Submitted to the Annals of Statistics
arXiv: arXiv:0000.0000

STRUCTURAL SIMILARITY AND DIFFERENCE TESTING ON MULTIPLE SPARSE GAUSSIAN GRAPHICAL MODELS

By Weidong Liu*

Shanghai Jiao Tong University

We present a new framework on inferring structural similarities and differences among multiple high-dimensional Gaussian graphical models (GGMs) corresponding to the same set of variables under distinct experimental conditions. The new framework adopts the partial correlation coefficients to characterize the potential changes of dependency strengths between two variables. A hierarchical method has been further developed to recover edges with different or similar dependency strengths across multiple GGMs. In particular, we first construct two-sample test statistics for testing the equality of partial correlation coefficients and conduct large-scale multiple tests to estimate the substructure of differential dependencies. After removing differential substructure from original GGMs, a follow-up multiple testing procedure is used to detect the substructure of similar dependencies among GGMs. In each step, false discovery rate is controlled asymptotically at a desired level. Power results are proved, which demonstrate that our method is more powerful on finding common edges than the common approach that separately estimates GGMs. The performance of the proposed hierarchical method is illustrated on simulated datasets.

1. Introduction. Gaussian graphical models (GGMs) are popular tools for studying dependency networks of multivariate random variables. In recent years, much interest has focused upon estimating high-dimensional sparse GGMs. There is a rich and growing literature on learning GGMs in various settings; see, for example, Meinshausen and Buhlmann (2006), Yuan and Lin (2007), Friedman, et al. (2008), d'Aspremont et al. (2008), Rothman, et al. (2008), Fan, et al. (2009), Ravikumar, et al. (2011), Yuan (2010), Zhang (2010), Cai, et al. (2011), Liu, et al. (2012), Xue and Zou (2012), Liu (2013) and Ren, et al. (2014).

*Research supported by NSFC, Grants No.11322107 and No.11431006, the Program for Professor of Special Appointment (Eastern Scholar) at Shanghai Institutions of Higher Learning, Shanghai Shuguang Program and 973 Program (2015CB856004).

MSC 2010 subject classifications: 62H12, 62H15

Keywords and phrases: common substructure, false discovery rate, Gaussian graphical model, high dimensional, structural similarity, structural difference

In many problems arising in social, biological, and other fields, observations are collected under distinct experimental conditions. It is not unusual for this type of dataset to have multiple different but related GGMs. For example, GGMs of normal tissue gene expression data and patients' gene expression data often possess different structures, and meanwhile they can also share some common edges. In genomic studies, it is of great interest to see how the network connected by gene pairs changes from one experimental condition to another. In fact, these changes may offer an important clue regarding an underlying biological process such as identification of pathways that correspond to such a change (Gill, Datta and Datta 2010, Chu, et al. 2011). The Pearson correlation coefficient and partial correlation coefficient are widely used to measure the strength of the association between two genes, and their differences can be used to quantify the change of genetic networks (Gill, Datta and Datta 2010, de la Fuente 2010, Schäfer and Strimmer 2005). Testing differences in GGMs is also useful in applications in brain connectivity analysis (Belilovsky, Varoquaux and Blaschko 2015). In addition to identifying network changes, it is of great interest to learn common nonzero edges with close weights in various networks (Hara and Washio 2013). For example, the common nonzero edges can help identify the invariant gene pathways across several conditions. In summary, it is important to conduct structural difference and similarity analysis for networks based on partial correlation under two or more experimental conditions.

More formally, let $\mathcal{G}_k = G(V, E_k)$, $1 \le k \le K$, denote K GGMs over a set of nodes under K distinct experimental conditions, where V represents a set of p nodes of interest. In each graph, $\{X_1^{(k)}, \ldots, X_p^{(k)}\} \sim N(\boldsymbol{\mu}^{(k)}, \boldsymbol{\Sigma}^{(k)})$ denotes a population of node states under the k-th experimental condition and E_k is a set of edges with

$$\mathsf{Cov}(X_i^{(k)}, X_j^{(k)} | X_m^{(k)}, m \neq i, j) \neq 0.$$

Under each experimental condition, independent and identically distributed random samples $\{\boldsymbol{X}_i^{(k)}, 1 \le i \le n_k\}$ are collected from $N(\boldsymbol{\mu}^{(k)}, \boldsymbol{\Sigma}^{(k)})$. In recent years, much attention has been focused on joint estimation of multiple GGMs. A majority of current investigations depend on three types of optimization methods. The first approach is a generalization of the well-known method called *graphical lasso*, which minimizes the regularized negative log-likelihood function

$$\min_{\Theta_1,\ldots,\Theta_K} \Big\{ \sum_{k=1}^{K} (\langle \hat{\boldsymbol{\Sigma}}_k, \Theta_k \rangle - \log |\Theta_k|) + P(\Theta_1, \ldots, \Theta_K) \Big\}, \tag{1.1}$$

where $\hat{\mathbf{\Sigma}}_k, 1 \le k \le K$, are the sample covariance matrices and $P(\Theta_1, \ldots, \Theta_K)$ is a penalty function. Guo, et al. (2011) proposed a hierarchical penalty for $P(\cdot)$. Danaher, et al. (2012) developed an algorithm for (1.1) with the group lasso and fused lasso penalties. The latter penalty was also considered in Yang, et al. (2013). Honorio and Samaras (2010) proposed a penalty function of the form $P(\Theta_1, \ldots, \Theta_K) = \sum_{1 \le i,j \le p} \max_{1 \le k \le K} |\theta_{ij,k}|$, where $\Theta_k = (\theta_{ij,k})_{1 \le i,j \le p}$. Hara and Washio (2013) considered a problem of estimating a common substructure $\{(i,j) : \omega_{ij,1} = \cdots = \omega_{ij,K} \ne 0, i \ne j\}$, where $\mathbf{\Omega}^{(k)} = (\mathbf{\Sigma}^{(k)})^{-1} = (\omega_{ij,k})_{1 \le i,j \le p}$. The second approach is the neighborhood selection method which solves

$$\min_{\beta_1, \ldots, \beta_K} \Big\{ \sum_{k=1}^{K} \|\mathbf{X}_i^{(k)} - \mathbf{X}_{-i}^{(k)} \beta_k\|_2^2 + P(\beta_1, \ldots, \beta_K) \Big\}, \tag{1.2}$$

where $\mathbf{X}^{(k)} = (\boldsymbol{X}_1^{(k)}, \ldots, \boldsymbol{X}_{n_k}^{(k)})' \in \mathbb{R}^{n_k \times p}$, $\mathbf{X}_i^{(k)}$ is the i-th column of $\mathbf{X}^{(k)}$ and $\mathbf{X}_{-i}^{(k)}$ denotes the remaining matrix with $\mathbf{X}_i^{(k)}$ being removed. In this approach, Zhang and Wang (2010) studied the case when $K = 2$ by using the fused lasso penalty. Chiquet, et al. (2011) introduced the graphical intertwined lasso penalty and cooperative-lasso penalty. The third approach is introduced by Zhao, et al. (2014). They proposed to estimate $\mathbf{\Omega}^{(1)} - \mathbf{\Omega}^{(2)}$ directly by the constrained l_1 regularization.

The structural similarities generated by these optimization-based methods rely on the equality of entries of precision matrices or regression coefficient vectors. It is inevitable that their changes might come from a variant of conditional variances. In other words, the above three classes of methods may capture spurious structural changes. A legitimate dependency measure to overcome this problem is the partial correlation coefficient, which is widely used in aforementioned genetic studies. In this paper, we will define a new framework for inferring structural similarities and differences based on partial correlation coefficients.

The false discovery rate (FDR), originally introduced for multiple testing (Benjamini and Hochberg, 1995), is particularly useful in evaluating the quality of an estimated genetic network, see Schäfer and Strimmer (2005), Zhu, et al. (2005), Li and Gui (2006), Ma, Gong and Bohnert (2007), Gill, Datta and Datta (2010). It has also been used in numerical study to measure the accuracy of differential edge estimation of multiple GGMs (Danaher, et al. 2014). However, as mentioned before, most existing methods on joint estimation of GGMs depend on optimization techniques. It is not easy to choose a tuning parameter to control a desired FDR while keeping nontrivial statistical power. In a single GGM estimation problem, Liu (2013) proposed

a procedure to control the FDR. However, the FDR control in the joint estimation of multiple GGMs remains an open problem. In this paper, we will propose a hierarchical method to estimate differential structures and similar structures under asymptotic control of FDR.

Now we introduce the framework on statistical inference of structural differences and similarities. Let $\rho_{ij\cdot,k}$ denote the partial correlation coefficient of $X_i^{(k)}$ and $X_j^{(k)}$ given $X_m^{(k)}$ with $m \neq i, j$. It is well-known that $\rho_{ij\cdot,k} = -\omega_{ij,k}/\sqrt{\omega_{ii,k}\omega_{jj,k}}$. Let

$$D_{ij}(\boldsymbol{\rho}) = \sqrt{\sum_{1\leq k<l\leq K} (\rho_{ij\cdot,k} - \rho_{ij\cdot,l})^2}.$$

The set

$$\mathcal{A}_1 = \{(i,j) : D_{ij}(\boldsymbol{\rho}) \neq 0, 1 \leq i < j \leq p\}$$

includes all pairs of nodes with different partial correlation coefficients across distinct experimental conditions, which is referred to as *differential substructure*. It typically contains useful information in the analysis of differential co-expression gene networks. The estimation of differential substructure can be cast as a multiple testing problem

$$H_{0ij} : \ D_{ij}(\boldsymbol{\rho}) = 0 \quad \text{versus} \quad H_{1ij} : \ D_{ij}(\boldsymbol{\rho}) \neq 0, \tag{1.3}$$

$1 \leq i < j \leq p$.

For a single GGM estimation, Liu (2013) proposed a new test statistic $T_{ij}^{(k)}$ for the testing problem of $\rho_{ij\cdot,k} = 0$ with the asymptotic distribution $N(\rho_{ij\cdot,k}, (1-\rho_{ij\cdot,k}^2)^2/n_k)$ (see Section 3). For the two-sample case $K = 2$, estimation of $\mathcal{A}_1$ can be done as in Liu (2013) with a parallel extension. It is noteworthy that the situation for $K \geq 3$ is technically more involved. In this case, we will use the test statistic $\sqrt{\sum_{1\leq k<l\leq K} w_{kl}(T_{ij}^{(k)} - T_{ij}^{(l)})^2}$ for H_{0ij}, where w_{kl} are some weights. Its asymptotic null distribution becomes a non-standard Chi-squared distribution and more novel theoretical techniques are needed to establish the FDR control result.

The complementary set of $\mathcal{A}_1$ can be further split into two parts

$$\begin{aligned} \mathcal{A}_2 &= \{(i,j) : \rho_{ij\cdot,1} = \cdots = \rho_{ij\cdot,K} \neq 0, 1 \leq i < j \leq p\}, \\ \mathcal{A}_3 &= \{(i,j) : \rho_{ij\cdot,1} = \cdots = \rho_{ij\cdot,K} = 0, 1 \leq i < j \leq p\}. \end{aligned}$$

Note that $\mathcal{A}_2$ is a set of common edges with nonzero equal partial correlation coefficients. In the analysis of gene networks, it is important to recover $\mathcal{A}_2$. Unfortunately, when nonzero $\rho_{ij\cdot,1}, \ldots, \rho_{ij\cdot,K}$ are extremely close

to each other, with a limited amount of samples, it is difficult to classify (i,j) into $\mathcal{A}_1$ or $\mathcal{A}_2$ correctly. In fact, it is more likely to classify (i,j) into $\mathcal{A}_2$ because in this case any test for (1.3) with small type I error rate is powerless. As a result, controlling the FDR in the estimation of $\mathcal{A}_2$ seems too ambitious and even impossible in the extreme case. Instead, we propose a hierarchical method to estimate a similar substructure. We first estimate the differential substructure by a random set $\hat{\mathcal{A}}_1$ including all (i,j) with $D_{ij}(\boldsymbol{\rho}) \geq C\sqrt{\log p/n}$ for some constant $C > 0$; see Section 2. Then the set of node pairs with nonzero partial correlation coefficients in $\hat{\mathcal{A}}_1^c$

$$\hat{\mathcal{A}}_2 = \{(i,j) \in \hat{\mathcal{A}}_1^c : \ (\rho_{ij\cdot,1}, \ldots, \rho_{ij\cdot,K}) \neq 0\}$$

is called *similar substructure* of multiple GGMs, which allows nearly equal $\rho_{ij\cdot,1}, \ldots, \rho_{ij\cdot,K}$. The similar substructure includes most of the common edges due to the small type I error in the recovery of the differential substructure. In this paper, we are interested in recovering $\hat{\mathcal{A}}_2$ instead of $\mathcal{A}_2$, which can be formulated into the following multiple testing problem,

$$\begin{aligned} H_{0ij} : \ & (\rho_{ij\cdot,1}, \ldots, \rho_{ij\cdot,K}) = 0 \\ \text{versus} \quad & H_{1ij} : (\rho_{ij\cdot,1}, \ldots, \rho_{ij\cdot,K}) \neq 0 \end{aligned} \tag{1.4}$$

for $(i,j) \in \hat{\mathcal{A}}_1^c$.

