Graph Estimation for Matrix-variate Gaussian Data

Xi Chen and Weidong Liu

Department of Information, Operations & Management Sciences, Stern School of Business
New York University
and Department of Mathematics, Institute of Natural Sciences and MOE-LSC
Shanghai Jiao Tong University

Supplementary Material

A Technical Lemmas

We introduce several technical lemmas, which will be used throughout the proofs of our propositions and theorems. In particular, let $\mathbf{\Psi} = (\psi_{ij})_{q\times q}$ and recall the definition of $\widehat{\mathbf{\Psi}} = (\widehat{\psi}_{ij})_{q\times q}$ from Section 3.2. We first prove a maximal concentration inequality on $\widehat{\psi}_{ij}$ in Lemma 1. Note that this result differs from the standard concentration inequality on the sample covariance matrix with *i.i.d.* samples since the "row samples" for constructing $\widehat{\mathbf{\Psi}}$ are correlated.

Lemma 1 *We have for any $M > 0$, there exists a constant C such that*

$$\mathbb{P}\Big(\max_{1\le i\le j\le q}\Big|\widehat{\psi}_{ij} - \frac{\mathrm{tr}(\mathbf{\Sigma})}{p}\psi_{ij}\Big| \ge C\sqrt{\frac{\log\max(q,np)}{np}}\Big) = O((q+np)^{-M}).$$

Proof. Recall that for any pair of $i \in [q]$ and $j \in [q]$

$$\begin{aligned}\widehat{\psi}_{ij} &= \frac{1}{(n-1)p}\sum_{k=1}^{n}\sum_{l=1}^{p}(X_{li}^{(k)} - \bar{X}_{li})(X_{lj}^{(k)} - \bar{X}_{lj}) \\ &= \frac{1}{(n-1)p}\sum_{l=1}^{p}\sum_{k=1}^{n}(X_{li}^{(k)} - \bar{X}_{li})(X_{lj}^{(k)} - \bar{X}_{lj}), \end{aligned} \tag{1}$$

where $\bar{X}_{li} = \frac{1}{n}\sum_{k=1}^{n} X_{li}^{(k)}$ and $\bar{X}_{lj} = \frac{1}{n}\sum_{k=1}^{n} X_{lj}^{(k)}$. Without loss of generality, we assume that $\boldsymbol{\mu} = 0$.

Let $\mathbf{A} \in \mathrm{R}^{n\times n}$ be an orthogonal matrix with the last row $(\frac{1}{\sqrt{n}},\ldots,\frac{1}{\sqrt{n}})$. Let $\mathbf{Y}_{li} = (Y_{li}^{(1)},\ldots,Y_{li}^{(n)})' = \mathbf{A}(X_{li}^{(1)},\ldots,X_{li}^{(n)})' \in \mathbb{R}^{n\times 1}$. So we have $\sqrt{n}\bar{X}_{li} = Y_{li}^{(n)}$ and

$$\sum_{k=1}^{n}(X_{li}^{(k)} - \bar{X}_{li})(X_{lj}^{(k)} - \bar{X}_{lj}) = \mathbf{Y}_{li}'\mathbf{Y}_{lj} - Y_{li}^{(n)}Y_{lj}^{(n)} = \sum_{k=1}^{n-1} Y_{li}^{(k)}Y_{lj}^{(k)}. \tag{2}$$

Since $(X_{li}^{(1)},\ldots,X_{li}^{(n)})^{'} \sim N(\mathbf{0},\sigma_{ll}\psi_{ii}\mathbf{I}_{n\times n})$, $(Y_{li}^{(1)},\ldots,Y_{li}^{(n-1)})^{'} \sim N(\mathbf{0},\sigma_{ll}\psi_{ii}\mathbf{I}_{(n-1)\times(n-1)})$. Let $\mathbf{Y}_k = (Y_{li}^{(k)})_{1\le l\le p,1\le i\le q}$ for $1 \le k \le n-1$. We have $\mathbf{Y}_k \sim N(\mathbf{0},\mathbf{\Sigma}\otimes\mathbf{\Psi})$ and $\mathbf{Y}_k$, $1 \le k \le n-1$, are independent. Let us define $\mathbf{Z}_k = \mathbf{\Sigma}^{-1/2}\mathbf{Y}_k \sim N(\mathbf{0},\mathbf{I}_{p\times p}\otimes\mathbf{\Psi})$. Let $\mathbf{Z}_{ki}$ be the i-th column of $\mathbf{Y}_k$. Then $(\mathbf{Z}_{ki},\mathbf{Z}_{kj}) \sim N(0,\mathbf{I}_{p\times p}\otimes\mathbf{\Psi}_{[i,j]})$, where $\mathbf{\Psi}_{[i,j]} = \begin{pmatrix} \varphi_{ii} & \varphi_{ij} \\ \varphi_{ji} & \varphi_{jj} \end{pmatrix}$.

Let the $\mathbf{U}^{'}\mathbf{D}\mathbf{U}$ be the eigenvalue decomposition of $\mathbf{\Sigma}$, where $\mathbf{U}$ is an orthogonal matrix and $\mathbf{D} = \mathrm{diag}(\lambda_1^{(1)},\ldots,\lambda_p^{(1)})$. Define $(\mathbf{W}_{ki},\mathbf{W}_{kj}) := (\mathbf{U}\mathbf{Z}_{ki},\mathbf{U}\mathbf{Z}_{kj}) \in \mathbb{R}^{p\times 2}$ where $\mathbf{W}_{ki} = (w_{ki,1},\ldots,w_{ki,p})^{'} \in \mathbb{R}^{p\times 1}$ and $\mathbf{W}_{kj} = (w_{kj,1},\ldots,w_{kj,p})^{'} \in \mathbb{R}^{p\times 1}$. Since $\mathbf{U}$ is an orthogonal matrix, $(\mathbf{W}_{ki},\mathbf{W}_{kj}) \sim N(\mathbf{0},\mathbf{I}_{p\times p}\otimes\mathbf{\Psi}_{[i,j]})$, which also implies that $(w_{ki,l},w_{kj,l})$ are independent for $1 \le l \le p$.

Now combining (1) and (2), we have,

$$\begin{aligned}
\widehat{\psi}_{ij} &= \frac{1}{(n-1)p}\sum_{l=1}^{p}\sum_{k=1}^{n-1}\mathbf{Y}_{ki}'\mathbf{Y}_{ki} \\
&= \frac{1}{(n-1)p}\sum_{k=1}^{n-1}(\mathbf{U}\mathbf{Z}_{ki})^{'}\mathbf{D}\mathbf{U}\mathbf{Z}_{kj} \\
&= \frac{1}{(n-1)p}\sum_{k=1}^{n-1}\sum_{l=1}^{p}\lambda_l^{(1)}w_{ki,l}w_{kj,l}. \qquad (3)
\end{aligned}$$

We further note that

$$\mathbb{E}\widehat{\psi}_{ij} = \frac{1}{(n-1)p}\sum_{k=1}^{n-1}\sum_{l=1}^{p}\lambda_l^{(1)}\psi_{ij} = \frac{\mathrm{tr}(\mathbf{\Sigma})}{p}\psi_{ij}.$$

