Joint Estimation of Multiple High-dimensional Precision Matrices

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

October 12, 2015

Author's Footnote:

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 Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, NC 27707 (Email: jichun.xie@duke.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.

Abstract

Motivated by 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, which can be effectively implemented by a second-order cone programming. Compared to separate estimation methods, the proposed joint estimation method leads to estimators converging to the true precision matrices faster. Under certain regularity conditions, the proposed procedure leads to an exact graph structure recovery with a probability tending to 1. Simulation studies show that the proposed joint estimation methods outperform other methods in graph structure recovery. The method is illustrated through an analysis of an ovarian cancer gene expression data. The results indicate that the patients with poor prognostic subtype lack some important links among the genes in the apoptosis pathway.

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

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)). They have been successfully applied to infer relationships between genes at transcriptional level (Schäfer and Strimmer (2005); Li and Gui (2006); Li, Hsu, Peng et al. (2013)). Gaussian graphical models are tightly linked to precision matrices. Suppose $\boldsymbol{X} = (X_1, \ldots, X_p)'$ follows a multivariate Gaussian distribution $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The precision matrix $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ describes the graphical structure of its corresponding Gaussian graph. If the (i, j)-th entry of the precision matrix ω_{ij} is equal to zero, then X_i and X_j are independent conditioning on all other variables X_k, $k \neq i, j$. Correspondingly, no edge exists between X_i and variable X_j in the graphical structure of Gaussian graphical model. If $\omega_{ij} \neq 0$, then X_i and X_j are conditionally dependent and they are therefore connected in the graphical structure. Define the support of $\boldsymbol{\Omega}$ by $\mathcal{S} = \{(i, j) : \ \omega_{ij} \neq 0\}$, which is also the set of edges in the Gaussian graphical model. If the maximum degree of $\boldsymbol{\Omega}$, $\max_i \sum_{j=1}^{p} I(\omega_{ij} \neq 0)$, is relatively small, we call $\boldsymbol{\Omega}$ 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 of a set of genes is expected to be sparse.

Many methods for estimating high-dimensional precision matrix or its Gaussian graphical model have been developed in the past decade. Meinshausen and Bühlmann (2006) introduced a neighborhood selection approach by regressing all other variables on each variable with an ℓ_1 penalty. The method consistently estimates the set of non-zero elements of the precision matrix. Efficient algorithms for exact maximization of the ℓ_1-penalized log-likelihood have also been proposed. Yuan and Lin (2007), Banerjee, Ghaoui and d'Aspremont (2008) and Dahl, Vandenberghe and Roychowdhury (2008) adopted an interior point optimization method to solve this problem. Based on the work of Banerjee, Ghaoui and d'Aspremont (2008) and a block-wise coordinate descent algorithm, Friedman, Hastie and Tibshirani (2008) developed the graphical Lasso (GLASSO) for sparse precision matrix estimation; it is computationally efficient even when the dimension is greater than the sample size. Yuan (2010) developed a linear programming procedure and obtained oracle inequalities for the estimation error in term of matrix operator norm. Cai, Liu and Luo (2011) developed a constrained ℓ_1 minimization approach (CLIME) to estimate sparse precision matrix. All

of these methods addressed the problem of estimating a single precision matrix or a single Gaussian graphical model.

In many applications, the problem of estimating multiple precision matrices arises when data are collected among multiple groups. For example, gene expression levels are often measured over multiple groups (tissues, environments, or subpopulations). Their precision matrices and the corresponding graphical structures imply gene regulatory mechanisms and are of great biological interest. Since the gene regulatory networks in different groups are often similar to each other, the graphical structures share many common edges. Estimating a single precision matrix group by group ignores the partial homogeneity in their graphical structures, which often leads to low power. To effectively utilize the shared graphical structures and to increase the estimation precision, it is important to estimate multiple precision matrices jointly.

Previous attempts to jointly estimate multiple precision matrices include Guo, Levina, Michailidis et al. (2011) and Danaher, Wang and Witten (2014). Guo, Levina, Michailidis et al. (2011) proposed a hierarchical penalized model to perserve the common graphical structure while allowing differences across groups. Their method achieves Frobenius norm convergence when $p\log(p)/n$ goes to zero, where p is the number of variables, and n is the total sample size. Unfortunately, for genomic applications, the number of genes often exceeds the total sample size and, as a result, invalidates the theoretical justification in Guo, Levina, Michailidis et al. (2011). Danaher, Wang and Witten (2014) proposed two algorithms of joint graphical lasso (FGL and GGL) to estimate precision matrices that share common edges. Their approach is based upon maximizing a penalized log likelihood with a fused Lasso or group Lasso penalty. The paper did not provide any theoretical justification on the statistical convergence rate of their estimators.

In this paper, we propose a weighted constrained ℓ_∞/ℓ_1 minimization method to estimate K sparse precision matrices (MPE) jointly. Different from Guo, Levina, Michailidis et al. (2011) and Danaher, Wang and Witten (2014), the proposed estimators converge to the true precision matrices even when $p = O\{\exp(n^a)\}$, for some $0 < a < 1$. In addition, when K is sufficiently large, compared to the estimators from separate estimation methods, our proposed estimators converge to the true precision matrices (under the entry-wise ℓ_∞ norm loss) faster. An additional thresholding step on the estimators with a carefully chosen threshold yields thresholded estimators with additional

theoretical properties. The thresholded estimators from our method converge to the true precision matrices under the matrix ℓ_1 norm. Finally, when the graphical structures across groups are the same, our method leads to the exact recovery of the graph structures with a probability tending to 1.

The rest of the paper is organized as follows. Section 2 presents the estimation method and the optimization algorithm. Theoretical properties of the proposed method 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 method is compared with other competing methods. The method is also illustrated by an analysis of an epithelial ovarian cancer gene expression study in Section 5. A brief discussion is given in Section 6 and proofs are presented in the Appendix.

2. METHODOLOGY

The following notations are used in the paper. For a vector $\mathbf{a} = (a_1, \ldots, a_p)^T \in \mathbb{R}^p$, define $|\mathbf{a}|_1 = \sum_{j=1}^p |a_j|$ and $|\mathbf{a}|_2 = (\sum_{j=1}^p a_j^2)^{1/2}$. The vector $\mathbf{a}_{-i}$ is the vector of $\mathbf{a}$ without the i-th entry. The support of $\mathbf{a}$ is defined as $\mathtt{supp}(\mathbf{a}) = \{i : a_i \neq 0\}$. For a matrix $\mathbf{A} = (a_{ij}) \in \mathbb{R}^{p\times q}$, its entrywise ℓ_∞ norm is denoted by $|\mathbf{A}|_\infty = \max_{i,j} |a_{ij}|$. Its matrix ℓ_1-norm is denoted by $\|A\|_1 = \max_{1\leq j\leq q} \sum_{i=1}^p |a_{ij}|$, and its spectral norm is denoted by $\|A\|_2$. The sub-matrix $\mathbf{A}_{-i,-i}$ is the matrix of $\mathbf{A}$ without the i-th row and the i-th column. Denote by $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ 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 $|a_n| \leq C|b_n|$ holds for all sufficiently large n, write $a_n = o(b_n)$ if $\lim_{n\to\infty} a_n/b_n = 0$. If $a_n = O(b_n)$ and $b_n = O(a_n)$, then $a_n \asymp b_n$.