Actually, under some regularity conditions, estimating the similar substructure $\hat{\mathcal{A}}_2$ is equivalent to estimating the common substructure $\mathcal{A}_2$. For example, assume that the differential substructure satisfies the lower bound condition $\min_{(i,j)\in\mathcal{A}_1} \sqrt{\sum_{1\leq k<l\leq K}(\rho_{ij\cdot,k} - \rho_{ij\cdot,l})^2} \geq C\sqrt{\log p/n}$ for some large $C > 0$, then Theorem 3.3 ensures that $\hat{\mathcal{A}}_1^c \subseteq \mathcal{A}_2$ with probability tending to one. Hence, any procedure for (1.4) with FDR at level α essentially estimates the common structure $\mathcal{A}_2$ with FDR at level α. Similar lower bound conditions on signals are frequently used in model selection consistency in high-dimensional regression and GGM estimations. We will give more discussions on what is estimated by (1.4) at the end of Section 2.

A challenge to solve (1.4) is that the indices of these hypothesis tests belong to a random set. The existing FDR control procedures typically require the indices to be non-random, otherwise, it will be hard to derive null distributions of test statistics. In fact, for a test statistic T_i and a random set $\mathcal{I}$, $\mathsf{P}(T_i \leq t | i \in \mathcal{I})$ can be different with $\mathsf{P}(T_i \leq t)$ when $\mathcal{I}$ and T_i are correlated. Therefore, a careful construction of test statistics is crucial to guarantee null distribution invariance following selection of indices for testing by the data at hand. Note that, without the restriction $(i,j) \in \hat{\mathcal{A}}_1^c$, the weighted sum squared test statistic $\sum_{k=1}^K n_k (T_{ij}^{(k)})^2$, which can be viewed as a likelihood

ratio test statistic under the approximation $T_{ij}^{(k)} \sim N(\rho_{ij\cdot,k}, (1-\rho_{ij\cdot,k}^2)^2/n_k)$, is a natural choice for (1.4). However, due to $(i,j) \in \hat{\mathcal{A}}_1^c$, the likelihood ratio test statistic turns to be $\sum_{k=1}^K n_k T_{ij}^{(k)}$, which is shown to be asymptotically independent of the sum squared type test statistic for (1.3); see Lemma 6.3. As a result, the disturbance from index randomness in hypothesis tests can be negligible, and a FDR control procedure will be developed.

Negahban and Wainwright (2011) studied simultaneous support recovery for multiple high-dimensional linear regression problems with the block l_1/l_∞ regularization. They showed that the performance of simultaneous estimation of supports using l_1/l_∞ regularization is superior to that of separate estimations with the naive Lasso-based approach only if the overlapping fraction between two supports κ is larger than 2/3. Similar properties were proved in Obozinski, et al. (2011) for high-dimensional multivariate regression. Note that simultaneous support recovery and our testing problems (1.3) and (1.4) are related but different. Problem (1.4) is focused on the similar/common substructure instead of the whole supports. We will show in Theorem 3.4 and Theorem 3.5 that, to detect the common edges, the related dependency strengths can be $1/\sqrt{K}$ times smaller than those when using separate estimations. This property allows all supports overlapping fraction to be any constant value $0 < \kappa \leq 1$ and is different from multiple high-dimensional linear regression problems with the block l_1/l_∞ regularization. Also, it is obvious that separate estimations cannot identify which part of GGMs has common edges with equal weights. Hence, our method is more powerful for detecting common edges than separate estimations. In addition, such advantage can also help improve the power on the estimation of whole supports, by simply adding the estimated common edges into separate estimations. Furthermore, we prove that the hierarchical method asymptotically controls the FDRs in the estimation of structural differences and similarities. In contrast, there is no FDR control result for optimization-based procedures.

The rest of the paper is organized as follows. In Section 2, we will propose FDR control procedures to estimate differential substructure and similar substructure. A detailed algorithm is summarized at the end of Section 2. Theoretical results on the FDR control and power analysis are given in Section 3. Section 4 provides numerical results. Some possible extensions are discussed in Section 5. The proofs of main results are given in Section 6. For any vector $\boldsymbol{x}$, define $|\boldsymbol{x}|_0 = \sum_{j=1}^p I\{x_j \neq 0\}$, $|\boldsymbol{x}|_1 = \sum_{j=1}^p |x_j|$ and $\|\boldsymbol{x}\| = \sqrt{\sum_{j=1}^p x_j^2}$. For a matrix $\boldsymbol{A} = (a_{ij}) \in \mathbf{R}^{p\times q}$, we define the spectral norm $\|\boldsymbol{A}\| = \sup_{\boldsymbol{x}\in\mathbf{R}^q, \|\boldsymbol{x}\|\leq 1} \|\boldsymbol{A}\boldsymbol{x}\|$. For any $p\times q$ matrix $\boldsymbol{A}$, let $\boldsymbol{A}_{i,-j}$ denote the i-th

row of $\boldsymbol{A}$ with its jth entry being removed and $\boldsymbol{A}_{-i,j}$ denote the j-th column of $\boldsymbol{A}$ with its ith entry being removed. $\boldsymbol{A}_{-i,-j}$ denotes a $(p-1)\times(q-1)$ matrix by removing the i-th row and j-th column of $\boldsymbol{A}$. Let $\lambda_{\max}(\boldsymbol{\Sigma})$ and $\lambda_{\min}(\boldsymbol{\Sigma})$ denote the largest eigenvalue and the smallest eigenvalue of $\boldsymbol{\Sigma}$ respectively. $\boldsymbol{I}_p$ denotes a $p \times p$ identity matrix. 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 $\lim_{n\to\infty} a_n/b_n = 0$, and write $a_n \asymp b_n$ if there are positive constants c and C such that $c \le a_n/b_n \le C$ for all $n \ge 1$. C is a constant which may be different in different places.

2. Methodology. In this section, we describe the proposed hierarchical approach. Assume that $\{\boldsymbol{X}_i^{(k)}, 1 \le i \le n_k\}$, $1 \le k \le K$, are independent. Write $\boldsymbol{X}_i^{(k)} = (X_{i1}^{(k)}, \ldots, X_{ip}^{(k)})'$ and

$$X_{ij}^{(k)} = \alpha_j^{(k)} + \boldsymbol{X}_{i,-j}^{(k)\prime}\boldsymbol{\beta}_j^{(k)} + \varepsilon_{ij}^{(k)},$$

where $\boldsymbol{X}_{i,-j}^{(k)}$ is a $p-1$ dimensional vector by removing the j-th entry of $\boldsymbol{X}_i^{(k)}$, $\varepsilon_{ij}^{(k)} \sim N(0, \sigma_{jj,k} - \boldsymbol{\Sigma}_{j,-j}^{(k)}(\boldsymbol{\Sigma}_{-j,-j}^{(k)})^{-1}\boldsymbol{\Sigma}_{-j,j}^{(k)})$ is independent of $\boldsymbol{X}_{i,-j}^{(k)}$, $\alpha_j^{(k)} = \mu_j^{(k)} - \boldsymbol{\Sigma}_{j,-j}^{(k)}(\boldsymbol{\Sigma}_{-j,-j}^{(k)})^{-1}\boldsymbol{\mu}_{-j}^{(k)}$ and $(\sigma_{ij,k})_{p\times p} = \boldsymbol{\Sigma}^{(k)}$; see Anderson (2003). The regression coefficients vector $\boldsymbol{\beta}_j^{(k)}$ and the error term $\varepsilon_{ij}^{(k)}$ satisfy

$$\boldsymbol{\beta}_j^{(k)} = -\omega_{jj,k}^{-1}\boldsymbol{\Omega}_{-j,j}^{(k)} \quad \text{and} \quad \mathsf{Cov}(\varepsilon_{ij_1}^{(k)}, \varepsilon_{ij_2}^{(k)}) = \frac{\omega_{j_1j_2,k}}{\omega_{j_1j_1,k}\omega_{j_2j_2,k}}.$$

It is reasonable to expect that GGM has a sparse structure in many applications and thus $\boldsymbol{\beta}_j^{(k)}$ is sparse.

We introduce the test statistic proposed by Liu (2013) for the null hypothesis $\rho_{ij\cdot,k} = 0$. Let $\hat{\boldsymbol{\beta}}_j^{(k)} = (\hat{\beta}_{1,j}^{(k)}, \ldots, \hat{\beta}_{p-1,j}^{(k)})'$ be any estimator of $\boldsymbol{\beta}_j^{(k)}$ satisfying

$$\max_{1\le j\le p} |\hat{\boldsymbol{\beta}}_j^{(k)} - \boldsymbol{\beta}_j^{(k)}|_1 = O_{\mathsf{P}}(a_{n1}) \tag{2.1}$$

and

$$\min\Big\{\lambda_{\max}^{1/2}(\boldsymbol{\Sigma}^{(k)}) \max_{1\le j\le p} \|\hat{\boldsymbol{\beta}}_j^{(k)} - \boldsymbol{\beta}_j^{(k)}\|, \max_{1\le j\le p} \sqrt{(\hat{\boldsymbol{\beta}}_j^{(k)} - \boldsymbol{\beta}_j^{(k)})'\hat{\boldsymbol{\Sigma}}_{-i,-i}^{(k)}(\hat{\boldsymbol{\beta}}_j^{(k)} - \boldsymbol{\beta}_j^{(k)})}\Big\} = O_{\mathsf{P}}(a_{n2}) \tag{2.2}$$

for some convergence rates a_{n1} and a_{n2}, where $\hat{\boldsymbol{\Sigma}}^{(k)} = \frac{1}{n_k}\sum_{i=1}^{n_k}(\boldsymbol{X}_i^{(k)} - \bar{\boldsymbol{X}}^{(k)})(\boldsymbol{X}_i^{(k)} - \bar{\boldsymbol{X}}^{(k)})'$ and $\bar{\boldsymbol{X}}^{(k)} = \frac{1}{n_k}\sum_{i=1}^{n_k}\boldsymbol{X}_i^{(k)}$. Define the residual by

$$\hat{\varepsilon}_{ij}^{(k)} = X_{ij}^{(k)} - \bar{X}_j^{(k)} - (\boldsymbol{X}_{i,-j}^{(k)} - \bar{\boldsymbol{X}}_{-j}^{(k)})'\hat{\boldsymbol{\beta}}_j^{(k)},$$

where $\bar{X}_j^{(k)} = \frac{1}{n_k}\sum_{i=1}^{n_k} X_{ij}^{(k)}$. Liu (2013) introduced the following test statistic for null hypothesis $\rho_{ij\cdot,k} = 0$,

$$T_{ij}^{(k)} = \sqrt{\frac{1}{\hat{r}_{ii}^{(k)}\hat{r}_{jj}^{(k)}}}T_{ij,0}^{(k)}, \tag{2.3}$$

where $\hat{r}_{ii}^{(k)} = \frac{1}{n_k}\sum_{m=1}^{n_k}(\hat{\varepsilon}_{mi}^{(k)})^2$ and

$$T_{ij,0}^{(k)} = \frac{1}{n_k}\Big(\sum_{m=1}^{n_k}\hat{\varepsilon}_{mi}^{(k)}\hat{\varepsilon}_{mj}^{(k)} + \sum_{m=1}^{n_k}(\hat{\varepsilon}_{mi}^{(k)})^2\hat{\beta}_{i,j}^{(k)} + \sum_{m=1}^{n_k}(\hat{\varepsilon}_{mj}^{(k)})^2\hat{\beta}_{j-1,i}^{(k)}\Big).$$

2.1. *Structural difference estimation .* We solve the estimation problem of structural differences by multiple tests

$$H_{0ij}:\ D_{ij}(\boldsymbol{\rho}) = 0 \quad \text{versus} \quad H_{1ij}:\ D_{ij}(\boldsymbol{\rho}) \neq 0, \tag{2.4}$$

$1 \le i < j \le p$. Under certain conditions, Proposition 3.1 shows that

$$\frac{\sqrt{n_k}(T_{ij}^{(k)} - \rho_{ij\cdot,k})}{1-\rho_{ij\cdot,k}^2} \Rightarrow N(0,1)$$

as $(n,p) \to \infty$. The partial correlation coefficient in the denominator is estimated by a thresholding estimator

$$\hat{\rho}_{ij\cdot,k} = T_{ij}^{(k)} I\Big\{|T_{ij}^{(k)}| \ge 2\sqrt{\frac{\log p}{n_k}}\Big\}.$$

Define the following two-sample test statistic

$$T_{ij}^{(k,l)} = \frac{T_{ij}^{(k)} - T_{ij}^{(l)}}{\sqrt{\frac{1}{n_k}(1-\hat{\rho}_{ij\cdot,k}^2)^2 + \frac{1}{n_l}(1-\hat{\rho}_{ij\cdot,l}^2)^2}},$$

and $\boldsymbol{T}_{ij} = (T_{ij}^{(k,l)}, 1 \le k < l \le K)'$. For each H_{0ij}, we use the sum squared test statistic

$$T_{ij,*} = \|\boldsymbol{T}_{ij}\|. \tag{2.5}$$

Note that, when $K = 2$, the asymptotic null distribution of $\boldsymbol{T}_{ij}$ is still normal. However, when $K \ge 3$, the distribution of $T_{ij,*}$ is changed to a non-standard Chi-squared distribution which is introduced below.