Put $w_{kij,l} = w_{ki,l}w_{kj,l} - \mathbb{E}w_{ki,l}w_{kj,l}$. We have for some $\eta > 0$ such that $\mathbb{E}e^{2\eta|\lambda_l^{(1)}w_{kij,l}|} \le K$ for some $K > 0$, uniformly in i,j,l,k. It implies that

$$\begin{aligned}
\sum_{k=1}^{n-1}\sum_{l=1}^{p}\mathbb{E}(\lambda_l^{(1)}w_{kij,l})^2e^{\eta|\lambda_l^{(1)}w_{kij,l}|} &\le K\sum_{k=1}^{n-1}\sum_{l=1}^{p}(\mathbb{E}(\lambda_l^{(1)}w_{kij,l})^4)^{1/2} \\
&= \sqrt{2}K(n-1)\|\mathbf{\Sigma}\|_F^2(\varphi_{ii}\varphi_{jj}+\varphi_{ij}^2).
\end{aligned}$$

By the exponential inequality in Lemma 1 in Cai and Liu (2011) and $\|\mathbf{\Sigma}\|_F^2/p \le \lambda_p^{(1)}$, for any $M > 0$, there exists a constant $C > 0$,

$$\mathbb{P}\Big(|\widehat{\psi}_{ij} - \mathbb{E}\widehat{\psi}_{ij}| \ge C\sqrt{\frac{\log\max(q,np)}{np}}\Big) = O((q+np)^{-M}).$$

This proves Lemma 1. ∎

The next concentration inequality involves the residuals. In particular, for $k \in [n]$, $l \in [p]$ and $j \in [q]$, we define:

$$\widetilde{\varepsilon}_{lj}^{(k)} = \varepsilon_{lj}^{(k)} - \frac{1}{n}\sum_{k=1}^{n}\varepsilon_{lj}^{(k)} =: \varepsilon_{lj}^{(k)} - \bar{\varepsilon}_{lj}, \quad \widehat{\sigma}_{jj,\varepsilon} = \frac{1}{(n-1)p}\sum_{k=1}^{n}\sum_{l=1}^{p}\Big(\widetilde{\varepsilon}_{lj}^{(k)}\Big)^2.$$

Lemma 2 *For any $M > 0$, there exists a constant C such that*

$$\mathbb{P}\Big(\max_{1\le i\le q} \max_{1\le h\le q, h\neq i} \Big| \frac{1}{np} \sum_{k=1}^{n} \sum_{l=1}^{p} \widetilde{\varepsilon}_{li}^{(k)} (X_{lh}^{(k)} - \bar{X}_{lh}) \Big| \ge C \sqrt{\frac{\log \max(q, np)}{np}} \Big) = O((q+np)^{-M})$$

and

$$\mathbb{P}\Big(\max_{1\le i\le p} \Big| \frac{1}{np} \sum_{k=1}^{n} \sum_{l=1}^{p} \widetilde{\varepsilon}_{li}^{(k)} (\boldsymbol{X}_{l,-i}^{(k)} - \bar{\boldsymbol{X}}_{l,-i}) \boldsymbol{\beta}_i \Big| \ge C \sqrt{\frac{\log \max(q, np)}{np}} \Big) = O((q+np)^{-M}).$$

Proof. Recall that

$$\varepsilon_{li}^{(k)} = X_{li}^{(k)} - \alpha_{li} - \mathbf{X}_{l,-i}^{(k)} \boldsymbol{\beta}_i.$$

Set $\boldsymbol{\varepsilon}_i^{(k)} = (\varepsilon_{1i}^{(k)}, \ldots, \varepsilon_{pi}^{(k)})'$ and $\mathbf{Y}_i^{(k)} = (\mathbf{X}_{1,-i}^{(k)} \boldsymbol{\beta}_i, \ldots, \mathbf{X}_{p,-i}^{(k)} \boldsymbol{\beta}_i)'$. It is easy to see that $\mathrm{Cov}(\boldsymbol{\varepsilon}_i^{(k)}) = \gamma_{ii}^{-1} \boldsymbol{\Sigma}$. Let $\mathbf{X}_h^{(k)}$ be the h-th column of $\mathbf{X}^{(k)}$. Since $\boldsymbol{\varepsilon}_i^{(k)}$ and $\mathbf{X}_h^{(k)}$ are independent for $h \neq i$, for the $p \times 2$ matrix $(\boldsymbol{\varepsilon}_i^{(k)}, \mathbf{X}_h^{(k)})$, we have $\mathrm{Cov}((\boldsymbol{\varepsilon}_i^{(k)}, \mathbf{X}_h^{(k)})) = \boldsymbol{\Sigma} \otimes \mathrm{diag}(\gamma_{ii}^{-1}, \psi_{ii})$ for $h \neq i$.

In addition, $\mathbf{X}_{\cdot,-i}^{(k)} \sim N(\mathbf{0}, \boldsymbol{\Sigma} \otimes \boldsymbol{\Psi}_{-i,-i})$ and $\boldsymbol{\beta}_i = -\frac{1}{\gamma_{ii}} \boldsymbol{\Gamma}_{-i,i}$, we have

$$\mathrm{Cov}(\mathbf{Y}_i^{(k)}) = \frac{1}{\gamma_{ii}^2} \mathrm{tr}(\boldsymbol{\Gamma}_{-i,i} \boldsymbol{\Gamma}_{i,-i} \boldsymbol{\Psi}_{-i,-i}) \boldsymbol{\Sigma}.$$

Further, by the fact that,

$$\mathrm{tr}(\boldsymbol{\Gamma}_{-i,i} \boldsymbol{\Gamma}_{i,-i} \boldsymbol{\Psi}_{-i,-i}) = \boldsymbol{\Gamma}_{i,-i} \boldsymbol{\Psi}_{-i,-i} \boldsymbol{\Gamma}_{-i,i} = -\gamma_{ii} \boldsymbol{\Psi}_{i,-i} \boldsymbol{\Gamma}_{-i,i} = -\gamma_{ii}(1 - \psi_{ii}\gamma_{ii}),$$

which further implies that,

$$\mathrm{Cov}(\mathbf{Y}_i^{(k)}) = \frac{\psi_{ii}\gamma_{ii} - 1}{\gamma_{ii}} \boldsymbol{\Sigma}.$$

Since $\boldsymbol{\varepsilon}_i^{(k)}$ and $\mathbf{Y}_i^{(k)}$ are independent, $\mathrm{Cov}((\boldsymbol{\varepsilon}_i^{(k)}, \mathbf{Y}_i^{(k)})) = \boldsymbol{\Sigma} \otimes \mathrm{diag}(\gamma_{ii}^{-1}, (\psi_{ii}\gamma_{ii} - 1)/\gamma_{ii})$. Following exactly the same proof of Lemma 1, where we replace $(X_{li}^{(k)}, X_{lj}^{(k)})$ by $(\varepsilon_{li}^{(k)}, X_{lh}^{(k)})$ or $(\varepsilon_{li}^{(k)}, \mathbf{X}_{l,-i}^{(k)} \boldsymbol{\beta}_i)$ and $\boldsymbol{\Psi}_{[i,j]}$ by $\mathrm{diag}(\gamma_{ii}^{-1}, \psi_{ii})$ or $\mathrm{diag}(\gamma_{ii}^{-1}, (\psi_{ii}\gamma_{ii} - 1)/\gamma_{ii})$, we can obtain Lemma 2 immediately. ∎

Lemma 3 *(i). We have, as $np \to \infty$,*

$$\frac{\sum_{k=1}^{n} \sum_{l=1}^{p} (\widetilde{\varepsilon}_{li}^{(k)} \widetilde{\varepsilon}_{lj}^{(k)} - \mathbb{E}\widetilde{\varepsilon}_{li}^{(k)} \widetilde{\varepsilon}_{lj}^{(k)})}{\sqrt{n_1 \|\boldsymbol{\Sigma}\|_F^2}} \to N\Big(0, \frac{1}{\gamma_{ii}\gamma_{jj}} + \frac{\gamma_{ij}^2}{(\gamma_{ii}\gamma_{jj})^2}\Big)$$

in distribution.