2.1 The Joint Estimation Method

We introduce an estimation method to jointly estimate K precision matrices with partial homogeneity in their graphical structures. The method uses a constrained ℓ_1 minimization approach, that has been successfully applied to high dimensional regression problems (Donoho, Elad and Temlyakov (2006); Candés and Tao (2007)) and signal precision matrix estimation problem (Cai, Liu and Luo (2011)) to recover the sparse vector or matrix.

For $1 \leq k \leq K$, let $\boldsymbol{X}^{(k)} \sim N(\boldsymbol{\mu}^{(k)}, \boldsymbol{\Sigma}^{(k)})$ be a p-dimensional random vector. The precision

matrix of $\boldsymbol{X}^{(k)}$, denoted by $\boldsymbol{\Omega}^{(k)} = (\omega_{ij}^{(k)})$, is the inverse of the covariance matrix $\boldsymbol{\Sigma}^{(k)}$. Suppose there are n_k identically and independently distributed random samples of $\boldsymbol{X}^{(k)}$: $\{\boldsymbol{X}_j^{(k)}, 1 \le j \le n_k\}$. The sample covariance matrix for the k-th group is

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

where $\bar{\boldsymbol{X}}^{(k)} = \sum_{j=1}^{n_k} \boldsymbol{X}_j^{(k)}/n_k$. Denote by $n = \sum_k n_k$ the total sample size and $w_k = n_k/n$ the weight of the k-th group. We estimate $\boldsymbol{\Omega}^{(k)} = (\omega_{ij}^{(k)})$ for $k = 1, \cdots, K$ by the constrained optimization

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

where $\lambda_n = C(\log p/n)^{1/2}$ 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 graphical structure.

Denote by $\hat{\boldsymbol{\Omega}}_1^{(k)}$ $(1 \le k \le K)$ the solution to (1). They are not symmetric in general. To make the solution symmetric, the estimator $\hat{\boldsymbol{\Omega}}^{(k)} = (\hat{\omega}_{ij}^{(k)})$ is constructed by comparing $\hat{\omega}_{1ij}^{(k)}$ and $\hat{\omega}_{1ji}^{(k)}$ and assigning the one with a smaller magnitude at both entries,

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

This symmetrizing procedure is not ad-hoc. The procedure assures that the final estimator $\hat{\boldsymbol{\Omega}}^{(k)}$ achieves the same entry-wise ℓ_∞ estimation error as $\hat{\boldsymbol{\Omega}}_1^{(k)}$. 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:

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

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

Lemma 1. *Suppose* $\hat{\mathbf{\Omega}}_1^{(k)}$ *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* $\hat{\mathbf{\Omega}}_1^{(k)} = \hat{\mathbf{B}}^{(k)}$ *for* $1 \leq k \leq K$.

Problem (2) can be solve by a second-order cone programming. The existing packages to solve (2) include the SDTP3 and the SeDuMi package in Matlab, and the CLSOCP package in R. CLSOCP uses a one-step smoothing Newton method of Liang, He and Hu (2009). This algorithm has good precision but works relatively slowly for high dimensional problems. SeDuMi and SDTP3 adopt 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 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 all the computations. 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

Choosing the tuning parameters in regularized estimation is in general a difficult problem. For linear regression models, Chen and Chen (2008); Wang, Li and Leng (2009); Wang and Zhu (2011) studied how to consistently choose the tuning parameters when $p = O(n^a)$ for some $a > 0$. Recently, Fan and Tang (2013) proposed a general information criterion (GIC) for choosing the tuning parameter for estimating the generalized linear model in ultra-high dimensional settings, $p = O(\exp(n^a))$ for some $a > 0$. The GIC criterion adopts a novel penalty on the degree freedom of the model so that it consistently chooses the proper tuning parameter under mild conditions. Unfortunately, the Gaussian graphical model is different from the generalized linear model, and therefore the justification of GIC does not apply to our problem. We propose a tuning parameter selection method based on BIC and a re-fitted precision matrix on the restricted model.

Since $\boldsymbol{X}_l^{(k)}$ $(1 \le l \le n_k)$ follows a multivariate Gaussian distribution $N_p(\mu^{(k)}, \boldsymbol{\Sigma}^{(k)})$, we have

$$X_{l,i}^{(k)} = \alpha_i^{(k)} + \boldsymbol{X}_{l,-i}^{(k)\prime}\boldsymbol{\beta}_i^{(k)} + \varepsilon_{l,i}^{(k)}, \tag{3}$$

where $\varepsilon_{l,i}^{(k)} \sim N(0, 1/\omega_{ii}^{(k)})$. The regression coefficients satisfy (Anderson (2003))

$$\alpha_i^{(k)} = \mu_i^{(k)} - \boldsymbol{\Sigma}_{i,-i}^{(k)}(\boldsymbol{\Sigma}_{-i,-i}^{(k)})^{-1}\boldsymbol{\mu}_{-i}^{(k)}, \quad \boldsymbol{\beta}_i^{(k)} = -(\omega_{ii}^{(k)})^{-1}\boldsymbol{\Omega}_{-i,i}^{(k)}. \tag{4}$$

Based on these results, we propose a tuning parameter selection procedure:

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

2. Define $\mathcal{S}_i^{(k)} = \{j : \hat{\omega}_{ij}^{(k)} \neq 0, j \neq i\}$, the set of non-zero off-diagonal elements of the i-th row of $\hat{\boldsymbol{\Omega}}^{(k)}$.

3. If $\texttt{Card}(\mathcal{S}_i^{(k)}) \ge n_k$, let the i-th column of $\hat{\boldsymbol{\Omega}}_2^{(k)}$ equal to the i-th column of $\hat{\boldsymbol{\Omega}}_{\cdot j}^{(k)}$, $\hat{\boldsymbol{\Omega}}_{2,\cdot i}^{(k)} = \hat{\boldsymbol{\Omega}}_{\cdot i}^{(k)}$. If $\texttt{Card}(\mathcal{S}_i^{(k)}) < n_k$, fit the regression model

$$X_{l,i}^{(k)} = \alpha_i^{(k)} + \sum_{j \in \mathcal{S}_i^{(k)}} X_{l,j}^{(k)\prime}\beta_{ij}^{(k)} + \epsilon_{l,i}^{(k)}. \tag{5}$$

   If $\mathcal{S}_i^{(k)}$ equals the true support $\mathcal{S}_{0,i}^{(k)} = \{j : \omega_{0,ij}^{(k)} \neq 0, j \neq i\}$, $\beta_{ij}^{(k)} = -\omega_{0,ij}^{(k)}/\omega_{0,ii}^{(k)}$ and $\mathsf{Var}(\epsilon_{l,i}^{(k)}) = 1/\omega_{0,ii}^{(k)}$. Suppose the solution to (5) is $\hat{\beta}_{ij}^{(k)}$. Define

$$\hat{\epsilon}_{l,i}^{(k)} = (X_{l,i}^{(k)} - \bar{X}_i^{(k)}) - \sum_{j \in \mathcal{S}_i^{(k)}} (X_{l,j}^{(k)} - \bar{X}_j^{(k)})'\hat{\beta}_{ij}^{(k)}.$$

   After fitting Model (5), let $\hat{\omega}_{2,ii}^{(k)} = n_k / \sum_{l=1}^{n_k}(\epsilon_{l,i}^{(k)})^2$, and $\hat{\omega}_{2,ij}^{(k)} = -\hat{\beta}_{ij}^{(k)}\hat{\omega}_{2,ii}^{(k)}$.