Let $\mathbf{D}_1, \ldots, \mathbf{D}_{K-1}$ be $K-1$ matrices and $\mathbf{D}_k =: (d_{i,j,k}) \in R^{(K-k)\times K}$, $1 \le k \le K-1$, where $d_{i,k,k} = (n_k^{-1} + n_{k+i}^{-1})^{-1/2}$ and $d_{i,i+k,k} = -(n_k^{-1} + n_{k+i}^{-1})^{-1/2}$ for $1 \le i \le K-k$, and $d_{i,j,k} = 0$ for other entries. Let $\boldsymbol{B} = (\mathbf{D}_1^{'}, \ldots, \mathbf{D}_{K-1}^{'})^{'}$ and $\mathbf{D} = \boldsymbol{B} diag(n_1^{-1}, \ldots, n_K^{-1}) \boldsymbol{B}^{'}$. The matrix $\mathbf{D}$ is the asymptotic covariance matrix of $\boldsymbol{T}_{ij}$ under H_{0ij}. Note that $\operatorname{rank}(\mathbf{D}) \le \min(K, K(K-1)/2)$. Write $\mathbf{D} = \boldsymbol{U}^{'} diag(\lambda_1, \ldots, \lambda_M, 0, \ldots, 0) \boldsymbol{U}$, where $M = \operatorname{rank}(\mathbf{D})$, $\boldsymbol{U}$ is an orthogonal matrix and $\lambda_1 \ge \cdots \ge \lambda_M > 0$. Let $M_1 \ge 1$ satisfy $\lambda_1 = \cdots = \lambda_{M_1} > \lambda_{M_1+1}$. Note that $\lambda_1, \ldots, \lambda_M$ are bounded, but they may still depend on $n_1, \ldots, n_K$. Throughout the paper, we assume that $\lambda_{M_1} - \lambda_{M_1+1} \ge c$ and $\lambda_M \ge c$ for some constant $c > 0$. (This holds trivially if $n_k/n_l \to \gamma_{kl}$ for some constants $\gamma_{kl} > 0$.) Let $Z_1, \ldots, Z_M$ be i.i.d. $N(0,1)$ random variables. We will show that $\mathsf{P}(T_{ij,*}^2 \le x | H_{0ij}) - \mathsf{P}(\sum_{i=1}^M \lambda_i Z_i^2 \le x) \to 0$ for $x > 0$.

Due to the correlation between test statistics, we use the FDR control procedure in Efron (2007). We first translate $T_{ij,*}$ into z-value. Let $\Phi(x)$ be a standard normal distribution. The z-value

$$T_{ij,1} = \Phi^{-1}(\Psi_0(T_{ij,*})), \quad \text{where } \Psi_0(t) = \mathsf{P}\Big(\sqrt{\textstyle\sum_{i=1}^M \lambda_i Z_i^2} \le t\Big). \tag{2.6}$$

By the monotonicity of $\Phi(t)$, $\{T_{ij,1} \ge t\}$ is equivalent to $\{T_{ij,*} \ge t^{'}\}$ for some $t^{'} > 0$. Following the notations in Efron (2007), define $A = (P_0 - \hat{P}_0)/Q_0$, where $P_0 = 2\Phi(1) - 1$, $\hat{P}_0 = \frac{2\sum_{1\le i<j\le p} I\{|T_{ij,1}| \le 1\}}{p^2-p}$ and $Q_0 = \sqrt{2}\varphi(1)$ with $\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$. Let $A(t) = (1 + |A| \frac{|t|\varphi(t)}{\sqrt{2}(1-\Phi(t))})^{-1}$. From (52) in Efron (2007), we use the following FDR control procedure to estimate $\mathcal{A}_1$:

Estimation of structural difference. For a given $0 < \alpha_1 < 1$, let

$$\hat{t}_1 = \inf\Big\{ t \in \mathbf{R} : 1 - \Phi(t) \le \frac{\alpha_1 A(t) \max(1, \sum_{1\le i<j\le p} I\{T_{ij,1} \ge t\})}{(p^2-p)/2} \Big\}. \tag{2.7}$$

Reject $H_{0ij} : D_{ij}(\boldsymbol{\rho}) = 0$ if $T_{ij,1} \ge \hat{t}_1$.

With procedure (2.7), we can obtain an estimator $\hat{\mathcal{A}}_1 = \{(i,j) : T_{ij,1} \ge \hat{t}_1,\ i \ne j\}$. Theorem 3.1 shows that, under some regularity conditions, the true FDR and FDP of procedure (2.7) will converge to α_1. Also, by the proof of Theorem 3.3, all entries in $\hat{\mathcal{A}}_1^c$ satisfy $D_{ij}(\boldsymbol{\rho}) = O(\sqrt{\log p/n})$ with high probability.

The factor $A(t)$ is used to control the influence of correlation between test statistics. Due to the sparsity in GGMs, we will show that $A(\hat{t}_1)$ is in fact close to 1. This indicates that procedure (2.7) is essentially equivalent

to the Benjamini and Hochberg method (Benjamini and Hochberg, 1995). However, because $A(t) < 1$, (2.7) typically achieves a more conservative FDR control.

2.2. *Structural similarity estimation .* Based on structural difference estimation, we can define the similar substructure between GGMs by

$$\hat{\mathcal{A}}_2 = \{(i,j) \in \hat{\mathcal{A}}_1^c : \ (\rho_{ij\cdot,1}, \ldots, \rho_{ij\cdot,K}) \neq 0, \ 1 \leq i < j \leq p\}.$$

To recover $\hat{\mathcal{A}}_2$, we consider the following multiple testing problem

$$H_{0ij} : \ (\rho_{ij\cdot,1}, \ldots, \rho_{ij\cdot,K}) = 0 \quad \text{versus} \quad H_{1ij} : (\rho_{ij\cdot,1}, \ldots, \rho_{ij\cdot,K}) \neq 0 \tag{2.8}$$

with $(i,j) \in \hat{\mathcal{A}}_1^c$. As mentioned in the introduction, we shall use the partial sum type test statistic

$$T_{ij,\star} = \frac{\sum_{k=1}^K n_k T_{ij}^{(k)}}{\sqrt{\sum_{k=1}^K n_k (1-\hat{\rho}_{ij\cdot,k}^2)^2}}, \quad (i,j) \in \hat{\mathcal{A}}_1^c. \tag{2.9}$$

The z-value is defined by

$$T_{ij,2} = \Phi^{-1}(2\Phi(|T_{ij,\star}|) - 1). \tag{2.10}$$

Let $\hat{P}_0' = |\hat{\mathcal{A}}_1^c|^{-1} \sum_{(i,j)\in\hat{\mathcal{A}}_1^c} I\{|T_{ij,2}| \leq 1\}$, $A' = (P_0 - \hat{P}_0')/Q_0$ and $A'(t) = (1 + |A'|\frac{|t|\varphi(t)}{\sqrt{2}(1-\Phi(t))})^{-1}$.

Estimation of structural similarity. For a given $0 < \alpha_2 < 1$, let

$$\hat{t}_2 = \inf\left\{t \in \mathbf{R} : 1 - \Phi(t) \leq \frac{\alpha_2 A'(t) \max(1, \sum_{(i,j)\in\hat{\mathcal{A}}_1^c} I\{T_{ij,2} \geq t\})}{|\hat{\mathcal{A}}_1^c|}\right\}. \tag{2.11}$$

Reject $H_{0ij} : (\rho_{ij\cdot,1}, \ldots, \rho_{ij\cdot,K}) = 0$ with $(i,j) \in \hat{\mathcal{A}}_1^c$ if $T_{ij,2} \geq \hat{t}_2$.

By (2.11), we obtain an estimator $\hat{\hat{\mathcal{A}}}_2 = \{(i,j) : \ T_{ij,2} \geq \hat{t}_2, \ (i,j) \in \hat{\mathcal{A}}_1^c, \ i \neq j\}$ for the similar substructure. We will show that $T_{ij,1}$ and $T_{ij,2}$ are asymptotically independent; see Lemma 6.3. Moreover, Theorem 3.2 shows that the FDR and FDP of procedure (2.11) converge to α_2 even when the index set $\hat{\mathcal{A}}_1^c$ is selected based on a first analysis step of the same dataset.

More discussions on the similar structure. As mentioned above, the similar structure is defined in a data-driven way. All node pairs in $\hat{\mathcal{A}}_1^c$

satisfy $D_{ij}(\boldsymbol{\rho}) = O(\sqrt{\log p/n})$, which means that they have nearly equal partial correlation coefficients across all graphs. Clearly, some of node pairs in $\hat{\mathcal{A}}_1^c$ can still have different edges. Based only on limited $\{n_k\}$ samples, it is difficult to detect these node pairs through statistical methods, due to the minimax rate of $\max_{1\le i,j\le p}|\hat{\omega}_{ij,k} - \omega_{ij,k}|$ is $O(\sqrt{\log p/n_k})$; see Ren, et al. (2015). Hence, in real data analysis, from the point of statistical significance, one may claim that node pairs in $\hat{\mathcal{A}}_1^c$ have the same partial coefficients, or more naturally, have similar partial coefficients. Consequently, $\hat{\mathcal{A}}_2$ is called to be the (unknown) similar structure across all graphs. The tests (1.4) in the second step aim to find the similar structure $\hat{\mathcal{A}}_2$ rather than the common structure $\mathcal{A}_2$. Once H_{0ij} in (1.4) is rejected, the pair (i,j) is interpreted to be of nonzero similar edges across all graphs. Note that (2.11) aims to control the false positives in the estimation of similar structure. The set $\hat{\hat{\mathcal{A}}}_2$ may contain node pairs from $\mathcal{A}_1$ with similar edges satisfying $D_{ij}(\boldsymbol{\rho}) = O(\sqrt{\log p/n})$ and all false positives in $\hat{\hat{\mathcal{A}}}_2$ come from $\mathcal{A}_3$.

For the convenience of the reader, we summarize the proposed procedures as follow.

Algorithm: structural difference/similarity estimation

Structural difference algorithm
1. Construct test statistics $T_{ij,1}$ in (2.6).
2. Perform multiple test procedure (2.7) and obtain $\hat{t}_1$.
3. The differential structure is estimated by $\hat{\mathcal{A}}_1 = \{(i,j):\ T_{ij,1} \ge \hat{t}_1,\ i \ne j\}$.

Structural similarity algorithm
4. Construct test statistics $T_{ij,2}$ in (2.10).
5. Perform multiple test procedure (2.11) and obtain $\hat{t}_2$.
6. The similar structure is estimated by $\hat{\hat{\mathcal{A}}}_2 = \{(i,j):\ T_{ij,2} \ge \hat{t}_2,\ (i,j) \in \hat{\mathcal{A}}_1^c,\ i \ne j\}$.

3. Theoretical results. In this section, we prove the theoretical results on the FDR/FDP control and power analysis. Furthermore, we will give a theoretical comparison between our procedure and separate estimations. Let $n = \sum_{k=1}^K n_k$. The following regularity condition is required to show the asymptotic normality for $T_{ij}^{(k)}$.

(C1). Suppose that $\max_{1\le k\le K}\max_{1\le i\le p}\sigma_{ii,k} \le c_0$, $\max_{1\le k\le K}\max_{1\le i\le p}\omega_{ii,k} \le c_0$ and $\max_{1\le k\le K}\max_{1\le i<j\le p}|\rho_{ij\cdot,k}| \le \rho$ for some $c_0 > 0$ and $0 < \rho < 1$. Assume that $\log p = o(\sqrt{n})$ and $c_1 \le n_i/n_j \le c_2$ for any $1 \le i,j \le K$ and some $c_1, c_2 > 0$.

PROPOSITION 3.1. *Suppose that (C1) holds and* $\hat{\boldsymbol{\beta}}_i^{(k)}$ *satisfies*

$$a_{n1} = o(1/\sqrt{\log p}) \quad \text{and} \quad a_{n2} = o(n^{-1/4}). \tag{3.1}$$

Then, for any $1 \le k \le K$, *we have as* $(n,p) \to \infty$,

$$\frac{\sqrt{n_k}(T_{ij}^{(k)} - \rho_{ij\cdot,k})}{1-\rho_{ij\cdot,k}^2} \Rightarrow N(0,1),$$

where the convergence in distribution is uniformly in $1 \le i < j \le p$.