(ii). For any $M > 0$, there exists a constant C such that

$$\mathbb{P}\Big(\max_{1\le i\le j\le q} \Big| \widehat{\sigma}_{ij,\varepsilon} - \frac{\mathrm{tr}(\boldsymbol{\Sigma})}{p} \frac{\gamma_{ij}}{\gamma_{ii}\gamma_{jj}} \Big| \ge C \sqrt{\frac{\log \max(q, np)}{np}} \Big) = O((q+np)^{-M}).$$

Proof. Note that $\mathrm{Cov}(\varepsilon_{li}^{(k)}, \varepsilon_{li}^{(k)}) = \frac{\gamma_{ij}}{\gamma_{ii}\gamma_{jj}}$. It is easy to show that $\mathrm{Cov}((\boldsymbol{\varepsilon}_i^{(k)}, \boldsymbol{\varepsilon}_j^{(k)})) = \boldsymbol{\Sigma} \otimes \Delta_{[i,j]}$, where $\Delta_{[i,j]} =$

$\begin{pmatrix} \frac{1}{\gamma_{ii}} & \frac{\gamma_{ij}}{\gamma_{ii}\gamma_{jj}} \\ \frac{\gamma_{ji}}{\gamma_{ii}\gamma_{jj}} & \frac{1}{\gamma_{jj}} \end{pmatrix}$. As in the proof of Lemma 1, we can write

$$\sum_{k=1}^{n}\sum_{l=1}^{p}\widetilde{\varepsilon}_{li}^{(k)}\widetilde{\varepsilon}_{lj}^{(k)} = \sum_{k=1}^{n-1}\sum_{l=1}^{p}\lambda_l^{(1)}\eta_{ki,l}\eta_{kj,l}, \tag{4}$$

where $(\eta_{ki,l},\eta_{kj,l})$, $1\le l\le p$, $1\le k\le n-1$, are i.i.d. $N(0,\Delta_{[i,j]})$ random vectors. Note that $\mathrm{Var}(\eta_{ki,l}\eta_{kj,l}) = \frac{1}{\gamma_{ii}\gamma_{jj}} + \frac{\gamma_{ij}^2}{(\gamma_{ii}\gamma_{jj})^2}$ and $\sum_{l=1}^{p}(\lambda_l^{(1)})^2 = \|\mathbf{\Sigma}\|_{\mathrm{F}}^2$. (i) follows from Lindeberg-Feller central limit theorem. (ii) follows from the exponential inequality in Lemma 1 in Cai and Liu (2011). ∎

B Proof of Proposition 1–3 on the properties of the proposed test statistics

Proof of Proposition 1. For notational simplicity, let $n_1 = n-1$ and $\zeta_{lji}^{(k)} = \widetilde{\varepsilon}_{lj}^{(k)} + (X_{li}-\bar{X}_{li})\beta_{i,j}$. Note that for all $k\in[n]$ and $i\neq j$,

$$\widehat{\varepsilon}_{lji}^{(k)} = \widetilde{\varepsilon}_{lj}^{(k)} + (X_{li}-\bar{X}_{li})\beta_{i,j} - (\mathbf{X}_{l,-j}^{(k)} - \bar{\mathbf{X}}_{l,-j})(\widehat{\boldsymbol{\beta}}_{j,\backslash i} - \boldsymbol{\beta}_{j,\backslash i}),$$

which implies that

$$\begin{aligned}\widehat{\varepsilon}_{lij}^{(k)}\widehat{\varepsilon}_{lji}^{(k)} &= \zeta_{lij}^{(k)}\zeta_{lji}^{(k)} - \zeta_{lij}^{(k)}(\mathbf{X}_{l,-j}^{(k)} - \bar{\mathbf{X}}_{l,-j})(\widehat{\boldsymbol{\beta}}_{j,\backslash i} - \boldsymbol{\beta}_{j,\backslash i}) \\ &\quad -\zeta_{lji}^{(k)}(\mathbf{X}_{l,-i}^{(k)} - \bar{\mathbf{X}}_{l,-i})(\widehat{\boldsymbol{\beta}}_{i,\backslash j} - \boldsymbol{\beta}_{i,\backslash j}) \\ &\quad +(\widehat{\boldsymbol{\beta}}_{i,\backslash j} - \boldsymbol{\beta}_{i,\backslash j})'(\mathbf{X}_{l,-i}^{(k)} - \bar{\mathbf{X}}_{l,-i})'(\mathbf{X}_{l,-j}^{(k)} - \bar{\mathbf{X}}_{l,-j})(\widehat{\boldsymbol{\beta}}_{j,\backslash i} - \boldsymbol{\beta}_{j,\backslash i}).\end{aligned} \tag{5}$$

Let $\sigma = \mathrm{tr}(\mathbf{\Sigma})/p$. By the assumption (C1), we have $c^{-1}\le\sigma\le c$. For the last term in (5), by Cauchy-Schwarz inequality, we have

$$\begin{aligned}&\left|\frac{1}{n_1 p}\sum_{k=1}^{n}\sum_{l=1}^{p}(\widehat{\boldsymbol{\beta}}_{i,\backslash j} - \boldsymbol{\beta}_{i,\backslash j})'(\mathbf{X}_{l,-i}^{(k)} - \bar{\mathbf{X}}_{l,-i})'(\mathbf{X}_{l,-j}^{(k)} - \bar{\mathbf{X}}_{l,-j})(\widehat{\boldsymbol{\beta}}_{j,\backslash i} - \boldsymbol{\beta}_{j,\backslash i})\right| \\ &\le \max_{1\le i,j\le p}\left|(\widehat{\boldsymbol{\beta}}_{i,\backslash j} - \boldsymbol{\beta}_{i,\backslash j})'\widehat{\mathbf{\Psi}}_{-i,-i}(\widehat{\boldsymbol{\beta}}_{i,\backslash j} - \boldsymbol{\beta}_{i,\backslash j})\right|.\end{aligned}$$

For any $i,j\in[q]$, we have

$$\begin{aligned}|(\widehat{\boldsymbol{\beta}}_{i,\backslash j} - \boldsymbol{\beta}_{i,\backslash j})'\widehat{\mathbf{\Psi}}_{-i,-i}(\widehat{\boldsymbol{\beta}}_{i,\backslash j} - \boldsymbol{\beta}_{i,\backslash j})| &\le |(\widehat{\boldsymbol{\beta}}_{i,\backslash j} - \boldsymbol{\beta}_{i,\backslash j})'(\widehat{\mathbf{\Psi}}_{-i,-i} - \sigma\mathbf{\Psi}_{-i,-i})(\widehat{\boldsymbol{\beta}}_{i,\backslash j} - \boldsymbol{\beta}_{i,\backslash j})| \\ &\quad +\sigma|(\widehat{\boldsymbol{\beta}}_{i,\backslash j} - \boldsymbol{\beta}_{i,\backslash j})'\mathbf{\Psi}_{-i,-i}(\widehat{\boldsymbol{\beta}}_{i,\backslash j} - \boldsymbol{\beta}_{i,\backslash j})|.\end{aligned}$$

