The Annals of Statistics
2013, Vol. 41, No. 6, 2948–2978
DOI: 10.1214/13-AOS1169

GAUSSIAN GRAPHICAL MODEL ESTIMATION WITH FALSE DISCOVERY RATE CONTROL

BY WEIDONG LIU[1]

Shanghai Jiao Tong University

This paper studies the estimation of a high-dimensional Gaussian graphical model (GGM). Typically, the existing methods depend on regularization techniques. As a result, it is necessary to choose the regularized parameter. However, the precise relationship between the regularized parameter and the number of false edges in GGM estimation is unclear. In this paper we propose an alternative method by a multiple testing procedure. Based on our new test statistics for conditional dependence, we propose a simultaneous testing procedure for conditional dependence in GGM. Our method can control the false discovery rate (FDR) asymptotically. The numerical performance of the proposed method shows that our method works quite well.

1. Introduction. Estimation of dependency networks for high-dimensional data sets is especially desirable in many scientific areas such as biology and sociology. The Gaussian graphical model (GGM) has proven to be a very powerful formalism to infer dependence structures of various data sets. GGM is an equivalent representation of conditional dependence of jointly Gaussian random variables. Inference on the structure of GGM is challenging when the dimension is greater than the sample size. Many classical methods do not work any more.

Let $\boldsymbol{X} = (X_1, \dots, X_p)'$ be a multivariate normal random vector with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. GGM is a graph $G = (V, E)$, where $V = \{X_1, \dots, X_p\}$ is the set of vertices and E is the set of edges between vertices. There is an edge between X_i and X_j if and only if X_i and X_j are conditional dependent given $\{X_k, k \neq i, j\}$. It is well known that estimating the structure of GGM is equivalent to recovering the support of precision matrix $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$; see Lauritzen (1996).

The typical way of GGM estimation depends on regularized optimizations. The past decade has witnessed significant developments on the regularization method for various statistical problems. For example, in the context of variable selection, Tibshirani (1996) introduced Lasso, which selects important variables in regression by solving the least squares optimization with the l_1 regularization. Graphical-Lasso, an extension of Lasso to GGM estimation, was introduced by Yuan and Lin

Received November 2012; revised September 2013.
[1]Supported in part by NSFC Grants 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.
MSC2010 subject classifications. 62H12, 62H15.
Key words and phrases. False discovery rate, Gaussian graphical model, multiple tests.

(2007), Friedman, Hastie and Tibshirani (2008) and d'Aspremont, Banerjee and El Ghaoui (2008). Graphical-Lasso estimates the support of precision matrix by an l_1 penalized likelihood method. Theoretical properties of Graphical-Lasso can be found in Rothman et al. (2008) and Ravikumar et al. (2011). Other methods, based on the l_1-minimization technique, can be found in Meinshausen and Buhlmann (2006), Yuan (2010), Zhang (2010), Cai, Liu and Luo (2011), Liu et al. (2012) and Xue and Zou (2012). The nonconvex penalties, such as the SCAD function penalty [Fan, Feng and Wu (2009)], have also been considered in the context of GGM estimation.

It is well known that regularization approaches often require the choice of tuning parameters. Large tuning parameters often lead to sparse networks and they are powerless on finding the edges with small weights. On the other hand, small tuning parameters will generate many false edges and result in high false discovery rates. The theory of the precise relationship between the number of false edges and the tuning parameter is very difficult to derive.

A different way of GGM estimation relies on simultaneous tests

$$H_{0ij} : \omega_{ij} = 0 \quad \text{versus} \quad H_{1ij} : \omega_{ij} \neq 0 \tag{1}$$

for $1 \leq i < j \leq p$, where $\mathbf{\Omega} =: (\omega_{ij})_{p\times p}$. An edge between X_i and X_j is included into the estimated network if and only if H_{0ij} is rejected. When the dimension p is fixed, Drton and Perlman (2004) proposed a multiple testing procedure to estimate GGM. They used the Fisher's z transformations of the sample partial correlation coefficients (SPCCs). A procedure on controlling the family-wise error was developed. However, when the dimension p is greater than the sample size, the sample partial correlation matrix is not even well defined. Hence, we do not have a natural pivotal estimator as SPCCs so that the asymptotic null distribution can be easily derived. In high-dimensional settings, it becomes very challenging to estimate GGM by tests on the entries of the precision matrix.

In the present paper, we study the estimation of GGM by multiple tests (1). We are particularly interested in high-dimensional settings. The false discovery rate (FDR) is a useful measure on evaluating the performance of GGM estimation. We will introduce a procedure called GGM estimation with FDR control (GFC).

A basic step in hypothesis tests is the construction of test statistics. The sample partial correlation coefficients are not well defined when $p > n$. Hence, we introduce new test statistics suitable for high-dimensional settings. The new test statistics are based on a bias correction version of the sample covariance coefficients of residuals. They are shown to be asymptotically normal distributed under some sparsity conditions on $\mathbf{\Omega}$. In addition to new test statistics, GFC carries out large-scale tests simultaneously. To this end, an adjustment for significance levels is necessary. In this paper, we develop a multiple testing procedure with an adjustment for significance levels and it controls the false discovery rate. The proposed procedure thresholds test statistics directly rather than p-values which were widely used [cf. Benjamini and Hochberg (1995)]. It is convenient for us to develop novel theoretical properties on FDR. We show that the GFC method controls both FDR and false discovery proportion (FDP) asymptotically.

In addition to its desirable theoretical properties, the GFC method is computationally very attractive for high-dimensional data. The computational cost is the same as the neighborhood selection method by Meinshausen and Buhlmann (2006) or the CLIME method by Cai, Liu and Luo (2011). We only need to solve p regression equations with the Lasso or the Dantzig selector. Numerical performance of GFC is investigated by simulated data. Results show that the procedure performs favorably in controlling FDR and FDP.

The rest of the paper is organized as follows. In Section 2.1 we introduce new test statistics for conditional dependence. The GFC procedure is introduced in Section 2.2. In Section 3 we give limiting distributions of our test statistics. Theoretical results on GFC are also stated. Since GFC needs initial estimations of regression coefficients, we provide their detailed implementations in Section 4. Numerical performance of the procedure is evaluated by simulation studies in Section 5. The proofs of the main results are delegated to Section 6.

2. Tests on conditional dependence. We begin this section by introducing basic notation. For any vector $\mathbf{x}$, let $\mathbf{x}_{-i}$ denote the $p-1$ dimensional vector by removing x_i from $\mathbf{x}=(x_1,\dots,x_p)'$. For any $p\times q$ matrix $\mathbf{A}$, let $\mathbf{A}_{i,-j}$ denote the ith row of $\mathbf{A}$ with its jth entry being removed and $\mathbf{A}_{-i,j}$ denote the jth column of $\mathbf{A}$ with its ith entry being removed. $\mathbf{A}_{-i,-j}$ denote a $(p-1)\times(q-1)$ matrix by removing the ith row and jth column of $\mathbf{A}$. Throughout, define $|\mathbf{x}|_0=\sum_{j=1}^{p} I\{x_j\neq 0\}$, $|\mathbf{x}|_1=\sum_{j=1}^{p}|x_j|$ and $|x|_2=\sqrt{\sum_{j=1}^{p}x_j^2}$. For a matrix $\mathbf{A}=(a_{ij})\in\mathbb{R}^{p\times q}$, we define the element-wise l_∞ norm $|\mathbf{A}|_\infty=\max_{1\le i\le p,1\le j\le q}|a_{ij}|$, the spectral norm $\|\mathbf{A}\|_2=\sup_{|\mathbf{x}|_2\le 1}|\mathbf{A}\mathbf{x}|_2$ and the matrix ℓ_1 norm $\|\mathbf{A}\|_{l_1}=\max_{1\le j\le q}\sum_{i=1}^{p}|a_{ij}|$. Let $\lambda_{\max}(\mathbf{\Sigma})$ and $\lambda_{\min}(\mathbf{\Sigma})$ denote the largest eigenvalue and the smallest eigenvalue of $\mathbf{\Sigma}$, respectively. $\mathbf{I}_p$ denotes a $p\times p$ identity matrix. Let $\mathcal{H}_0=\{(i,j):\omega_{ij}=0, 1\le i<j\le p\}$ and $\mathcal{H}_1=\{(i,j):\omega_{ij}\neq 0, 1\le i<j\le p\}$.

It is well known that, for $\boldsymbol{X}=(X_1,\dots,X_p)'\sim N(\boldsymbol{\mu},\mathbf{\Sigma})$, we can write

$$X_i=\alpha_i+\boldsymbol{X}_{-i}'\boldsymbol{\beta}_i+\varepsilon_i, \tag{2}$$

where $\varepsilon_i\sim N(0,\sigma_{ii}-\mathbf{\Sigma}_{i,-i}\mathbf{\Sigma}_{-i,-i}^{-1}\mathbf{\Sigma}_{-i,i})$ is independent of $\boldsymbol{X}_{-i}$, $\alpha_i=\mu_i-\mathbf{\Sigma}_{i,-i}\mathbf{\Sigma}_{-i,-i}^{-1}\boldsymbol{\mu}_{-i}$ and $(\sigma_{ij})_{p\times p}=\mathbf{\Sigma}$; see Anderson (2003). The regression coefficients vector $\boldsymbol{\beta}_i$ and the error terms ε_i satisfy

$$\boldsymbol{\beta}_i=-\omega_{ii}^{-1}\mathbf{\Omega}_{-i,i}\quad\text{and}\quad \mathsf{Cov}(\varepsilon_i,\varepsilon_j)=\frac{\omega_{ij}}{\omega_{ii}\omega_{jj}}.$$

We estimate GGM by recovering the support of $\mathbf{\Sigma}_\varepsilon$, the covariance matrix of $(\varepsilon_1,\dots,\varepsilon_p)'$.

2.1. *Test statistics for H_{0ij}.* In this subsection we introduce new test statistics for H_{0ij}. Let $\mathbf{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)'$, where $\boldsymbol{X}_k = (X_{k1}, \ldots, X_{kp})'$, $1 \le k \le n$, are independent and identically distributed random samples from $\boldsymbol{X}$. By (2), we can write

$$X_{ki} = \alpha_i + \mathbf{X}_{k,-i}\boldsymbol{\beta}_i + \varepsilon_{ki}, \qquad 1 \le k \le n,$$

where $\mathbf{X}_{k,-i}$ is the kth row of $\mathbf{X}$ with its ith entry being removed and ε_{ki} is independent with $\mathbf{X}_{k,-i}$. Let $\hat{\boldsymbol{\beta}}_i = (\hat{\beta}_{1,i}, \ldots, \hat{\beta}_{p-1,i})'$ be any estimators of $\boldsymbol{\beta}_i$ satisfying

$$\max_{1\le i\le p} |\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i|_1 = O_{\mathsf{P}}(a_{n1}) \tag{3}$$

and

$$\min\Big\{\lambda_{\max}^{1/2}(\boldsymbol{\Sigma}) \max_{1\le i\le p} |\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i|_2, \max_{1\le i\le p} \sqrt{(\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i)'\hat{\boldsymbol{\Sigma}}_{-i,-i}(\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i)}\Big\} = O_{\mathsf{P}}(a_{n2}) \tag{4}$$

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

$$\hat{\varepsilon}_{ki} = X_{ki} - \bar{X}_i - (\mathbf{X}_{k,-i} - \bar{\mathbf{X}}_{-i})\hat{\boldsymbol{\beta}}_i$$

and the sample covariance coefficients between the residuals by

$$\hat{r}_{ij} = \frac{1}{n}\sum_{k=1}^n \hat{\varepsilon}_{ki}\hat{\varepsilon}_{kj}, \tag{5}$$

where $\bar{X}_i = \frac{1}{n}\sum_{k=1}^n X_{ki}$ and $\bar{\mathbf{X}}_{-i} = \frac{1}{n}\sum_{k=1}^n \mathbf{X}_{k,-i}$. Our test statistics are based on a bias correction of $\hat{r}_{ij}$. To this end, for $1 \le i < j \le p$, define

$$T_{ij} := \frac{1}{n}\Bigg(\sum_{k=1}^n \hat{\varepsilon}_{ki}\hat{\varepsilon}_{kj} + \sum_{k=1}^n \hat{\varepsilon}_{ki}^2\hat{\beta}_{i,j} + \sum_{k=1}^n \hat{\varepsilon}_{kj}^2\hat{\beta}_{j-1,i}\Bigg). \tag{6}$$

It should be noted that the index is $j-1$ in $\hat{\beta}_{j-1,i}$ and $\hat{\boldsymbol{\beta}}_i$ is a $p-1$ dimensional vector. Let

$$b_{nij} = \omega_{ii}\hat{\sigma}_{ii,\varepsilon} + \omega_{jj}\hat{\sigma}_{jj,\varepsilon} - 1, \tag{7}$$

where $(\hat{\sigma}_{ij,\varepsilon})_{1\le i,j\le p} = \frac{1}{n}\sum_{k=1}^n (\boldsymbol{\varepsilon}_k - \bar{\boldsymbol{\varepsilon}})(\boldsymbol{\varepsilon}_k - \bar{\boldsymbol{\varepsilon}})'$, $\boldsymbol{\varepsilon}_k = (\varepsilon_{k1}, \ldots, \varepsilon_{kp})'$ and $\bar{\boldsymbol{\varepsilon}} = \frac{1}{n}\sum_{k=1}^n \boldsymbol{\varepsilon}_k$. We will prove that

$$T_{ij} = -b_{nij}\frac{\omega_{ij}}{\omega_{ii}\omega_{jj}} + \frac{\sum_{k=1}^n (\varepsilon_{ki}\varepsilon_{kj} - \mathsf{E}\varepsilon_{ki}\varepsilon_{kj})}{n} + O_{\mathsf{P}}\bigg(\lambda_{\max}(\boldsymbol{\Sigma})a_{n2}^2 + a_{n1}\sqrt{\frac{\log p}{n}} + \frac{\log p}{n}\bigg).$$

And under

$$a_{n2} = o(n^{-1/4}) \quad \text{and} \quad a_{n1} = o(1/\sqrt{\log p}), \tag{8}$$

we will prove that

$$\sqrt{\frac{n}{\hat{r}_{ii}\hat{r}_{jj}}}\Big(T_{ij} + b_{nij}\frac{\omega_{ij}}{\omega_{ii}\omega_{jj}}\Big) \Rightarrow N\Big(0, 1 + \frac{\omega_{ij}^2}{\omega_{ii}\omega_{jj}}\Big). \tag{9}$$

Note that, under H_{0ij}, the limiting distribution in (9) does not depend on any unknown parameter. Also, $b_{nij} \to 1$ in probability, uniformly in $1 \le i \le j \le p$. Hence, for the hypothesis test H_{0ij}, we shall use the following test statistic:

$$\hat{T}_{ij} = \sqrt{\frac{n}{\hat{r}_{ii}\hat{r}_{jj}}} T_{ij}. \tag{10}$$

The estimators $\hat{\boldsymbol{\beta}}_i$, $1 \le i \le p$, can be Lasso estimators or Dantzig selectors [Candès and Tao (2007)]. Theoretical results on the convergence rates in (8) have been proved by many papers under various conditions. For example, for the Dantzig selector, it can be proved by (46) and (47) that, under (C1) in Section 3, (8) is satisfied when $\max_{1\le i\le p} |\boldsymbol{\beta}_i|_0 = o(\lambda_{\min}(\boldsymbol{\Sigma})\frac{\sqrt{n}}{\log p})$. The same conclusion holds for the Lasso estimators. The detailed choices of $\hat{\boldsymbol{\beta}}_i$ will be given in Section 4.

