, the structures across different groups become more different, the joint estimation methods gradually lose advantages. Our method has the best performance in terms of graph structure recovery among all the methods. Even when $\rho = 1$, it still performs significantly better than the separate estimation methods. Our method also has the smallest L_1 error norms. Its L_2 error norms are comparable to other joint estimation methods. In general, it has comparable Frobenius error norms to separate estimating procedures but has slightly larger Frobenius error norms than other joint estimation methods.

Since the tuning parameter selection may affect the performance of the methods, we plot in Figure 1 the receiver operating characteristic (ROC) curves averaged over 100 repetitions with false positive rate controlled under 10%. The methods proposed by Danaher et al. (2013) have two tuning parameters. For each sparsity tuning parameter, we first choose an optimal similarity tuning parameters from a grid of candidates by BIC criterion (4), and then plot the ROC curves based on the a sequence of sparsity tuning parameters and their corresponding optimal similarity tuning parameters. In practice, these methods are slower to implement than our method since it involves choosing two tuning parameters. Figure 1 shows that our method consistently outperforms the other methods in support recovery.

5. EPITHELIAL OVARIAN CANCER DATA ANALYSIS

Epithelial ovarian cancer is a molecularly diverse disease lack of effective personalized therapy. Tothill et al. (2008) identified six molecular subtypes of ovarian cancer, labeled as C1–C6, where C1 subtype was characterized by a significant differential expression of genes associated with stromal

Table 1: Simulation results for data generated based on the Barabási and Albert graph with different ratios of the number of individual-specific edges to the number of shared edges. Results are based on 100 replications. The numbers in the brackets are standard errors. CLIME: method of Cai et al. (2011); GLASSO: graphical Lasso; JEMGM: method of Guo et al. (2011); FGLand GGL: methods of Danaher et al. (2013); MPE: proposed method.

Model(ρ)	Method	Performance					
		L_1	L_2	L_{F}	SEN	SPE	MCC
BA(0)	CLIME	18.87(0.14)	6.31(0.06)	28.76(0.72)	0.26(0.02)	0.99(0.00)	0.25(0.01)
	GLASSO	18.48(1.23)	6.42(0.31)	22.22(0.68)	0.11(0.03)	1.00(0.00)	0.28(0.02)
	JEMGM	13.34(1.36)	5.39(0.36)	22.06(3.15)	0.24(0.08)	1.00(0.00)	0.40(0.03)
	FGL	17.79(0.69)	6.23(0.21)	21.30(0.48)	0.14(0.09)	0.99(0.04)	0.28(0.04)
	GGL	18.11(0.81)	6.39(0.19)	22.09(0.44)	0.15(0.03)	1.00(0.00)	0.34(0.03)
	MPE	17.11(1.27)	5.54(0.10)	27.68(1.38)	0.53(0.05)	0.99(0.08)	0.60(0.06)
BA(1/4)	CLIME	18.45(0.17)	6.41(0.06)	30.15(0.76)	0.23(0.02)	0.99(0.00)	0.24(0.02)
	GLASSO	18.25(0.61)	6.47(0.22)	23.93(0.61)	0.07(0.02)	1.00(0.00)	0.21(0.01)
	JEMGM	13.55(0.81)	5.27(0.30)	22.19(2.38)	0.22(0.07)	1.00(0.00)	0.36(0.03)
	FGL	18.16(0.78)	6.43(0.29)	23.29(0.65)	0.12(0.13)	0.99(0.06)	0.22(0.03)
	GGL	18.01(0.42)	6.42(0.16)	23.89(0.39)	0.11(0.02)	1.00(0.00)	0.26(0.02)
	MPE	16.80(1.85)	5.69(0.10)	30.06(1.95)	0.46(0.06)	0.99(0.09)	0.57(0.06)
BA(1)	CLIME	21.97(0.24)	7.18(0.07)	35.38(0.75)	0.18(0.02)	0.99(0.00)	0.21(0.01)
	GLASSO	21.66(0.54)	7.27(0.12)	29.12(0.30)	0.04(0.00)	1.00(0.00)	0.19(0.01)
	JEMGM	18.50(1.46)	6.73(0.45)	34.36(3.91)	0.07(0.02)	1.00(0.00)	0.23(0.02)
	FGL	21.85(0.81)	7.13(0.50)	28.11(1.31)	0.11(0.20)	0.97(0.09)	0.18(0.02)
	GGL	21.72(0.83)	7.30(0.19)	28.95(0.33)	0.05(0.01)	1.00(0.00)	0.20(0.02)
	MPE	19.39(0.28)	6.31(0.09)	34.53(0.72)	0.33(0.02)	1.00(0.00)	0.47(0.02)

Table 2: Simulation results for data generated based on the Erdös and Rényi graph with different ratios of the number of individual-specific edges to the number of shared edges. Results are based on 100 replications. The numbers in the brackets are standard errors. CLIME: method of Cai et al. (2011); GLASSO: graphical Lasso; JEMGM: method of Guo et al. (2011); FGLand GGL: methods of Danaher et al. (2013); MPE: proposed method.