By Lemma 1,

$$\max_{1\le i,j\le q}|(\widehat{\boldsymbol{\beta}}_{i,\backslash j} - \boldsymbol{\beta}_{i,\backslash j})'(\widehat{\mathbf{\Psi}}_{-i,-i} - \sigma\mathbf{\Psi}_{-i,-i})(\widehat{\boldsymbol{\beta}}_{i,\backslash j} - \boldsymbol{\beta}_{i,\backslash j})| = O_{\mathbb{P}}\Big(a_{n1}^2\sqrt{\frac{\log\max(q,np)}{np}}\Big).$$

Moreover,

$$|(\widehat{\boldsymbol{\beta}}_{i,\setminus j}-\boldsymbol{\beta}_{i,\setminus j})^{'}\boldsymbol{\Psi}_{-i,-i}(\widehat{\boldsymbol{\beta}}_{i,\setminus j}-\boldsymbol{\beta}_{i,\setminus j})|=O_{\mathbb{P}}(\lambda_{\max}(\boldsymbol{\Psi})|\widehat{\boldsymbol{\beta}}_i-\boldsymbol{\beta}_i|_2^2)$$

uniformly in $i\in[q]$. Combining the above arguments,

$$\begin{aligned}&\left|\frac{1}{n_1p}\sum_{k=1}^{n}\sum_{l=1}^{p}(\widehat{\boldsymbol{\beta}}_{i,\setminus j}-\boldsymbol{\beta}_{i,\setminus j})^{'}(\mathbf{X}_{l,-i}^{(k)}-\bar{\mathbf{X}}_{l,-i})^{'}(\mathbf{X}_{l,-j}^{(k)}-\bar{\mathbf{X}}_{l,-j})(\widehat{\boldsymbol{\beta}}_{j,\setminus i}-\boldsymbol{\beta}_{j,\setminus i})\right|\\&=O_{\mathbb{P}}\Big(a_{n2}^2+a_{n1}^2\sqrt{\frac{\log\max(q,np)}{np}}\Big).\end{aligned}\tag{6}$$

Under the null $H_{0ij}:\gamma_{ij}=0$, we have $\zeta_{lji}^{(k)}=\widetilde{\varepsilon}_{lj}^{(k)}$. Note that

$$(\mathbf{X}_{l,-j}^{(k)}-\bar{\mathbf{X}}_{l,-j})(\widehat{\boldsymbol{\beta}}_{j,\setminus i}-\boldsymbol{\beta}_{j,\setminus i})=(\mathbf{X}_{l,-\{i,j\}}^{(k)}-\bar{\mathbf{X}}_{l,-\{i,j\}})(\widehat{\boldsymbol{\beta}}_{j,-i}-\boldsymbol{\beta}_{j,-i}).\tag{7}$$

So

$$\zeta_{lij}^{(k)}(\mathbf{X}_{l,-j}^{(k)}-\bar{\mathbf{X}}_{l,-j})(\widehat{\boldsymbol{\beta}}_{j,\setminus i}-\boldsymbol{\beta}_{j,\setminus i})\quad=\quad\sum_{h\neq i,j}\widetilde{\varepsilon}_{li}^{(k)}(X_{lh}^{(k)}-\bar{X}_{lh})(\widehat{\beta}_{h,j}-\beta_{h,j}),$$

where $\widehat{\boldsymbol{\beta}}_j=(\widehat{\beta}_{1,j},\ldots,\widehat{\beta}_{p-1,j})^{'}$ and we set $\widehat{\beta}_{p,j}=0$. By Lemma 2 (i),

$$\begin{aligned}&\max_{1\le i\le j\le q}\Big|\sum_{h\neq i,j}\frac{1}{n_1p}\sum_{k=1}^{n}\sum_{l=1}^{p}\widetilde{\varepsilon}_{li}^{(k)}(X_{lh}^{(k)}-\bar{X}_{lh})(\widehat{\beta}_{h,j}-\beta_{h,j})\Big|\\\le\;&\max_{1\le i\le j\le q}\max_{h\neq i,j}\left|\frac{1}{n_1p}\sum_{k=1}^{n}\sum_{l=1}^{p}\widetilde{\varepsilon}_{li}^{(k)}(X_{lh}^{(k)}-\bar{X}_{lh})\right||\widehat{\boldsymbol{\beta}}_j-\boldsymbol{\beta}_j|_1\\=\;&O_{\mathbb{P}}\Big(a_{n1}\sqrt{\frac{\log\max(q,np)}{np}}\Big).\end{aligned}\tag{8}$$

A similar inequality holds for the third term on the right hand side of (5). Therefore, for $i\neq j$, under $\gamma_{ij}=0$,

$$\begin{aligned}\frac{1}{n_1p}\sum_{k=1}^{n}\sum_{l=1}^{p}\widehat{\varepsilon}_{lij}^{(k)}\widehat{\varepsilon}_{lji}^{(k)}\quad=\quad&\frac{1}{n_1p}\sum_{k=1}^{n}\sum_{l=1}^{p}\widetilde{\varepsilon}_{li}^{(k)}\widetilde{\varepsilon}_{lj}^{(k)}\\&+O_{\mathbb{P}}\Big((a_{n1}^2+a_{n1})\sqrt{\frac{\log\max(q,np)}{np}}+a_{n2}^2\Big)\end{aligned}\tag{9}$$

uniformly in $1\le i\neq j\le q$. By (6) and (8) with $i=j$, we obtain that

$$\widehat{r}_{ii}=\frac{1}{n_1p}\sum_{k=1}^{n}\sum_{l=1}^{p}(\widetilde{\varepsilon}_{li}^{(k)})^2+O_{\mathbb{P}}\Big((a_{n1}^2+a_{n1})\sqrt{\frac{\log\max(q,np)}{np}}+a_{n2}^2\Big)\tag{10}$$

uniformly in $1\le i\le q$. The proof of Proposition 1 is complete by Lemma 3. ∎

Proof of Proposition 2. By Lemma 1, it is easy to show that

$$\frac{1}{n_1 p}\sum_{k=1}^{n}\sum_{l=1}^{p}(X_{li}-\bar{X}_{li})(\mathbf{X}_{l,-i}^{(k)}-\bar{\mathbf{X}}_{l,-i})(\widehat{\boldsymbol{\beta}}_{i,\backslash j}-\boldsymbol{\beta}_{i,\backslash j})=O_{\mathbb{P}}(a_{n1}) \tag{11}$$

uniformly in i,j. Also, by (7) and (8),

$$\frac{1}{n_1 p}\sum_{k=1}^{n}\sum_{l=1}^{p}\widetilde{\varepsilon}_{li}^{(k)}(\mathbf{X}_{l,-j}^{(k)}-\bar{\mathbf{X}}_{l,-j})(\widehat{\boldsymbol{\beta}}_{j,\backslash i}-\boldsymbol{\beta}_{j,\backslash i})=O_{\mathbb{P}}\Big(a_{n1}\sqrt{\frac{\log\max(q,np)}{np}}\Big). \tag{12}$$