REMARK 1. There are a number of recent papers in the regression context where bias correction is used to derive p-values or confidence intervals for the regression coefficients in the high-dimensional case; see Zhang and Zhang (2011), Bühlmann (2013), van de Geer et al. (2013) and Javanmard and Montanari (2013). When applying their methods in GGM estimation, we briefly discuss the difference between our method and theirs. The theoretical results in Javanmard and Montanari (2013) require the standard Gaussian designs or the covariance matrix of the covariates is known. Also, the simulation in Javanmard and Montanari (2013) shows that the method in Bühlmann (2013) is very conservative. These existing methods are computation-intensive and time-consuming. For every i, to get the p-values for the components of $\boldsymbol{\beta}_i$, they need to estimate the $(p-1)\times(p-1)$ precision matrix of $\boldsymbol{X}_{-i}$. So, to derive the p-values for all of the components of $\boldsymbol{\beta}_i$, $1 \le i \le p$, estimators of p precision matrices with dimension $(p-1)\times(p-1)$ are required. Sun and Zhang (2012a) and Ren et al. (2013) also developed a different and interesting residual-based estimator to construct confidence intervals for ω_{ij}. Their method needs to solve two $(p-1)$-dimensional regressions for each entry in $\boldsymbol{\Omega}$, and $(p^2 - p)$ high-dimensional regressions in total for all of the entries in $\boldsymbol{\Omega}$. This requires a huge computational cost. Our method only needs the initial estimators for $\boldsymbol{\beta}_i$. No additional precision matrix estimator or regression coefficient estimator is required. Moreover, the theoretical results in

Zhang and Zhang (2011), Sun and Zhang (2012a) and Ren et al. (2013) require $c_1 \le \lambda_{\min}(\mathbf{\Sigma}) \le \lambda_{\max}(\mathbf{\Sigma}) \le c_2$ for some $c_1, c_2 > 0$. In contrast, our method with the initial estimator Lasso or the Dantzig selector does not need the boundedness condition on $\lambda_{\max}(\mathbf{\Sigma})$ and allows $\lambda_{\min}(\mathbf{\Sigma}) \to 0$; see Propositions 4.1 and 4.2.

It is difficult to give a comprehensive comparison between these existing methods with ours. For example, Sun and Zhang (2012a) and Ren et al. (2013) developed the asymptotic confidence intervals for ω_{ij}. Our method is focused on the testing problem, although with a little more effort, the asymptotic confidence interval result can be easily proved. It is also unknown whether these existing methods can be used to control FDR due to the complicated dependence between the test statistics.

2.2. *GGM estimation with FDR control.* With the new test statistic $\hat{T}_{ij}$, we can carry out $(p^2 - p)/2$ tests (1) simultaneously and control FDR as follows. Let t be the threshold level such that H_{0ij} is rejected if $|\hat{T}_{ij}| \ge t$. The false discovery rate and false discovery proportion are defined by

$$\mathrm{FDP}(t) = \frac{\sum_{(i,j)\in\mathcal{H}_0} I\{|\hat{T}_{ij}| \ge t\}}{\max\{\sum_{1\le j<j\le p} I\{|\hat{T}_{ij}| \ge t\}, 1\}}, \qquad \mathrm{FDR}(t) = \mathsf{E}[\mathrm{FDP}(t)].$$

A "good" threshold level t makes many true alternative hypotheses be rejected and remains that the FDR/FDP be controlled at a pre-specified level $0 < \alpha < 1$. So an ideal choice of t is

$$\hat{t}_o = \inf\left\{0 \le t \le 2\sqrt{\log p} : \frac{\sum_{(i,j)\in\mathcal{H}_0} I\{|\hat{T}_{ij}| \ge t\}}{\max\{\sum_{1\le j<j\le p} I\{|\hat{T}_{ij}| \ge t\}, 1\}} \le \alpha\right\},$$

where $\mathcal{H}_0 = \{(i,j) : \omega_{ij} = 0, 1 \le i < j \le p\}$. In the definition of $\hat{t}_o$, t is restricted to $[0, 2\sqrt{\log p}]$ because $\mathsf{P}(\max_{(i,j)\in\mathcal{H}_0} |\hat{T}_{ij}| \ge 2\sqrt{\log p}) \to 0$ by the proof in Section 6. Since $\mathcal{H}_0$ is unknown, we shall use an estimator of $\sum_{(i,j)\in\mathcal{H}_0} I\{|\hat{T}_{ij}| \ge t\}$. As we will prove in Section 6, an accurate approximation for $\sum_{(i,j)\in\mathcal{H}_0} I\{|\hat{T}_{ij}| \ge t\}$ is $2(1 - \Phi(t))|\mathcal{H}_0|$, where $\Phi(t) = \mathsf{P}(N(0,1) \le t)$. Moreover, $\mathcal{H}_0$ can be estimated by $(p^2 - p)/2$ due to the sparsity of $\mathbf{\Omega}$. This leads to the following procedure:

GFC PROCEDURE. Calculate test statistics $\hat{T}_{ij}$ in (10). Let $0 < \alpha < 1$ and

$$\hat{t} = \inf\left\{0 \le t \le 2\sqrt{\log p} : \frac{G(t)(p^2 - p)/2}{\max\{\sum_{1\le i<j\le p} I\{|\hat{T}_{ij}| \ge t\}, 1\}} \le \alpha\right\}, \tag{11}$$

where $G(t) = 2 - 2\Phi(t)$. If $\hat{t}$ in (11) does not exist, then let $\hat{t} = 2\sqrt{\log p}$. For $1 \le i < j \le p$, we reject H_{0ij} if $|\hat{T}_{ij}| \ge \hat{t}$.

In GFC procedure, the estimators $\hat{\boldsymbol{\beta}}_i$, $1 \le i \le p$, are needed. As mentioned earlier, we can use the Lasso estimators or the Dantzig selectors. Both of them require the choice of tuning parameters. In Section 4 we will propose a method on the choice of tuning parameters, which is particularly suitable for our multiple testing problem.

For general multiple testing problems, Liu and Shao (2012) developed a procedure that controls the false discovery rate. They proposed to threshold test statistics directly rather than the true p-values as in Benjamini and Hochberg (1995), because the true p-values are unknown in practice. Additionally, to control FDR, the Benjamini–Hochberg method requires the independence or some kind of positive regression dependency between p-values. Our test statistics do not meet such conditions. By thresholding the test statistics directly as in Liu and Shao (2012), we shall show that $\mathrm{FDR}(\hat{t}) \to \alpha$ and $\mathrm{FDP}(\hat{t}) \to \alpha$ in probability. It should be pointed out that Liu and Shao (2012) imposed the dependence condition among the test statistics. In GGM estimation, it is more natural to impose the dependence condition on the precision matrix. To this end, we need many novel techniques in the proof.

Meinshausen and Bühlmann (2010) proposed the stability selection for variable selection and Gaussian graphical modeling. They established the bound for the expectation of falsely selected variables. It is unknown whether their procedure can be used to control FDR.

3. Theoretical results. In this section we will show that GFC procedure can control the false discovery rate asymptotically at any pre-specified level.

(C1) Let $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Suppose that $\max_{1\le i\le p} \sigma_{ii} \le c_0$ and $\max_{1\le i\le p} \omega_{ii} \le c_0$ for some constant $c_0 > 0$. Assume that $\log p = o(n)$.

Since $\sigma_{ii}\omega_{ii} \ge 1$, (C1) implies that $\min_{1\le i\le p} \omega_{ii} \ge c_0^{-1}$ and $\min_{1\le i\le p} \sigma_{ii} \ge c_0^{-1}$. We give the asymptotic distribution of $\hat{T}_{ij}$, which is useful in testing a single $H_{0ij} : \omega_{ij} = 0$.

PROPOSITION 3.1. *Suppose that* (C1) *holds. Let* $\hat{\boldsymbol{\beta}}_i$ *be any estimator satisfying* (3), (4) *and* (8). *Then, we have*

$$\sqrt{\frac{n}{\hat{r}_{ii}\hat{r}_{jj}}}\left(T_{ij} + b_{nij}\frac{\omega_{ij}}{\omega_{ii}\omega_{jj}}\right) \Rightarrow N\left(0, 1 + \frac{\omega_{ij}^2}{\omega_{ii}\omega_{jj}}\right)$$

as $(n, p) \to \infty$, *where the convergence in distribution is uniformly in* $1 \le i < j \le p$.

Let the false discovery proportion and false discovery rate of GFC be defined by

$$\mathrm{FDP} = \frac{\sum_{(i,j)\in\mathcal{H}_0} I\{|\hat{T}_{ij}| \ge \hat{t}\}}{\max(\sum_{1\le i<j\le p} I\{|\hat{T}_{ij}| \ge \hat{t}\}, 1)}, \qquad \mathrm{FDR} = \mathsf{E}(\mathrm{FDP}).$$

Recall that $\mathcal{H}_0 = \{(i, j): \omega_{ij} = 0, 1 \le i < j \le p\}$. Let $q_0 = \mathrm{Card}(\mathcal{H}_0)$ be the cardinality of $\mathcal{H}_0$ and $q = (p^2 - p)/2$. For a constant $\gamma > 0$ and $1 \le i \le p$, define

$$\mathcal{A}_i(\gamma) = \{j : 1 \le j \le p, j \ne i, |\omega_{ij}| \ge (\log p)^{-2-\gamma}\}.$$

Theorem 3.1 shows that GFC controls FDP and FDR at level α asymptotically.

THEOREM 3.1. *Let $p \le n^r$ for some $r > 0$. Suppose that*

$$\mathrm{Card}\left\{(i, j): 1 \le i < j \le p, \frac{|\omega_{ij}|}{\sqrt{\omega_{ii}\omega_{jj}}} \ge 4\sqrt{\log p/n}\right\} \ge \sqrt{\log\log p}. \tag{12}$$

Assume that $q_0 \ge cp^2$ for some $c > 0$ and $\hat{\boldsymbol{\beta}}_i$ satisfies (3), (4) *and*

$$a_{n1} = o(1/\log p) \quad \text{and} \quad a_{n2} = o\big((n \log p)^{-1/4}\big). \tag{13}$$

Under (C1) *and* $\max_{1\le i\le p} \mathrm{Card}(\mathcal{A}_i(\gamma)) = O(p^\rho)$ *for some $\rho < 1/2$ and $\gamma > 0$, we have*

$$\lim_{(n,p)\to\infty} \frac{\mathrm{FDR}}{\alpha q_0/q} = 1 \quad \text{and} \quad \frac{\mathrm{FDP}}{\alpha q_0/q} \to 1 \qquad \text{in probability}$$

as $(n, p) \to \infty$.

The dimension p can be much larger than the sample size because r can be arbitrarily large. Note that $q_0 \ge cp^2$ is a natural condition. If $q_0 = o(p^2)$, then almost all of ω_{ij} are nonzero. Hence, rejecting all the hypothesis tests leads to FDR $\to 0$. The condition $\max_{1\le i\le p} \mathrm{Card}(\mathcal{A}_i(\gamma)) = O(p^\rho)$ is also mild. For example, if $p \ge n^\delta$ for some $\delta > 1$ and $\boldsymbol{\Omega}$ is a $s_{n,p}$-sparse matrix with $s_{n,p} = O(\sqrt{n})$ (i.e., the number of nonzero entries in each row is no more than $s_{n,p}$), then this condition holds. The sparsity $s_{n,p} = O(\sqrt{n})$ is often imposed in the literature on precision matrix estimation.

The technical condition (12) is used to ensure $|\mathcal{H}_0|G(\hat{t}) \to \infty$ which is almost necessary for

$$\sum_{(i,j)\in\mathcal{H}_0} I\{|\hat{T}_{ij}| \ge \hat{t}\}/\big(|\mathcal{H}_0|G(\hat{t})\big) \to 1 \tag{14}$$

in probability. We believe (14) is nearly necessary for the false discovery proportion $\frac{\mathrm{FDP}}{\alpha q_0/q} \to 1$ in probability. On the other hand, the condition for controlling FDR may be weaker than that for controlling FDP. Even if (14) is violated, the false discovery rate may still be controlled at level α. Hence, it is possible that (12) is not needed for FDR results. In addition, (12) is not strong because the total number of hypothesis tests is $(p^2 - p)/2$ and we only require a few standardized off-diagonal entries of $\boldsymbol{\Omega}$ having magnitudes exceeding $4\sqrt{\log p/n}$.