Model(ρ)	Method	Performance					
		L_1	L_2	L_{F}	SEN	SPE	MCC
ER(0)	CLIME	19.62(0.32)	7.72(0.08)	49.48(0.83)	0.25(0.01)	0.99(0.00)	0.34(0.01)
	GLASSO	20.57(0.38)	7.91(0.12)	46.81(0.68)	0.03(0.02)	1.00(0.00)	0.11(0.02)
	JEMGM	16.21(0.47)	6.82(0.05)	39.24(0.20)	0.31(0.01)	0.99(0.00)	0.46(0.01)
	FGL	20.26(0.72)	7.89(0.09)	46.89(0.62)	0.05(0.02)	1.00(0.00)	0.12(0.02)
	GGL	20.15(0.83)	7.87(0.10)	46.42(0.64)	0.05(0.03)	1.00(0.00)	0.13(0.02)
	MPE	18.33(0.38)	7.32(0.10)	48.27(0.98)	0.47(0.02)	1.00(0.00)	0.63(0.01)
ER(1/4)	CLIME	20.60(0.27)	8.62(0.09)	57.13(0.86)	0.20(0.02)	0.98(0.00)	0.27(0.01)
	GLASSO	21.48(0.56)	8.22(0.19)	51.82(1.47)	0.17(0.05)	0.97(0.01)	0.19(0.03)
	JEMGM	19.17(0.42)	7.65(0.06)	47.07(0.21)	0.25(0.01)	0.99(0.00)	0.37(0.01)
	FGL	21.72(0.59)	8.51(0.05)	54.21(0.39)	0.11(0.01)	0.98(0.00)	0.16(0.01)
	GGL	21.28(0.36)	8.56(0.09)	54.37(0.65)	0.10(0.03)	0.99(0.00)	0.16(0.03)
	MPE	19.47(0.29)	8.36(0.10)	55.93(0.88)	0.32(0.01)	0.99(0.00)	0.49(0.01)
ER(1)	CLIME	28.73(0.30)	10.86(0.17)	72.98(1.68)	0.11(0.03)	0.98(0.01)	0.17(0.01)
	GLASSO	35.47(48.52)	11.40(10.62)	74.19(75.65)	0.11(0.08)	0.97(0.08)	0.13(0.02)
	JEMGM	28.25(0.62)	9.91(0.07)	62.74(0.19)	0.12(0.00)	0.99(0.00)	0.22(0.01)
	FGL	29.96(0.56)	10.56(0.08)	68.89(0.67)	0.05(0.02)	0.99(0.00)	0.09(0.01)
	GGL	30.25(0.66)	10.58(0.06)	68.71(0.51)	0.05(0.01)	0.99(0.00)	0.09(0.01)
	MPE	28.43(0.32)	10.60(0.09)	70.83(0.85)	0.19(0.01)	0.99(0.00)	0.34(0.01)

Table 3: Simulation results for data generated based on the geometric random graph with different ratios of the number of individual-specific edges to the number of shared edges. Results are based on 100 replications. The numbers in the brackets are standard errors. CLIME: method of Cai et al. (2011); GLASSO: graphical Lasso; JEMGM: method of Guo et al. (2011); FGLand GGL: methods of Danaher et al. (2013); MPE: proposed method.

Model(ρ)	Method	Performance					
		L_1	L_2	L_{F}	SEN	SPE	MCC
GR(0)	CLIME	5.48(0.11)	3.84(0.11)	19.43(0.89)	0.49(0.10)	1.00(0.00)	0.56(0.05)
	GLASSO	5.56(0.16)	3.83(0.21)	15.47(0.55)	0.19(0.10)	1.00(0.00)	0.30(0.03)
	JEMGM	4.98(0.17)	3.16(0.17)	12.40(0.52)	0.71(0.10)	0.99(0.00)	0.63(0.07)
	FGL	5.79(0.20)	4.00(0.16)	15.00(0.57)	0.30(0.14)	1.00(0.00)	0.39(0.04)
	GGL	5.52(0.13)	3.80(0.12)	15.41(0.29)	0.24(0.06)	1.00(0.00)	0.40(0.03)
	MPE	4.77(0.24)	3.57(0.18)	18.32(1.05)	0.81(0.11)	1.00(0.00)	0.81(0.02)
GR(1/4)	CLIME	6.92(0.13)	4.54(0.09)	22.88(0.71)	0.41(0.06)	0.99(0.00)	0.45(0.04)
	GLASSO	7.05(0.13)	4.51(0.18)	19.11(0.68)	0.19(0.06)	1.00(0.00)	0.27(0.02)
	JEMGM	6.19(0.19)	3.64(0.14)	15.45(0.69)	0.55(0.08)	0.99(0.00)	0.52(0.02)
	FGL	7.16(0.15)	4.46(0.10)	18.86(0.36)	0.24(0.05)	1.00(0.00)	0.31(0.02)
	GGL	7.14(0.10)	4.59(0.11)	19.27(0.38)	0.22(0.05)	1.00(0.00)	0.32(0.02)
	MPE	6.28(0.11)	4.15(0.07)	21.94(0.54)	0.76(0.02)	1.00(0.00)	0.76(0.02)
GR(1)	CLIME	7.84(0.14)	4.66(0.09)	26.75(0.74)	0.38(0.05)	0.99(0.00)	0.39(0.02)
	GLASSO	8.98(0.39)	4.52(0.12)	23.07(0.77)	0.19(0.07)	0.99(0.00)	0.21(0.04)
	JEMGM	7.61(0.24)	3.83(0.18)	19.05(0.89)	0.52(0.08)	0.99(0.00)	0.46(0.01)
	FGL	9.16(0.29)	4.56(0.08)	23.27(0.49)	0.16(0.05)	0.99(0.00)	0.19(0.03)
	GGL	8.95(0.27)	4.59(0.06)	23.73(0.30)	0.14(0.03)	0.99(0.00)	0.18(0.02)
	MPE	7.04(0.13)	4.31(0.08)	25.31(0.53)	0.69(0.02)	1.00(0.00)	0.72(0.01)

Table 4: Simulation results for data generated based on the Watt-Strogatz graph with different ratios of the number of individual-specific edges to the number of shared edges. Results are based on 100 replications. The numbers in the brackets are standard errors. CLIME: method of Cai et al. (2011); GLASSO: graphical Lasso; JEMGM: method of Guo et al. (2011); FGLand GGL: methods of Danaher et al. (2013); MPE: proposed method.