By (6), (10), (11) and (12), it suffices to prove that

$$\frac{p}{\mathrm{tr}(\boldsymbol{\Sigma})}\frac{\sqrt{\gamma_{ii}\gamma_{jj}}}{n_1 p}\sum_{k=1}^{n}\sum_{l=1}^{p}\zeta_{lij}^{(k)}\zeta_{lji}^{(k)}-(1-\gamma_{ij}\psi_{ij})\rho_{ij\cdot}^{\boldsymbol{\Gamma}}\to 0 \tag{13}$$

in probability. We have

$$\mathbb{E}\Big[(\varepsilon_{lj}^{(k)}+X_{li}\beta_{i,j})(\varepsilon_{li}^{(k)}+X_{lj}\beta_{i,j})\Big]=\sigma_{ll}(-\frac{\gamma_{ij}}{\gamma_{ii}\gamma_{jj}}+\psi_{ij}\frac{\gamma_{ij}^2}{\gamma_{ii}\gamma_{jj}}).$$

Now, as the proof of Lemma 1, we can write

$$\sum_{k=1}^{n}\sum_{l=1}^{p}\zeta_{lij}^{(k)}\zeta_{lji}^{(k)}=\sum_{k=1}^{n-1}\sum_{l=1}^{p}\lambda_l^{(1)}\xi_{ki,l}\xi_{kj,l}, \tag{14}$$

where $(\xi_{ki,l},\xi_{kj,l})$, $1\le l\le p$, are i.i.d. with $\mathbb{E}\xi_{ki,l}\xi_{kj,l}=(-\frac{\gamma_{ij}}{\gamma_{ii}\gamma_{jj}}+\psi_{ij}\frac{\gamma_{ij}^2}{\gamma_{ii}\gamma_{jj}})$. This proves (13). ∎

Proof of Proposition 3. Let $\phi_{ij}=\frac{\mathrm{tr}(\boldsymbol{\Gamma})}{q}\sigma_{ij}$ and define $\widetilde{\lambda}=\lambda\sqrt{\frac{\log\max(p,nq)}{nq}}$. We have

$$\sum_{j=1}^{p}\widehat{\sigma}_{ij,\lambda}^2=\sum_{j=1}^{p}(\widehat{\sigma}_{ij}^2-\phi_{ij}^2)I\{|\widehat{\sigma}_{ij}|\ge\widetilde{\lambda}\}+\sum_{j=1}^{p}\phi_{ij}^2 I\{|\widehat{\sigma}_{ij}|\ge\widetilde{\lambda}\}.$$

Also by Lemma 1, with probability tending to one,

$$\sum_{j=1}^{p}\phi_{ij}^2 I\{|\widehat{\sigma}_{ij}|<\widetilde{\lambda}\}\le\sum_{j=1}^{p}\phi_{ij}^2 I\{|\phi_{ij}|<2\widetilde{\lambda}\}=O(\widetilde{\lambda}^{2-\tau}s(p))$$

uniformly in $1\le i\le p$ by the assumption (C2). So, for the last term,

$$\sum_{j=1}^{p}\phi_{ij}^2 I\{|\widehat{\sigma}_{ij}|\ge\widetilde{\lambda}\}=\sum_{j=1}^{p}\phi_{ij}^2-\sum_{j=1}^{p}\phi_{ij}^2 I\{|\widehat{\sigma}_{ij}|<\widetilde{\lambda}\}=(1+O_{\mathbb{P}}(\widetilde{\lambda}^{2-\tau}s(p)))\sum_{j=1}^{p}\phi_{ij}^2$$

uniformly in $1\le i\le p$. Moreover, with probability tending to one,

$$\sum_{j=1}^{p}|\widehat{\sigma}_{ij}^2-\phi_{ij}^2|I\{|\widehat{\sigma}_{ij}|\ge\widetilde{\lambda}\}\quad=\quad\sum_{j=1}^{p}|\widehat{\sigma}_{ij}+\phi_{ij}||\widehat{\sigma}_{ij}-\phi_{ij}|I\{|\widehat{\sigma}_{ij}|\ge\widetilde{\lambda}\}$$

$$\begin{aligned} &\leq C\sum_{j=1}^{p}|\phi_{ij}||\widehat{\sigma}_{ij}-\phi_{ij}|I\{|\phi_{ij}|\geq\widetilde{\lambda}/2\} \\ &\leq C\widetilde{\lambda}^{-(\tau-1)\vee 0}\sum_{j=1}^{p}|\phi_{ij}|^{\tau}|\widehat{\sigma}_{ij}-\phi_{ij}| \\ &\leq C\widetilde{\lambda}^{(2-\tau)\wedge 1}s(p), \end{aligned}$$

where the last inequality following from $\max_{1\leq i\leq j\leq q}|\widehat{\sigma}_{ij}-\phi_{ij}|=O_{\mathbb{P}}(\widetilde{\lambda})$ by Lemma 1. It implies that $\max_{1\leq i\leq q}|\sum_{j=1}^{p}\widehat{\sigma}_{ij,\lambda}^2-\sum_{j=1}^{p}\phi_{ij}^2|=O_{\mathbb{P}}(\widetilde{\lambda}^{(2-\tau)\wedge 1}s(p))$ and hence

$$\frac{\|\widehat{\mathbf{\Sigma}}_{\lambda}\|_F^2}{\|\mathbf{\Sigma}\|_F^2}=O_{\mathbb{P}}(\widetilde{\lambda}^{(2-\tau)\wedge 1}s(p)).$$

By Lemma 1, we have $\mathrm{tr}(\widehat{\mathbf{\Sigma}}_{\lambda})/\mathrm{tr}(\mathbf{\Sigma})=O_{\mathbb{P}}(\widetilde{\lambda})$. This implies this proposition holds. ∎

C Proof of Theorems 1–4 on FDR control and power analysis

It is easy to see that the Benjamini and Hochberg (BH) method is equivalent to reject H_{0ij} if $|\widehat{T}_{ij}|\geq\widehat{t}$, where

$$\widehat{t}=\inf\Big\{t\geq 0:\frac{G(t)(q^2-q)/2}{\max\{\sum_{1\leq i<j\leq q}I\{|\widehat{T}_{ij}|\geq t\},1\}}\leq\alpha\Big\}, \tag{15}$$

where $G(t):=2-2\Phi(t)$. We first give some key lemmas which are the generalization of Lemmas 6,1 and 6.2 in Liu (2013) from i.i.d. case to independent case (but not necessarily identically distributed).

Let $\xi_1,\ldots,\xi_n$ be independent d-dimensional random vectors with mean zero. Define $|\cdot|_{(d)}$ by $|\mathbf{z}|_{(d)}=\min\{|z_i|;1\leq i\leq d\}$ for $\mathbf{z}=(z_1,\ldots,z_d)'$. Let (p,n) be a sequence of positive integers and the constants c,r,b,γ,K,C mentioned below do not depend on (p,n).

Lemma 4 *Suppose that $p\leq cn^r$ and $\max_{1\leq k\leq n}\mathbb{E}|\xi_k|_2^{bdr+2+\epsilon}\leq K$ for some fixed $c>0$, $r>0$, $b>0$, $K>0$ and $\epsilon>0$. Assume that $\|\frac{1}{n}\mathrm{Cov}(\sum_{k=1}^{n}\xi_k)-\mathbf{I}_d\|\leq C(\log p)^{-2-\gamma}$ for some $\gamma>0$ and $C>0$. Then we have*

$$\sup_{0\leq t\leq\sqrt{b\log p}}\Big|\frac{\mathbb{P}(|\sum_{k=1}^{n}\xi_k|_{(d)}\geq t\sqrt{n})}{(G(t))^d}-1\Big|\leq C(\log p)^{-1-\gamma_1}$$

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

Let $\boldsymbol{\eta}_k=(\eta_{k1},\eta_{k2})'$, $1\leq k\leq n$, are independent 2-dimensional random vectors with mean zero.