This proposition is a refined version of Proposition 3.1 in Liu (2013). Condition (3.1) can be satisfied by many popular methods such as Lasso (Tibshirani, 1996) and Dantzig selector (Candès and Tao, 2007) under the sparsity condition $\max_{1\le i\le p}|\boldsymbol{\beta}_i^{(k)}|_0 = o\Big(\lambda_{\min}(\boldsymbol{\Sigma}^{(k)})\frac{\sqrt{n}}{(\log p)^{3/2}}\Big)$; see Section 4 in Liu (2013). Note that the Lasso method and Dantzig selector require the selection of tuning parameters. A data-driven approach to determine tuning parameters is provided in Section 4.1.

3.1. *FDR control results.* We now study the FDR/FDP control for procedures (2.7) and (2.11) in Section 2. Define the FDR/FDP of structural difference testing by

$$\text{FDP}_1 = \frac{\sum_{(i,j)\in\hat{\mathcal{A}}_1} I\{D_{ij}(\boldsymbol{\rho}) = 0\}}{\max(1, |\hat{\mathcal{A}}_1|)} \quad \text{and} \quad \text{FDR}_1 = \mathsf{E}[\text{FDP}_1]. \tag{3.2}$$

Similarly, define the FDR/FDP of structural similarity testing by

$$\text{FDP}_2 = \frac{\sum_{(i,j)\in\hat{\hat{\mathcal{A}}}_2} I\{(\rho_{ij\cdot,1},\ldots,\rho_{ij\cdot,K}) = 0\}}{\max(1, |\hat{\hat{\mathcal{A}}}_2|)} \quad \text{and} \quad \text{FDR}_2 = \mathsf{E}[\text{FDP}_2]. \tag{3.3}$$

We shall introduce some dependence conditions. For a constant $\gamma > 0$ and $1 \le i \le p$, define

$$\mathcal{A}_i^{(k)}(\gamma) = \{j : 1 \le j \le p,\ j \ne i,\ |\rho_{ij\cdot,k}| \ge (\log p)^{-3-\gamma}\}.$$

Let $\mathcal{H}_1 = \{(i,j) : (\rho_{ij\cdot,1},\ldots,\rho_{ij\cdot,K}) \ne 0, 1 \le i < j \le p\}$.

(C2). Suppose that, for some $\gamma > 0$ and $0 < \xi < \min\{(1-\rho)/(1+\rho), 1/3\}$, we have $\max_{1\le i\le p} Card(\mathcal{A}_i^{(k)}(\gamma)) = O(p^\xi)$ for all $1 \le k \le K$. Assume that $Card(\mathcal{H}_1) = o(p^2/\log p)$.

Let $\boldsymbol{\rho}_{ij} = (\rho_{ij\cdot,kl}, 1 \le k < l \le K)$, where

$$\rho_{ij\cdot,kl} = \frac{\rho_{ij\cdot,k} - \rho_{ij\cdot,l}}{\sqrt{\frac{1}{n_k}(1-\rho^2_{ij\cdot,k})^2 + \frac{1}{n_l}(1-\rho^2_{ij\cdot,l})^2}}.$$

THEOREM 3.1. *Let $p \le n^r$ for some $r > 0$. Suppose that $\hat{\boldsymbol{\beta}}_i^{(k)}$ satisfies*

$$a_{n1} = o(1/\log p) \quad \text{and} \quad a_{n2} = o((n\log p)^{-1/4}). \tag{3.4}$$

Assume that

$$Card\Big\{(i,j) : 1 \le i < j \le p, \quad \|\boldsymbol{\rho}_{ij}\| \ge \theta\sqrt{\lambda_1 \log p}\Big\} \to \infty \tag{3.5}$$

for some $\theta > 2$. Under (C1) and (C2), we have for any $\varepsilon > 0$,

$$P(FDP_1 \le \alpha_1 + \varepsilon) \to 1$$

as $(n,p) \to \infty$. Consequently, $limsup_{(n,p)\to\infty} FDR_1 \le \alpha_1$. Furthermore, if

$$a_{n1} = o(1/(\log p)^{3/2}) \quad \text{and} \quad a_{n2} = o((n(\log p)^2)^{-1/4}), \tag{3.6}$$

then $A(\hat{t}_1) \to 1$ in probability, $FDP_1 \to \alpha_1$ in probability and $\lim_{(n,p)\to\infty} FDR_1 = \alpha_1$.

We note that the condition (C2) is a weak dependence condition on the partial correlation coefficients. It is quite mild for sparse GGMs. For example, in high-dimensional precision matrix estimation, it is often assumed that $\boldsymbol{\Omega}^{(k)}$ is a $\sqrt{n}$-sparse matrix. The condition (C2) is clearly much weaker than such sparsity assumption when $p \ge n^{1/(2\xi)}$. Condition (3.5) requires the number of true alternatives tends to infinite. It is nearly necessary for the FDP control; see Proposition 2.1 in Liu and Shao (2014) for general multiple testing problems.

We now state the FDR and FDP control result for procedure (2.11).

THEOREM 3.2. *Let $p \le n^r$ for some $r > 0$. Suppose that (C1), (C2) and (3.4) hold. Assume that*

$$Card\Big\{(i,j) : 1 \le i < j \le p, \quad \rho_{ij\cdot,1} = \cdots = \rho_{ij\cdot,K}, \ |\rho_{ij\cdot,1}| \ge \theta\sqrt{\log p/n}\Big\} \to \infty$$

for some $\theta > 2$. We have for any $\varepsilon > 0$,

$$P(FDP_2 \le \alpha_2 + \varepsilon) \to 1$$

as $(n,p) \to \infty$. Consequently, $limsup_{(n,p)\to\infty} FDR_2 \le \alpha_2$. Furthermore, if (3.6) holds, then $A'(\hat{t}_2) \to 1$ in probability, $FDP_2 \to \alpha_2$ in probability and $\lim_{(n,p)\to\infty} FDR_2 = \alpha_2$.

3.2. *Power analysis.* In this section, we analyze statistical powers for the proposed method. Assume that $|\mathcal{A}_1| \asymp p^{\theta_1}$ and $|\mathcal{A}_2| \asymp p^{\theta_2}$ for some $0 < \theta_1, \theta_2 < 2$. The power of $\hat{\mathcal{A}}_1$ is defined to be

$$\text{power}_1 = \mathsf{E}\Big(\frac{\sum_{(i,j)\in\mathcal{A}_1} I\{(i,j) \in \hat{\mathcal{A}}_1\}}{|\mathcal{A}_1|}\Big).$$

For simplicity, we assume the signal sizes in $\mathcal{A}_1$ satisfy

$$\|\boldsymbol{\rho}_{ij}\| = \delta\sqrt{\lambda_1 \log p}, \quad (i,j) \in \mathcal{A}_1, \quad \text{for some } \delta > 0. \tag{3.7}$$

THEOREM 3.3. *Let $p \le n^r$ for some $r > 0$, (3.4), (3.7), (C1) and (C2) hold. If $\delta > \sqrt{4 - 2\theta_1}$, then power$_1 \to 1$ as $(n,p) \to \infty$. If $\delta < \sqrt{4 - 2\theta_1}$, then power$_1 \to 0$ as $(n,p) \to \infty$.*

Theorem 3.3 shows $\sqrt{4 - 2\theta_1}$ is the critical level of signal sizes for power$_1 \to 1$. Let $\mathcal{A}_{11} = \{(i,j) : \ \|\boldsymbol{\rho}_{ij}\| \ge \delta\sqrt{\lambda_1 \log p}\}$ for some $\delta > 4$. From the proof of Theorem 3.1, we can also obtain that $\mathsf{P}(\mathcal{A}_{11} \subseteq \hat{\mathcal{A}}_1) \to 1$. This indicates that the similar structure has nearly equal partial correlation coefficients across GGMs.

We now compare statistical powers on common edge detection between our procedure and the separate estimation approach. Define the power of $\hat{\hat{\mathcal{A}}}_2$ for common edge detection by

$$\text{power}_2 = \mathsf{E}\Big(\frac{\sum_{(i,j)\in\mathcal{A}_2} I\{(i,j) \in \hat{\hat{\mathcal{A}}}_2\}}{|\mathcal{A}_2|}\Big).$$

Assume that

$$|\rho_{ij\cdot,1}| = \cdots = |\rho_{ij\cdot,K}| = \delta\sqrt{\log p/n}, \ (i,j) \in \mathcal{A}_2, \text{ for some } \delta > 0. \tag{3.8}$$

THEOREM 3.4. *Let $p \le n^r$ for some $r > 0$, (3.4), (3.8) (C1) and (C2) hold. If $\delta > \sqrt{4 - 2\theta_2}$, then power$_2 \to 1$ as $(n,p) \to \infty$.*

We next state the power result of separate estimations. To this end, we use the FDR control procedure in Liu (2013) to estimate $\mathcal{G}_1$ based only on $\{\boldsymbol{X}_1^{(1)}, \ldots, \boldsymbol{X}_{n_1}^{(1)}\}$. Consider

$$H_{0ij} : \ \rho_{ij\cdot,1} = 0 \quad \text{versus} \quad H_{1ij} : \ \rho_{ij\cdot,1} \neq 0$$

for $1 \le i < j \le p$. Let $\hat{G}_1$ be the set of edges estimated by the method in Liu (2013), i.e., $\hat{G}_1 = \{(i,j) : \text{there is an edge between } X_i^{(1)} \text{ and } X_j^{(1)} \text{ in the}$

estimated graph}. The power on common edge detection by $\hat{G}_1$ is

$$\text{power}_{sepa} = \mathsf{E}\Big(\frac{\sum_{(i,j)\in\mathcal{A}_2} I\{(i,j)\in\hat{G}_1\}}{|\mathcal{A}_2|}\Big).$$

Assume that

$$|\rho_{ij\cdot,1}| = \delta\sqrt{\log p/n_1}, \quad (i,j)\in\mathcal{A}_2, \text{ for some } \delta>0. \tag{3.9}$$

The overlapping fraction of edges among GGMs is defined to be $\kappa := \kappa_p = \frac{|\mathcal{A}_2|}{|\mathcal{A}_1|+|\mathcal{A}_2|}$. For $K=2$, $\kappa=0$ means that two graphs are totally different; while $\kappa=1$ means that two graphs are exactly the same.

THEOREM 3.5. *Let $p\le n^r$ for some $r>0$, (3.4), (3.9), (C1) and (C2) hold. Assume that κ_p is bounded away from zero. If $\delta<\sqrt{4-2\theta_2}$, then $power_{sepa}\to 0$ as $(n,p)\to\infty$.*

In Theorem 3.5, we assume that $\kappa>0$, i.e., the number of common edges is comparable to or larger than the number of differential edges. This condition is quite mild in many applications in which only a relatively small part of edges may be changed. It is clearly that $\sqrt{\log p/n}<\sqrt{\log p/n_1}$ as $n=\sum_{k=1}^K n_k$. Hence, the lower bound of δ for $\text{power}_2\to 1$ is strictly smaller than that for $\text{power}_{sepa}\to 1$. Especially, when $n_1=\cdots=n_K$, the signal size in (3.8) can be as small as $1/\sqrt{K}$ times of that in (3.9). This indicates that, it is easier for our method to detect common edges across GGMs than separate estimations.

4. Numerical results. In this section, we demonstrate the performance of the proposed hierarchical method on similar/different structures recovery via simulation experiments.

4.1. *Selection of initial estimators.* In our experiment, we use the Lasso estimator for constructing $\hat{\boldsymbol{\beta}}_j^{(k)}$. Other estimators such as Dantzig selector and square-root lasso can also be used. Let $\mathbf{D}_i^{(k)} = diag(\hat{\boldsymbol{\Sigma}}_{-i,-i}^{(k)})$ and

$$\lambda_{ni}^{(k)}(\delta) = \delta\sqrt{\frac{\hat{\sigma}_{ii,k}\log p}{n_k}} \quad \text{for } \delta>0,$$

where $\hat{\sigma}_{ii,k}=\frac{1}{n_k}\sum_{m=1}^{n_k}(X_{mi}^{(k)}-\bar{X}_i^{(k)})^2$. Let

$$\hat{\alpha}_{i,k}(\delta) = \arg\min_{\alpha\in\mathbf{R}^{p-1}}\Big\{\frac{1}{2n_k}\sum_{j=1}^{n_k}(X_{ji}^{(k)}-\bar{X}_i^{(k)}-(\boldsymbol{X}_{j,-i}^{(k)}-\bar{\boldsymbol{X}}_{-i}^{(k)})(\mathbf{D}_i^{(k)})^{-1/2}\alpha)^2+\lambda_{ni}^{(k)}(\delta)|\alpha|_1\Big\}.$$

The initial estimator $\hat{\boldsymbol{\beta}}_i^{(k)}$ is taken to be $\hat{\boldsymbol{\beta}}_i^{(k)}(\delta) = (\mathbf{D}_i^{(k)})^{-1/2}\hat{\alpha}_{i,k}(\delta)$. The tuning parameter δ is selected as in Liu (2013), by

$$\hat{\delta} = \hat{j}/20, \quad \hat{j} = \arg\min_{0\le j\le 40} \sum_{k=3}^{9} \Big\{ \Big(\frac{\sum_{1\le i\ne j\le p} I\{|T_{ij,1}(j/N)| \ge \Phi^{-1}(1-\frac{k}{20})\}}{k(p^2-p)/10} - 1 \Big)^2 + \Big(\frac{\sum_{1\le i\ne j\le p} I\{|T_{ij,2}(j/N)| \ge \Phi^{-1}(1-\frac{k}{20})\}}{k(p^2-p)/10} - 1 \Big)^2 \Big\}, \tag{4.1}$$

where $T_{ij,1}(\delta)$ and $T_{ij,2}(\delta)$ are test statistics in (2.6) and (2.10) with the initial estimators $\hat{\boldsymbol{\beta}}_i^{(k)}(\delta)$.