Model(ρ)	Method	Performance					
		L_1	L_2	L_{F}	SEN	SPE	MCC
WS(0)	CLIME	29.80(0.23)	13.05(0.19)	87.71(1.96)	0.11(0.03)	0.99(0.00)	0.25(0.02)
	GLASSO	29.55(0.25)	12.35(0.08)	79.20(0.47)	0.08(0.01)	0.99(0.00)	0.20(0.01)
	JEMGM	29.59(0.46)	11.89(0.27)	74.32(2.50)	0.11(0.04)	1.00(0.00)	0.27(0.03)
	FGL	29.45(0.33)	12.43(0.13)	80.10(1.22)	0.10(0.02)	0.99(0.00)	0.23(0.03)
	GGL	29.65(0.23)	12.65(0.11)	80.53(0.67)	0.08(0.02)	1.00(0.00)	0.22(0.02)
	MPE	28.99(0.22)	12.72(0.12)	84.35(1.11)	0.16(0.01)	1.00(0.00)	0.34(0.01)
WS(1/4)	CLIME	42.70(0.35)	14.80(0.19)	102.03(2.08)	0.08(0.02)	0.98(0.01)	0.15(0.01)
	GLASSO	42.79(0.58)	14.25(0.07)	95.38(0.49)	0.04(0.00)	0.99(0.00)	0.09(0.00)
	JEMGM	43.06(0.60)	13.44(0.17)	84.35(1.75)	0.12(0.02)	0.98(0.00)	0.21(0.01)
	FGL	43.33(0.59)	14.31(0.06)	96.24(0.49)	0.06(0.01)	0.98(0.00)	0.11(0.01)
	GGL	42.84(0.54)	14.30(0.11)	95.95(0.95)	0.05(0.01)	0.99(0.00)	0.11(0.01)
	MPE	42.15(0.39)	14.60(0.11)	100.13(1.16)	0.10(0.00)	0.99(0.00)	0.23(0.01)
WS(1)	CLIME	63.27(0.35)	18.64(0.19)	128.83(1.94)	0.04(0.01)	0.99(0.00)	0.08(0.01)
	GLASSO	62.94(0.31)	17.85(0.11)	120.11(1.01)	0.02(0.01)	0.99(0.00)	0.04(0.01)
	JEMGM	63.67(0.85)	16.82(0.23)	107.53(2.20)	0.07(0.01)	0.98(0.00)	0.12(0.01)
	FGL	63.11(0.28)	17.86(0.07)	120.82(0.66)	0.02(0.00)	0.99(0.00)	0.04(0.00)
	GGL	63.13(0.35)	17.87(0.12)	120.31(1.05)	0.03(0.01)	0.99(0.01)	0.04(0.01)
	MPE	62.89(2.57)	18.18(0.47)	123.98(2.71)	0.08(0.09)	0.98(0.09)	0.15(0.01)