Lemma 5 *Suppose that $p\leq cn^r$ and $\max_{1\leq k\leq n}\mathbb{E}|\boldsymbol{\eta}_k|_2^{2br+2+\epsilon}<\infty$ for some fixed $c>0$, $r>0$, $b>0$ and $\epsilon>0$. Assume that $\sum_{k=1}^{n}\mathrm{Var}(\eta_{k1})=\sum_{k=1}^{n}\mathrm{Var}(\eta_{k2})=n$ and $|\frac{1}{n}\sum_{k=1}^{n}\mathrm{Cov}(\eta_{k1},\eta_{k2})|\leq\delta$ for some $0\leq\delta<1$. Then we have*

$$\mathbb{P}\Big(|\sum_{k=1}^{n}\eta_{k1}|\geq t\sqrt{n},|\sum_{k=1}^{n}\eta_{k2}|\geq t\sqrt{n}\Big)\leq C(t+1)^{-2}\exp(-t^2/(1+\delta))$$

uniformly for $0\leq t\leq\sqrt{b\log p}$, where C only depends on c,b,r,ϵ,δ.

The proofs of Lemmas 4 and 5 are the same as those of Lemma 6.1 and 6.2 in Liu (2013).

Recall $\eta_{ki,l}$ in (4) and $\xi_{ki,l}$ in (14). For $1 \le i < j \le q$, let

$$U_{ij} = \frac{\sum_{k=1}^{n-1}\sum_{l=1}^{p}\lambda_l^{(1)}(\eta_{ki,l}\eta_{kj,l} - \mathbb{E}\eta_{ki,l}\eta_{kj,l})(\gamma_{ii}\gamma_{jj})^{1/2}}{\sqrt{(n-1)pE_p}},$$
$$V_{ij} = \frac{\sum_{k=1}^{n-1}\sum_{l=1}^{p}\lambda_l^{(1)}(\xi_{ki,l}\xi_{kj,l} - \mathbb{E}\xi_{ki,l}\xi_{kj,l})(\gamma_{ii}\gamma_{jj})^{1/2}}{\sqrt{(n-1)pE_p}}, \tag{16}$$

where $E_p = p^{-1}\sum_{l=1}^{p}(\lambda_l^{(1)})^2$. Note that $\lambda_l^{(1)}$ are bounded away from zero and infinity. Also, $\mathrm{Var}(\eta_{ki,l}\eta_{kj,l}) = (\gamma_{ii}\gamma_{jj})^{-1}(1+\gamma_{ij}^2(\gamma_{ii}\gamma_{jj})^{-1})$, $\mathrm{Var}(U_{ij}) = 1+\gamma_{ij}^2(\gamma_{ii}\gamma_{jj})^{-1}$ and $Corr(U_{ij},U_{kl}) = Corr(\eta_{1i,1}\eta_{1j,1},\eta_{1k,1}\eta_{1l,1})$. By Lemma 4 with $d=1$, we have

$$\max_{1\le i,j\le q}\sup_{0\le t\le 4\sqrt{\log q}}\left|\frac{\mathbb{P}(|U_{ij}| \ge t\sqrt{1+\gamma_{ij}^2(\gamma_{ii}\gamma_{jj})^{-1}})}{G(t)} - 1\right| \le C(\log q)^{-1-\epsilon},$$
$$\max_{1\le i,j\le q}\sup_{0\le t\le 4\sqrt{\log q}}\left|\frac{\mathbb{P}(|V_{ij}| \ge t\sqrt{\mathrm{Var}(\xi_{1i,1}\xi_{1j,1})\gamma_{ii}\gamma_{jj}})}{G(t)} - 1\right| \le C(\log q)^{-1-\epsilon},$$

for some $\epsilon > 0$. Therefore, $\max_{1\le i,j\le q}|U_{ij}| = O_{\mathbb{P}}(\sqrt{\log q})$ and $\max_{1\le i,j\le q}|V_{ij}| = O_{\mathbb{P}}(\sqrt{\log q})$. This, together with (10), Lemma 3, Proposition 3, the proof of Proposition 2, (4.27) and (9), implies that

$$\max_{1\le i<j\le q}\left|\widehat{T}_{ij} + b_{nij}\sqrt{\frac{(n-1)p}{A_p}}(1-\gamma_{ij}\psi_{ij})\rho_{ij.}^{\mathbf{\Gamma}} - V_{ij}\right| = o_{\mathbb{P}}((\log q)^{-1/2}) \tag{17}$$

as $np, q \to \infty$, where b_{nij} satisfies $\max_{1\le i<j\le q}|b_{nij}-1| \to 0$ in probability. Note that under the null $\gamma_{ij} = 0$, $U_{ij} = V_{ij}$. Now Theorem 1 follows from the proof of Theorem 3.1 in Liu (2013) step by step, by using Lemmas 4 and 5 and replacing U_{ij} in Liu (2013) by U_{ij} in (16) and the sample size in Liu (2013) by $(n-1)p$. The proof of Theorem 2 is similar. Theorem 3 follows from the formula of FDP and Theorems 1 and 2.

Under (4.33), we have $\mathrm{Var}(\xi_{1i,1}\xi_{1j,1})\gamma_{ii}\gamma_{jj} = 1+o(1)$ uniformly in i,j. Hence $\max_{1\le i<j\le q}|V_{ij}| \le (2+\delta)\sqrt{\log q}$ for some $\delta > 0$ with probability tending to one. This shows that

$$\mathbb{P}\Big(\min_{(i,j)\in\mathcal{H}_1}|\widehat{T}_{ij}| \ge (2+\delta')\sqrt{\log q}\Big) \to 1.$$

for some $\delta' > 0$. By the definition of $\widehat{t}$ in (15), we have $\widehat{t} \le 2\sqrt{\log q}$ as $q \to \infty$. Thus, $\mathbb{P}(\mathcal{H}_1 \subseteq \widehat{\mathrm{supp}(\mathbf{\Gamma})}) \to 1$. Similarly, we can show that $\mathbb{P}(\mathcal{H}_1' \subseteq \widehat{\mathrm{supp}(\mathbf{\Omega})}) \to 1$. This finishes the proof of Theorem 4. ∎

D Proof of Proposition 4 on the convergence rate of $\widehat{\boldsymbol{\beta}}_j$

Proof of Proposition 4. Define

$$\widehat{\mathbf{a}}_j = \frac{1}{(n-1)p}\sum_{k=1}^{n}\sum_{l=1}^{p}(\mathbf{X}_{l,-j}^{(k)} - \bar{\mathbf{X}}_{l,-j})'(X_{lj}^{(k)} - \bar{X}_{lj}).$$

D. PROOF OF PROPOSITION 4 ON THE CONVERGENCE RATE OF $\widehat{\boldsymbol{\beta}}_J$

We let θ_{nj} and $\boldsymbol{\alpha}_j$ denote $\theta_{nj}(\delta) = \delta\sqrt{\frac{\widehat{\psi}_{jj}\log q}{np}}$ and $\boldsymbol{\alpha}_j(\delta)$ (defined in (3.18)), respectively. By the Karush-Kuhn-Tucker (KKT) condition, we have

$$\left|\mathbf{D}_j^{-1/2}\widehat{\boldsymbol{\Psi}}_{-j,-j}\widehat{\boldsymbol{\beta}}_j - \mathbf{D}_j^{-1/2}\widehat{\mathbf{a}}_j\right|_\infty \le \theta_{nj}. \tag{18}$$