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

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

We use $\hat{\boldsymbol{\Omega}}_3^{(k)} = (\hat{\omega}_{3,ij}^{(k)})$, $k = 1, \ldots, K$ as the estimators corresponding to the tuning parameter λ. The optimal tuning parameter can be selected by Bayesian information criterion (BIC),

$$\text{BIC}(\lambda) = \sum_{k=1}^{K} \left\{ n_k \texttt{tr}\left(\hat{\boldsymbol{\Sigma}}^{(k)}\hat{\boldsymbol{\Omega}}_3^{(k)}\right) - n_k \log(\det \hat{\boldsymbol{\Omega}}_3^{(k)}) + \log(n_k) s_k \right\}, \tag{6}$$

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

Since the refitted estimator can potentially reduce some bias introduced in the optimization due to the penalty term, using it in BIC (6) improves the tuning parameter selection in the numerical studies. However, using $\hat{\mathbf{\Omega}}_3^{(k)}$ as an estimator of $\mathbf{\Omega}^{(l)}$ is not recommended because when $\texttt{Card}(\mathcal{S}_i^{(k)})$ is large, the re-fitted estimator might lead to overfitting. Overfitting does not severely affect the tuning parameter selection, because BIC puts penalties on complicated models that are less likely to be chosen.

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

(C3). If $e_{ij} = I(i=j)$, $Z_{ij}^{(k)} = (\boldsymbol{X}^{(k)}\boldsymbol{X}^{(k)\prime}\mathbf{\Omega}^{(k)})_{ij} - e_{ij}$, $Z_{ij} = (Z_{ij}^{(1)}, \ldots, Z_{ij}^{(K)})'$, and the largest eigenvalue of $\texttt{Cov}(Z_{ij})$ is $\lambda_{\max,ij}$, $\sup_{ij}\lambda_{\max,ij} \leq M_1$.

(C4). $n_1 \asymp n_2 \asymp \cdots \asymp n_K \asymp n/K$.

Condition (C1) allows p to grow exponentially fast as n. It also allows the number of groups K to grow slowly with p and n. For example, when $\log p = O(n^r)$ and $K = O(n^b)$ for $r + 2b < 1$ and $3b < r$, (C1) holds. Condition (C3) allows $\boldsymbol{X}^{(k)}$ to be dependent across groups. When $\boldsymbol{X}^{(k)}$ are independent, $\texttt{Cov}(Y_{ij}) = \mathbf{I}_K$, and thus $\max_{ij}\lambda_{\max,ij} = 1$.

Let $M_n = \sup_{1\leq k\leq K}\max_j\sum_{i=1}^p|\omega_{ij}^{(k)}| = \sup_{1\leq k\leq K}\|\mathbf{\Omega}^{(k)}\|_1$ be the maximum matrix ℓ_1 norms of the K matrices. A 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 $C_0 > \sqrt{2M_0+2}$. If (C1)-(C4) hold,*

$$\sup_{i,j}\Big\{\sum_{k=1}^{K} w_k|(\hat{\mathbf{\Omega}}^{(k)} - \mathbf{\Omega}^{(k)})_{i,j}|^2\Big\}^{1/2} \le C_1 M_n\left(\frac{\log K \cdot \log p}{n}\right)^{1/2} \tag{7}$$

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 use the tuning parameter selection method, such as BIC in (6), to choose λ_n. The details are discussed in Section 2.3.

Remark 2: Theorem 1 (and Theorems 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 $\mathbf{\Omega}^{(k)}$, $k = 1, \ldots, K$.

Define

$$\mathcal{U}(M) = \Big\{(\mathbf{\Omega}^{(1)}, \ldots, \mathbf{\Omega}^{(K)}) : \ \lambda_{\max}(\mathbf{\Omega}^{(k)})/\lambda_{\min}(\mathbf{\Omega}^{(k)}) \le M, \\ \|\mathbf{\Omega}^{(k)}\|_1 \asymp M_n = o\{(n/\log p)^{1/2}\}, \ k = 1, \ldots, K\Big\}.$$

Proposition 1. *Let $\tilde{\mathcal{U}}$ 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 $\{\boldsymbol{X}_j^{(k)}; 1 \le j \le n_k\}$. Assume the samples are independent across K groups. Under (C4), there exists a constant $\alpha > 0$, such that*

$$\inf_{i,j}\inf_{(\tilde{\mathbf{\Omega}}^{(1)},\ldots,\tilde{\mathbf{\Omega}}^{(K)})\in\tilde{\mathcal{U}}}\sup_{(\mathbf{\Omega}^{(1)},\ldots,\mathbf{\Omega}^{(K)})\in\mathcal{U}} P\left[\Big\{\sum_{k=1}^{K} w_k|(\tilde{\mathbf{\Omega}}^{(k)} - \mathbf{\Omega}^{(k)})_{i,j}|^2\Big\}^{1/2} \ge \alpha M_n\left(\frac{K\cdot\log p}{n}\right)^{1/2}\right] \ge \alpha^K,$$

for sufficiently large n.

Theorem 1 shows that the convergence rate of $\Big\{\sum_{k=1}^{K} w_k|(\tilde{\mathbf{\Omega}}^{(k)} - \mathbf{\Omega}^{(k)})_{i,j}|^2\Big\}^{1/2}$ from our joint estimation method is less than or equal to $C_1 M_n(\log K \log p/n)^{1/2}$ with a probability tending to 1. When K is bounded, Proposition 1 shows that with a non-vanishing probability α^K, the minimax convergence rate of any separate estimation method is at least $\alpha M_n(K\log p/n)^{1/2}$. For bounded but sufficiently large K, $C_1(\log K)^{1/2} \le \alpha K^{1/2}$. Therefore, the convergence upper bound $C_1 M_n\left(\frac{\log K\cdot\log p}{n}\right)^{1/2}$ in (7) for the joint estimation method is less than the convergence lower bound $\alpha M_n\left(\frac{K\cdot\log p}{n}\right)^{1/2}$ in Proposition 1 for the separate estimation method. In other words, compared to separate estimation methods, our joint estimation method leads to estimators with a faster convergence.