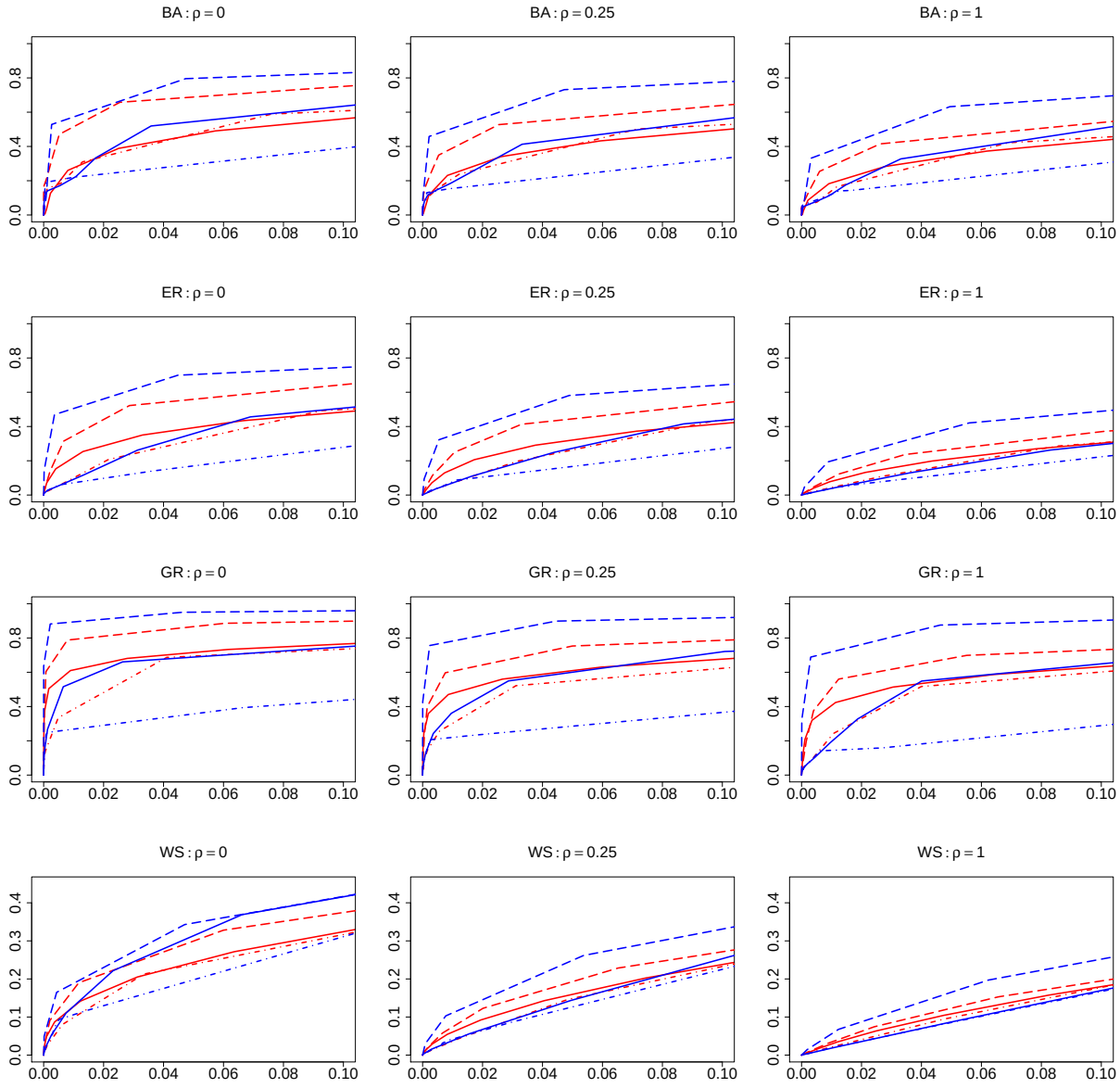

Figure 1: *Receiver operator characteristic curves for graph structure recovery for the simulated Barabási and Albert graphs (first row), the Erdős and Rényi graphs (second row), the geometric random graphs (the third row), and the Watts-Strogatz graphs (the fourth row). The x-axis and y-axis of each panel are average false positive rate and average sensitivity across $K = 3$ groups. Red solid line:* CLIME*; red dot-dashed line:* GLASSO*; red long-dashed line:* JEMGM*; blue solid line:* FGL*; blue dot-dashed line:* GGL*; blue long-dashed line:* MPE.

and immune cell types. The patients in the C1 subtype group has shown to have a lower survival rate compared to the patients from other subtypes. The data they used contain RNA expression data collected from $n = 78$ patients of C1 subtype and $n = 113$ patients from the other subtypes. We are interested to see how the wiring (conditional dependency) of the genes at the transcription levels differs among molecular subgroups of ovarian cancer. We focused on the apoptosis pathway from the KEGG database (Orgata et al., 1999; Kanehisa et al., 2012) to see whether the genes related to this pathway ($p = 87$) are differentially wired (conditionally dependent) between the C1 and other subtypes.

To stabilize the graph structure selection, we bootstrapped the samples 100 times within each of the two groups. At each time, I_{ik} is sampled uniformly taking values in $i = \{1, \ldots, n_k\}$, with $k = 1, 2$. Let $\tilde{\mathbf{x}}_i^{(k)} = \mathbf{x}_{I_{ik}}^{(k)}$, where $\mathbf{x}_{I_{ik}}^{(k)}$ is the p-dimensional gene expression data for the I_{ik}-th patient in the kth subtype group. The bootstrap sample is $\tilde{\mathbf{X}}^{(k)} = (\tilde{\mathbf{x}}_1^{(k)}, \ldots, \tilde{\mathbf{x}}_{n_k}^{(k)})$, with $k = 1, 2$. We then apply our proposed method and its competitors to each of the bootstrapped samples to obtain the estimators of the precision matrix $\hat{\mathbf{\Omega}}^{(k)}$. The support of the estimators are recorded so that $\tilde{\mathbf{\Omega}}^{(k)} = (I(\hat{\omega}_{ij}^{(k)} \neq 0))$. We then add $\tilde{\mathbf{\Omega}}^{(k)}$ up for all bootstrap samples and get the total frequency of each edge being selected. Those edges that were selected in more than 50 times out of 100 bootstrap samples were finally selected as important edges. This type of bootstrap aggregation methods has been studied by Meinshausen and Bühlmann (2010) and Li et al. (2013a). They found that thresholding the selection frequency can lead to better selection stability for precision matrix.

Table 5 lists the number of edges selected by the bootstrap aggregation of our proposed method and its competitors. The separate estimation methods (CLIME and GLASSO) resulted in graphs that share fewer edges in the precision matrices of the two cancer subtype groups. JEMGM resulted in most shared edges, followed by GGL and our method (MPE). Overall, FGL and GGL selected a lot more linked genes than other methods. Figure 2 displays the Gaussian graphical model estimated by these six different methods. FGL, GGL and MPE selected more unique edges among the gene expression levels for the C2-C6 subtype cancer than those of the C1 subtype. This suggests that the patients of the poor prognostic subtype (C1) lack some important edges among these Apoptosis genes.

We further define those nodes with degrees equal or larger than five based on the union of

the estimated graphs of two subtypes as the central nodes. FGL and GGL yielded estimators with most of the central nodes completely unlinked in the estimated graph for C1 cancer subtype. The estimators by MPE had several edges between the central nodes shared by both subtype groups, while also displayed some edges unique to each group. The central nodes identified by MPE are: FASLG, CASP10, CSF2RB, IL1B, MYD88, NFKB1, NFKBIA, PIK3CA, IKBKG and PIK3R5. Among these, CASP10, PIK3CA, IL1B and NFKb1 have been implicated in ovarian cancer risk or progression. In particular, PIK3CA has been implicated as an oncogene in ovarian cancer(Shayesteh et al., 1999), indicating the importance of these central genes in ovarian cancer progression.

Table 5: Number of edges selected by the proposed method and its competitors. "C1 unique" counts the number of edges that only appear in the precision matrix of the gene expression levels in C1 cancer subtype; "Other unique" counts the number of edges that only appear in C2-C6 cancer subtypes; and "Common" counts the number of edges shared by both precision matrices.