By Lemma 1, we have $c^{-1} \le \min_{1\le j\le q-1}\mathbf{D}_j \le \max_{1\le j\le q-1}\mathbf{D}_j \le c$ for some $c>0$ with probability tending to one. This, together with Lemma 2, implies that, for sufficiently large δ,

$$\left|\frac{1}{(n-1)p}\mathbf{D}_j^{-1/2}\sum_{k=1}^{n}\sum_{l=1}^{p}(\mathbf{X}_{l,-j}^{(k)} - \bar{\mathbf{X}}_{l,-j})'\widetilde{\varepsilon}_{lj}^{(k)}\right|_\infty \le \frac{1}{2}\theta_{nj} \tag{19}$$

uniformly in $1\le j\le q$, with probability tending to one. Note that $\widetilde{\varepsilon}_{lj}^{(k)} = X_{lj}^{(k)} - \bar{X}_{lj} - (\mathbf{X}_{l,-j}^{(k)} - \bar{\mathbf{X}}_{l,-j})\boldsymbol{\beta}_j$. Therefore,

$$\left|\mathbf{D}_j^{-1/2}\widehat{\boldsymbol{\Psi}}_{-j,-j}\boldsymbol{\beta}_j - \mathbf{D}_j^{-1/2}\widehat{\mathbf{a}}_j\right|_\infty \le \frac{1}{2}\theta_{nj} \tag{20}$$

uniformly in $1\le j\le q$. Note that inequalities (18) and (20) imply that

$$\left|\mathbf{D}_j^{-1/2}\widehat{\boldsymbol{\Psi}}_{-j,-j}(\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j)\right|_\infty \le \frac{3}{2}\theta_{nj}. \tag{21}$$

Define $\boldsymbol{\Lambda} = \mathrm{diag}(\boldsymbol{\Psi})^{-1/2}\boldsymbol{\Psi}\mathrm{diag}(\boldsymbol{\Psi})^{-1/2}$. For any subset $T\subset\{1,2,\cdots,q-1\}$ and $\boldsymbol{\nu}\in\mathbb{R}^{q-1}$ with $|T| = o\Big(\sqrt{\frac{np}{\log\max(q,np)}}\Big)$ and $|\boldsymbol{\nu}_{T^c}|_1 \le c|\boldsymbol{\nu}_T|_1$ for some $c>0$, by Lemma 1 and the conditions in Proposition 4, we have

$$\boldsymbol{\nu}'\mathbf{D}_j^{-1/2}\widehat{\boldsymbol{\Psi}}_{-j,-j}\mathbf{D}_j^{-1/2}\boldsymbol{\nu} \ge \lambda_{\min}(\boldsymbol{\Lambda}_{-j,-j})|\boldsymbol{\nu}|_2^2 - O_{\mathbb{P}}\Big(\sqrt{\frac{\log\max(q,np)}{np}}\Big)|\boldsymbol{\nu}|_1^2 \ge |\boldsymbol{\nu}|_2^2/C, \tag{22}$$

for some constant $C>0$, where the first inequality follows from the fact

$$|\boldsymbol{\nu}'(\mathbf{D}_j^{-1/2}\widehat{\boldsymbol{\Psi}}_{-j,-j}\mathbf{D}_j^{-1/2} - \boldsymbol{\Lambda}_{-j,-j})\boldsymbol{\nu}| \le |\mathbf{D}_j^{-1/2}\widehat{\boldsymbol{\Psi}}_{-j,-j}\mathbf{D}_j^{-1/2} - \boldsymbol{\Lambda}_{-j,-j}|_\infty|\boldsymbol{\nu}|_1^2$$

and the second inequality follows from the fact $|\boldsymbol{\nu}|_1^2 \le (1+c)^2|\boldsymbol{\nu}_T|_1^2 \le (1+c)^2|T||\boldsymbol{\nu}|_2^2$.

Now let T be the support of $\boldsymbol{\beta}_j$, $\boldsymbol{\alpha}_j = \mathbf{D}_j^{1/2}\boldsymbol{\beta}_j$ and $\boldsymbol{\nu} = \mathbf{D}_j^{1/2}(\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j) = \widehat{\boldsymbol{\alpha}}_j - \boldsymbol{\alpha}_j$. We first show that $|\boldsymbol{\nu}_{T^c}|_1 \le 3|\boldsymbol{\nu}_T|_1$ uniformly in $1\le j\le q$ with probability tending to one. Define

$$\begin{aligned} Q(\boldsymbol{\alpha}_j) &= \frac{1}{2(n-1)p}\sum_{k=1}^{n}\sum_{l=1}^{p}(X_{lj}^{(k)} - \bar{X}_{lj} - (\mathbf{X}_{l,-j}^{(k)} - \bar{\mathbf{X}}_{l,-j})\mathbf{D}_j^{-1/2}\boldsymbol{\alpha}_j)^2, \\ S(\boldsymbol{\alpha}_j) &= \mathbf{D}_j^{-1/2}\widehat{\mathbf{a}}_j - \mathbf{D}_j^{-1/2}\widehat{\boldsymbol{\Psi}}_{-j,-j}\boldsymbol{\beta}_j. \end{aligned}$$

Note that $S(\boldsymbol{\alpha}_j)$ is the gradient of $Q(\boldsymbol{\alpha}_j)$. By the definition of $\widehat{\boldsymbol{\alpha}}_j$, we have

$$Q(\widehat{\boldsymbol{\alpha}}_j) - Q(\boldsymbol{\alpha}_j) \le \theta_{nj}(\delta)|\boldsymbol{\alpha}_j|_1 - \theta_{nj}|\widehat{\boldsymbol{\alpha}}_j|_1 \le \theta_{nj}(|\boldsymbol{\nu}_T|_1 - |\boldsymbol{\nu}_{T^c}|_1),$$

and by (20), with probability tending to one,

$$Q(\widehat{\boldsymbol{\alpha}}_j) - Q(\boldsymbol{\alpha}_j) \geq S^{'}(\boldsymbol{\alpha}_j)\boldsymbol{\nu} \geq -\frac{1}{2}\theta_{nj}|\boldsymbol{\nu}|_1 = -\frac{1}{2}\theta_{nj}(|\boldsymbol{\nu}_T|_1 + |\boldsymbol{\nu}_{T^c}|_1)$$

uniformly in $1 \leq j \leq q$. It follows from the above two inequalities that $|\boldsymbol{\nu}_{T^c}|_1 \leq 3|\boldsymbol{\nu}_T|_1$. So by (21) and (22) we have

$$\begin{aligned} |\boldsymbol{\nu}|_2^2 &\leq C\boldsymbol{\nu}^{'}\mathbf{D}_j^{-1/2}\widehat{\boldsymbol{\Psi}}_{-j,-j}\mathbf{D}_j^{-1/2}\boldsymbol{\nu} \\ &\leq C|\mathbf{D}_j^{-1/2}\widehat{\boldsymbol{\Psi}}_{-j,-j}\mathbf{D}_j^{-1/2}\boldsymbol{\nu}|_\infty|\boldsymbol{\nu}|_1 \\ &\leq \frac{3}{2}C\theta_{nj}(|\boldsymbol{\nu}_T|_1 + |\boldsymbol{\nu}_{T^c}|_1) \\ &\leq 6C\theta_{nj}|\boldsymbol{\nu}_T|_1 \\ &\leq 6C\theta_{nj}\sqrt{|\boldsymbol{\beta}_j|_0}|\boldsymbol{\nu}_T|_2 \end{aligned}$$