An additional thresholding step on the estimators with a careful chosen threshold leads to new estimators, that converge to the true precision matrices under the matrix operator norm. Define the thresholded estimator $\breve{\mathbf{\Omega}}^{(k)} = (\breve{\omega}_{ij}^{(k)})$ as follows,

$$\breve{\omega}_{ij}^{(k)} = \hat{\omega}_{ij}^{(k)} I\left\{ \left(\sum_{k=1}^{K} w_k (\hat{\omega}_{ij}^{(k)})^2 \right)^{1/2} > C_1 M_n \left(\frac{\log K \cdot \log p}{n} \right)^{1/2} \right\}$$

with C_1 the constant defined in (7). Let $\mathcal{S}_j^{(k)} = \{(i,j) : \ \omega_{ij}^{(k)} \neq 0, \ i < j\}$ and $\mathcal{S}_j = \cup_{k=1}^{k} S_j^{(k)}$. Define $s_0(p) = \max_{1 \leq j \leq p} \mathrm{Card}(\mathcal{S}_j)$ as the union sparsity.

Theorem 2. *Suppose that (C1)-(C4) hold. Then with a high probability converging to 1,*

$$\max_j \sum_{i=1}^{p} \Big\{ \sum_{k=1}^{K} w_k (\breve{\mathbf{\Omega}}^{(k)} - \mathbf{\Omega}^{(k)})_{ij}^2 \Big\}^{1/2} \leq C_1 M_n s_0(p) \left(\frac{\log K \cdot \log p}{n} \right)^{1/2}. \tag{8}$$

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, $s_0(p) = \max_{1 \leq j \leq p} \mathtt{Card}(\mathcal{S}_j^{(k)})$ for all $k = 1, \ldots, K$. When the number of shared edges in the graphical structures increases, the union sparsity $s_0(p)$ decreases, and consequently $\breve{\mathbf{\Omega}}^{(k)}$ converges to $\mathbf{\Omega}^{(k)}$ faster.

Let $\hat{a}_{ij} = \sqrt{\sum_{k=1}^{K} w_k (\breve{\mathbf{\Omega}}^{(k)} - \mathbf{\Omega}^{(k)})_{ij}^2}$. Matrix $\hat{\mathbf{A}} = (\hat{a}_{ij})_{p \times p}$ measures the overall estimation errors among the entries of $\breve{\mathbf{\Omega}}^{(k)}$ and $\mathbf{\Omega}^{(k)}$ for $k = 1, ..., K$.

Corollary 1. *Suppose that the conditions in Theorem 2 hold. Then with a high probability converging to 1,*

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

3.2 Graphical Structure Recovery

Theoretical analysis for graphical structure recovery is very complicated when the graphical structures of the precision matrices are different across the K groups since the results depend on the structures of the shared edges. Here we focus on the case in which the K precision matrices have a common support. Let $\mathcal{S}_k = \{(i,j) : \omega_{ij}^{(k)} \neq 0\}$ be the support for the k-th precision matrix, and the common support be $\mathcal{S} = \cap_k \mathcal{S}_k$. When $\mathcal{S}_1 = \ldots = \mathcal{S}_K$, $\mathcal{S} = \mathcal{S}_k$, $k = 1, \ldots, K$, 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 a result on support recovery.

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

$$\theta_n > 2C_1 M_n \left(\frac{\log K \cdot \log p}{n} \right)^{1/2}. \tag{9}$$

Then $\hat{\mathcal{S}} = \mathcal{S}$ with a high probability converging to 1.

When the graphical structures are the same across all K groups, the lower bound condition (9) is weaker than the lower bound condition needed for graphical structure recovery by separate estimation methods. Based on Proposition 1 and its proof, to fully recover the shared graphical structure by separate estimation methods, a necessary condition is

$$\inf_{(i,j)\in\mathcal{S}} \inf_{k} |\omega_{ij}^{(k)}| > 2\alpha M_n (K \log p/n)^{1/2}.$$

When K is sufficiently large, this condition is stronger than (9).

4. SIMULATION STUDIES

4.1 Data generation

We evaluated the numerical performance of the proposed method and other competitive methods, including the separate precision matrix estimation procedures proposed by Friedman, Hastie and Tibshirani (2008) and Cai, Liu and Luo (2011) and the joint estimation method proposed by Guo, Levina, Michailidis et al. (2011) and Danaher, Wang and Witten (2014). The separate precision matrix estimation methods were applied to each group, and therefore ignored the partial homogeneity in graphical structures among groups. In all numerical studies, we set $p = 200$, $K = 3$ and $(n_1, n_2, n_3) = (80, 120, 150)$. The simulated observations were generated in each group independently from a multivariate Gaussian distribution $\mathrm{N}\{0, (\mathbf{\Omega}^{(k)})^{-1}\}$, where $\mathbf{\Omega}^{(k)}$ is the precision matrix in the k-th group. For each model, 100 replications were performed.

We present results for two different types of graphical models: the Erdös and Rényi (ER) model (Erdös and Rényi (1960)) and the Watts-Strogatz (WS) model (Watts and Strogatz (1998)). For the

ER model, the graph contains p vertices and each pair of vertices are connected with a probability 0.05. For the WS model, first a ring lattice of p vertex is created; one vertex is connected with its neighbors within order distance of 15, and then the edges of the lattice are rewired uniformly and randomly with a probability 0.01. These graph models have several topological properties such as sparcity and a "small world" property often observed in true biological gene regulatory networks. See Fell and Wagner (2000); Jeong, Tombor, Albert et al. (2000); Vendrascolo, Dokholyan, Paci et al. (2002); Greene and Higman (2003).

Based on the ER model or the WS model, a common graph structure was generated. Let M be the number of edges in the common graph structure. Then, $\lfloor \rho M \rfloor$ random edges were added to the common graph structure to generate graph structures for each group, where the parameter ρ was the ratio between the number of group-specific edges and common edges. We considered $\rho = 0$, $1/4$, and 1. The first setting represents the scenario where the precision matrices in all groups share the same graph structure. After the graph structure of each group was determined, the values of non-zero off-diagonal entries were generated independently as uniform in $[-1, -0.5] \cup [0.5, 1]$. The diagonal values were assigned to a constant so that the condition number of each precision matrix was equal to p.