Method	C1 unique	Other unique	Common
CLIME	40	43	20
GLASSO	11	11	7
JEMGM	23	22	77
FGL	8	112	23
GGL	14	148	44
MPE	13	38	42

6. CONCLUDING REMARKS

We have developed a weighted constrained ℓ_∞/ℓ_1 minimization for jointly estimating multiple precision matrices. It was shown that when the precision matrices share a common support, the proposed method leads to more accurate estimation of the precision matrices and better recovery of the corresponding graph structures. Different from the penalized likelihood approaches proposed in literature (Guo et al., 2011), our approach is based on the constrained ℓ_∞/ℓ_1 minimization of the precision matrices. It can be regarded as an extension of the constrained ℓ_1 minimization procedure for single precision matrix (Cai et al., 2011). For support recovery, we showed that the

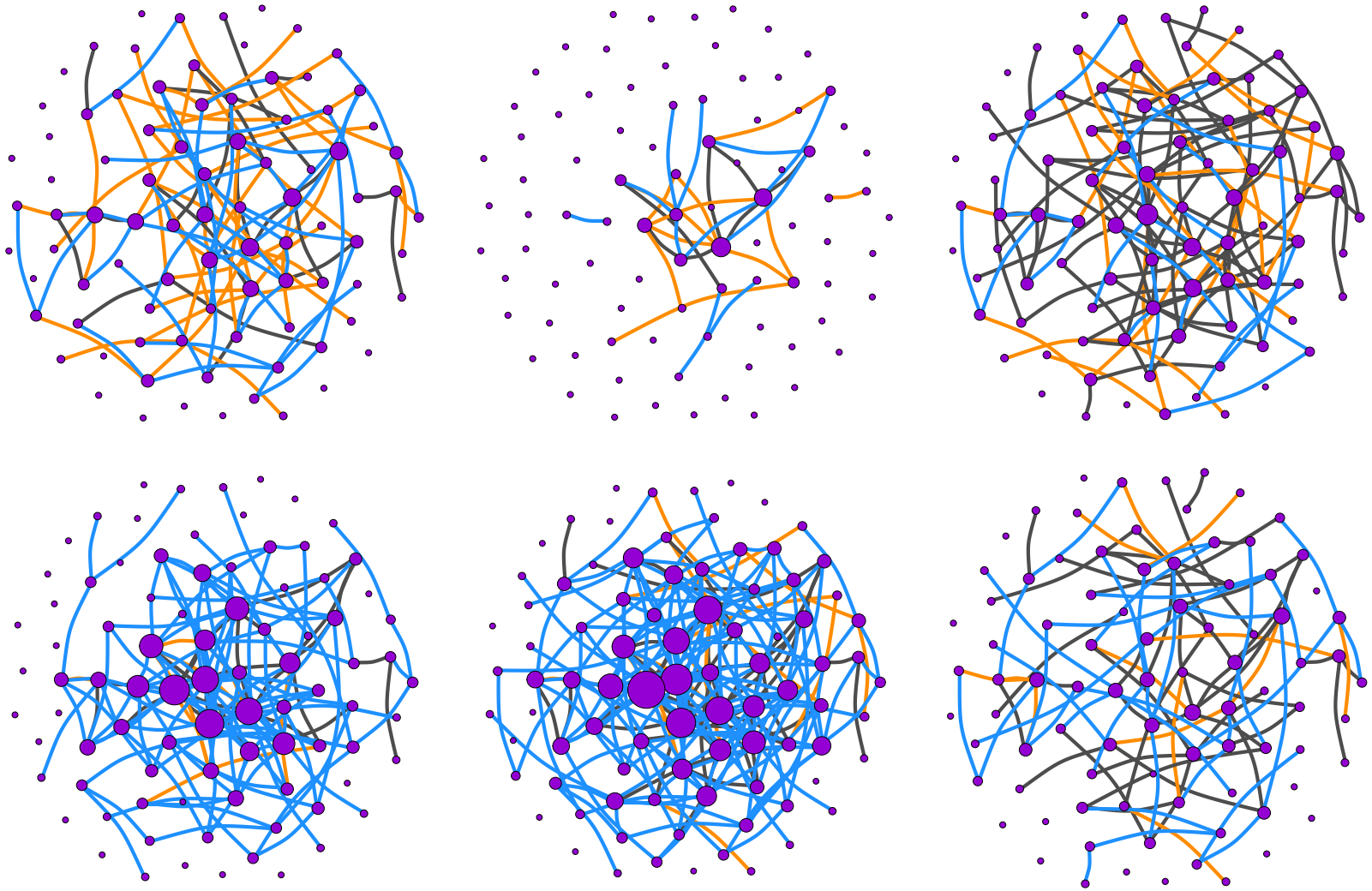

Figure 2: Estimated Gaussian Graphical model by the proposed method and its competitors.The orangle edges are edges unique to the precision matrix estimator of the C1 subtype, the blue edges are unique to that of other subtypes, and the dark grey edges are shared by both estimators. The size of node is a linear function of its degree. Upper left panel: CLIME; upper middle panel: GLASSO; upper right panel: JEMGM; bottom left panel: FGL; bottom middle panel: GGL; bottom right panel: MPE.

proposed method can recover the graph structure exactly whenever the minimum signal level is $2CM_n(\log K \log p/n)^{1/2}$, in contrast to the method of (Guo et al., 2011) and the method of Danaher, Wang and Witten, both of which require that the minimum signal level to be at least a constant. In addition, our method allows the observations of different groups to be dependent, while existing literature focuses on the case that all the observations are independent from each other.

APPENDIX: PROOFS OF THEOREMS

We first state a lemma which follows from Theorem 1 in Zaïtsev (1987).