4.2. *Simulation results.* In this section, we conduct simulation studies. Note that there are some joint estimation methods on multiple GGMs as introduced in Section 1. However, these methods define the similar/different structures by the equality of $\omega_{ij,k}$ or $\beta_{ij,k} = -\omega_{ij,k}/\omega_{ii,k}$, $1 \le k \le K$, which is different from the definition in this paper. Moreover, these methods depend on the choice of tuning parameters and it is still unknown how to control the FDR by a data-driven choice of tuning parameters. For these reasons, we only compare our method to the separate estimations on the power of recovering similar structures. We use the method in Liu (2013) to estimate GGMs separately and calculate its power on recovering common structure $\mathcal{A}_2$.

4.2.1. *Performance of the proposed method.* In this section, we illustrate the performance of the proposed method by simulated data. We will report two important measures on evaluating the performance of our method. The first one is the empirical FDR which is related to false positives. The second one is the empirical power which is related to false negatives. We let $K = 2$ and consider three model settings for $\boldsymbol{\Omega}^{(1)}$ and $\boldsymbol{\Omega}^{(2)}$. Let $\mathbf{H}_1 = diag(\mathbf{H}_{1,1}, \ldots, \mathbf{H}_{1,p/2})$ and $\mathbf{H}_2 = diag(\mathbf{H}_{2,1}, \ldots, \mathbf{H}_{2,p/2})$, where

$$\mathbf{H}_{1i} = \begin{bmatrix} 1, & 0.5 \\ 0.5, & 1 \end{bmatrix}, \quad \mathbf{H}_{2i} = \begin{bmatrix} 1, & -0.5 \\ -0.5, & 1 \end{bmatrix}, \quad 1 \le i \le p/2.$$

To construct $\boldsymbol{\Omega}^{(1)}$ and $\boldsymbol{\Omega}^{(2)}$, we first define three graphical models.

- **Model 1 (Band graph)**. $\boldsymbol{\Omega} = (\omega_{ij})$, where $\omega_{ii} = 1$, $\omega_{i,i+1} = \omega_{i+1,i} = 0.6$ and $\omega_{ij} = 0$ for $|i-j| \ge 2$.
- **Model 2 (Hub graph)**. There are $p/10$ rows with sparsity 11. The rest every row has sparsity 2. To this end, we let $\boldsymbol{\Omega} = (\omega_{ij})$, $\omega_{ij} = \omega_{ji} = 0.5$ for $i = 10(k-1)+1$ and $10(k-1)+2 \le j \le 10(k-1)+10$, $1 \le k \le p/10$. The diagonal $\omega_{ii} = 1$ and others entries are zero.

- **Model 3 (Erdös-Rényi random graph)**. There is an edge between each pair of nodes with probability $\min(0.05, 5/p)$ independently. Let $\mathbf{\Omega} = (\omega_{ij})$, where $\omega_{ij} = u_{ij} * \delta_{ij}$, $u_{ij} \sim U(0.2, 0.6)$ is a uniform random variable and δ_{ij} is a Bernoulli random variable with success probability $\min(0.05, 5/p)$. u_{ij} and δ_{ij} are independent.

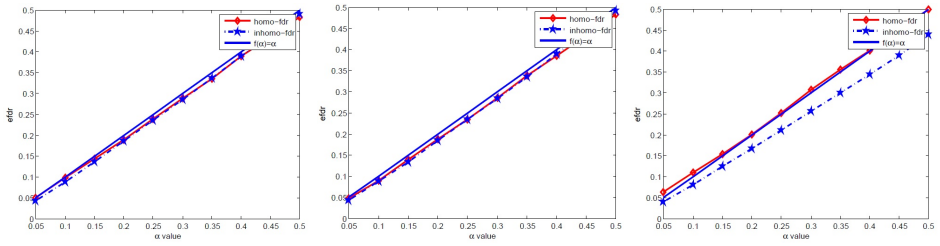
The band graph and Erdös-Rényi random graph are commonly used in the simulation for GGM estimation (see Yuan and Lin 2007; Fan, et al. 2009; Cai, et al. 2011). The hub graph is related to networks in some real applications (e.g. gene networks) where a hub node is related to a hub gene.

Let $\mathbf{\Omega}_1^* = \mathbf{\Omega} + \mathbf{H}_1$ and $\mathbf{\Omega}_2^* = \mathbf{\Omega} + \mathbf{H}_2$. ξ_1 and ξ_2 are the smallest eigenvalues of $\mathbf{\Omega}_1^*$ and $\mathbf{\Omega}_2^*$, respectively. Let $\mathbf{\Omega}^{(1)} = \mathbf{\Omega}_1^* + (\max(\xi_1^-, \xi_2^-) + 0.01)\boldsymbol{I}_p$ and $\mathbf{\Omega}^{(2)} = \mathbf{\Omega}_2^* + (\max(\xi_1^-, \xi_2^-) + 0.01)\boldsymbol{I}_p$. Note that $\mathbf{\Omega}^{(1)} - \mathbf{\Omega}^{(2)} = \mathbf{H}_1 - \mathbf{H}_2$. For each model, we generate $n_1 = n_2 = 100$ random samples from $N(0, (\mathbf{\Omega}^{(1)})^{-1})$ and $N(0, (\mathbf{\Omega}^{(2)})^{-1})$, respectively. The dimension is taken to be $p = 50, 100, 200$. The simulation is replicated 100 times. In each replication, we obtain a differential substructure estimator $\hat{\mathcal{A}}_1$ and a similar substructure estimator $\hat{\hat{\mathcal{A}}}_2$. We then calculate FDP_1 and FDP_2 in (3.2) and (3.3). The empirical FDR_1 of $\hat{\mathcal{A}}_1$ and empirical FDR_2 of $\hat{\hat{\mathcal{A}}}_2$ are calculated by the average of FDPs of 100 replications. The empirical power_1 and power_2 are obtained in a similar way.

The target FDRs in (2.7) and (2.11) are taken uniformly to $\alpha_1 = \alpha_2 = \alpha = i/20$, $1 \le i \le 10$. We plot the empirical FDR curves and empirical power curves for Models 1-3. For the reason of space, we put the numerical results for $p = 50, 100$ in the supplementary material. The "homo-fdr/power" curves denote the values of empirical $\text{FDR}_1/\text{power}_1$, and the "inhomo-fdr/power" curves denote the values of empirical $\text{FDR}_2/\text{power}_2$. The solid line is the curve of function $f(\alpha) = \alpha$. As we can see from Figure 1, the curves of "homo-fdr" and "inhomo-fdr" are always close to or slightly below the solid line. This means that our method controls the FDR quite well. The power curves are plotted in Figure 2. As we can see, for Model 1, the powers on the estimation of $\mathcal{A}_1$ and $\mathcal{A}_2$ are close to 1. The powers for Models 2 and 3 are also reasonably good. Note that homo-power and inhomo-power are different in three models. This is because that homo-power depends on the value of $\|\boldsymbol{\rho}_{ij}\|$ and inhomo-power depends on the value of $\frac{\sum_{k=1}^K n_k \rho_{ij\cdot,k}}{\sqrt{\sum_{k=1}^K n_k}}$. In Models 1-3, these values are different.

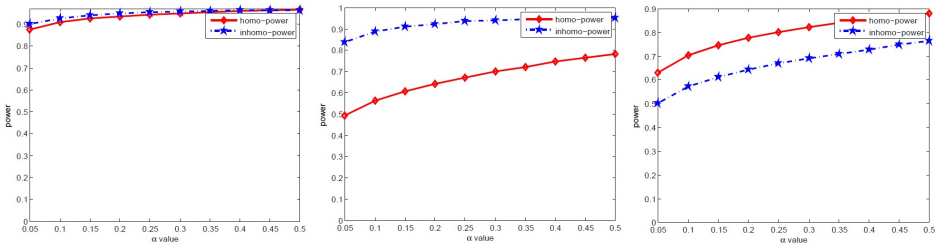
We now consider the case with $\mathcal{A}_2 = \emptyset$ and check how many discoveries in procedure (2.11). To this end, we let $\mathbf{\Omega}_1^* = \mathbf{\Omega}$ and $\mathbf{\Omega}_2^* = \mathbf{\Omega} + 0.01 * (I\{\omega_{ij} \neq 0, i \neq j\})$, and then define $\mathbf{\Omega}^{(1)}$ and $\mathbf{\Omega}^{(2)}$ as above. Note that nonzero partial correlation coefficients between two graphs are different but quite close. We

plot empirical FDR_2 in Figure 3 for Models 1-3 with $p = 100$ and $n = 100$. We can see that FDR_2 is still well controlled below α for all models. We now plot the curve of RS= $\frac{|\hat{\hat{\mathcal{A}}}_2 \cap \mathcal{A}_3^c|}{|\mathcal{A}_3^c|}$, which is the ratio of true similar edges estimated by $\hat{\hat{\mathcal{A}}}_2$ among $\mathcal{A}_3^c$. As we can see from Figure 3, for band graph and hub graph, the ratios of true similar edges (RS) are close to one for all α. This means that most of edges are estimated to be the similar structure. For the ER graph, the diagonal entries in $\mathbf{\Omega}^{(1)}$ and $\mathbf{\Omega}^{(2)}$ are larger than 1.95 so that nonzero partial correlation coefficients in this model is in [0.1,0.3]. Hence, a part of edges with small values cannot be detected by (2.11).



(a) Band graph (b) Hub graph (c) E-R graph

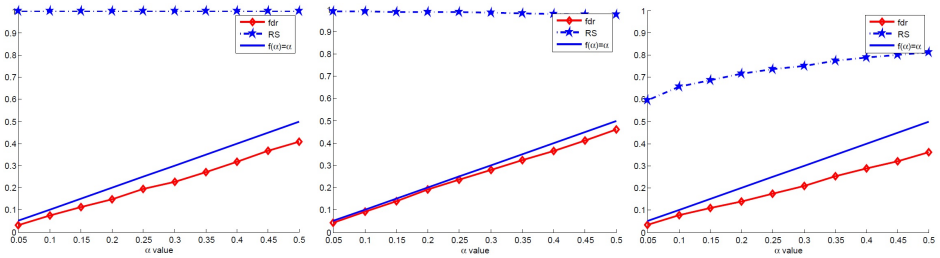
Fig 1: $p = 200$. *The x-axis denotes the α value and the y-axis denotes the empirical FDR.*



(a) Band graph (b) Hub graph (c) E-R graph

Fig 2: $p = 200$. *The x-axis denotes the α value and the y-axis denotes the empirical power.*

4.2.2. *Comparison with the separate estimation approach.* In this section, we compare our method to the separate estimation approach on recovering the common structure $\mathcal{A}_2$. We first carry out the comparison on three



(a) Band graph (b) Hub graph (c) E-R graph

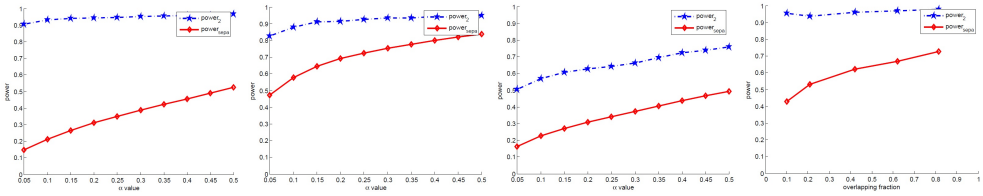
Fig 3: $p = 100$. *The x-axis denotes the* α *value. The red curve* $-\diamond-$ *denotes the empirical FDR, the blue curve* $-\star-$ *denotes RS*$= \frac{|\hat{\mathcal{A}}_2 \cap \mathcal{A}_3^c|}{|\mathcal{A}_3^c|}$ *and the blue solid line denotes the curve of* $f(\alpha) = \alpha$.

models in Section 4.2.1. For the sake of space, we only present the result for the dimension $p = 200$. The results for $p = 50$ and $p = 100$ are similar. The separate estimation method in Liu (2013) is used to estimate Gaussian graphical models defined by $\mathbf{\Omega}^{(1)}$, given in Section 4.2.1. The power of recovering the common substructure $\mathcal{A}_2$ by separate estimation is calculated from 100 average of power$_{sepa}$ that is defined in Section 3.2. The results for three graphical models are plotted in Figure 4 (a-c). We also plot the empirical power curves of procedure (2.11), i.e., average of power$_2$ over 100 replications in Section 3.2. As we can see from Figure 4, procedure (2.11) significantly outperforms separate estimation on detecting common edges. Note that the empirical FDR of similar structure estimation, FDR$_2$, can be found in Figure 1. In the supplementary material, we also plot the curve of empirical FDR of the separate estimation method.