The condition (13) is stronger than (8). In large scale multiple tests, the result on convergence in distribution [i.e., $\mathsf{P}(|\hat{T}_{ij}| \ge t) \to 2 - 2\Phi(t)$] is not enough to ensure the accuracy. Because the threshold level $\hat{t}$ typically tends to infinity, we often need the Cramér type moderate deviation result such as

$$\max_{(i,j)\in\mathcal{H}_0}\sup_{0\le t\le 2\sqrt{\log p}}\left|\frac{\mathsf{P}(|\hat{T}_{ij}|\ge t)}{2-2\Phi(t)}-1\right|\to 0,$$

which requires a stronger condition (13).

4. Data-driven choice of $\hat{\boldsymbol{\beta}}_i$. GFC requires to choose the estimators of $\boldsymbol{\beta}_i$. There is much literature on the estimation of high-dimensional regression coefficients. In this paper we use the popular Dantzig selector and Lasso estimator. Some other recent procedures such as the scaled-Lasso [Sun and Zhang (2012b)] and the Square-root Lasso [Belloni, Chernozhukov and Wang (2011)] can also be used and similar theoretical results such as Propositions 4.1 and 4.2 can be established.

Dantzig selector for $\hat{\boldsymbol{\beta}}_i$. The Dantzig selector estimates $\boldsymbol{\beta}_i$ by solving the following optimization problems:

$$\hat{\boldsymbol{\beta}}_i(\delta)=\arg\min\{|\boldsymbol{\omega}|_1 \text{ subject to } |\mathbf{D}_i^{-1/2}\hat{\boldsymbol{\Sigma}}_{-i,-i}\boldsymbol{\omega}-\mathbf{D}_i^{-1/2}\hat{\mathbf{a}}|_\infty\le\lambda_{ni1}(\delta)\}\tag{15}$$

for $1\le i\le p$, where $\mathbf{D}_i=\mathrm{diag}(\hat{\boldsymbol{\Sigma}}_{-i,-i})$, $\hat{\mathbf{a}}=\frac{1}{n}\sum_{k=1}^n(\mathbf{X}_{k,-i}-\mathbf{X}_{-i})'(X_{k,i}-\bar{X}_i)$ and

$$\lambda_{ni1}(\delta)=\delta\sqrt{\frac{\hat{\sigma}_{ii,X}\log p}{n}}$$

for $\delta>0$, where $\hat{\sigma}_{ii,X}=\frac{1}{n}\sum_{k=1}^n(X_{ki}-\bar{X}_i)^2$. Note that $\mathsf{Var}(\varepsilon_i)=\omega_{ii}^{-1}$. Hence, the choice of $\lambda_{ni1}(\delta)$ in the original version of the Dantzig selector should be $\delta\sqrt{\frac{\omega_{ii}^{-1}\log p}{n}}$. As ω_{ii}^{-1} is unknown, $\hat{\sigma}_{ii,X}$ is used in place of ω_{ii}^{-1} by the inequality $\sigma_{ii}\omega_{ii}\ge 1$. We can let $\delta=2$, which is fully specified and has theoretical interest. For finite sample sizes, we will propose a more useful data-driven choice for δ in (19).

PROPOSITION 4.1. *Suppose that* (C1) *holds and* $\max_{1\le i\le p}|\boldsymbol{\beta}_i|_0=o(\lambda_{\min}(\boldsymbol{\Sigma})\frac{\sqrt{n}}{(\log p)^{3/2}})$. *For* $\delta=2$ *in* (15), $\hat{\boldsymbol{\beta}}_i(2)$, $1\le i\le p$, *satisfy* (3), (4) *and* (13).

Lasso estimator for $\hat{\boldsymbol{\beta}}_i$. The coefficients $\boldsymbol{\beta}_i$ can be estimated by the Lasso as follows:

$$\hat{\boldsymbol{\beta}}_i(\delta)=\mathbf{D}_i^{-1/2}\hat{\boldsymbol{\alpha}}_i(\delta),\tag{16}$$

where

$$\hat{\boldsymbol{\alpha}}_i(\delta)=\mathop{\arg\min}_{\boldsymbol{\alpha}\in\mathbf{R}^{p-1}}\left\{\frac{1}{2n}\sum_{k=1}^n(X_{ki}-\bar{X}_i-(\mathbf{X}_{k,-i}-\bar{\mathbf{X}}_{-i})\mathbf{D}_i^{-1/2}\boldsymbol{\alpha})^2+\lambda_{ni1}(\delta)|\boldsymbol{\alpha}|_1\right\}.$$

The following proposition shows that for any $\delta > 2$, (13) is satisfied. The data-driven choice for δ is given in (19).

PROPOSITION 4.2. *Suppose that* (C1) *holds and* $\max_{1\le i\le p}|\boldsymbol{\beta}_i|_0 = o(\lambda_{\min}(\boldsymbol{\Sigma})\frac{\sqrt{n}}{(\log p)^{3/2}})$. *For any* $\delta > 2$ *in* (16), $\hat{\boldsymbol{\beta}}_i(\delta)$, $1 \le i \le p$, *satisfy* (3), (4) *and* (13).

Data-driven choice of δ. As in many regularization approaches, the choice $\delta \ge 2$ is often large. Hence, in this paper, we propose to select δ adaptively by data. We let $\hat{\boldsymbol{\beta}}_i(\delta)$ be the solution to (15) or (16) and then obtain the statistics $\hat{T}_{ij}(\delta)$, $1 \le i < j \le p$. As noted in Section 2.2, GFC works because for good estimators $\hat{\boldsymbol{\beta}}_i(\delta)$, $1 \le i \le p$, $\sum_{(i,j)\in\mathcal{H}_0} I\{|\hat{T}_{ij}(\delta)| \ge t\}$ will be close to $|\mathcal{H}_0|G(t)$. Hence, an oracle choice of δ can be

$$\hat{\delta}_o = \mathop{\arg\min}_{0\le\delta\le2} \int_{\tau_p}^{1}\left(\frac{\sum_{(i,j)\in\mathcal{H}_0} I\{|\hat{T}_{ij}(\delta)| \ge \Phi^{-1}(1-\alpha/2)\}}{\alpha|\mathcal{H}_0|} - 1\right)^2 d\alpha, \tag{17}$$

where $\tau_p = G(2\sqrt{\log p})$. $\mathcal{H}_0$ is unknown, however. Since $\boldsymbol{\Omega}$ is sparse, $|\mathcal{H}_0|$ is close to $(p^2-p)/2$. So a good choice of δ should minimize the following error:

$$\int_{\tau}^{1}\left(\frac{\sum_{1\le i\ne j\le p} I\{|\hat{T}_{ij}(\delta)| \ge \Phi^{-1}(1-\alpha/2)\}}{\alpha(p^2-p)} - 1\right)^2 d\alpha, \tag{18}$$

where $\tau > 0$ is a fixed number bounded away from zero. The constraint $\alpha \ge \tau$ aims to ensure the nonzero entries part $\sum_{(i,j)\in\mathcal{H}_1} I\{|\hat{T}_{ij}(\delta)| \ge \Phi^{-1}(1-\frac{\alpha}{2})\} = o(\alpha(p^2-p))$. In our choice, we let $\tau = 0.3$. This leads to the final choice of δ by discretizing the integral as follows:

$$\begin{gathered}\hat{\delta} = \hat{j}/N,\\ \hat{j} = \mathop{\arg\min}_{0\le j\le 2N}\sum_{k=3}^{9}\left(\frac{\sum_{1\le i\ne j\le p} I\{|\hat{T}_{ij}(j/N)| \ge \Phi^{-1}(1-k/20)\}}{k(p^2-p)/10} - 1\right)^2,\end{gathered} \tag{19}$$

where N is an integer number that can be pre-specified. Finally, we use $\hat{\boldsymbol{\beta}}_i(\hat{\boldsymbol{\delta}})$ as the estimator of $\boldsymbol{\beta}_i$. Deriving theoretical properties for $\hat{\delta}$ is important. We leave this as a future work.

5. Numerical results. In this section we carry out simulations to examine the performance of GFC by the following graphs:

- *Band graph*. $\boldsymbol{\Omega} = (\omega_{ij})$, where $\omega_{i,i+1} = \omega_{i+1,i} = 0.6$, $\omega_{i,i+2} = \omega_{i+2,i} = 0.3$, $\omega_{ij} = 0$ for $|i-j| \ge 3$. $\boldsymbol{\Omega}$ is a 5-sparse matrix.
- *Hub graph*. There are $p/10$ rows with sparsity 11. The rest of the rows have sparsity 2. To this end, we let $\boldsymbol{\Omega}_1 = (\omega_{ij})$, $\omega_{ij} = \omega_{ji} = 0.5$ for $i = 10(k-1)+1$

TABLE 1
Empirical false discovery rates

	$\alpha = 0.1$				$\alpha = 0.2$			
p	**50**	**100**	**200**	**400**	**50**	**100**	**200**	**400**
				GFC-Dantzig				
Band	0.0899	0.1085	0.1160	0.1168	0.1738	0.1991	0.2103	0.2035
Hub	0.0722	0.0599	0.0557	0.0459	0.1651	0.1415	0.1369	0.1154
E–R	0.1174	0.0887	0.0747	0.0892	0.2099	0.1738	0.1516	0.1703
				GFC-Lasso				
Band	0.0849	0.0768	0.0801	0.0842	0.1759	0.1650	0.1707	0.1718
Hub	0.0917	0.0835	0.0766	0.0708	0.1937	0.1852	0.1693	0.1560
E–R	0.1038	0.0967	0.1011	0.1180	0.2149	0.1963	0.2083	0.2297

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. Finally, we let $\mathbf{\Omega} = \mathbf{\Omega}_1 + (|\min(\lambda_{\min})| + 0.05)\mathbf{I}_p$ to make the matrix be positive definite.

- *Erdős–Rényi random graph*. There is an edge between each pair of nodes with probability $\min(0.05, 5/p)$ independently. Let $\omega_{ij} = u_{ij} * \delta_{ij}$, where $u_{ij} \sim U(0.4, 0.8)$ is the uniform random variable and δ_{ij} is the Bernoulli random variable with success probability $\min(0.05, 5/p)$. u_{ij} and δ_{ij} are independent. Finally, we let $\mathbf{\Omega} = \mathbf{\Omega}_1 + (|\min(\lambda_{\min})| + 0.05)\mathbf{I}_p$ such that the matrix is positive definite.

For each model, we generate $n = 100$ random samples with $\boldsymbol{X}_k \sim N(\boldsymbol{\mu}, \mathbf{\Sigma})$, $\mathbf{\Sigma} = \mathbf{\Omega}^{-1}$ and $p = 50, 100, 200, 400$. We use the Dantzig selector and Lasso to estimate $\boldsymbol{\beta}_i$ in GFC and denote the corresponding procedures by GFC-Dantzig and GFC-Lasso. The tuning parameter $\lambda_{ni1}(\hat{\delta})$ is given in Section 4 with $N = 20$. The simulation results are based on 100 replications. As we can see from Table 1, the FDRs of the GFC-Dantzig for Band graph and Erdős–Rényi (E–R) random graph are close to α. The FDRs for Hub graph are somewhat smaller than α. For all three graphs, the FDRs can be effectively controlled below the level α. Similarly, GFC-Lasso can control FDR at the level α. The FDPs of GFC-Dantzig in 100 replications are plotted in Figure 1 with $p = 200$. For the reason of space, we give the other figures for $p = 50, 100, 400$ and GFC-Lasso in the supplemental material [Liu (2013)]. We can see from these figures that most of the FDPs are concentrated around the FDRs.

In Figure 2 we plot the FDPs for all GFC-Dantizg estimators with $p = 200$, $\alpha = 0.2$ and $\hat{\boldsymbol{\beta}}_i(j/20)$, $1 \le j \le 40$. The histograms of $\hat{j}$ are plotted in Figure 3. We use $\widehat{\mathrm{FDR}}(j)$ to denote the false discovery rates for GFC-Dantzig with $\hat{\boldsymbol{\beta}}(j/20)$. As we can see from Figure 2, there always exist several j such that $\widehat{\mathrm{FDR}}(j)$ are well controlled at level $\alpha = 0.2$. From the histograms of $\hat{j}$ in Figure 3, we see that