4.2 Simulation results

Each method was evaluated for a range of tuning parameters under each model. The optimal tuning parameter was chosen by BIC (6). Several measures are used to compare the performances of these estimators. The estimation error was evaluated in terms of average matrix ℓ_1 norm, ℓ_2 norm (spectral norm), and Frobenius norm:

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

The graph structure recovery results were evaluated by average sensitivity (SEN), specificity (SPE), and Matthews correlation coefficient (MCC). For a true precision matrix $\mathbf{\Omega}_0 = (\omega_{0,ij})$ with

support set $\mathcal{S}_0 = \{(i,j) : \omega_{0,ij} \neq 0 \text{ and } i \neq j\}$, suppose 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:

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

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

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

We compared $\hat{\mathbf{\Omega}}^{(k)}$ 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 and 2. It shows that when $\rho = 0$, 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, and the joint estimation methods gradually lose their advantages. Our method has the best performance in graph structure recovery among all the methods. Even when $\rho = 1$, it still performs significantly better than the separate estimation methods. Our method has the best performance in graph structure recovery and it achieves the highest Matthews correlation coefficient. Its estimation error measured under the matrix ℓ_1, ℓ_2, and the Frobenius norm are comparable to 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 the false positive rate controlled under 10%. The methods proposed by Danaher, Wang and Witten (2014) have two tuning parameters. For each sparsity tuning parameter, we first chose an optimal similarity tuning parameter from a grid of candidates by BIC and then plotted the ROC curves based on a sequence of sparsity tuning parameters and their corresponding optimal similarity tuning parameters. Because these methods involve choosing two tuning parameters, they are slower than our method in computation (The computation of FGL and GGL is based on the R package "JGL"

Table 1: 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, Liu and Luo (2011); GLASSO: graphical Lasso; JEMGM: method of Guo, Levina, Michailidis et al. (2011); FGLand GGL: methods of Danaher, Wang and Witten (2014); 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 2: 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, Liu and Luo (2011); GLASSO: graphical Lasso; JEMGM: method of Guo, Levina, Michailidis et al. (2011); FGLand GGL: methods of Danaher, Wang and Witten (2014); 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)

contributed by Danaher (2013)). Figure 1 shows that our method consistently outperforms the other methods in graph structure recovery.

5. EPITHELIAL OVARIAN CANCER DATA ANALYSIS

Epithelial ovarian cancer is a molecularly diverse cancer that lacks effective personalized therapy. Tothill, Tinker, George et al. (2008) identified six molecular subtypes of ovarian cancer, labeled as C1–C6, where the C1 subtype was characterized by significant differential expressions of genes associated with stromal and immune cell types. The patients in C1 subtype group have shown to have a lower survival rate compared to the patients from other subtypes. The data set includes 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 focus on the apoptosis pathway from the KEGG database (Orgata, Goto, Sato et al. (1999); Kanehisa, Goto, Sato et al. (2012)) to see whether the genes related to this pathway ($p = 87$) are differentially wired between the C1 and other subtypes.

To stabilize the graph structure selection, we bootstraped the samples 100 times within each group. At each time, I_{ik} was 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 applied our proposed method and its competitors to each of the bootstrapped samples to obtain the estimators of the two precision matrices $\hat{\mathbf{\Omega}}^{(k)}, k = 1, 2$. The supports of the estimators were recorded so that $\tilde{\mathbf{\Omega}}^{(k)} = (I(\hat{\omega}_{ij}^{(k)} \neq 0))$. We then added $\tilde{\mathbf{\Omega}}^{(k)}$ up for all bootstrap samples and got the final frequency of each edge being selected. Those edges that were selected in more than 50 of the 100 bootstrap samples were finally selected as important edges. This type of bootstrap aggregation methods has been commonly used in recovering the sparse graphical structures (Meinshausen and Bühlmann (2010); Li, Hsu, Peng et al. (2013)), which often leads to better selection stability for sparse precision matrix.

Table 3 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

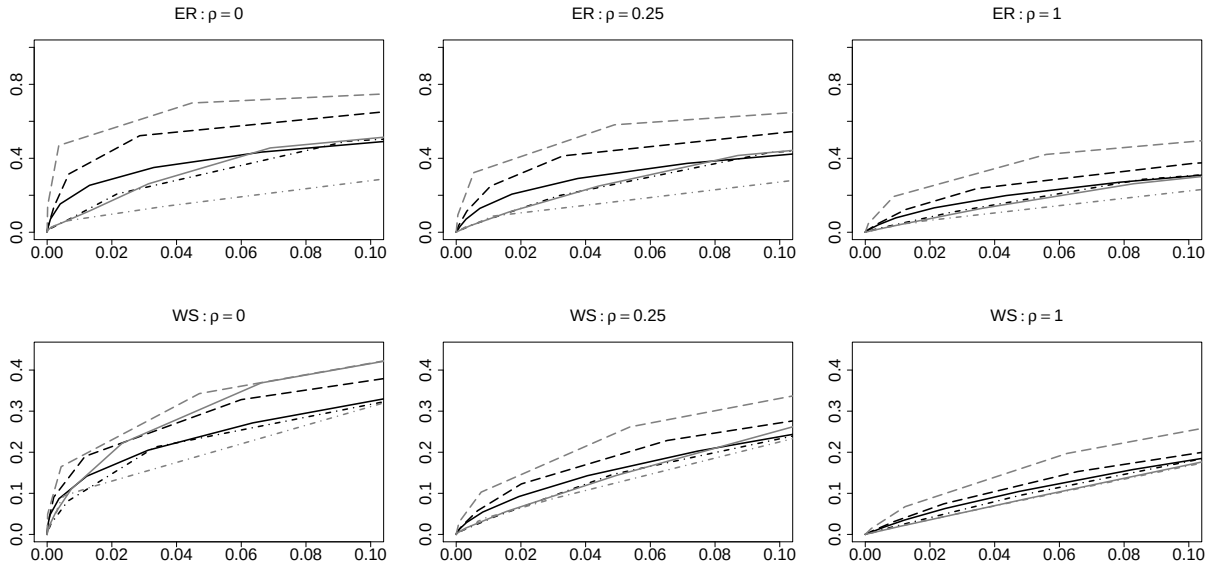

Figure 1: *Receiver operator characteristic curves for graph structure recovery for the simulated Erdős and Rényi graphs (the first row), and the Watts-Strogatz graphs (the second row). The x-axis and y-axis of each panel are average false positive rate and average sensitivity across* $K = 3$ *groups.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

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 shows the Gaussian graphs estimated by these six methods. FGL, GGL and MPE selected more unique edges among the gene expression levels for the C2-C6 subtype cancer than those for the C1 subtype. This suggests that the patients with poor prognostic subtype (C1) lack some important links among the Apoptosis genes.

We further defined the nodes with degrees equal or larger than five based on the union of the estimated graphs of two subtypes as the hub nodes. FGL and GGL yielded estimators with most of the hub nodes completely unlinked in the estimated graph for C1 cancer subtype. The estimators by MPE had several edges between the hub nodes shared by both subtype groups, while also displaying some links unique to each group. The hub nodes identified by MPE were 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, Lu, Kuo et al. (1999)), indicating the importance of these hub genes in ovarian cancer progression.