We next compare the power on common edge detection as the overlapping fraction between two supports κ decreases. To this end, we fix $p = 100$, $\alpha_1 = \alpha_2 = \alpha = 0.1$ and take the following model:

- Model 4. Let $\mathbf{\Omega}_{11}$ be the matrix $\mathbf{\Omega}$ in Hub graph with dimension p_1. Let $\mathbf{\Omega}_{22} = (\omega_{i,j})_{(p-p_1)\times(p-p_1)}$, where $\omega_{i,i+1} = \omega_{i+1,i} = 0.5$ and $\omega_{i,j} = 0$ for all $|j - i| \geq 2$. Define $\mathbf{\Omega}_1^* = diag(\boldsymbol{I}_{p_1\times p_1}, \mathbf{\Omega}_{22})$ and $\mathbf{\Omega}_2^* = diag(\mathbf{\Omega}_{11}, \mathbf{\Omega}_{22})$. Finally, $\mathbf{\Omega}^{(1)} = \mathbf{\Omega}_1^* + (\max(\xi_1^-, \xi_2^-) + 0.01)\boldsymbol{I}_p$ and $\mathbf{\Omega}^{(2)} = \mathbf{\Omega}_2^* + (\max(\xi_1^-, \xi_2^-) + 0.01)\boldsymbol{I}_p$.

Note that as p_1 grows from 10 to 80, the overlapping fraction κ decreases from 0.8144 to 0.1. The power result is given in Figure 4 (d). We can see that the power of our joint estimation on common edges is quite stable to the overlapping fraction κ. Procedure (2.11) still significantly outperforms



(a) Band graph (b) Hub graph (c) E-R graph (d) Model 4

Fig 4: *Power comparison with separate estimation.* $power_{sepa}$ $(-\diamond-)$ *and* $power_2$ $(-\star-)$ *denote the powers of separate estimation and joint estimation procedure (2.11), respectively. The x-axis denotes the* α *value and the y-axis denotes the empirical power.*

separate estimation for small overlapping fraction.

5. Discussion. In the estimation of a single GGM, Liu (2013) proposed an asymptotically normally distributed test statistic. Based on this statistic, we further develop two new asymptotically independent test statistics and propose a hierarchical method to estimate the differential/similar structure among multiple GGMs. Theoretical results using new proof techniques are developed, which show that the false discovery rates in estimators of differential/similar structure can be controlled at a nominal level. Power results are further established, which show that the proposed method can be more powerful than separate estimation on detecting the similar structure.

The method and proof techniques developed in this paper may be extended to other settings. For example, in the covariance graph models (Cox and Wermuth, 1996), the partial correlation coefficient is replaced by Pearson correlation coefficient. We note that the sample correlation coefficient has a similar property as Proposition 3.1 (Anderson, 2003) and thus it is easy to extend the proposed hierarchical method to detect differential/similar structure among multiple covariance graphs. Theoretical results can be established by the proof techniques in this paper.

Some recent works have studied the estimation of high dimensional nonparanormal graphical models (Liu, et al. 2012, Xue and Zou 2012). For nonparanormal graphical models, Gu, et al. (2015) established asymptotic normality results for estimators of entries in precision matrix. It would be interesting to study the differential/similar structure estimation between multiple nonparanormal graphical models. One might replace $T_{ij}^{(k)}$ by the estimators in Gu, et al. (2015) and construct similar test statistics as in (2.5) and (2.9). The development of related theoretical results is left for

future work. It would also be interesting to study this problem for other exponential family graphical models such as Ising models, Poisson graphical models and exponential graphical models. As we can see from the construction of test statistics in Section 2, one possible pivotal tool is an estimator with the asymptotic distribution for the weight in each graph. However, this is still an open and challenging problem.

6. Proof. In this section, we give the proofs of Theorem 3.1 and Theorem 3.2. The proofs of Proposition 3.1, Theorems 3.3-3.5 and other technical lemmas are given in the supplementary material Liu (2016).

6.1. *Proof of Theorem 3.1.* It is easy to see that

$$1-\Phi(\hat{t}_1)=\frac{\alpha_1 A(\hat{t}_1)\max(1,\sum_{1\le i<j\le p} I\{T_{ij,1}\ge \hat{t}_1\})}{(p^2-p)/2}. \tag{6.1}$$

Let $\mathcal{A}_0=\{(i,j): D_{ij}(\boldsymbol{\rho})=0, 1\le i<j\le p\}$. We first prove that

$$\sup_{t\le b_p}\Big|\frac{\sum_{(i,j)\in\mathcal{A}_0} I\{T_{ij,1}\ge t\}}{|\mathcal{A}_0|G(t)}-1\Big|\to 0 \quad \text{in probability} \tag{6.2}$$

as $(n,p)\to\infty$, where $G(t)=1-\Phi(t)$ and b_p satisfies $p^2G(b_p)\to\infty$. Note that (6.2) is equivalent to

$$\sup_{0\le t\le b_p'}\Big|\frac{\sum_{(i,j)\in\mathcal{A}_0} I\{T_{ij,*}\ge t\}}{|\mathcal{A}_0|(1-\Psi_0(t))}-1\Big|\to 0 \quad \text{in probability,} \tag{6.3}$$

where $b_p'=\Psi_0^{-1}(\Phi(b_p))$. Recall that $\lambda_{M_1}-\lambda_{M_1+1}\ge c>0$. By the result in Zolotarev (1961), the density function of $\sum_{i=1}^M \lambda_i Z_i^2$ satisfies

$$p_{\boldsymbol{W}}(x)=\frac{K_1}{(2\lambda_1^2)^{M_1/2}\Gamma(M_1/2)}x^{M_1/2-1}e^{-\frac{x}{2\lambda_1}}(1+\varepsilon(x)), \tag{6.4}$$

where $\varepsilon(x)\to 0$ uniformly in $\{n_1,\ldots,n_K\}$ as $x\to\infty$, $K_1=\prod_{r=2}^M(1-\lambda_r/\lambda_1)^{-M_r/2}$ and M_r is the multiplicity of λ_r. By the integration by parts,

$$1-\Psi_0(t)\sim\frac{K_1}{2^{M_1/2-1}\lambda_1^{M_1-1}\Gamma(\frac{M_1}{2})}t^{M_1-2}\exp\Big(-\frac{t^2}{2\lambda_1}\Big) \tag{6.5}$$

as $t\to\infty$. Let $\boldsymbol{U}_{ij}=(U_{ij}^{(k,l)}, 1\le k<l\le p)$, where

$$U_{ij}^{(k,l)} = \frac{\frac{1}{n_k}\sum_{h=1}^{n_k}\varepsilon_{hij}^{(k)}-\frac{1}{n_l}\sum_{h=1}^{n_l}\varepsilon_{hij}^{(l)}}{\sqrt{\frac{1}{n_k}(1-\rho_{ij\cdot,k}^2)^2+\frac{1}{n_l}(1-\rho_{ij\cdot,l}^2)^2}},$$

$$\begin{aligned}\boldsymbol{\varepsilon}_{mij}^{(k)} &= \sqrt{\omega_{ii,k}\omega_{jj,k}}(\varepsilon_{mi}^{(k)}\varepsilon_{mj}^{(k)} - \mathsf{E}\varepsilon_{mi}^{(k)}\varepsilon_{mj}^{(k)}) \\ &\quad + \frac{1}{2}\rho_{ij\cdot,k}(\omega_{ii,k}((\varepsilon_{mi}^{(k)})^2 - 1) + \frac{1}{2}\rho_{ij\cdot,k}(\omega_{jj,k}((\varepsilon_{mj}^{(k)})^2 - 1).\end{aligned}$$

Note that $\mathsf{Var}(\boldsymbol{\varepsilon}_{mij}^{(k)}) = (1-\rho_{ij\cdot,k}^2)^2$. By the proof of Proposition 3.1, we have

$$T_{ij}^{(k)} - \rho_{ij\cdot,k} = \frac{1}{n_k}\sum_{m=1}^{n_k}\boldsymbol{\varepsilon}_{mij}^{(k)} + O_{\mathsf{P}}\Big(a_{n1}\sqrt{\log p/n} + a_{n2}^2 + \frac{\log p}{n}\Big) \tag{6.6}$$

uniformly in $1 \le i, j \le p$, which implies that $\max_{(i,j)\in\mathcal{A}_0}\|\boldsymbol{T}_{ij} - \boldsymbol{U}_{ij}\| = o_{\mathsf{P}}\Big(\frac{1}{\sqrt{\log p}}\Big)$. Put $\mathrm{U}_{ij} = \|\boldsymbol{U}_{ij}\|$. By (6.5), it suffices to prove that

$$\sup_{0\le t\le b_p'}\Big|\frac{\sum_{(i,j)\in\mathcal{A}_0} I\{\mathrm{U}_{ij}\ge t\}}{q_0(1-\Psi_0(t))} - 1\Big| \to 0 \quad \text{in probability} \tag{6.7}$$

as $(n,p)\to\infty$, where $q_0 = |\mathcal{A}_0|$. Put

$$f_{ij}(t) = I\{\mathrm{U}_{ij}\ge t\} - \mathsf{P}(\mathrm{U}_{ij}\ge t).$$

Following Lemma 6.1 and the proof of Lemma 6.3 in Liu (2013), we only need to show that for any $\varepsilon > 0$,

$$\sup_{0\le t\le b_p'}\mathsf{P}\Big(\Big|\frac{\sum_{(i,j)\in\mathcal{A}_0} f_{ij}(t)}{q_0(1-\Psi_0(t))}\Big| \ge \varepsilon\Big) = o(1), \tag{6.8}$$

and

$$\int_0^{b_p'}\mathsf{P}\Big(\Big|\frac{\sum_{(i,j)\in\mathcal{A}_0} f_{ij}(t)}{q_0(1-\Psi_0(t))}\Big| \ge \varepsilon\Big)dt = o(1/\sqrt{\log p}). \tag{6.9}$$

We only prove (6.9) because (6.8) follows from the proof of (6.9) directly.

Define

$$\begin{aligned}&\mathcal{B}_i = \{j : \sum_{k=1}^{K}|\rho_{ij\cdot,k}| \ge K(\log p)^{-3-\gamma}\},\ \mathcal{S} = \{(i,j) : 1\le i\le p, j\in\mathcal{B}_i\}, \\ &\mathcal{A}_{01} = \mathcal{A}_0\cap\mathcal{S}, \quad \mathcal{A}_{02} = \mathcal{A}_0\cap\mathcal{S}^c.\end{aligned}$$

By (C2), we have $\mathrm{Card}(\mathcal{A}_{01}) \le Cp^{1+\xi}$ with $\xi < 1/3$. Recall that $q_0 \ge cp^2$ for some $c > 0$. By Lemma 6.1, uniformly in $0\le t\le b_p'$,

$$\mathsf{E}\Big|\frac{\sum_{(i,j)\in\mathcal{A}_{01}} f_{ij}(t)}{q_0(1-\Psi_0(t))}\Big| \le C\frac{p^{1+\xi}(1-\Psi_0(t))}{q_0(1-\Psi_0(t))} = O(p^{-1+\xi}). \tag{6.10}$$

Note that for $(i,j) \in \mathcal{A}_{02}$ and $(k,l) \in \mathcal{A}_{02}$,

$$Corr(\varepsilon_{1ij}^{(m)}, \varepsilon_{1kl}^{(m)}) = \rho_{ik\cdot,m}\rho_{jl\cdot,m} + \rho_{il\cdot,m}\rho_{kj\cdot,m} + O((\log p)^{-3-\gamma}).$$

For some large constant $C > 0$, define

$$\begin{aligned}
\mathcal{A}_{4m} &= \{(i,j,k,l) : \ (i,j) \in \mathcal{A}_{02}, \ (k,l) \in \mathcal{A}_{02}, \\
&\qquad |Corr(\varepsilon_{1ij}^{(m)}, \varepsilon_{1kl}^{(m)})| \le C(\log p)^{-3-\gamma}\}, \\
\mathcal{A}_{5m} &= \{(i,j,k,l) \notin \mathcal{A}_{4m} : \ (i,j) \in \mathcal{A}_{02}, \ (k,l) \in \mathcal{A}_{02}, \\
&\qquad |Corr(\varepsilon_{1ij}^{(m)}, \varepsilon_{1kl}^{(m)})| \le \rho + C(\log p)^{-3-\gamma}\}, \\
\mathcal{A}_{6m} &= \{(i,j,k,l) \notin \mathcal{A}_{4m} \cup \mathcal{A}_{5m} : \ (i,j) \in \mathcal{A}_{02}, \ (k,l) \in \mathcal{A}_{02}\}.
\end{aligned}$$