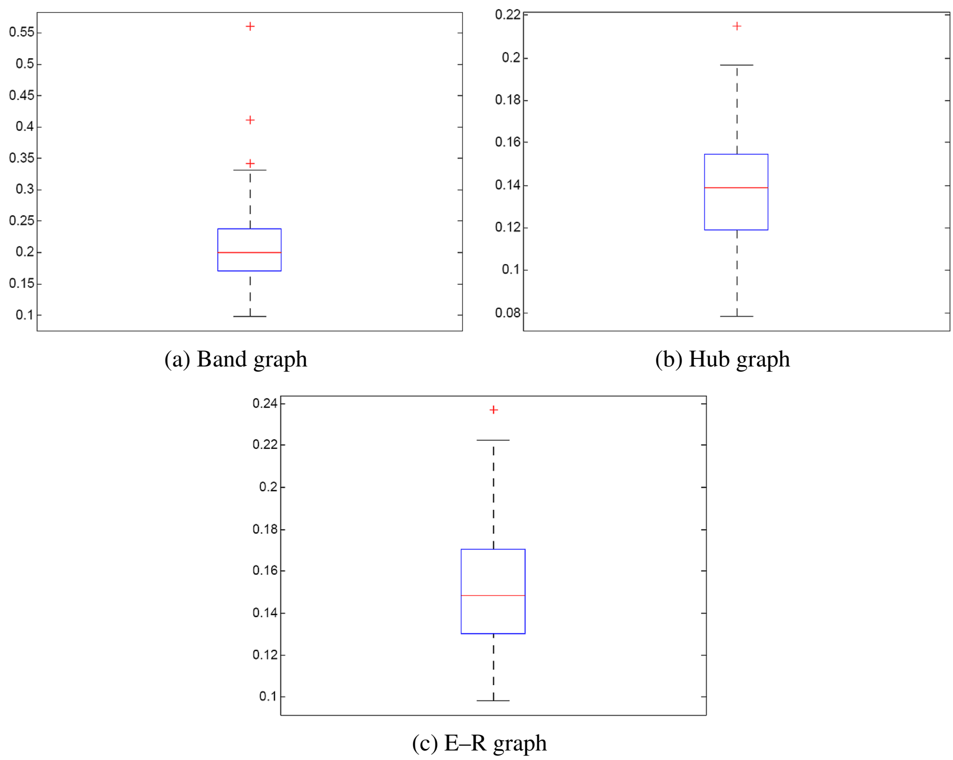


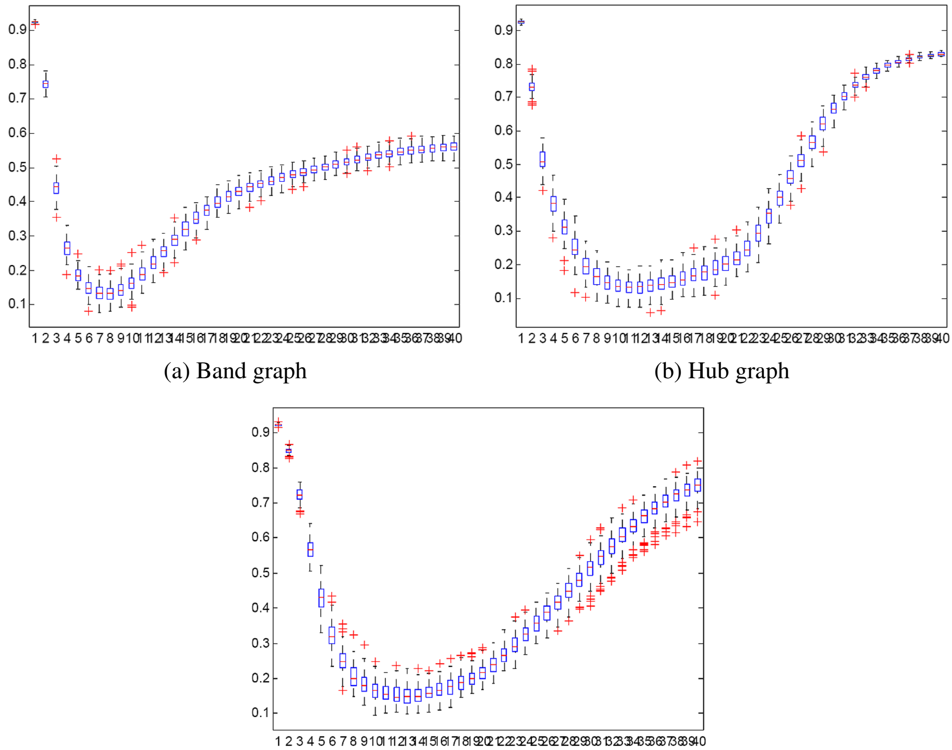
FIG. 1. *FDP* (*GFC-Dantzig*, $p = 200$ *and* $\alpha = 0.2$).

$\hat{j}$ in Section 4 can always take the values of these j's for all three graphs. Similar phenomenon can be observed in GFC-Lasso; see the supplemental material [Liu (2013)]. We can also see from Figure 2 that, when $\delta = 2$ (i.e., $j = 40$), the FDPs are much higher than α. This indicates that the choice of $\delta = 2$ is too big to have a good performance when the sample size is small.

We examine the power of GFC on controlling FDR. Based on 100 replications, the average powers are defined by

$$\text{Average}\left\{\frac{\sum_{(i,j)\in\mathcal{H}_1} I\{|\hat{T}_{ij}| \geq \hat{t}\}}{\text{Card}(\mathcal{H}_1)}\right\}.$$

We state the numerical results in Table 2. The power increases when α increases. For the Hub graph, the powers are close to one. For the Band graph, GFC-Dantzig can also effectively detect the edges and GFC-Lasso is more powerful than GFC-Dantzig. For the Erdős–Rényi random graph, GFC has nontrivial powers when $p = 50$, 100 and 200. The powers are low when $p = 400$. This is mainly due to the very small magnitude of ω_{ij}. Actually, all of $\frac{\omega_{ij}}{\sqrt{\omega_{ii}\omega_{jj}}}$ belong to the interval $(0.1275, 0.255)$ when $p = 400$. So it is very difficult to detect such small nonzero entries.



(a) Band graph (b) Hub graph

(c) E–R graph

FIG. 2. *FDP for* $j = 1, \ldots, 40$ (*GFC-Dantzig*, $p = 200$ *and* $\alpha = 0.2$).

Finally, we compare GFC with the Graphical Lasso (Glasso) which estimates the graph by solving the following optimization problem:

$$\hat{\mathbf{\Omega}}(\lambda_n) := \arg\min_{\mathbf{\Omega} \succ 0} \{\langle \mathbf{\Omega}, \hat{\mathbf{\Sigma}}_n \rangle - \log\det(\mathbf{\Omega}) + \lambda_n \|\mathbf{\Omega}\|_1\}.$$

As in Rothman et al. (2008), Fan, Feng and Wu (2009) and Cai, Liu and Luo (2011), the tuning parameter λ_n is selected by the popular cross-validation method. To this end, we generate another $n = 100$ training samples from $\mathbf{X}$ and let $\hat{\mathbf{\Sigma}}_{\text{train}}$ be the sample covariance matrix from the training samples. We choose the following tuning parameter:

$$\lambda_n = \hat{k}/50, \qquad \hat{k} = \arg\min_{1 \le k \le 200} \{\langle \hat{\mathbf{\Omega}}(k/50), \hat{\mathbf{\Sigma}}_{\text{train}} \rangle - \log\det(\hat{\mathbf{\Omega}}(k/50))\}.$$

The empirical false discovery rates and the standard deviations are stated in Table 3. We can see that for all three graphs the FDRs of Glasso are quite close to 1. This indicates that Glasso with the cross-validation method fails to control the false discovery rate. We next examine the power of Glasso. Since the power of Glasso depends on the choice of λ_n, we plot all of the FDRs and the average powers for $\hat{\mathbf{\Omega}}(\lambda_n)$ with $\lambda_n = \frac{1}{50}, \frac{2}{50}, \ldots, \frac{200}{50}$ in Figure 4 with $p = 200$. Other figures

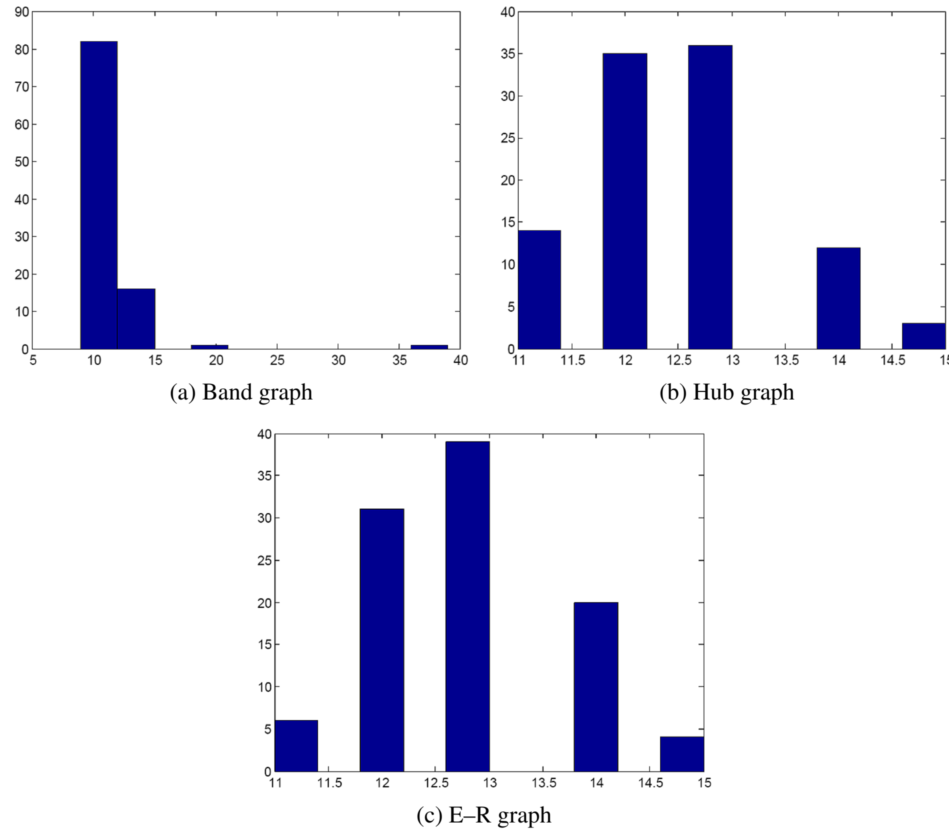


FIG. 3. *Histogram for* $\hat{j}$ (*GFC-Dantzig*, $p = 200$ *and* $\alpha = 0.2$).

for $p = 50, 100, 400$ are given in the supplemental material [Liu (2013)]. As we can see from these figures, for the Band graph and ER graph, the powers are quite low (≤ 0.05) if the FDRs ≤ 0.2. Hence, for these two graphs, GFC significantly outperforms Glasso even if we know the oracle choice of the tuning parameter for Glasso. It is also interesting to see that, for the Hub graph, the power of Glasso is close to one even when the FDRs are small. This phenomenon is similar to that of GFC, which also performs quite well for the Hub graph.

6. Proof.

6.1. *Proof of Proposition* 3.1. Put $\tilde{\varepsilon}_{ki} = \varepsilon_{ki} - \bar{\varepsilon}_i$, where $(\bar{\varepsilon}_1, \dots, \bar{\varepsilon}_p)' = \bar{\boldsymbol{\varepsilon}}$. Recall the definitions of $\mathbf{X}_{k,-j}$ and $\bar{\mathbf{X}}_{-j}$ in Section 2.1. Note that

$$\begin{aligned}\hat{\varepsilon}_{ki}\hat{\varepsilon}_{kj} = {} & \tilde{\varepsilon}_{ki}\tilde{\varepsilon}_{kj} - \tilde{\varepsilon}_{ki}(\mathbf{X}_{k,-j} - \bar{\mathbf{X}}_{-j})(\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j) \\ & - \tilde{\varepsilon}_{kj}(\mathbf{X}_{k,-i} - \bar{\mathbf{X}}_{-i})(\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i) \\ & + (\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i)'(\mathbf{X}_{k,-i} - \bar{\mathbf{X}}_{-i})'(\mathbf{X}_{k,-j} - \bar{\mathbf{X}}_{-j})(\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j).\end{aligned} \tag{20}$$

TABLE 2
Power of GFC (SD)

	$\alpha = 0.1$				$\alpha = 0.2$			
p	**50**	**100**	**200**	**400**	**50**	**100**	**200**	**400**
				GFC-Dantzig				
Band	0.7934 (0.0447)	0.7182 (0.0368)	0.6688 (0.0255)	0.6265 (0.0151)	0.8547 (0.0430)	0.7937 (0.0409)	0.7399 (0.0283)	0.6865 (0.0157)
Hub	0.9607 (0.0503)	0.9767 (0.0208)	0.9776 (0.0140)	0.9778 (0.0087)	0.9767 (0.0384)	0.9877 (0.0139)	0.9873 (0.0096)	0.9868 (0.0074)
E–R	0.7319 (0.0652)	0.3596 (0.0445)	0.2623 (0.0249)	0.1416 (0.0140)	0.7943 (0.0551)	0.4693 (0.0448)	0.3505 (0.0240)	0.2051 (0.0177)
				GFC-Lasso				
Band	0.8814 (0.0365)	0.8489 (0.0244)	0.8027 (0.0215)	0.7491 (0.0149)	0.9227 (0.0306)	0.8939 (0.0234)	0.8490 (0.0172)	0.7955 (0.0155)
Hub	0.9224 (0.0647)	0.9202 (0.0389)	0.9202 (0.0323)	0.9327 (0.0181)	0.9553 (0.0456)	0.9531 (0.0308)	0.9513 (0.0218)	0.9570 (0.0132)
E–R	0.7629 (0.0561)	0.4178 (0.0429)	0.3014 (0.0266)	0.1596 (0.0149)	0.8265 (0.0550)	0.5294 (0.0412)	0.4063 (0.0258)	0.2390 (0.0168)

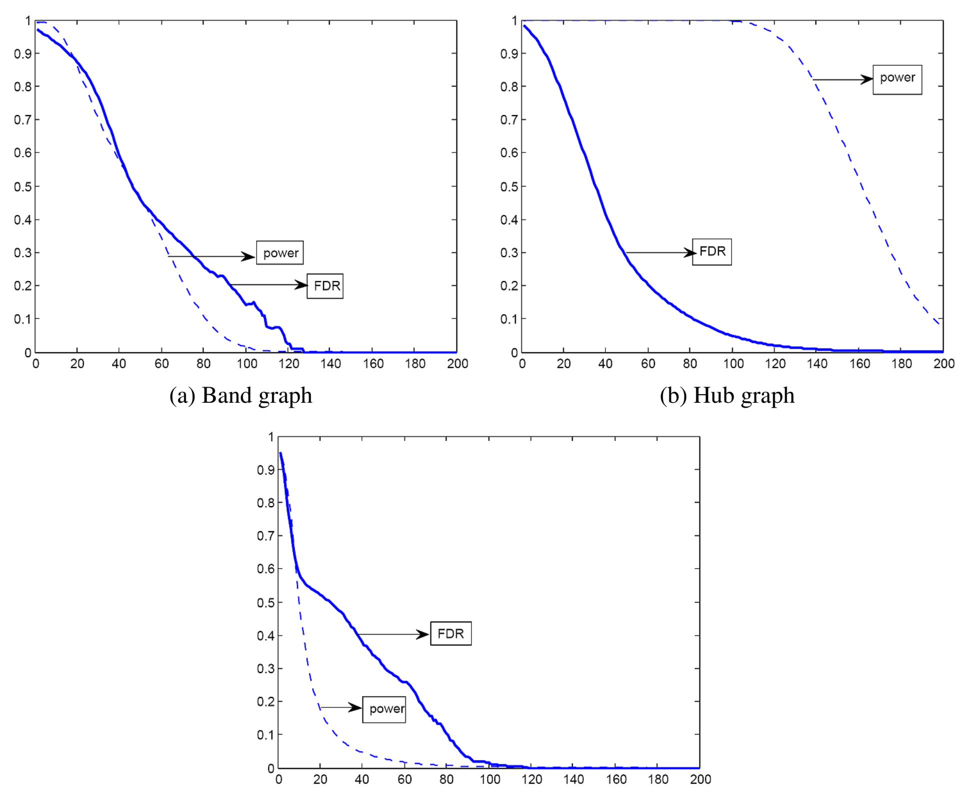
TABLE 3

Empirical false discovery rates (SD) for Glasso

p	50	100	200	400
Band	0.8449 (0.0073)	0.8887 (0.0035)	0.9156 (0.0022)	0.9354 (0.0020)
Hub	0.8622 (0.0101)	0.9074 (0.0055)	0.9333 (0.0013)	0.9509 (0.0010)
E–R	0.8513 (0.0154)	0.8257 (0.0042)	0.8564 (0.0253)	0.8692 (0.0024)

For the last term in (20), we have

$$\begin{aligned}\left|(\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i)' \hat{\boldsymbol{\Sigma}}_{-i,-j} (\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j)\right| \le{}& \left|(\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i)' (\hat{\boldsymbol{\Sigma}}_{-i,-j} - \boldsymbol{\Sigma}_{-i,-j}) (\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j)\right| \\ &+ \left|(\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i)' \boldsymbol{\Sigma}_{-i,-j} (\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j)\right|.\end{aligned}$$



(a) Band graph (b) Hub graph

(c) E–R graph

FIG. 4. *FDR curve and power curve for graphical lasso* ($p = 200$).