Lemma 2 *Let $|\cdot|_K$ denotes the Euclidean norm of K dimensional vector. Suppose $X_1, \ldots, X_n$ be independent K-dimensional random vectors satisfying $EX_i = 0$ and $|X_i|_K \le M$ for $1 \le i \le n$. We have for any $\delta > 0$ and $x > \delta$*

$$P\Big(\Big|\sum_{k=1}^n X_k\Big|_K \ge x\Big) \le P\Big\{|N|_K \ge (x-\delta)/\lambda_{\max}^{1/2}\Big\} + c_1 K^{5/2}\exp(-c_2 K^{-5/2}\delta/M),$$

where $\lambda_{\max}$ is the largest eigenvalue of $Cov(\sum_{k=1}^n X_k)$, N is a d-dimensional standard normal random vector and c_1, c_2 are absolute positive constants.

Proof of Theorem 1. Suppose that the true $\mathbf{\Omega}^{(k)}$ belong to the above feasible set, that is

$$\max_{ij}\Big\{\sum_{k=1}^K w_k|(\hat{\Sigma}^{(k)}\mathbf{\Omega}^{(k)} - I)_{ij}|^2\Big\}^{1/2} \le \lambda_n. \tag{A.1}$$

We have

$$\begin{aligned}
&\max_{i,j}\Big\{\sum_{k=1}^K w_k|(\hat{\mathbf{\Omega}}_1^{(k)} - \mathbf{\Omega}^{(k)})_{ij}|^2\Big\}^{1/2} \\
&= \max_{i,j}\Big[\sum_{k=1}^K w_k|\{(\mathbf{\Omega}^{(k)}\hat{\Sigma}^{(k)} - I)\hat{\mathbf{\Omega}}_1^{(k)} - \mathbf{\Omega}^{(k)}(\hat{\Sigma}^{(k)}\hat{\mathbf{\Omega}}_1^{(k)} - I)\}_{ij}|^2\Big]^{1/2} \\
&\le \max_{i,j}\Big[\sum_{k=1}^K w_k|\{(\mathbf{\Omega}^{(k)}\hat{\Sigma}^{(k)} - I)\hat{\mathbf{\Omega}}_1^{(k)}\}_{ij}|^2\Big]^{1/2} + \max_{i,j}\Big[\sum_{k=1}^K w_k|\{\mathbf{\Omega}^{(k)}(\hat{\Sigma}^{(k)}\hat{\mathbf{\Omega}}_1^{(k)} - I)\}_{ij}|^2\Big]^{1/2} \\
&=: I_1 + I_2.
\end{aligned}$$

Note that

$$\{(\mathbf{\Omega}^{(k)}\hat{\Sigma}^{(k)} - I)\hat{\mathbf{\Omega}}_1^{(k)}\}_{ij} = \delta_{i\cdot}^{(k)}\hat{\omega}_{1\cdot j}^{(k)},$$

where $\delta_{i\cdot}^{(k)} =: (\delta_{i1}^{(k)}, \ldots, \delta_{ip}^{(k)})$ is the i-th row of $\mathbf{\Omega}^{(k)}\hat{\Sigma}^{(k)} - I$ and $\hat{\omega}_{1\cdot j}^{(k)} = (\hat{\omega}_{11j}^{(k)}, \ldots, \hat{\omega}_{1pj}^{(k)})^T$ is the j-th column of $\hat{\mathbf{\Omega}}_1^{(k)}$. We have

$$\begin{aligned} I_1 &\leq \max_{i,j}\Big(\sum_{k=1}^{K} w_k \sum_{1\leq l,m\leq p} \delta_{il}^{(k)}\delta_{im}^{(k)}\hat{\omega}_{1lj}^{(k)}\hat{\omega}_{1mj}^{(k)}\Big)^{1/2} \\ &\leq \max_{i,j}\Big(\sum_{1\leq l,m\leq p}\sum_{k=1}^{K} w_k|\delta_{il}^{(k)}\delta_{im}^{(k)}||\hat{\omega}_{1lj}^{(k)}\hat{\omega}_{1mj}^{(k)}|\Big)^{1/2}. \end{aligned}$$

Without loss of generality, we can assume that $w_K|\delta_{il}^{(K)}\delta_{im}^{(K)}| \leq \cdots \leq w_1|\delta_{il}^{(1)}\delta_{im}^{(1)}|$. Since by (A.1),

$$\sum_{k=1}^{K} w_k|\delta_{il}^{(k)}\delta_{im}^{(k)}| \leq 2^{-1}\sum_{k=1}^{K} w_k(|\delta_{il}^{(k)}|^2 + |\delta_{im}^{(k)}|^2) \leq \max_{i,j}\Big(\sum_{k=1}^{K} w_k|\delta_{ij}^{(k)}|^2\Big) \leq \lambda_n^2,$$

we have

$$\max_{i,l,m} w_k|\delta_{il}^{(k)}\delta_{im}^{(k)}| \leq k^{-1}\max_{i,l,m}\sum_{j=1}^{k} w_j|\delta_{il}^{(j)}\delta_{im}^{(j)}| \leq \lambda_n^2/k.$$

Therefore

$$\begin{aligned} I_1 &\leq \max_{i,j}\Big(\sum_{1\leq l,m\leq p}\sum_{k=1}^{K} k^{-1}|\hat{\omega}_{1lj}^{(k)}\hat{\omega}_{1mj}^{(k)}|\Big)^{1/2}\lambda_n \\ &\leq \Big(\sum_{k=1}^{K} k^{-1}\hat{M}_n^2\Big)^{1/2}\lambda_n \leq (\log K)^{1/2}\hat{M}_n\lambda_n, \end{aligned} \tag{A.2}$$

where $\hat{M}_n = \max_{1\leq k\leq K}\|\hat{\mathbf{\Omega}}_1^{(k)}\|_{l_1}$. Similarly, we can show that

$$I_2 \leq (\log K)^{1/2}M_n\lambda_n. \tag{A.3}$$

By the definition of $\hat{\mathbf{\Omega}}_1^{(k)}$, we have $\hat{M}_n \leq M_n$.