Let $\mathcal{A}_4 = \cap_{m=1}^K \mathcal{A}_{4m}$, $\mathcal{A}_6 = \cup_{m=1}^K \mathcal{A}_{6m}$ and $\mathcal{A}_5 = (\cup_{m=1}^K \mathcal{A}_{5m}) \setminus \mathcal{A}_6$. It can be shown that $Card(\mathcal{A}_5) = O(p^{2+2\xi})$ and $Card(\mathcal{A}_6) = O(p^{1+3\xi}+p^2)$. Moreover, for $(i,j,k,l) \in \mathcal{A}_5$, we have for all $1 \le m \le K$,

$$|Corr(\varepsilon_{1ij}^{(m)}, \varepsilon_{1kl}^{(m)})| \le \rho + C(\log p)^{-3-\gamma}.$$

Set $f_{ijkl}(t) = \mathsf{P}(\mathrm{U}_{ij} \ge t, \mathrm{U}_{kl} \ge t) - \mathsf{P}(\mathrm{U}_{ij} \ge t)\mathsf{P}(\mathrm{U}_{kl} \ge t)$. Then

$$\mathsf{E}\Big[\sum_{(i,j)\in\mathcal{A}_{02}} f_{ij}(t)\Big]^2 = \sum_{(i,j,k,l)\in\mathcal{A}_4} f_{ijkl}(t) + \sum_{(i,j,k,l)\in\mathcal{A}_5} f_{ijkl}(t) + \sum_{(i,j,k,l)\in\mathcal{A}_6} f_{ijkl}(t).$$

By Lemma 6.1, for any $b > 0$, we have uniformly in $0 \le t \le b\sqrt{\log p}$ that

$$\Big|\frac{\sum_{(i,j,k,l)\in\mathcal{A}_4} f_{ijkl}(t)}{q_0^2(1-\Psi_0(t))^2}\Big| \le C(\log p)^{-2-\gamma} \tag{6.11}$$

$$\Big|\frac{\sum_{(i,j,k,l)\in\mathcal{A}_5} f_{ijkl}(t)}{q_0^2(1-\Psi_0(t))^2}\Big| \le \frac{C}{p^{2-2\xi-\delta}[(1-\Psi_0(t))]^{(2\rho)/(1+\rho)}} \tag{6.12}$$

and

$$\Big|\frac{\sum_{(i,j,k,l)\in\mathcal{A}_6} f_{ijkl}(t)}{q_0^2(1-\Psi_0(t))^2}\Big| \le \frac{C}{p^{3-3\xi}(1-\Psi_0(t))} + \frac{C}{p^2(1-\Psi_0(t))} \tag{6.13}$$

for some $\gamma > 0$ and any $\delta > 0$. By (6.5) and some elementary calculations,

$$\int_0^{b_p'} \Big[\frac{1}{p^{2-2\xi-\delta}[(1-\Psi_0(t))]^{(2\rho)/(1+\rho)}} + \frac{1}{p^2(1-\Psi_0(t))}\Big]dt = o(1/\sqrt{\log p}).$$

This, together with (6.10)-(6.13), implies that (6.9) holds.

Note that $(1/\log p) = O(A(c\sqrt{\log p}))$ for any $c > 0$. By the definition of $\hat{t}_1$ and the tail probability of normal distributions, we have $\hat{t}_1 \le (2+\epsilon)\sqrt{\log p}$ for any $\epsilon > 0$ as $(n,p) \to \infty$. Let $\mathcal{D} := \Big\{(i,j) : 1 \le i < j \le p, \quad \|\boldsymbol{\rho}_{ij}\| \ge \theta\sqrt{\lambda_1 \log p}\Big\}$ for $\theta > 2$ in (3.5). By (6.6) and Proposition 3.1, we have

$$\mathsf{P}(T_{ij,1} \ge \hat{t}_1) \to 1 \quad \text{as } (n,p) \to \infty, \text{ uniformly in } (i,j) \in \mathcal{D}.$$

By Markov's inequality, we obtain that as $(n,p) \to \infty$,

$$\frac{\sum_{(i,j)\in\mathcal{D}} I\{T_{ij,1} \ge \hat{t}_1\}}{|\mathcal{D}|} \to 1 \quad \text{in probability.} \tag{6.14}$$

Let b_p in (6.2) satisfy $p^2 G(b_p) \to \infty$ and $p^2 G(b_p)/|\mathcal{D}| \to 0$. By Lemma 6.1,

$$\mathsf{E}(\mathrm{FDP}_1 I\{\hat{t}_1 \ge b_p\}) \le 2\frac{\sum_{(i,j)\in\mathcal{A}_0} \mathsf{P}(T_{ij,1} \ge b_p)}{|\mathcal{D}|} + o(1) = o(1). \tag{6.15}$$

By (6.1) and (6.2), we can get, for any $\varepsilon > 0$,

$$\mathsf{P}(\mathrm{FDP}_1 I\{\hat{t}_1 \le b_p\} \ge \alpha_1 + \varepsilon) \to 0 \quad \text{as } (n,p) \to \infty. \tag{6.16}$$

It yields that $\mathsf{P}(\mathrm{FDP}_1 \le \alpha_1 + \varepsilon) \to 1$ and $limsup_{(n,p)\to\infty} \mathrm{FDR}_1 \le \alpha_1$.

Suppose that (3.6) holds. By (6.10)-(6.13), we have for any bounded t,

$$\mathsf{P}\Big(\Big|\frac{\sum_{(i,j)\in\mathcal{A}_0} f_{ij}(t)}{|\mathcal{A}_0|}\Big| \le (\log p)^{-1-\gamma/3}\Big) \to 1 \tag{6.17}$$

as $(n,p) \to \infty$. Note that the volume of $\{(z_1,\ldots,z_M) : t \le (\sum_{i=1}^M \lambda_i z_i^2)^{1/2} \le t + o(1/\log p)\}$ is of order of $o(1/\log p)$. So we have $\Psi_0(t + o(1/\log p)) - \Psi_0(t) = o(1/\log p)$. By Lemma 6.1,

$$\mathsf{P}(\mathrm{U}_{ij} \ge t + o(1/\log p)) - \mathsf{P}(\mathrm{U}_{ij} \ge t)) = o(1/\log p)$$

uniformly in $(i,j) \in \mathcal{A}_0$ and any bounded t. By (3.6) and (6.6),

$$\max_{(i,j)\in\mathcal{A}_0} \|\boldsymbol{T}_{ij} - \boldsymbol{U}_{ij}\| = o_{\mathsf{P}}\Big(1/\log p\Big).$$

The above inequalities together with $Card(\mathcal{H}_1) = o(p^2/\log p)$ imply that $P_0 - \hat{P}_0 = o_{\mathsf{P}}(1/\log p)$. Since $\hat{t}_1 \le (2+\epsilon)\sqrt{\log p}$, we have $A(\hat{t}_1) \to 1$ in probability. By (6.1), the fact $|\mathcal{D}| \to \infty$ and (6.14), we can obtain that $p^2 G(\hat{t}_1) \to \infty$ in probability. Hence, there exists a sequence $\{b_p\}$ such that $p^2 G(b_p) \to \infty$ and $\mathsf{P}(\hat{t}_1 \le b_p) \to 1$ as $(n,p) \to \infty$. This, together with (6.2), yields that

$$\Big|\frac{\sum_{(i,j)\in\mathcal{A}_0} I\{T_{ij,1} \ge \hat{t}_1\}}{|\mathcal{A}_0| G(\hat{t}_1)} - 1\Big| \to 0 \quad \text{in probability.} \tag{6.18}$$

By (6.1), we prove $\mathrm{FDP}_1 \to \alpha_1$ in probability and $\mathrm{FDR}_1 \to \alpha_1$. □

6.2. *Proof of Theorem 3.2.* As in the proof of Theorem 3.1, we first show that, for any b_p satisfying $p^2G(b_p) \to \infty$,

$$\sup_{t\le b_p}\Big|\frac{\sum_{(i,j)\in\hat{\mathcal{A}}_{10}^c} I\{T_{ij,2}\ge t\}}{|\hat{\mathcal{A}}_{10}^c|G(t)}-1\Big|\to 0 \quad \text{in probability} \tag{6.19}$$

where $\hat{\mathcal{A}}_{10}^c=\{(i,j)\in\hat{\mathcal{A}}_1^c:\ (\rho_{ij\cdot,1},\dots,\rho_{ij\cdot,K})=0\}$. Let

$$V_{ij}^{(k)}=\frac{1}{n_k}\sum_{h=1}^{n_k}\varepsilon_{hij}^{(k)},\quad V_{ij}=\frac{\sum_{k=1}^K n_kV_{ij}^{(k)}}{\sqrt{\sum_{k=1}^K n_k(1-\rho_{ij\cdot,k}^2)^2}},\quad (i,j)\in\hat{\mathcal{A}}_1^c.$$

By (6.6), it suffices to show that

$$\sup_{0\le t\le b_p'}\Big|\frac{\sum_{(i,j)\in\hat{\mathcal{A}}_{10}^c} I\{|V_{ij}|\ge t\}}{|\hat{\mathcal{A}}_{10}^c|(2-2\Phi(t))}-1\Big|\to 0 \quad \text{in probability} \tag{6.20}$$

as $(n,p)\to\infty$, where $b_p'=\Psi_1^{-1}(\Phi(b_p))$ and $\Psi_1(x)=2\Phi(x)-1$. Since $\hat{t}_1\le(2+\epsilon)\sqrt{\log p}$ and $(1/\log p)=O(A(c\sqrt{\log p}))$ for any $c>0$, we have

$$\{\sum_{1\le i<j\le p} I\{T_{ij,1}\ge\hat{t}_1\}\ge p^2/(\log p)^3\}\subseteq\{p^2G(\hat{t}_1)\ge p\}=\{\hat{t}_1\le G^{-1}(1/p)\}.$$

By (6.16), for any $\varepsilon>0$,

$$\mathsf{P}\Big(\mathrm{FDP}_1 I\Big\{\sum_{1\le i<j\le p} I\{T_{ij,1}\ge\hat{t}_1\}\ge p^2/(\log p)^3\Big\}\ge\alpha_1+\varepsilon\Big)\to 0$$

as $(n,p)\to\infty$. This, together with $Card(\mathcal{H}_1)=o(p^2/\log p)$, implies that $\sum_{1\le i<j\le p} I\{T_{ij,1}\ge\hat{t}_1\}=o_{\mathsf{P}}(p^2/\log p)$ and $|\hat{\mathcal{A}}_{10}^c|/((p^2-p)/2)\to 1$ in probability. By (6.1), $1-\Phi(\hat{t}_1)=o_{\mathsf{P}}(1/\log p)$. So we can choose a sequence $\{c_n\}$ that satisfies $\mathsf{P}(\hat{t}_1\ge c_n)\to 1$ and $1-\Phi(c_n)=o(1/\log p)$. We now show that

$$\sup_{0\le t\le b_p'}\Big|\frac{\sum_{(i,j)\in\mathcal{H}_0} I\{T_{ij,1}\ge c_n,|V_{ij}|\ge t\}}{p^2(1-\Phi(t))}\Big|\to 0 \quad \text{in probability,} \tag{6.21}$$

where $\mathcal{H}_0=\{(i,j):\rho_{ij\cdot,1}=\cdots=\rho_{ij\cdot,K}=0,1\le i<j\le p\}$. Following the proof of Lemma 6.3 in Liu (2013), we only need to show that, for any $\varepsilon>0$,

$$\int_0^{b_p'}\mathsf{P}\Big(\Big|\frac{\sum_{(i,j)\in\mathcal{H}_0} I\{T_{ij,1}\ge c_n,|V_{ij}|\ge t\}}{p^2(1-\Phi(t))}\Big|\ge\varepsilon\Big)dt=o(1/\sqrt{\log p}) \tag{6.22}$$

and

$$\sup_{0\le t\le b'_p} \mathsf{P}\Big(\Big|\frac{\sum_{(i,j)\in\mathcal{H}_0} I\{T_{ij,1}\ge c_n, |V_{ij}|\ge t\}}{p^2(1-\Phi(t))}\Big|\ge \varepsilon\Big) = o(1). \tag{6.23}$$

By Lemma 6.3 and Markov's inequality, we have

$$\mathsf{P}\Big(\Big|\frac{\sum_{(i,j)\in\mathcal{H}_0} I\{T_{ij,1}\ge c_n, |V_{ij}|\ge t\}}{p^2(1-\Phi(t))}\Big|\ge \varepsilon\Big) = o(1/\log p). \tag{6.24}$$

This implies (6.22) and (6.23). To prove (6.20), by (6.21), it is enough to show that

$$\sup_{0\le t\le b'_p} \Big|\frac{\sum_{(i,j)\in\mathcal{H}_0} I\{|V_{ij}|\ge t\}}{|\mathcal{H}_0|(2-2\Phi(t))} - 1\Big| \to 0 \quad \text{in probability} \tag{6.25}$$

as $(n,p)\to\infty$. This follows from Lemma 6.2 and the proof of (6.7).