It follows from Lemma 1 in Cai and Liu (2011) that, for any $M > 0$, there exists $C > 0$ such that

$$\mathsf{P}\Big(\max_{1\le i<j\le p}|\hat{\sigma}_{ij}-\sigma_{ij}|\ge C\sqrt{\log p/n}\Big)=O\big(p^{-M}\big). \tag{21}$$

Hence,

$$\max_{i,j}\big|(\hat{\boldsymbol{\beta}}_i-\boldsymbol{\beta}_i)'(\hat{\boldsymbol{\Sigma}}_{-i,-j}-\boldsymbol{\Sigma}_{-i,-j})(\hat{\boldsymbol{\beta}}_j-\boldsymbol{\beta}_j)\big|=O_{\mathsf{P}}\big(a_{n1}^2(\log p/n)^{1/2}\big).$$

Moreover,

$$\big|(\hat{\boldsymbol{\beta}}_i-\boldsymbol{\beta}_i)'\boldsymbol{\Sigma}_{-i,-j}(\hat{\boldsymbol{\beta}}_j-\boldsymbol{\beta}_j)\big|=O_{\mathsf{P}}\big(\lambda_{\max}(\boldsymbol{\Sigma})|\hat{\boldsymbol{\beta}}_i-\boldsymbol{\beta}_i|_2^2\big)$$

uniformly in $1\le i\le j\le p$. By the Cauchy–Schwarz inequality, we have

$$\begin{aligned}&\left|\frac{1}{n}\sum_{k=1}^n(\hat{\boldsymbol{\beta}}_i-\boldsymbol{\beta}_i)'(\mathbf{X}_{k,-i}-\bar{\mathbf{X}}_{-i})'(\mathbf{X}_{k,-j}-\bar{\mathbf{X}}_{-j})(\hat{\boldsymbol{\beta}}_j-\boldsymbol{\beta}_j)\right|\\&\quad\le\max_{1\le i\le p}(\hat{\boldsymbol{\beta}}_i-\boldsymbol{\beta}_i)'\hat{\boldsymbol{\Sigma}}_{-i,-i}(\hat{\boldsymbol{\beta}}_i-\boldsymbol{\beta}_i).\end{aligned}$$

Combining the above arguments,

$$\begin{aligned}&\left|\frac{1}{n}\sum_{k=1}^n(\hat{\boldsymbol{\beta}}_i-\boldsymbol{\beta}_i)'(\mathbf{X}_{k,-i}-\bar{\mathbf{X}}_{-i})'(\mathbf{X}_{k,-j}-\bar{\mathbf{X}}_{-j})(\hat{\boldsymbol{\beta}}_j-\boldsymbol{\beta}_j)\right|\\&\quad=O_{\mathsf{P}}\big(a_{n2}^2+a_{n1}^2(\log p/n)^{1/2}\big).\end{aligned}$$

We now estimate the second term on the right-hand side of (20). For $1\le i\le j\le p$, write

$$\begin{aligned}\tilde{\varepsilon}_{ki}(\mathbf{X}_{k,-j}-\bar{\mathbf{X}}_{-j})(\hat{\boldsymbol{\beta}}_j-\boldsymbol{\beta}_j)&=\tilde{\varepsilon}_{ki}(X_{ki}-\bar{X}_i)(\hat{\beta}_{i,j}-\beta_{i,j})I\{i\ne j\}\\&\quad+\sum_{l\ne i,j}\tilde{\varepsilon}_{ki}(X_{kl}-\bar{X}_l)(\hat{\beta}_{l,j}-\beta_{l,j}),\end{aligned}$$

where $\hat{\boldsymbol{\beta}}_j=(\hat{\beta}_{1,j},\dots,\hat{\beta}_{p-1,j})'$ and we set $\hat{\beta}_{p,j}=0$. Recall that ε_{ki} is independent with $\mathbf{X}_{k,-j}$. Then it can be proved that, for any $M>0$, there exists $C>0$ such that

$$\mathsf{P}\left(\max_{1\le i\le p}\max_{1\le l\le p,l\ne i}\left|\frac{1}{n}\sum_{k=1}^n\tilde{\varepsilon}_{ki}(X_{kl}-\bar{X}_l)\right|\ge C\sqrt{\frac{\log p}{n}}\right)=O\big(p^{-M}\big).$$

This implies that

$$\max_{1\le i\le j\le p}\left|\sum_{l\ne i,j}\frac{1}{n}\sum_{k=1}^n\tilde{\varepsilon}_{ki}(X_{kl}-\bar{X}_l)(\hat{\beta}_{l,j}-\beta_{l,j})\right|=O_{\mathsf{P}}\big(a_{n1}\sqrt{\log p/n}\big).$$

A similar inequality holds for the third term on the right-hand side of (20). Therefore,

$$\begin{aligned}\frac{1}{n}\sum_{k=1}^{n}\hat{\varepsilon}_{ki}\hat{\varepsilon}_{kj} &= \frac{1}{n}\sum_{k=1}^{n}\tilde{\varepsilon}_{ki}\tilde{\varepsilon}_{kj} - \frac{1}{n}\sum_{k=1}^{n}\tilde{\varepsilon}_{ki}(X_{ki}-\bar{X}_i)(\hat{\beta}_{i,j}-\beta_{i,j})I\{i\neq j\} \\ &\quad - \frac{1}{n}\sum_{k=1}^{n}\tilde{\varepsilon}_{kj}(X_{kj}-\bar{X}_j)(\hat{\beta}_{j-1,i}-\beta_{j-1,i})I\{i\neq j\} \qquad (22)\\ &\quad + O_{\mathsf{P}}\big((a_{n1}^2+a_{n1})\sqrt{\log p/n}+a_{n2}^2\big)\end{aligned}$$

uniformly in $1\le i\le j\le p$. By (2), we have

$$\frac{1}{n}\sum_{k=1}^{n}\tilde{\varepsilon}_{ki}(X_{ki}-\bar{X}_i) = \frac{1}{n}\sum_{k=1}^{n}\tilde{\varepsilon}_{ki}^2 + \frac{1}{n}\sum_{k=1}^{n}\tilde{\varepsilon}_{ki}(\mathbf{X}_{k,-i}-\bar{\mathbf{X}}_{-i})\boldsymbol{\beta}_i. \qquad (23)$$

By (C1), we have $\mathsf{Var}(\mathbf{X}_{k,-i}\boldsymbol{\beta}_i)=(\sigma_{ii}\omega_{ii}-1)/\omega_{ii}\le C$. It follows that

$$\mathsf{P}\left(\max_{1\le i\le p}\left|\frac{1}{n}\sum_{k=1}^{n}\tilde{\varepsilon}_{ki}(\mathbf{X}_{k,-i}-\bar{\mathbf{X}}_{-i})\boldsymbol{\beta}_i\right|\ge C\sqrt{\frac{\log p}{n}}\right)=O(p^{-M}).$$

By (22) and (23), we have, uniformly in $1\le i\le p$,

$$\begin{aligned}\frac{1}{n}\sum_{k=1}^{n}\tilde{\varepsilon}_{ki}(X_{ki}-\bar{X}_i) &= \frac{1}{n}\sum_{k=1}^{n}\tilde{\varepsilon}_{ki}^2 + O_{\mathsf{P}}(\sqrt{\log p/n}) \\ &= \frac{1}{n}\sum_{k=1}^{n}\hat{\varepsilon}_{ki}^2 + O_{\mathsf{P}}(\sqrt{\log p/n}), \qquad (24)\\ &\quad + O_{\mathsf{P}}\big((a_{n1}^2+a_{n1})\sqrt{\log p/n}+a_{n2}^2\big),\end{aligned}$$

where the last equation follows from (22) with $i=j$. So, by (22), (24) and $\max_{i,j}|\hat{\beta}_{i,j}-\beta_{i,j}|=O_{\mathsf{P}}(a_{n1})=o_{\mathsf{P}}(1)$, for $1<i<j\le p$,

$$\begin{aligned}\frac{1}{n}\sum_{k=1}^{n}\hat{\varepsilon}_{ki}\hat{\varepsilon}_{kj} &= \frac{1}{n}\sum_{k=1}^{n}\tilde{\varepsilon}_{ki}\tilde{\varepsilon}_{kj} - \frac{1}{n}\sum_{k=1}^{n}\hat{\varepsilon}_{ki}^2(\hat{\beta}_{i,j}-\beta_{i,j}) \\ &\quad - \frac{1}{n}\sum_{k=1}^{n}\hat{\varepsilon}_{kj}^2(\hat{\beta}_{j-1,i}-\beta_{j-1,i}) \\ &\quad + O_{\mathsf{P}}\big((a_{n1}^2+a_{n1})\sqrt{\log p/n}+a_{n2}^2\big).\end{aligned}$$

By (22), we have uniformly in $1\le i\le p$,

$$\frac{1}{n}\sum_{k=1}^{n}\hat{\varepsilon}_{ki}^2 = \frac{1}{n}\sum_{k=1}^{n}\tilde{\varepsilon}_{ki}^2 + O_{\mathsf{P}}\big((a_{n1}^2+a_{n1})\sqrt{\log p/n}+a_{n2}^2\big). \qquad (25)$$

So, by (25) and $\max_{i,j}|\beta_{i,j}| \le C$ for some constant $C > 0$,

$$
\begin{aligned}
&\frac{1}{n}\sum_{k=1}^{n}\hat{\varepsilon}_{ki}\hat{\varepsilon}_{kj} + \frac{1}{n}\sum_{k=1}^{n}\hat{\varepsilon}_{ki}^2\hat{\beta}_{i,j} + \frac{1}{n}\sum_{k=1}^{n}\hat{\varepsilon}_{kj}^2\hat{\beta}_{j-1,i} \\
&\quad = \frac{1}{n}\sum_{k=1}^{n}\tilde{\varepsilon}_{ki}\tilde{\varepsilon}_{kj} + \frac{1}{n}\sum_{k=1}^{n}\hat{\varepsilon}_{ki}^2\beta_{i,j} + \frac{1}{n}\sum_{k=1}^{n}\hat{\varepsilon}_{kj}^2\beta_{j-1,i} \\
&\qquad + O_{\mathsf{P}}\big((a_{n1}^2 + a_{n1})\sqrt{\log p/n} + a_{n2}^2\big) \qquad (26) \\
&\quad = -b_{nij}\frac{\omega_{ij}}{\omega_{ii}\omega_{jj}} + \frac{\sum_{k=1}^{n}(\varepsilon_{ki}\varepsilon_{kj} - \mathsf{E}\varepsilon_{ki}\varepsilon_{kj})}{n} \\
&\qquad + O_{\mathsf{P}}\bigg(a_{n1}\sqrt{\log p/n} + a_{n2}^2 + \frac{\log p}{n}\bigg)
\end{aligned}
$$

uniformly in $1 \le i < j \le p$. The proposition is proved by (C1) and the central limit theorem.

6.2. *Proof of Theorem* 3.1. To prove Theorem 3.1, we need some lemmas. Let $\xi_1, \dots, \xi_n$ be independent and identically distributed d-dimensional random vectors with mean zero. Let $G(t) = 2 - 2\Phi(t)$ and define $|\cdot|_{(d)}$ by $|\mathbf{z}|_{(d)} = \min\{|z_i|; 1 \le i \le d\}$ for $\mathbf{z} = (z_1, \dots, z_d)'$.

LEMMA 6.1. *Suppose that $p \le cn^r$ and $\mathsf{E}|\xi_1|_2^{bdr+2+\epsilon} < \infty$ for some fixed $c > 0$, $r > 0$, $b > 0$ and $\epsilon > 0$. Assume that $\|\mathsf{Cov}(\xi_1) - \mathbf{I}_d\|_2 \le C(\log p)^{-2-\gamma}$ for some $\gamma > 0$. Then we have*

$$
\sup_{0\le t\le\sqrt{b\log p}}\bigg|\frac{\mathsf{P}(|\sum_{k=1}^{n}\xi_k|_{(d)} \ge t\sqrt{n})}{(G(t))^d} - 1\bigg| \le C(\log p)^{-1-\gamma_1}
$$

for $\gamma_1 = \min\{\gamma, 1/2\}$.

REMARK 2. In the application of Lemma 6.1, only $d = 2$ is needed.