So it suffices to prove (A.1) holds with probability greater than $1 - O(p^{-\epsilon})$. Without loss of generality, we assume that $X_l^{(k)} \sim N(0, \Sigma^{(k)})$. Let $Y_{lij}^{(k)} = w_k^{1/2}\{(n_k-1)^{-1}[(X_l^{(k)}X_l^{(k)'}\mathbf{\Omega}^{(k)})_{ij} - e_{ij}]\}$ and $Y_{lij} = (Y_{lij}^{(1)}, \ldots, Y_{lij}^{(K)})$. When $l \geq n_k$, we set $Y_{lij}^{(k)} = 0$. Let $|\cdot|_K$ denotes the Euclidean norm of K dimensional vector. Then we have

$$\Big\{\sum_{k=1}^{K} w_k|(\hat{\Sigma}^{(k)}\beta^{(k)} - e_j)_i|^2\Big\}^{1/2} =^d \Big|\sum_{l=1}^{n} Y_{lij}\Big|_K.$$

For $1 \leq l \leq n$,$1 \leq k \leq K$ and $1 \leq i,j \leq p$, let

$$\hat{Y}_{lij}^{(k)} = Y_{lij}^{(k)}I\Big\{|Y_{lij}^{(k)}| \leq (n\log p)^{-1/2}K^{1/2-a}\Big\} - \mathsf{E}Y_{lij}^{(k)}I\Big\{|Y_{lij}^{(k)}|(n\log p)^{-1/2}K^{1/2-a}\Big\}$$

and $\hat{Y}_{lij} = (\hat{Y}_{lij}^{(1)}, \ldots, \hat{Y}_{lij}^{(K)})$. Note that $n \max_{i,j} \left| \mathsf{E}(Y_{lij} - \hat{Y}_{lij}) \right|_K = o(1)\lambda_n$. We have for any $\delta > 0$,

$$\mathsf{P}\Big(\Big| \sum_{l=1}^n Y_{lij} \Big|_K \geq \lambda_n \Big) \leq \mathsf{P}\Big(\Big| \sum_{l=1}^n \hat{Y}_{lij} \Big|_K \geq (1-\delta)\lambda_n \Big) + (\max_k n_k) K \max_{1 \leq k \leq K} \mathsf{P}\Big\{ |Y_{lij}^{(k)}| \geq \left(\frac{K^{1-2a}}{n \log p} \right)^{1/2} \Big\}. \quad (A.4)$$

Let $Z_{lij}^{(k)} = (X_l^{(k)} X_l^{(k)'} \mathbf{\Omega}^{(k)})_{ij} - e_{ij}$. We have for some constant $\eta > 0$,

$$\begin{aligned} & (\max_k n_k) K \max_{1 \leq k \leq K} \mathsf{P}\Big\{ |Y_{lij}^{(k)}| \geq \left(\frac{K^{1-2a}}{n \log p} \right)^{1/2} \Big\} \\ \leq & C n \max_{1 \leq k \leq K} \mathsf{P}\left\{ |Z_{lij}^{(k)}| \geq \left(\frac{n}{K^{2a} \log p} \right)^{1/2} \right\} \\ \leq & C \exp\left\{ \log n - \eta \left(\frac{n}{K^{2a} \log p} \right)^{1/2} \right\} = o(1) \end{aligned}$$

It is easy to show that

$$\lambda_{\max} \left\{ \sum_{l=1}^n \mathsf{Cov}(\hat{Y}_{lij}) \right\} \leq \{1 + o(1)\}(M_0 + 1)/n$$

uniformly for $1 \leq i, j \leq p$. Therefore it follows from (C1), Lemma 2, the tail probability of Chi-squared distribution and some tedious calculations that

$$\mathsf{P}\Big\{ \Big| \sum_{l=1}^n \hat{Y}_{lij} \Big|_K \geq (1-\delta)\lambda_n \Big\} \leq C \exp\{-C(\log p - K)\} + C \exp\left\{ \frac{5}{2} \log K - C_2 K^{a-4} (\log p) \right\} = o(1). \quad (A.5)$$

Combining (A.4)-(A.5), we prove that (A.1) holds.

Proof of Theorem 3. Suppose that

$$\max_{i,j} \Big\{ \sum_{k=1}^K w_k |(\hat{\mathbf{\Omega}}^{(k)} - \mathbf{\Omega}^{(k)})_{ij}|^2 \Big\}^{1/2} \leq C M_n \left(\frac{\log K \cdot \log p}{n} \right)^{1/2}.$$

For $i \in S_j^c$, $\max_{i,j} \Big\{ \sum_{k=1}^K w_k |(\hat{\mathbf{\Omega}}^{(k)})_{ij}|^2 \Big\}^{1/2} \leq C M_n (\log K \log p/n)^{1/2}$. Thus $(\breve{\mathbf{\Omega}}^{(k)})_{ij} = 0$ for $i \in S_j^c$. It yields that

$$\begin{aligned} \sum_{i=1}^p \Big\{ \sum_{k=1}^K w_k (\breve{\mathbf{\Omega}}^{(k)} - \mathbf{\Omega}^{(k)})_{ij}^2 \Big\}^{1/2} & \leq \sum_{i \in S_j} \Big\{ \sum_{k=1}^K w_k (\breve{\mathbf{\Omega}}^{(k)} - \mathbf{\Omega}^{(k)})_{ij}^2 \Big\}^{1/2} + \sum_{i \in S_j^c} \Big\{ \sum_{k=1}^K w_k (\breve{\mathbf{\Omega}}^{(k)})_{ij}^2 \Big\}^{1/2} \\ & \leq C M_n s_0(p) \left(\frac{\log K \cdot \log p}{n} \right)^{1/2}. \end{aligned}$$

Theorem 2 then follows from Theorem 1.