Table 3: 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

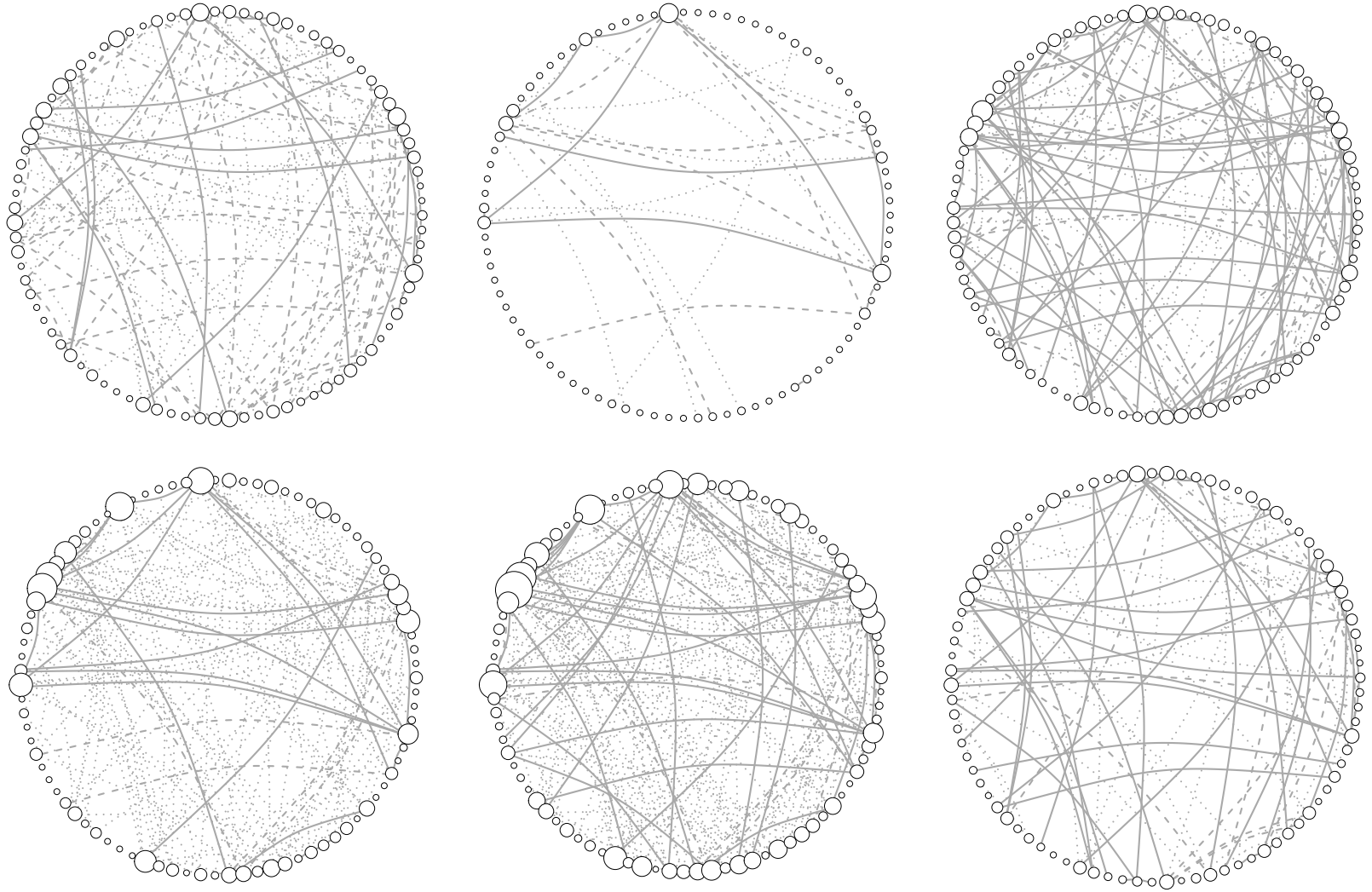

Figure 2: Estimated Gaussian Graphs by the proposed method and its competitors. The dashed edges are links unique to the precision matrix estimator of the C1 subtype, the dotted edges are unique to that of other subtypes, and the solid edges are shared by both estimators. The size of node is a linear 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.

6. DISCUSSION

It is of interest to discuss the connection and difference between the problem considered in this paper and the problem of estimating matrix graphical models (Leng and Tang (2012); Yin and Li (2012); Zhou (2014)). Matrix graphical models consider the random matrix variate $\mathbf{X}$ following the distribution $MN_{p\times q}(\mathbf{M};\mathbf{U},\mathbf{V})$ with the probability density function (pdf)

$$p(\mathbf{X}\mid\mathbf{M},\mathbf{U},\mathbf{V})=(2\pi)^{-pq/2}|\mathbf{U}^{-1}|^{q/2}|\mathbf{V}^{-1}|^{q/2}\exp\left\{-\mathtt{tr}\left[(\mathbf{X}-\mathbf{M})'\mathbf{V}^{-1}(\mathbf{X}-\mathbf{M})'\mathbf{U}^{-1}/2\right]\right\}.$$

Here $\mathbf{M}\in\mathbb{R}^{p\times q}$ is the mean matrix, $\mathbf{U}\in\mathbb{R}^{p\times p}$ is the row covariance matrix, and $\mathbf{V}\in\mathbb{R}^{q\times q}$ is the column covariance matrix. Thus, each row and each column of $\mathbf{X}$ share the same covariance matrices $\mathbf{U}$ and $\mathbf{V}$. This distribution implies that $\mathrm{vec}(\mathbf{X})$ follows a vector multivariate Gaussian distribution $N_{pq}(\mathrm{vec}(\mathbf{M}),\mathbf{U}\otimes\mathbf{V})$, where "vec" is the vectorization operator and "$\otimes$" is the Kronecker product.

Our model assumes $X^{(k)}$ follows the vector multivariate Gaussian distribution $N_p(\boldsymbol{\mu},\boldsymbol{\Sigma}^{(k)})$. If $n=n_1=\ldots=n_K$, $\mathbf{X}=(\boldsymbol{X}^{(1)},\ldots,\boldsymbol{X}^{(K)})\in\mathbb{R}^{p\times K}$ is a random matrix variate. However, each column of $\mathbf{X}$ has its own covariance matrix $\boldsymbol{\Sigma}^{(k)}$, and each row of $\mathbf{X}$ may also have its own. Therefore, the covariance matrix of $\mathrm{vec}(\mathbf{X})$ cannot be expressed as the Kronecker product of two positive definite matrices. In general, the degree of freedom of $\mathsf{Cov}(\mathrm{vec}(\mathbf{X}))$ is larger than that of $\mathbf{U}\otimes\mathbf{V}$ mentioned above.

7. SUPPLEMENTARY MATERIAL

In the supplementary material, we provide additional simulation studies to compare the proposed methods and other competitive methods. We also include the proofs of the theorems.

ACKNOWLEDGMENTS

T. Tony Cai's research was supported in part by NSF Grants DMS-1208982 and DMS-1403708, and NIH Grant R01 CA127334. Hongzhe Li's research was supported in part by NHI Grants R01 CA127334 and R01 GM097505. Weidong Liu's research was supported in part by NSFC, Grants No. 11201298, No. 11322107, the Program for Professor of Special Appointment (Eastern Scholar) at Shanghai Institutions of Higher Learning Shanghai Pujiang Program, Shanghai Shuguang Program and 973 Program (2015CBB56004). Jichun Xie's research was supported in part by NIH Grant UL1 TR001117.

We thank the Editor, an associate editor and the reviewers for their insightful comments and suggestions that have helped us substantially improve the quality of the paper.