PROOF OF LEMMA 6.1. For $1 \le i \le p$, put

$$
\begin{aligned}
\hat{\xi}_i &= \xi_i I\{|\xi_i|_2 \le \sqrt{n}/(\log p)^4\} - \mathsf{E}\xi_i I\{|\xi_i|_2 \le \sqrt{n}/(\log p)^4\},\\
\tilde{\xi}_i &= \xi_i - \hat{\xi}_i.
\end{aligned}
$$

We have

$$
\begin{aligned}
\mathsf{P}\bigg(\bigg|\sum_{k=1}^{n}\xi_k\bigg|_{(d)} \ge t\sqrt{n}\bigg) &\le \mathsf{P}\bigg(\bigg|\sum_{k=1}^{n}\hat{\xi}_k\bigg|_{(d)} \ge t\sqrt{n} - \sqrt{n}/(\log p)^2\bigg)\\
&\quad + \mathsf{P}\bigg(\bigg|\sum_{k=1}^{n}\tilde{\xi}_k\bigg|_2 \ge \sqrt{n}/(\log p)^2\bigg).
\end{aligned}
$$

Note that

$$\sum_{i=1}^{n} \mathsf{E}|\xi_i|_2 I\{|\xi_i|_2 > \sqrt{n}/(\log p)^4\} = o\big(\sqrt{n}/(\log p)^2\big).$$

We have by condition $\mathsf{E}|\xi_1|_2^{bdr+2+\epsilon} < \infty$,

$$\mathsf{P}\left(\left|\sum_{k=1}^{n} \tilde{\xi}_k\right|_2 \geq \sqrt{n}/(\log p)^2\right) \leq n\mathsf{P}\big(|\xi_1|_2 \geq \sqrt{n}/(\log p)^4\big) \leq C(\log p)^{-3/2}\big(G(t)\big)^d$$

uniformly in $0 \leq t \leq \sqrt{b \log p}$. Similarly, we have

$$\begin{aligned}\mathsf{P}&\left(\left|\sum_{k=1}^{n} \xi_k\right|_{(d)} \geq t\sqrt{n}\right) \\ &\geq \mathsf{P}\left(\left|\sum_{k=1}^{n} \hat{\xi}_k\right|_{(d)} \geq t\sqrt{n} + \sqrt{n}/(\log p)^2\right) - C(\log p)^{-3/2}\big(G(t)\big)^d.\end{aligned}$$

So it suffices to prove

$$\sup_{0\leq t\leq\sqrt{b\log p}} \left|\frac{\mathsf{P}(|\sum_{k=1}^{n} \hat{\xi}_k|_{(d)} \geq (t \pm (\log p)^{-2})\sqrt{n})}{(G(t))^d} - 1\right| \leq C(\log p)^{-1-\gamma_1}.$$

By Theorem 1 in Zaïtsev (1987), we have

$$\begin{aligned}\mathsf{P}&\left(\left|\sum_{k=1}^{n} \hat{\xi}_k\right|_{(d)} \geq \big(t - (\log p)^{-2}\big)\sqrt{n}\right) \\ &\leq \mathsf{P}\big(|\mathbf{W}|_{(d)} \geq t - 2(\log p)^{-2}\big) + c_{1,d}\exp\big(-c_{2,d}(\log p)^2\big), \\ \mathsf{P}&\left(\left|\sum_{k=1}^{n} \hat{\xi}_k\right|_{(d)} \geq \big(t + (\log p)^{-2}\big)\sqrt{n}\right) \\ &\geq \mathsf{P}\big(|\mathbf{W}|_{(d)} \geq t + 2(\log p)^{-2}\big) - c_{1,d}\exp\big(-c_{2,d}(\log p)^2\big),\end{aligned}$$

where $c_{1,d}$ and $c_{2,d}$ are positive constants depending only on d, and $\mathbf{W}$ is a multivariate normal vector with mean zero and covariance matrix $\mathsf{Cov}(\sum_{i=1}^{n} \hat{\xi}_i/\sqrt{n})$. By $\mathsf{E}|\xi_1|_2^{bdr+2+\epsilon} < \infty$,

$$\left\|\mathsf{Cov}\left(\sum_{i=1}^{n} \hat{\xi}_i/\sqrt{n}\right) - \mathbf{I}_d\right\|_2 \leq C(\log p)^{-2-\gamma}.$$

So, with the density of the multivariate normal random variable, it is easy to show that

$$\mathsf{P}\big(|\mathbf{W}|_{(d)} \geq t - 2(\log p)^{-2}\big) \leq \big(1 + C(\log p)^{-1-\gamma}\big)\big(G(t)\big)^d$$

uniformly in $0 \le t \le \sqrt{b\log p}$. By noting that $c_{1,d}\exp(-c_{2,d}(\log p)^2) \le C(\log p)^{-1-\gamma_1}(G(t))^d$ for $0 \le t \le \sqrt{b\log p}$, we obtain that

$$\mathsf{P}\left(\left|\sum_{k=1}^{n}\hat{\xi}_k\right|_{(d)} \ge \bigl(t-(\log p)^{-2}\bigr)\sqrt{n}\right) \le \bigl(1+C(\log p)^{-1-\gamma_1}\bigr)\bigl(G(t)\bigr)^d$$

uniformly in $0 \le t \le \sqrt{b\log p}$. Similarly, we can prove that

$$\mathsf{P}\left(\left|\sum_{k=1}^{n}\hat{\xi}_k\right|_{(d)} \ge \bigl(t-(\log p)^{-2}\bigr)\sqrt{n}\right) \ge \bigl(1-C(\log p)^{-1-\gamma_1}\bigr)\bigl(G(t)\bigr)^d.$$

This finishes the proof. □

Let $\boldsymbol{\eta}_k = (\eta_{k1}, \eta_{k2})'$ be independent and identically distributed 2-dimensional random vectors with mean zero.

LEMMA 6.2. *Suppose that $p \le cn^r$ and $\mathsf{E}|\boldsymbol{\eta}_1|_2^{2br+2+\epsilon} < \infty$ for some fixed $c > 0$, $r > 0$, $b > 0$ and $\epsilon > 0$. Assume that $\mathsf{Var}(\eta_{11}) = \mathsf{Var}(\eta_{12}) = 1$ and $|\mathsf{Cov}(\eta_{11}, \eta_{12})| \le \delta$ for some $0 \le \delta < 1$. Then we have*

$$\mathsf{P}\left(\left|\sum_{k=1}^{n}\eta_{k1}\right| \ge t\sqrt{n}, \left|\sum_{k=1}^{n}\eta_{k2}\right| \ge t\sqrt{n}\right) \le C(t+1)^{-2}\exp\bigl(-t^2/(1+\delta)\bigr)$$

uniformly for $0 \le t \le \sqrt{b\log p}$, where C only depends on $c, b, r, \epsilon, \delta$.

PROOF. The proof is similar to that of Lemma 6.1. Actually, following the proof of Lemma 6.1, we only need to prove

$$\mathsf{P}\bigl(|\mathbf{W}|_{(2)} \ge t - 2(\log p)^{-2}\bigr) \le C(t+1)^{-2}\exp\bigl(-t^2/(1+\delta)\bigr), \tag{27}$$

where $\mathbf{W}$ is a two-dimensional normal vector with mean zero and covariance matrix $\mathsf{Cov}(\sum_{i=1}^n \hat{\boldsymbol{\eta}}_i/\sqrt{n})$ and

$$\hat{\boldsymbol{\eta}}_i = \boldsymbol{\eta}_i I\{|\boldsymbol{\eta}_i|_2 \le \sqrt{n}/(\log p)^4\} - \mathsf{E}\boldsymbol{\eta}_i I\{|\boldsymbol{\eta}_i|_2 \le \sqrt{n}/(\log p)^4\}.$$

By $\mathsf{E}|\boldsymbol{\eta}_1|_2^{2br+2+\epsilon} < \infty$, we have

$$\left\|\mathsf{Cov}\left(\sum_{i=1}^{n}\hat{\boldsymbol{\eta}}_i/\sqrt{n}\right) - \mathsf{Cov}(\boldsymbol{\eta}_1)\right\|_2 \le C(\log p)^{-2-\gamma}.$$

This, together with Lemma 2 in Berman (1962) and some tedious calculations, implies (27). □

We now start to prove Theorem 3.1. Let $\rho_{ij,\omega} = \omega_{ij}/\sqrt{\omega_{ii}\omega_{jj}}$. Put

$$\sigma_{ii,\varepsilon} = \mathsf{Var}(\varepsilon_i) \quad \text{and} \quad U_{ij} = \frac{\sum_{k=1}^n(\varepsilon_{ki}\varepsilon_{kj} - \mathsf{E}\varepsilon_{ki}\varepsilon_{kj})}{\sqrt{n}\sigma_{ii,\varepsilon}^{1/2}\sigma_{jj,\varepsilon}^{1/2}}.$$

Note that $\mathsf{Var}(\varepsilon_{ki}\varepsilon_{kj}) = \sigma_{ii,\varepsilon}\sigma_{jj,\varepsilon}(1+\rho^2_{ij,\omega})$. By letting $b = 16$ in Lemma 6.1,

$$\max_{i,j} \sup_{0\le t\le 4\sqrt{\log p}} \left| \frac{\mathsf{P}(|U_{ij}| \ge t\sqrt{1+\rho^2_{ij,\omega}})}{G(t)} - 1 \right| \le C(\log p)^{-1-\gamma_1}. \tag{28}$$

By (22), it is easy to see that

$$\max_{1\le i\le p} |\hat{r}_{ii} - \sigma_{ii,\varepsilon}| = O_{\mathsf{P}}\left(\sqrt{\frac{\log p}{n}}\right).$$

By (13) and (26), we have

$$\max_{1\le i<j\le p} \left| \sqrt{\frac{n}{\hat{r}_{ii}\hat{r}_{jj}}} \left(T_{ij} + b_{nij}\frac{\omega_{ij}}{\omega_{ii}\omega_{jj}} \right) - U_{ij} \right| = o_{\mathsf{P}}\big((\log p)^{-1/2}\big).$$

This implies that

$$\mathsf{P}\left(\max_{1\le i<j\le p} \sqrt{\frac{n}{\hat{r}_{ii}\hat{r}_{jj}(1+\rho^2_{ij,\omega})}} \left| T_{ij} + b_{nij}\frac{\omega_{ij}}{\omega_{ii}\omega_{jj}} \right| \ge \left(2 - O\left(\frac{1}{\log p}\right)\right)\sqrt{\log p} \right) \to 0.$$

Under the conditions of Theorem 3.1 and noting that $\max_{1\le i\le j\le p} |\mathbf{b}_{nij} - 1| = O_{\mathsf{P}}(\sqrt{\log p/n})$, we have

$$\sum_{1\le i<j\le p} I\{|\hat{T}_{ij}| \ge 2\sqrt{\log p}\} \ge \max(c_p, d_p)$$

with probability tending to one, where

$$c_p = \sqrt{\log\log p} \quad \text{and} \quad d_p = \frac{1}{2}\max_{1\le i\le p} \operatorname{Card}(\mathcal{A}_i(\gamma)).$$

Hence,

$$\frac{(p^2-p)/2}{\max\{\sum_{1\le i<j\le p} I\{|\hat{T}_{ij}| \ge 2\sqrt{\log p}\}, 1\}} \le \frac{p^2-p}{2}\frac{1}{\max(c_p, d_p)} \tag{29}$$

with probability tending to one. For $0 < \theta < (1-\rho)/(1+\rho)$, let

$$\Lambda(\theta) = \left\{ 1 \le i \le p : \exists j \ne i, s.t. \frac{|\omega_{ij}|}{\sqrt{\omega_{ii}\omega_{jj}}} \ge \theta \right\}.$$

If $\operatorname{Card}(\Lambda(\theta)) \ge p/(\log p)^6$, then

$$\sum_{1\le i<j\le p} I\{|\hat{T}_{ij}| \ge 2\sqrt{\log p}\} \ge 2^{-1}p/(\log p)^6$$

with probability tending to one and the upper bound in (29) can be replaced by $Cp(\log p)^6$. Set $d_p = \max_{1\le i\le p} \mathrm{Card}(\mathcal{A}_i(\gamma))$. We let

$$b_p = G^{-1}\big(p^{-2}\alpha \max\{c_p, d_p\}\big) \quad \text{and} \quad \theta_1 = \theta$$

if $\mathrm{Card}(\Lambda(\theta)) < p/(\log p)^6$;

$$b_p = \sqrt{2\log p + 14\log\log p} \quad \text{and} \quad \theta_1 = 1$$

if $\mathrm{Card}(\Lambda(\theta)) \ge p/(\log p)^6$. Note that

$$1 - \Phi(b_p) \sim \frac{1}{\sqrt{2\pi}\, b_p} \exp\big(-b_p^2/2\big).$$

Hence, by the definition of $\hat{t}$, we have $\mathsf{P}(0 \le \hat{t} \le b_p) \to 1$. By the continuity of $G(t)$ and the monotonicity of the indicator function, we can obtain that, for $0 \le \hat{t} < 2\sqrt{\log p}$,

$$\frac{G(\hat{t})(p^2 - p)/2}{\max\{\sum_{1\le i<j\le p} I\{|\hat{T}_{ij}| \ge \hat{t}\}, 1\}} = \alpha.$$

To prove Theorem 3.1, by $\mathsf{P}(0 \le \hat{t} \le b_p) \to 1$, it is enough to show that

$$\sup_{0\le t\le b_p} \left| \frac{\sum_{(i,j)\in\mathcal{H}_0} I\{|\hat{T}_{ij}| \ge t\}}{q_0 G(t)} - 1 \right| \to 0 \tag{30}$$

in probability, where $q_0 = \mathrm{Card}(\mathcal{H}_0)$. To prove (30), we need the following lemma.