REFERENCES

Banerjee, O., Ghaoui, L. E., and d'Aspremont, A. (2008), "Model selection through sparse maximum likelihood estimation for Multivariate Gaussian or Binary Data," *J. Machine Learning Research*, 9, 485–516.

Barabási, A. and Albert, R. (1999), "Emergence of Scaling in Random Networks," *Science*, 286, 509–512.

Cai, T., Liu, W., and Luo, X. (2011), "A constrained ℓ_1 minimization approach to sparse precision matrix estimation," *Journal of American Statistical Association*, 106, 594–607.

Cai, T. T., Liu, W., and Zhou, H. (2012), "Estimating sparse precision matrix: Optimal rates of convergence and adaptive estimation," Tech. rep., University of Pennsylvania.

Candés, E. and Tao, T. (2007), "The Dantzig Selector: Statistical Estimation When p is Much Larger than n," *The Annals of Statistics*, 35, 2313–2351.

Dahl, J., Vandenberghe, L., and Roychowdhury, V. (2008), "Covariance selection for non-chordal graphs via chordal embedding," *Optimization Methods and Software*, 23, 501–420.

Danaher, P., Wang, P., and Witten, D. (2013), "The joint graphical lasso for inverse covariance estimation across multiple classes," *Journal of the Royal Statistical Society, Series B*, To appear.

Donoho, D., Elad, M., and Temlyakov, V. (2006), "Stable Recovery of Sparse Overcomplete Representations in the Presence of Noise," *IEEE Transactions on Information Theory*, 52, 6–18.

Erdös, P. and Rényi, A. (1960), "On the evolution of random graphs," *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*, 5, 17–61.

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

Greene, L. and Higman, V. (2003), "Uncovering network systems within protein structures," *Journal of Molecular Biology*, 334, 781–791.

Guo, J., Levina, E., Michailidis, G., and Zhu, J. (2011), "Joint Estimation of Mulitple Graphical Models," *Biometrika*, 98, 1–15.

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., and Tanabe, M. (2012), "KEGG for integration and interpretation of large-scale molecular data sets," *Nucleic Acids Research*, D109–D114.

Lauritzen, S. L. (1996), *Graphical Models*, Oxford: Clarendon Press.

Li, H. and Gui, J. (2006), "Gradient directed regularization for sparse Gaussian concentration graphs, with applications to inference of genetic networks," *Biostatistics*, 7, 302–317.

Li, S., H., L., Peng, J., and Wang, P. (2013a), "Bootstrap inference for network construction with an application to a breast cancer microarray study," *The Annals of Applied Statistics*, 391–417.

Li, S., Hsu, L., Wang, P., and Peng, J. (2013b), "Bootstrap Inference for Network Construction With an Application to a Breast Cancer Microarray Study," *Annals of Applied Statistics*, In press.

Liang, F., He, G., and Hu, Y. (2009), "A new smoothing Newton-type method for second-order cone programming problems," *Applied Mathematics and Compuation*, 215, 1020–1029.

Meinshausen, N. and Bühlmann, P. (2006), "High-dimensional graphs and variable selection with the Lasso," *Annals of Statistics*, 34, 1436–1462.

Meinshausen, N. and Bühlmann, P. (2010), "Stability selection," *Journal of Royal Statistics Society, Series B*, 417–473.

Newsterov, Y. and Todd, M. (1998), "Primal-Dual Interior-Point Methods for Self-Scaled Cones," *SIAM Journal on Optimization*, 8, 324–364.

Orgata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., and Kanehisa, M. (1999), "KEGG: Kyoto encyclopedia of genes and genomes," *Nucleic Acids Research*, 29–34.

Penrose, M. (2003), *Random geometric graphs*, Oxford University Press, U.S.A.

Schäfer, J. and Strimmer, K. (2005), "An empirical Bayes approach to inferring large-scale gene association networks," *Bioinformatics*, 21, 754–764.

Shayesteh, L., Lu, Y., Kuo, W., Baldocchi, R., Godfrey, T., Collins, C., Pinkel, D., Powell, B., Mills, G., and Gray, J. (1999), "PIK3CA is implicated as an oncogene in ovarian cancer," *Nature Genetics*, 21, 99–102.

Tothill, R., Tinker, A., George, J., Brown, R., Fox, S., Lade, S., Johnson, D., Trivett, J., Etemadmoghadam, D., Locandro, B., Traficante, N., Fereday, S., Hung, J., Chiew, Y., Haviv, I., Group, A. O. C. S., Gertig, D., deFazio, A., and Bowtell, D. (2008), "Novel Molecular Subtypes of Serous and Endometrioid Ovarian Cancer Linked to Clinical Outcome," *Clinical Cancer Research*, 14, 5198–5208.

Vendrascolo, M., Dokholyan, N., Paci, E., and Karplus, M. (2002), "Small-world view of the amino acids that play a key role in protein folding," *Physical Review E*, 65.

Watts, D. and Strogatz, S. (1998), "Collective dynamics of "small world" networks," *Nature*, 393, 440–442.

Whittaker, J. (1990), *Graphical Models in Applied Multivariate Analysis*, Wiley.

Yuan, M. (2010), "Sparse inverse covariance matrix estimation via linear programming," *Journal of Machine Learning Research*, 11, 2261–2286.

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

Zaïtsev, A. (1987), "On the Gaussian approximation of convolutions under multidimensional analogues of S.N. Bernstein's inequality conditions," *Probability Theory and Related Fields*, 74, 535–566.