REFERENCES

Anderson, T. W. (2003) *An introduction to multivariate statistical analysis.* Wiley-Interscience.

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.

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.

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.

Chen, J. and Chen, Z. (2008) Extended bayesian information criterion for model selection with large model space. *Biometrika* 95, 232–253.

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. (2013) *JGL: Performs the Joint Graphical Lasso for sparse inverse covariance estimation on multiple classes.* R package version 2.3.

Danaher, P., Wang, P., and Witten, D. (2014) The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society, Series B* 76-2, 373–397.

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.

Fan, Y. and Tang, C. (2013) Tuning parameter selection in high dimensional penalized likelihood. *Journal of the Royal Statistical Society* 75, 531–552.

Fell, D. and Wagner, A. (2000) The small world of metabolism. *Nature Biotechnology* 18-11, 1121–1122.

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.

Jeong, H., Tombor, B., Albert, R., Oltvai, Z., and Barabási, A. (2000) The large-scale organization of metabolic networks. *Nature* 407-6804, 651–654.

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* 40, D109–D114.

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

Leng, C. and Tang, C. (2012) Sparse matrix graphical models. *Journal of American Statistical Association* 107-499, 1187–2012.

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., Hsu, L., Peng, J., and Wang, P. (2013) Bootstrap inference for network construction with an application to a breast cancer microarray study. *The Annals of Applied Statistics* 7-1, 391–417.

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* 72, 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* 27, 29–34.

Ren, Z., Sun, T., Zhang, C.-H., and Zhou, H. (2014) Asymptotic normality and optimality in estimation of large gaussian graphical model. *Annals of Statistics* To appear.

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.

Wang, H., Li, B., and Leng, C. (2009) Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of Royal Statistical Society, Series B* 71, 671–683.

Wang, T. and Zhu, L. (2011) Consistent tuning parameter selection in high dimensional sparse linear regression. *Journal of Multivariate Analysis* 102, 1141–1151.

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.

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. (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.

Zhou, S. (2014) Gemini: Graph estimation with matrix variate normal instances. *Annals of Statistics* 42, 532–562.

Joint Estimation of Multiple High-dimensional Precision Matrices

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

In this supplementary file, we include additional simulations studies and the proofs of the theorems in the paper.

S.1. ADDITIONAL SIMULATION STUDIES

In addition to two graph models (ER model and WS model) discussed in Section 4, we generate common graph structures based on another two commonly-used graph models, Barabási and Albert model and geometric random graph model. 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 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.

Based on each graph model (ER model or WS model), a common graph structure is generated. The following data generation and evaluation procedure is the same as described in Section 4. Figure S1 shows the ROC curves of all the methods. Clearly, our joint estimation method (MPE) outperforms all other methods. We also applied tuning parameter selection procedure on all these methods and compared the performance of the resulting estimators. Table 1 and Table 2 list the results. our joint estimation method (MPE) has the highest MCC, so that it can recover the graphical structure better.

S.2. PROOFS OF THEOREMS

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

Lemma 2. *Let $|\cdot|_K$ denote the Euclidean norm of K dimensional vector. Suppose $X_1, \ldots, X_n$ are 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 $x > \delta$*

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

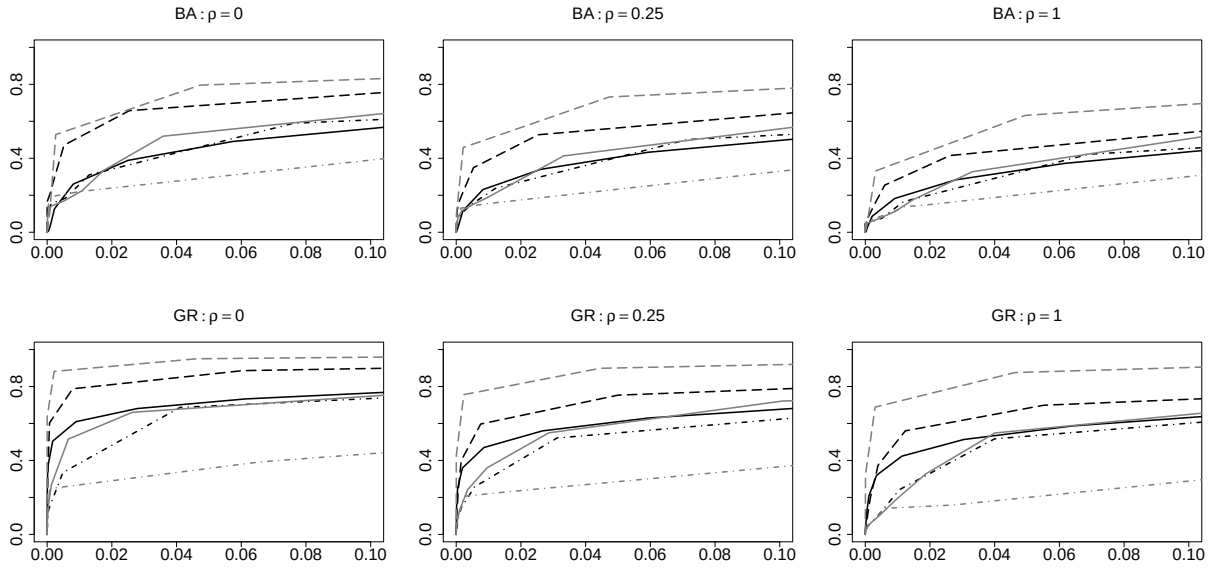

Figure S1: *Receiver operator characteristic curves for graph structure recovery for the simulated Barabási and Albert graphs (first row), and the geometric random graphs (the second row). The x-axis and y-axis of each panel are average false positive rate and average sensitivity across* $K = 3$ *groups. Black solid line:* CLIME*; black dot-dashed line:* GLASSO*; black long-dashed line:* JEMGM*; grey solid line:* FGL*; grey dot-dashed line:* GGL*; grey long-dashed line:* MPE.

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, Liu and Luo (2011); GLASSO: graphical Lasso; JEMGM: method of Guo, Levina, Michailidis et al. (2011); FGLand GGL: methods of Danaher, Wang and Witten (2014); 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 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, Liu and Luo (2011); GLASSO: graphical Lasso; JEMGM: method of Guo, Levina, Michailidis et al. (2011); FGLand GGL: methods of Danaher, Wang and Witten (2014); 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)

where $\lambda_{\max}$ is the largest eigenvalue of $\mathsf{Cov}(\sum_{k=1}^n X_k)$, N is a K-dimensional standard Gaussian 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

$$\max_{ij}\Big\{\sum_{k=1}^K w_k|(\hat{\Sigma}^{(k)}\mathbf{\Omega}^{(k)}-I)_{ij}|^2\Big\}^{1/2}\le\lambda_n. \tag{S1}$$

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

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

Assume that $w_K|\delta_{il}^{(K)}\delta_{im}^{(K)}|\le\cdots\le w_1|\delta_{il}^{(1)}\delta_{im}^{(1)}|$. Since by (S1),