Define

$$\mathcal{H}_{11} = \Big\{(i,j): 1\le i<j\le p, \quad |\rho_{ij\cdot,1} = \cdots = \rho_{ij\cdot,K}| \ge \theta\sqrt{\log p/n}\Big\}.$$

By $\hat{t}_1\to\infty$ in probability and Proposition 3.1 (or Lemma 6.1), we have $\min_{(i,j)\in\mathcal{H}_{11}} \mathsf{P}(T_{ij,1}\ge \hat{t}_1)\to 0$. By Markov's inequality,

$$\frac{\sum_{(i,j)\in\mathcal{H}_{11}} I\{T_{ij,1}\ge \hat{t}_1\}}{|\mathcal{H}_{11}|} \to 0 \quad \text{in probability.} \tag{6.26}$$

Also, we can show that $\hat{t}_2 \le (2+\epsilon)\sqrt{\log p}$ for any $\epsilon>0$ when n and p are large. It follows from $\theta>2$ and Proposition 3.1 (or Lemma 6.2) that $\mathsf{P}(T_{ij,2}\ge \hat{t}_2)\to 1$ uniformly in $(i,j)\in\mathcal{H}_{11}$. This, together with (6.26), implies that

$$\frac{\sum_{(i,j)\in\mathcal{H}_{11}} I\{T_{ij,2}\ge \hat{t}_2\}}{|\mathcal{H}_{11}|}\to 1 \quad \text{and} \quad \frac{\sum_{(i,j)\in\mathcal{H}_{11}\cap\hat{\mathcal{A}}_1^c} I\{T_{ij,2}\ge \hat{t}_2\}}{|\mathcal{H}_{11}|}\to 1$$

in probability. Now using the same arguments as in the proof of (6.14) and (6.15), we can show that $\mathsf{P}(\mathrm{FDP}_2\le \alpha_2+\varepsilon)\to 1$ for any $\varepsilon>0$.

As the proof of (6.17), we have for any bounded t,

$$\mathsf{P}\Big(\Big|\frac{\sum_{(i,j)\in\mathcal{H}_0}[I\{V_{ij}\ge t\} - \mathsf{P}(V_{ij}\ge t)]}{|\mathcal{H}_0|}\Big| \le (\log p)^{-1-\gamma/3}\Big)\to 1 \tag{6.27}$$

as $(n,p)\to\infty$. By (6.24), for any $\varepsilon>0$,

$$\mathsf{P}\Big(\Big|\frac{\sum_{(i,j)\in\mathcal{H}_0} I\{T_{ij,1}\ge c_n, |V_{ij}|\ge t\}}{p^2}\Big|\ge \frac{\varepsilon}{\log p}\Big)=o(1). \tag{6.28}$$

It follows from (6.27) and (6.28) that

$$\frac{\sum_{(i,j)\in\hat{\mathcal{A}}_{10}^c}[I\{T_{i,j,2}\ge t\}-\mathsf{P}(T_{ij,2}\ge t)]}{|\hat{\mathcal{A}}_{10}^c|}=o_{\mathsf{P}}(1/\log p).$$

This, together with (6.6), Lemma 6.2 and $Card(\mathcal{H}_1)=o(p^2/\log p)$, implies that $P_0-\hat{P}_0^{'}=o_{\mathsf{P}}(1/\log p)$ and $A^{'}(\hat{t}_2)\to 1$ in probability. By (6.19), we prove that $\mathrm{FDP}_2\to\alpha_2$ in probability and $\lim_{(n,p)\to\infty}\mathrm{FDR}_2=\alpha_2$. □

6.3. *Technical lemmas.* In this section, we give some technical lemmas and their proofs are given in the supplementary material Liu (2016).

LEMMA 6.1. *(i). For any $b>0$,*

$$\max_{(i,j)\in\mathcal{A}_0}\sup_{0\le t\le b\sqrt{\log p}}\Big|\frac{P(\|\boldsymbol{U}_{ij}\|\ge t)}{1-\Psi_0(t)}-1\Big|\le C(\log p)^{-3}. \tag{6.29}$$

(ii). For any $\delta>0$ and $b>0$,

$$P(\|\boldsymbol{U}_{ij}\|\ge t,\|\boldsymbol{U}_{kl}\|\ge t)\le C\exp\Big(-\frac{t^2}{\lambda_1(1+\rho+\delta)}\Big) \tag{6.30}$$

uniformly for $(i,j,k,l)\in\mathcal{A}_5$ and $0\le t\le b\sqrt{\log p}$.

(iii).For any $b>0$,

$$P(\|\boldsymbol{U}_{ij}\|\ge t,\|\boldsymbol{U}_{kl}\|\ge t)=(1+A_n)(1-\Psi_0(t))^2 \tag{6.31}$$

uniformly for $(i,j,k,l)\in\mathcal{A}_4$ and $0\le t\le b\sqrt{\log p}$, where $|A_n|\le(\log p)^{-2-\gamma}$ for some $\gamma>0$.

In Lemma 6.2, define $\mathcal{A}_4$, $\mathcal{A}_5$ and $\mathcal{A}_6$ as in the proof of Theorem 3.1 with $\mathcal{A}_0$ being replaced by $\mathcal{H}_0$.

LEMMA 6.2. *(i). For any $b>0$,*

$$\max_{1\le i<j\le p}\sup_{0\le t\le b\sqrt{\log p}}\Big|\frac{P(|V_{ij}|\ge t)}{2-2\Phi(t)}-1\Big|\le C(\log p)^{-3}. \tag{6.32}$$

(ii). For any $\delta > 0$ and $b > 0$,

$$P(|V_{ij}| \geq t, |V_{kl}| \geq t) \leq C \exp\Big(- \frac{t^2}{\lambda_1(1+\rho+\delta)} \Big) \tag{6.33}$$

uniformly for $(i,j,k,l) \in \mathcal{A}_5$ and $0 \leq t \leq b\sqrt{\log p}$.

(iii). For any $b > 0$,

$$P(|V_{ij}| \geq t, |V_{kl}| \geq t) = (1 + A_n)(2 - 2\Phi(t))^2 \tag{6.34}$$

uniformly for $(i,j,k,l) \in \mathcal{A}_4$ and $0 \leq t \leq b\sqrt{\log p}$, where $|A_n| \leq (\log p)^{-2-\gamma}$ for some $\gamma > 0$.

LEMMA 6.3. *We have for any $b > 0$,*

$$P(\|\boldsymbol{U}_{ij}\| \geq t_1, |V_{ij}| \geq t_2) = (1 + o(1))(1 - \Psi_0(t_1))(2 - 2\Phi(t_2))$$

uniformly in $(i,j) \in \mathcal{H}_0$ and $0 \leq t_1, t_2 \leq b\sqrt{\log p}$.

Acknowledgements. The author would like to thank the associate editor, three referees and Xi Chen for their helpful constructive comments which have helped to improve quality and presentation of the paper.

Supplementary material: Structural similarity and difference testing on multiple sparse Gaussian graphical models

The supplementary material includes the proofs of Proposition 3.1, Theorems 3.3-3.5 and Lemmas 6.1-6.3. Also, a part of numerical results in Section 4 are included.

References.

[1] Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis.* Third edition. Wiley-Interscience.

[2] d'Aspremont, A., Banerjee, O., and El Ghaoui, L. (2008). First-order methods for sparse covariance selection. *SIAM Journal on Matrix Analysis and its Applications*, 30, 56-66.

[3] Belilovsky, E., Varoquaux, G. and Blaschko, M.B. (2015). Hypothesis testing for differences in Gaussian graphical models: applications to brain connectivity. Technical report. http://arxiv.org/abs/1512.08643

[4] Belloni, A., Chernozhukov, V. and Wang, L. (2011). Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, 98, 791-806.

[5] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 57, 289-300.

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

[7] Candès, E. and Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n. *Annals of Statistics*, 35, 2313-2351.

[8] Chiquet, J., Grandvalet, Y. and Ambroise, C. (2011). Inferring multiple graphical structures. *Statistics and Computing*, 21, 537-553.

[9] Chu, J.H., Lazarus, R., Carey, V.J. and Raby, B.A. (2011). Quantifying differential gene connectivity between disease states for objective identification of disease-relevant genes. *BMC Systems Biology*, 5:89.

[10] Cox, D.R. and Wermuth, N. (1996). Multivariate Dependencies. Chapman and Hall, London.

[11] 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, 373-397.

[12] de la Fuente, A. (2010). From "differential expression" to "differential networking"- identification of dysfunctional regulatory networks in diseases. *Trends in Genetics* 26: 326-333.

[13] Efron, B. (2007). Correlation and large-scale simultaneous significance testing. *Journal of the American Statistical Association*, 102, 93-103.

[14] Fan, J., Feng, Y. and Wu, Y. (2009). Network exploration via the adaptive lasso and SCAD penalties. *Annals of Applied Statistics*, 2, 521-541.

[15] Friedman, J., Hastie, T. and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9, 432-441.

[16] Gill, R., Datta, S. and Datta, S. (2010). A statistical framework for differential network analysis from microarray data. *BMC Bioinformatics*, 11:95.

[17] Gu, Q., Cao, Y., Ning, Y. and Liu, H. (2015). Local and global inference for high dimensional nonparanormal graphical models. Technical report. http://arxiv.org/abs/1502.02347.

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

[19] Hara, S. and Washio, T. (2013). Learning a common substructure of multiple graphical Gaussian Models. *Neural Networks*, 38, 23-38.

[20] Honorio, J. and Samaras, D. (2010). Multi-task learning of Gaussian graphical models. *Proceedings of the 27 th International Conference on Machine Learning, Haifa, Israel, 2010.*

[21] 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.

[22] Liu, H., Han, F., Yuan, M., Lafferty, J. and Wasserman, L. (2012). High dimensional semiparametric Gaussian copula graphical models. *Annals of Statistics*, 40, 2293-2326.

[23] Liu, W. (2013). Gaussian graphical model estimation with false discovery rate control. *Annals of Statistics*, 41, 2948-2978.

[24] Liu, W. (2016). Supplement to "Structural similarity and difference testing on multiple sparse Gaussian graphical models".

[25] Liu, W. and Shao, Q.M. (2014). Phase transition and regularized bootstrap in large-scale t-tests with false discovery rate control. *Annals of Statistics*, 42, 2003-2025.

[26] Ma, S., Gong, Q. and Bohnert, H.J. (2007). An Arabidopsis gene network based on the graphical Gaussian model. *Genome Research*, 17, 1614-1625.

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

[28] Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society, Series B*, 72, 417-473.

[29] Negahban, S.N. and Wainwright, M.J. (2011). Simultaneous support recovery in high dimensions: benefits and perils of block l_1-regularization. *IEEE Transactions on Information Theory*, 57, 3841-3863.

[30] Obozinski, G., Wainwright, M. and Jordan, M.I. (2011). Support union recovery in high-dimensional multivariate regression. *The Annals of Statistics*, 39, 1-47.

[31] Ravikumar, P., Wainwright, M., Raskutti, G. and Yu, B. (2011). High-dimensional covariance estimation by minimizing l_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5: 935-980.

[32] Ren, Z., Sun, T., Zhang, C.H. and Zhou, H.H. (2015). Asymptotic normality and optimalities in estimation of large Gaussian graphical model. *Annals of Statistics*, 43, 991-1026.

[33] Rothman, A., Bickel, P., Levina, E. and Zhu, J. (2008). Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2, 494-515.

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

[35] Spira, A., Beane, J., Shah, V., Steiling, K., Liu, G., Schembri, F., Gilman, S., Dumas, Y., Calner, P., Sebastiani, P., Sridhar, S., Beamis, J., Lamb, C., Anderson, T., Gerry, N., Keane, J., Lenburg, M. and Brody, J. (2007). Airway epithelial gene expression in the diagnostic evaluation of smokers with suspect lung cancer. *Nature Medicine*, 13, 361-366.

[36] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58, 267-288.

[37] Xue, L. and Zou, H. (2012). Regularized rank-based estimation of high-dimensional nonparanormal graphical models. *Annals of Statistics*, 40, 2541-2571.

[38] Yang, S., Lu, Z., Shen, X., Wonka, P. and Ye, J. (2014). Fused multiple graphical lasso. Technical report.

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

[40] Yuan, M. and Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94, 19-35.

[41] Zhang, B. and Wang, Y. (2010). Learning structural changes of Gaussian graphical models in controlled experiments, in *Proc. of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence (UAI 2010)*, Avalon, CA: 701-708, July 2010.

[42] Zhang, C. (2010). Estimation of large inverse matrices and graphical model selection. Technical Report, Rutgers University, Department of Statistics and Biostatistics.

[43] Zhao, S.D., Cai, T.T. and Li, H. (2014). Direct estimation of differential networks. *Biometrika*, 101, 253-268.

[44] Zhu, D., Hero, A.O., Qin, Z.S. and Swaroop, A. (2005). High throughput screening of co-expressed gene pairs with controlled false discovery rate (FDR) and minimum acceptable strength (MAS). *Journal of Computational Biology*, 12, 1029-1045.

Department of Mathematics
Institute of Natural Sciences and MOE-LSC
Shanghai Jiao Tong University
Shanghai
E-mail: weidongl@sjtu.edu.cn