LEMMA 6.3. *Suppose that for any $\varepsilon > 0$,*

$$\sup_{0\le t\le b_p} \mathsf{P}\left(\left| \frac{\sum_{(i,j)\in\mathcal{H}_0} [I\{|U_{ij}| \ge t\} - \mathsf{P}(|U_{ij}| \ge t)]}{2q_0(1 - \Phi(t))} \right| \ge \varepsilon\right) = o(1) \tag{31}$$

and

$$\int_0^{b_p} \mathsf{P}\left(\left| \frac{\sum_{(i,j)\in\mathcal{H}_0} [I\{|U_{ij}| \ge t\} - \mathsf{P}(|U_{ij}| \ge t)]}{2q_0(1 - \Phi(t))} \right| \ge \varepsilon\right) dt = o(v_p), \tag{32}$$

where $v_p = 1/\sqrt{(\log p)(\log\log p)^{1/2}}$. Then (30) holds.

Let us first finish the proof of Theorem 3.1. By Lemma 6.3, it suffices to prove (31) and (32). Define

$$\mathcal{S}_1 = \begin{cases} \{(i,j) : i \in \Lambda(\theta), j \ge i\}, & \text{if } \mathrm{Card}(\Lambda(\theta)) < p/(\log p)^6, \\ \varnothing, & \text{if } \mathrm{Card}(\Lambda(\theta)) \ge p/(\log p)^6, \end{cases}$$

$$\mathcal{S}_2 = \{(i,j) : 1 \le i \le p, j \in \mathcal{A}_i(\gamma)\},$$

$$\mathcal{H}_{01} = \mathcal{H}_0 \cap \{\mathcal{S}_1 \cup \mathcal{S}_2\}, \qquad \mathcal{H}_{02} = \mathcal{H}_0 \cap \{\mathcal{S}_1 \cup \mathcal{S}_2\}^c.$$

Note that ε_i and ε_j have strong correlations for $(i, j) \in \mathcal{H}_{01}$. However, because the cardinality of $\mathcal{H}_{01}$ is $O(p^{1+\rho} + p^2/(\log p)^6)$, the terms in $\mathcal{H}_{01}$ can be neglected. Actually, by (28) and the fact that $q_0 \geq cp^2$,

$$\begin{aligned} &\mathsf{E}\left|\frac{\sum_{(i,j)\in\mathcal{H}_{01}}[I\{|U_{ij}| \geq t\} - \mathsf{P}(|U_{ij}| \geq t)]}{q_0 G(t)}\right| \\ &\quad \leq C\frac{(p^{1+\rho} + p^2/(\log p)^6)G(t)}{p^2 G(t)} \\ &\quad = O\big((\log p)^{-6}\big) \end{aligned} \tag{33}$$

uniformly for $0 \leq t \leq 2\sqrt{\log p}$. On the other hand, since $|\mathcal{H}_{02}| \sim p^2/2$, we need to calculate the variance of the sum $\sum_{(i,j)\in\mathcal{H}_{02}}[\cdots]$ as follows:

$$\begin{aligned} &\mathsf{E}\left[\frac{\sum_{(i,j)\in\mathcal{H}_{02}}\{I\{|U_{ij}| \geq t\} - \mathsf{P}(|U_{ij}| \geq t)\}}{q_0 G(t)}\right]^2 \\ &\quad = \frac{\sum_{(i,j)\in\mathcal{H}_{02}}\sum_{(k,l)\in\mathcal{H}_{02}}\{\mathsf{P}(|U_{ij}| \geq t, |U_{kl}| \geq t) - \mathsf{P}(|U_{ij}| \geq t)\mathsf{P}(|U_{kl}| \geq t)\}}{q_0^2 G^2(t)}. \end{aligned} \tag{34}$$

To estimate the sums with four indices i, j, k, l in (34), we split the set $\mathcal{H}_{02}$ into two subsets as in Cai, Liu and Xia (2013). Let $G_{abcd} = (V_{abcd}, E_{abcd})$ be a graph, where $V_{abcd} = \{a, b, c, d\}$ is the set of vertices and E_{abcd} is the set of edges. There is an edge between $i \neq j \in \{a, b, c, d\}$ if and only if $|\omega_{ij}| \geq (\log p)^{-2-\gamma}$. If the number of different vertices in V_{abcd} is 3, then we call G_{abcd} a three vertices graph (3-G). Similarly, G_{abcd} is a four vertices graph (4-G) if the number of different vertices in V_{abcd} is 4. A vertex in G_{abcd} is said to be *isolated* if there is no edge connected to it. Note that for any $(i, j) \in \mathcal{H}_{02}$, $(k, l) \in \mathcal{H}_{02}$ and $(i, j) \neq (k, l)$, G_{ijkl} is 3-G or 4-G. We say a graph $\mathcal{G} := G_{ijkl}$ satisfies $(\star)$ if

$$\begin{aligned} &\text{If } \mathcal{G} \text{ is 4-G, then there is at least one isolated vertex in } \mathcal{G}; \\ &\text{otherwise } \mathcal{G} \text{ is 3-G and } E_{ijkl} = \varnothing. \end{aligned} \tag{$\star$}$$

Note that for any integers $1 \leq i, j, k, l \leq p$,

$$\mathsf{E}[\varepsilon_i\varepsilon_j\varepsilon_k\varepsilon_l] = \frac{\omega_{ij}\omega_{kl} + \omega_{ik}\omega_{jl} + \omega_{il}\omega_{jk}}{\omega_{ii}\omega_{jj}\omega_{kk}\omega_{ll}}.$$

Hence, for any G_{ijkl} satisfying $(\star)$,

$$\big|\mathsf{E}[\varepsilon_i\varepsilon_j\varepsilon_k\varepsilon_l]\big| = O\big((\log p)^{-2-\gamma}\big), \tag{35}$$

where $O(1)$ is uniformly for i, j, k, l. By the above definition, we further divide the indices set in (34) into

$$\mathcal{H}_{020} = \big\{(i, j) \in \mathcal{H}_{02}, (k, l) \in \mathcal{H}_{02} : (i, j) = (k.l)\big\};$$

$$\mathcal{H}_{021} = \big\{(i, j) \in \mathcal{H}_{02}, (k, l) \in \mathcal{H}_{02} : (i, j) \neq (k.l), \mathcal{G}_{ijkl} \text{ satisfies } (\star)\big\};$$

$$\mathcal{H}_{022} = \big\{(i, j) \in \mathcal{H}_{02}, (k, l) \in \mathcal{H}_{02} : (i, j) \neq (k.l), \mathcal{G}_{ijkl} \text{ does not satisfy } (\star)\big\}.$$

For the indices in $\mathcal{H}_{020} \cup \mathcal{H}_{022}$, U_{ij} and U_{kl} may have strong correlations, but the cardinalities of $\mathcal{H}_{020}$ and $\mathcal{H}_{022}$ are small compared to p^2. For these two subsets, we will use (28) and Lemma 6.2 to estimate the joint tail probabilities of U_{ij} and U_{kl}. On the other hand, it follows from (35) that the correlation between U_{ij} and U_{kl} in $\mathcal{H}_{021}$ is weak so that their joint tail probabilities can be estimated by Lemma 6.1. Thus, the sums in the three subsets can be further bounded in the following way. For the indices in $\mathcal{H}_{020}$, we have by (28),

$$\left|\frac{\sum_{\{(i,j),(k,l)\}\in\mathcal{H}_{020}}\{\mathsf{P}(|U_{ij}|\geq t, |U_{kl}|\geq t) - \mathsf{P}(|U_{ij}|\geq t)\mathsf{P}(|U_{kl}|\geq t)\}}{q_0^2 G^2(t)}\right| \leq \frac{C}{p^2 G(t)}. \tag{36}$$

It is easy to show that Card($\mathcal{H}_{022}$)$\leq Cp^2 d_p^2$. We say the graph $\mathcal{G}_{ijkl}$ is aG-bE if G_{ijkl} is a-G and there are b edges in E_{ijkl} for $a = 3, 4$ and $b = 0, 1, 2, 3, 4$. Note that for any $(i, j) \in \mathcal{H}_0$, the vertices i and j are not connected. So we can divide $\mathcal{H}_{022}$ into two parts:

$$\mathcal{H}_{022,1} = \{\{(i,j),(k,l)\} \in \mathcal{H}_{022} : \mathcal{G}_{ijkl} \text{ is 3G-1E or 4G-2E}\},$$
$$\mathcal{H}_{022,2} = \{\{(i,j),(k,l)\} \in \mathcal{H}_{022} : \mathcal{G}_{ijkl} \text{ is 4G-3E or 4G-4E}\}.$$

It can be shown that Card$(\mathcal{H}_{022,2}) = O(pd_p^3)$ and Card$(\mathcal{H}_{022,1}) = O(p^2 d_p^2)$. Then, by (28),

$$\left|\frac{\sum_{\{(i,j),(k,l)\}\in\mathcal{H}_{022,2}}\{\mathsf{P}(|U_{ij}|\geq t, |U_{kl}|\geq t) - \mathsf{P}(|U_{ij}|\geq t)\mathsf{P}(|U_{kl}|\geq t)\}}{q_0^2 G^2(t)}\right| \leq \frac{Cd_p^3}{p^3 G(t)}. \tag{37}$$

It remains for us to estimate the terms in $\mathcal{H}_{022,1}$ and $\mathcal{H}_{021}$. To this end, we need the following lemma.

LEMMA 6.4. *We have*

$$\max_{\{(i,j),(k,l)\}\in\mathcal{H}_{021}} \mathsf{P}(|U_{ij}|\geq t, |U_{kl}|\geq t) = (1 + A_n)G^2(t) \tag{38}$$

and

$$\max_{\{(i,j),(k,l)\}\in\mathcal{H}_{022,1}} \mathsf{P}(|U_{ij}|\geq t, |U_{kl}|\geq t) \leq C(t+1)^{-1}\exp(-t^2/(1+\theta_1)) \tag{39}$$

uniformly in $0 \leq t \leq b_p$, *where* $A_n \leq C(\log p)^{-1-\gamma_1}$.

PROOF. It can be proved that, uniformly for $\{(i,j),(k,l)\} \in \mathcal{H}_{021}$,

$$\|\mathrm{Corr}((U_{ij}, U_{kl})) - \mathbf{I}_2\|_2 = O((\log p)^{-2-\gamma}),$$

and uniformly for $\{(i,j),(k,l)\} \in \mathcal{H}_{022,1}$,

$$\big|\mathrm{Corr}(U_{ij}, U_{kl})\big| \le \theta_1 + O\big((\log p)^{-2-\gamma}\big).$$

The proof is complete by Lemmas 6.1 and 6.2. □

By Lemma 6.4, we have

$$\left|\frac{\sum_{\{(i,j),(k,l)\}\in\mathcal{H}_{021}}\{\mathsf{P}(U_{ij}\ge t, U_{kl}\ge t) - \mathsf{P}(U_{ij}\ge t)\mathsf{P}(U_{kl}\ge t)\}}{q_0^2 G^2(t)}\right| \le C(\log p)^{-1-\gamma_1} \tag{40}$$

and

$$\left|\frac{\sum_{\{(i,j),(k,l)\}\in\mathcal{H}_{022,1}}\{\mathsf{P}(U_{ij}\ge t, U_{kl}\ge t) - \mathsf{P}(U_{ij}\ge t)\mathsf{P}(U_{kl}\ge t)\}}{q^2 G^2(t)}\right| \le C p^{-2} d_p^2 \big[G(t)\big]^{-2\theta_1/(1+\theta_1)}. \tag{41}$$

Using some elementary calculations,

$$\int_0^{b_p}\left[\frac{1}{p^2 G(t)} + \frac{d_p^3}{p^3 G(t)} + \frac{d_p^2}{p^2[G(t)]^{2\theta_1/(1+\theta_1)}}\right]dt = o(v_p).$$

Combining (33), (37), (40), (41) and the fact $d_p = O(p^{\rho})$, we prove (32). The proof of (31) is exactly the same with that of (32) and hence is omitted.

PROOF OF LEMMA 6.3. Recall the definition of b_p in the proof of Theorem 3.1. Let $0 = t_0 < t_1 < \cdots < t_m = b_p$ satisfy $t_i - t_{i-1} = v_p$ for $1 \le i \le m-1$ and $t_m - t_{m-1} \le v_p$. So $m \sim b_p/v_p$. For any $t_{j-1} \le t \le t_j$, we have

$$\frac{\sum_{(i,j)\in\mathcal{H}_0} I\{|\hat{T}_{ij}| \ge t\}}{q_0 G(t)} \le \frac{\sum_{(i,j)\in\mathcal{H}_0} I\{|\hat{T}_{ij}| \ge t_{j-1}\}}{q_0 G(t_{j-1})}\frac{G(t_{j-1})}{G(t_j)} \tag{42}$$

and

$$\frac{\sum_{(i,j)\in\mathcal{H}_0} I\{|\hat{T}_{ij}| \ge t\}}{q_0 G(t)} \ge \frac{\sum_{(i,j)\in\mathcal{H}_0} I\{|\hat{T}_{ij}| \ge t_j\}}{q_0 G(t_j)}\frac{G(t_j)}{G(t_{j-1})}. \tag{43}$$

In view of (42), (43) and $G(t_j)/G(t_{j-1}) \to 1$, we only need to prove

$$\max_{0\le j\le m}\left|\frac{\sum_{(i,j)\in\mathcal{H}_0}[I\{|\hat{T}_{ij}| \ge t_j\} - G(t_j)]}{q_0 G(t_j)}\right| \to 0$$

in probability. We have

$$\max_{1\le i<j\le p}|\hat{T}_{ij} - U_{ij}| = O_{\mathsf{P}}\big(a_{n1}\sqrt{\log p} + \sqrt{n}a_{n2}^2 + (\log p)/\sqrt{n}\big).$$