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

we have

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

Therefore

$$\begin{aligned} I_1 &\leq \max_{i,j}\Big(\sum_{1\leq l,m\leq p}\sum_{k=1}^{K}k^{-1}|\hat{\omega}_{1lj}^{(k)}\hat{\omega}_{1mj}^{(k)}|\Big)^{1/2}\lambda_n \\ &\leq \Big(\sum_{k=1}^{K}k^{-1}\hat{M}_n^2\Big)^{1/2}\lambda_n \leq (\log K)^{1/2}\hat{M}_n\lambda_n, \end{aligned} \tag{S2}$$

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

$$I_2 \leq (\log K)^{1/2}M_n\lambda_n. \tag{S3}$$

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

So it suffices to prove (S1) holds with probability greater than $1-O(p^{-\epsilon})$. Assume a new set of random variables $\tilde{X}_l^{(k)}$ independently follows $N(0,\Sigma^{(k)})$. Let $Y_{lij}^{(k)} = w_k^{1/2}n_k^{-1}\{(\tilde{X}_l^{(k)}\tilde{X}_l^{(k)'}\mathbf{\Omega}^{(k)})_{ij} - \tilde{e}_{ij}\}$ and $Y_{lij} = (Y_{lij}^{(1)},\ldots,Y_{lij}^{(K)})$, where $\tilde{e}_{ij} = e_{ij}n_k/(n_k-1)$. When $l \geq n_k-1$, we set $Y_{lij}^{(k)} = 0$. Let $|\cdot|_K$ denotes the Euclidean norm of K dimensional vector. Because $\sum_{l=1}^{n_k}(X_l^{(k)}-\bar{X}^{(k)})(X_l^{(k)}-\bar{X}^{(k)})'$ follows the same Wishart distribution as $\sum_{l=1}^{n_k-1}\tilde{X}_l^{(k)}\tilde{X}_l^{(k)'}$, we have $\Big\{\sum_{k=1}^{K}w_k|(\hat{\Sigma}^{(k)}\beta^{(k)}-e_j)_i|^2\Big\}^{1/2}$ follows the same distribution as $\Big|\sum_{l=1}^{n}Y_{lij}\Big|_K$. For $1\leq l\leq n$,$1\leq k\leq K$ and $1\leq i,j\leq p$, let

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

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

$$\mathsf{P}\Big(\Big|\sum_{l=1}^{n}Y_{lij}\Big|_K \geq \lambda_n\Big) \leq \mathsf{P}\Big(\Big|\sum_{l=1}^{n}\hat{Y}_{lij}\Big|_K \geq (1-\delta)\lambda_n\Big) + (\max_k n_k)K\max_{1\leq k\leq K}\mathsf{P}\Big\{|Y_{lij}^{(k)}| \geq \Big(\frac{K^{1-2a}}{n\log p}\Big)^{1/2}\Big\}. \tag{S4}$$

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

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

By Condition (C3), it is easy to show that

$$\lambda_{\max}\left\{\sum_{l=1}^{n}\mathsf{Cov}(\hat{Y}_{lij})\right\} \leq \{1+o(1)\}(M_1+1)/n$$

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

$$\mathsf{P}\Big\{\Big|\sum_{l=1}^{n}\hat{Y}_{lij}\Big|_K \ge (1-\delta)\lambda_n\Big\} \le C\exp\{-C(\log p - K)\} + C\exp\left\{\frac{5}{2}\log K - C_2K^{a-4}(\log p)\right\} = o(1). \tag{S5}$$

Combining (S4)-(S5), we prove that (S1) holds. □

Proof of Proposition 1. Consider $K = 1$ first. Define $\mathcal{V}(M, M_n) = \{\mathbf{\Omega} : \lambda_{\max}(\mathbf{\Omega})/\lambda_{\min}(\mathbf{\Omega}) \le M,\ \|\mathbf{\Omega}\|_1 \asymp M_n = o\{(n/\log p)^{1/2}\}\}$. The proof of Proposition 1 follows the proof of Theorem 5 in Ren, Sun, Zhang et al. (2014). Construct $\mathbf{\Omega}_0$ and $\mathbf{\Omega}_m$ in the same way. In the proof, $\max_{1\le j\le p}\sum_{i\ne j} I(\omega_{ij} \ne 0) \le k_{n,p}$. Let $k_{n,p}$ satisfy $k_{n,p}(\log p/n_1)^{1/2} \to \infty$ and $k_{n,p} = o(n/\log p)$. Then $M_n \asymp \|\mathbf{\Omega}_m\|_1 \asymp (\log p/n_1)^{1/2}k_{n,p}$. Following the proof of Theorem 5, Eqn. (59) in Ren, Sun, Zhang et al. (2014) will lead to

$$\inf_{1\le m\le m^*}|\omega_{11}^{(m)} - \omega_{11}^{(0)}| \ge C_3M_n(\log p/n_1)^{1/2}.$$
$$\inf_{1\le m\le m^*}|\omega_{12}^{(m)} - \omega_{12}^{(0)}| \ge C_4M_n(\log p/n_1)^{1/2}.$$

Following the rest of the proof, for some constant $c_1 > 0$ and $\alpha_1 > 0$, we have

$$\inf_{i,j}\inf_{\hat{\omega}_{ij}}\sup_{\mathbf{\Omega}\in\mathcal{V}(M,M_n)}\mathsf{P}\left\{|\hat{\omega}_{ij} - \omega_{ij}| \ge c_1M_n(\log p/n_1)^{1/2}\right\} \ge \alpha_1. \tag{S6}$$

Now consider the case $K \ge 1$. For each $\mathbf{\Omega}^{(k)} \in \mathcal{V}(M, M_n)$ and separate estimation method, (S6) holds. Therefore, for some c_0 satisfying $c_0K\log p/n \le \inf_k \log p/n_k$,

$$\begin{aligned}
&\inf_{i,j}\inf_{\tilde{\mathcal{U}}}\sup_{\mathcal{U}}\mathsf{P}\left\{\left\{\sum_{k=1}^{K}w_k|(\tilde{\mathbf{\Omega}}^{(k)} - \mathbf{\Omega}^{(k)})_{i,j}|^2\right\}^{1/2} \ge c_0M_n(K\log p/n)^{1/2}\right\}\\
\ge\ &\inf_{i,j}\inf_{\tilde{\mathcal{U}}}\sup_{\mathcal{U}}\mathsf{P}\left\{\inf_k|\hat{\omega}_{ij}^{(k)} - \omega_{ij}^{(k)}| \ge c_0M_n(K\log p/n)^{1/2}\right\}\\
\ge\ &\inf_{i,j}\inf_{\tilde{\mathcal{U}}}\sup_{\mathcal{U}}\prod_k\mathsf{P}\left\{|\hat{\omega}_{ij}^{(k)} - \omega_{ij}^{(k)}| \ge c_1M_n(\log p/n_k)^{1/2}\right\}\\
\ge\ &\alpha_1^K,
\end{aligned}$$

which leads to Proposition 1.

□

Proof of Theorem 2. Suppose that

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

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

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

Theorem 2 then follows from Theorem 1. □