Since

$$\frac{G(t+o(\sqrt{1/\log p}))}{G(t)} = 1 + o(1)$$

uniformly in $0 \le t \le 2\sqrt{\log p}$, by (13), it suffices to show that

$$\max_{0\le j\le m}\left|\frac{\sum_{(i,j)\in\mathcal{H}_0}[I\{|U_{ij}|\ge t_j\} - G(t_j)]}{q_0 G(t_j)}\right| \to 0$$

in probability. We have

$$\begin{aligned}
&\mathsf{P}\left(\max_{1\le j\le m}\left|\frac{\sum_{(i,j)\in\mathcal{H}_0}[I\{|U_{ij}|\ge t_j\} - G(t_j)]}{q_0 G(t_j)}\right| \ge \varepsilon\right)\\
&\quad\le \sum_{j=1}^{m}\mathsf{P}\left(\left|\frac{\sum_{(i,j)\in\mathcal{H}_0}[I\{|U_{ij}|\ge t_j\} - G(t_j)]}{q_0 G(t_j)}\right| \ge \varepsilon\right)\\
&\quad\le \frac{1}{v_p}\int_0^{b_p}\mathsf{P}\left(\frac{\sum_{(i,j)\in\mathcal{H}_0} I\{|U_{ij}|\ge t\}}{q_0 G(t)} \ge 1+\varepsilon/2\right)dt\\
&\qquad + \frac{1}{v_p}\int_0^{b_p}\mathsf{P}\left(\frac{\sum_{(i,j)\in\mathcal{H}_0} I\{|U_{ij}|\ge t\}}{q_0 G(t)} \le 1-\varepsilon/2\right)dt\\
&\qquad + \sum_{j=m-1}^{m}\mathsf{P}\left(\left|\frac{\sum_{(i,j)\in\mathcal{H}_0}[I\{|U_{ij}|\ge t_j\} - G(t_j)]}{q_0 G(t_j)}\right| \ge \varepsilon\right).
\end{aligned}$$

So it suffices to prove

$$\int_0^{b_p}\mathsf{P}\left(\left|\frac{\sum_{(i,j)\in\mathcal{H}_0} I\{|U_{ij}|\ge t\} - G(t)}{q_0 G(t)}\right| \ge \varepsilon\right)dt = o(v_p)$$

and

$$\sum_{k=m-1}^{m}\mathsf{P}\left(\left|\frac{\sum_{(i,j)\in\mathcal{H}_0}[I\{|U_{ij}|\ge t_k\} - G(t_k)]}{q_0 G(t_k)}\right| \ge \varepsilon\right) = o(1),$$

which are the conditions of Lemma 6.3. □

6.3. *Proof of Propositions* 4.1 *and* 4.2.

PROOF OF PROPOSITION 4.1. We first show that the true $\boldsymbol{\beta}_i$ belongs to the region

$$|\mathbf{D}_i^{-1/2}\hat{\boldsymbol{\Sigma}}_{-i,-i}\boldsymbol{\beta}_i - \mathbf{D}_i^{-1/2}\hat{\mathbf{a}}|_\infty \le \lambda_{ni}(2) \tag{44}$$

with probability tending to one. Without loss of generality, we assume $\mathsf{E}\mathbf{X}_k = 0$. It suffices to prove that

$$\left|\frac{1}{n}\sum_{k=1}^{n}(X_{kj} - \bar{X}_j)\left\{\sum_{l\neq i}(X_{kl} - \bar{X}_l)\beta_l - X_{ki} + \bar{X}_i\right\}\right|$$

$$= \left|\frac{1}{n}\sum_{k=1}^{n}(X_{kj} - \bar{X}_j)\varepsilon_{ki}\right| \le \sqrt{\hat{\sigma}_{jj}}\lambda_{ni}(2),$$

uniformly in $1 \le i \ne j \le p$, with probability tending to one. By the independence between $\{\varepsilon_{ki}\}$ and $\{X_{k,j}, j \ne i\}$, we have

$$\mathsf{P}\left(\max_{i\neq j}\frac{1}{\sqrt{n\hat{\sigma}_{jj}\,\mathsf{Var}(\varepsilon_i)}}\left|\sum_{k=1}^{n}(X_{kj} - \bar{X}_j)\varepsilon_{ki}\right| \ge \left(2 + O\left((\log p)^{-1/2}\right)\right)\sqrt{\log p}\right)$$

$$\le C(\log p)^{-1/2}.$$

Since $\mathsf{Var}(\varepsilon_i) = 1/\omega_{ii} \le \sigma_{ii}$, we prove (44). By the definition of $\hat{\boldsymbol{\beta}}_i$,

$$\left|\mathbf{D}_i^{-1/2}\hat{\boldsymbol{\Sigma}}_{-i,-i}\hat{\boldsymbol{\beta}}_i - \mathbf{D}_i^{-1/2}\hat{\mathbf{a}}\right|_\infty \le \lambda_{ni}(2).$$

Then it follows that

$$\left|\mathbf{D}_i^{-1/2}\hat{\boldsymbol{\Sigma}}_{-i,-i}(\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i)\right|_\infty \le 2\lambda_{ni}(2)$$

with probability tending to one. We next prove the restricted eigenvalue (RE) assumption in Bickel, Ritov and Tsybakov (2009), page 1710, holds with $\kappa(s,1) \ge c\lambda_{\min}(\boldsymbol{\Sigma})^{1/2}$ for some $c > 0$. Actually, the RE assumption follows from

$$\max_{1\le i\le p}|\boldsymbol{\beta}_i|_0 = o\left(\lambda_{\min}(\boldsymbol{\Sigma})\sqrt{\frac{n}{\log p}}\right)$$

and the inequality

$$\boldsymbol{\delta}'\hat{\boldsymbol{\Sigma}}_{-i,-i}\boldsymbol{\delta} \ge \lambda_{\min}(\boldsymbol{\Sigma}_{-i,-i})|\boldsymbol{\delta}|_2^2 - O_{\mathsf{P}}\left(\sqrt{\frac{\log p}{n}}\right)|\boldsymbol{\delta}|_1^2 \tag{45}$$

for any $\delta \in \mathbf{R}^p$. By the proof of Theorem 7.1 in Bickel, Ritov and Tsybakov (2009), we obtain that

$$\max_{1\le i\le p}(\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i)'\hat{\boldsymbol{\Sigma}}_{-i,-i}(\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i) = O_{\mathsf{P}}\left(\frac{\max_{1\le i\le p}|\boldsymbol{\beta}_i|_0\log p}{\lambda_{\min}(\boldsymbol{\Sigma})n}\right) \tag{46}$$

and

$$\max_{1\le i\le p}|\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i|_1 = O_{\mathsf{P}}\left(\max_{1\le i\le p}|\boldsymbol{\beta}_i|_0\lambda_{\min}(\boldsymbol{\Sigma})^{-1}\sqrt{\frac{\log p}{n}}\right). \tag{47}$$

This implies Proposition 4.1. □

PROOF OF PROPOSITION 4.2. By the proof of Proposition 4.1, we have for any $\delta > 2$ and some $1 < c < \delta/2$,

$$\max_{i \neq j} \frac{1}{\sqrt{n\hat{\sigma}_{jj}}} \left| \sum_{k=1}^{n} (X_{kj} - \bar{X}_j)\varepsilon_{ki} \right| \leq \frac{1}{c}\lambda_{ni1}(\delta) \tag{48}$$

with probability tending to one. For a vector $\mathbf{a} = (a_1, \ldots, a_p)'$ and an index set $T \subseteq \{1, 2, \ldots, p\}$, let $\mathbf{a}_T$ be the vector with $(\mathbf{a}_T)_i = a_i$ for $i \in T$ and $(\mathbf{a}_T)_i = 0$ for $i \in T^c$. Let T_i be the support of $\boldsymbol{\beta}_i$. Then by the proof of Theorem 1 in Belloni, Chernozhukov and Wang (2011), we can get $|(\hat{\boldsymbol{\alpha}}_i(\delta) - \mathbf{D}_i^{1/2}\boldsymbol{\beta}_i)_{T_i^c}|_1 \leq \bar{c}|(\hat{\boldsymbol{\alpha}}_i(\delta) - \mathbf{D}_i^{1/2}\boldsymbol{\beta}_i)_{T_i}|_1$ for $\bar{c} = (c+1)/(c-1)$. Also,

$$\left|\mathbf{D}_i^{-1/2}\hat{\boldsymbol{\Sigma}}_{-i,-i}\mathbf{D}_i^{-1/2}(\hat{\boldsymbol{\alpha}}_i - \mathbf{D}_i^{1/2}\boldsymbol{\beta}_i)\right|_\infty \leq 2\lambda_{ni}(\delta)$$

with probability tending to one. By the proof of Theorem 7.1 in Bickel, Ritov and Tsybakov (2009), we can get that (46) and (47) hold for $\hat{\boldsymbol{\beta}}_i = \hat{\boldsymbol{\beta}}_i(\delta)$. □

Acknowledgments. The author would like to thank the Associate Editor and three referees for their helpful constructive comments which have helped to improve quality and presentation of the paper.

SUPPLEMENTARY MATERIAL

Supplement to "Gaussian graphical model estimation with false discovery rate control" (DOI: 10.1214/13-AOS1169SUPP; .pdf). This supplemental material includes additional numerical results for GFC-Dantzig and GFC-Lasso.

REFERENCES

ANDERSON, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*, 3rd ed. Wiley, Hoboken, NJ. MR1990662

BELLONI, A., CHERNOZHUKOV, V. and WANG, L. (2011). Square-root lasso: Pivotal recovery of sparse signals via conic programming. *Biometrika* **98** 791–806. MR2860324

BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **57** 289–300. MR1325392

BERMAN, S. M. (1962). A law of large numbers for the maximum in a stationary Gaussian sequence. *Ann. Math. Statist.* **33** 93–97. MR0133856

BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732. MR2533469

BÜHLMANN, P. (2013). Statistical significance in high-dimensional linear models. *Bernoulli* **19** 1212–1242. MR3102549

CAI, T. and LIU, W. (2011). Adaptive thresholding for sparse covariance matrix estimation. *J. Amer. Statist. Assoc.* **106** 672–684. MR2847949

CAI, T., LIU, W. and LUO, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *J. Amer. Statist. Assoc.* **106** 594–607. MR2847973

CAI, T. T., LIU, W. and XIA, Y. (2013). Two-sample covariance matrix testing and support recovery in high-dimensional and sparse settings. *J. Amer. Statist. Assoc.* **108** 265–277.

CANDES, E. and TAO, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.* **35** 2313–2351. MR2382644

D'ASPREMONT, A., BANERJEE, O. and EL GHAOUI, L. (2008). First-order methods for sparse covariance selection. *SIAM J. Matrix Anal. Appl.* **30** 56–66. MR2399568

DRTON, M. and PERLMAN, M. D. (2004). Model selection for Gaussian concentration graphs. *Biometrika* **91** 591–602. MR2090624

FAN, J., FENG, Y. and WU, Y. (2009). Network exploration via the adaptive lasso and SCAD penalties. *Ann. Appl. Stat.* **3** 521–541. MR2750671

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9** 432–441.

JAVANMARD, A. and MONTANARI, A. (2013). Hypothesis testing in high-dimensional regression under the gaussian random design model: Asymptotic theory. Technical report. Available at arXiv:1301.4240.

LAURITZEN, S. L. (1996). *Graphical Models*. *Oxford Statistical Science Series* **17**. Oxford Univ. Press, New York. MR1419991

LIU, W. (2013). Supplement to "Gaussian graphical model estimation with false discovery rate control." DOI:10.1214/13-AOS1169SUPP.

LIU, W. and SHAO, Q. M. (2012). A robust and powerful approach on control of false discovery rate under dependence. Technical report.

LIU, H., HAN, F., YUAN, M., LAFFERTY, J. and WASSERMAN, L. (2012). High-dimensional semiparametric Gaussian copula graphical models. *Ann. Statist.* **40** 2293–2326. MR3059084

MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34** 1436–1462. MR2278363

MEINSHAUSEN, N. and BÜHLMANN, P. (2010). Stability selection. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **72** 417–473. MR2758523

RAVIKUMAR, P., WAINWRIGHT, M. J., RASKUTTI, G. and YU, B. (2011). High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electron. J. Stat.* **5** 935–980. MR2836766

REN, Z., SUN, T., ZHANG, C. H. and ZHOU, H. H. (2013). Asymptotic normality and optimalities in estimation of large Gaussian graphical model. Technical report. Available at http://www.stat.yale.edu/~hz68/InferenceGGM.pdf.

ROTHMAN, A. J., BICKEL, P. J., LEVINA, E. and ZHU, J. (2008). Sparse permutation invariant covariance estimation. *Electron. J. Stat.* **2** 494–515. MR2417391

SUN, T. and ZHANG, C.-H. (2012a). Comment: "Minimax estimation of large covariance matrices under ℓ_1-norm" [MR3027084]. *Statist. Sinica* **22** 1354–1358. MR3027086

SUN, T. and ZHANG, C.-H. (2012b). Scaled sparse linear regression. *Biometrika* **99** 879–898. MR2999166

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **58** 267–288. MR1379242

VAN DE GEER, S., BÜHLMANN, P. RITOV, Y. and DEZEURE, R. (2013). On asymptotically optimal confidence regions and tests for high-dimensional models. Technical report. Available at arXiv:1303.0518.

XUE, L. and ZOU, H. (2012). Regularized rank-based estimation of high-dimensional nonparanormal graphical models. *Ann. Statist.* **40** 2541–2571. MR3097612

YUAN, M. (2010). High dimensional inverse covariance matrix estimation via linear programming. *J. Mach. Learn. Res.* **11** 2261–2286. MR2719856

YUAN, M. and LIN, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika* **94** 19–35. MR2367824

ZAÏTSEV, A. Y. (1987). On the Gaussian approximation of convolutions under multidimensional analogues of S. N. Bernstein's inequality conditions. *Probab. Theory Related Fields* **74** 535–566. MR0876255

ZHANG, C. (2010). Estimation of large inverse matrices and graphical model selection. Technical report. Dept. Statistics and Biostatistics, Rutgers Univ.

ZHANG, C. H. and ZHANG, S. S. (2011). Confidence intervals for low-dimensional parameters with highdimensional data. Technical report. Available at arXiv:1110.2563.

DEPARTMENT OF MATHEMATICS
INSTITUTE OF NATURAL SCIENCES AND MOE-LSC
SHANGHAI JIAO TONG UNIVERSITY
SHANGHAI
CHINA
E-MAIL: weidongl@sjtu.edu.cn