uniformly in $1 \leq j \leq q$ with probability tending to one. By noting that $c^{-1} \leq \min_{1\leq j\leq q-1}\mathbf{D}_j \leq \max_{1\leq j\leq q-1}\mathbf{D}_j \leq c$ with probability tending to one, we have $|\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j|_2 \leq c|\boldsymbol{\nu}|_2$. Hence, by the conditions in Proposition 4, we have $a_{n2} = o_{\mathbb{P}}\Big((np\log q)^{-1/4}\Big)$. Note that $|\boldsymbol{\nu}|_1 \leq 4|\boldsymbol{\nu}_T|_1 \leq 4\sqrt{|\boldsymbol{\beta}_j|_0}|\boldsymbol{\nu}_T|_2 = o((\log\max(q,np))^{-1})$ uniformly in $1 \leq j \leq q-1$ with probability tending to one. This proves Proposition 4 holds. ∎

E Additional Experiments

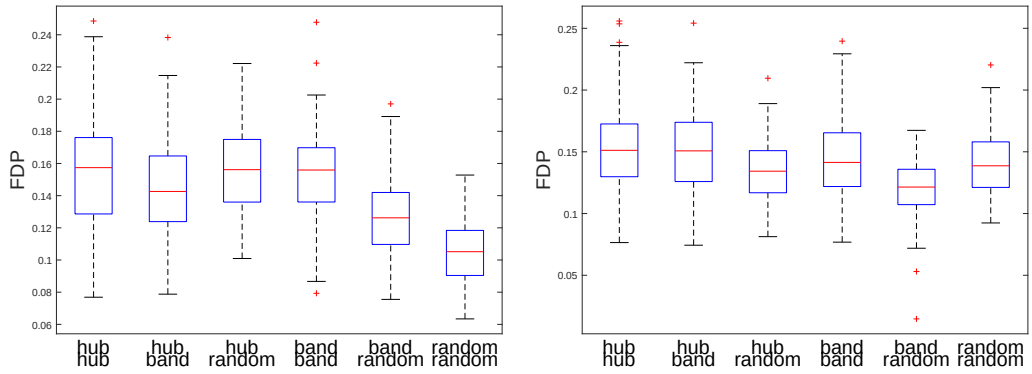
In this section, we present some additional simulation studies and real data analysis. We first note that for the choice of tuning parameters, our theoretical results will hold for any large enough constants λ in (3.15) for estimating $\widehat{A}_p$ (see Proposition 3) and $\delta > 0$ in (3.19) for $\widehat{\boldsymbol{\beta}}_j(\delta)$ (see Proposition 4). In our experiment, we will adopt a data-driven parameter-tuning strategy from Liu (2013). In particular, λ and δ are selected by

$$(\widehat{\lambda},\widehat{\delta}) = \arg\min_{\lambda,\delta}\sum_{k=3}^{9}\left(\frac{\sum_{1\leq i\neq j\leq q} I\{|\widehat{T}_{ij}(\lambda,\delta)| \geq \Phi^{-1}(1-\frac{k}{20})\}}{k(q^2-q)/10} - 1\right)^2, \tag{23}$$

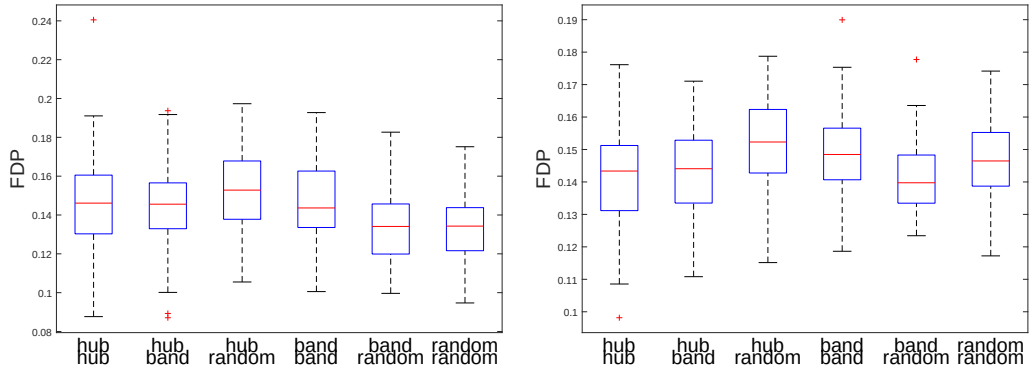
where $\widehat{T}_{ij}(\lambda,\delta)$ is the test statistic in (3.14) with an initial estimator $\widehat{\boldsymbol{\beta}}_j(\delta)$ and $\widehat{A}_p$ (depending on the threshold λ). The choice of (λ,δ) in (23) makes the distributions of $\widehat{T}_{ij}$, on average, close to the standard normal distribution. We note that although the parameter searching is conducted on a two-dimensional grid on λ and δ, the main computational cost is the construction of $\widehat{\boldsymbol{\beta}}_j(\delta)$, which is irrelevant of λ. Therefore, the computational cost of the parameter searching is moderate.

E.1 Boxplots of FDPs

We present the boxplots of FDPs when $n = 100$ over 100 replications in Figure 1 for different p, q, and precision matrix structures. As we can see from Figure 1, FDPs are well concentrated, which suggests that the performance of the proposed estimator is quite stable.

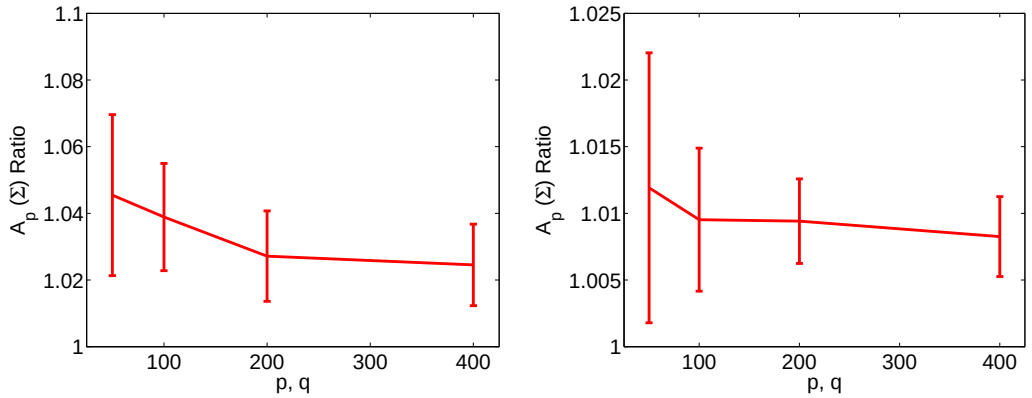


(a) $p = 100$, $q = 100$ (b) $p = 200$, $q = 50$



(c) $p = 200$, $q = 200$ (d) $p = 400$, $q = 400$

Figure 1: Boxplots for FDP when $n = 100$ and $\alpha = 0.1$.



(a) $n = 20$ (b) $n = 100$

Figure 2: The ratio $\widehat{A}_p/A_p$ for different n and $p = q$ when $\mathbf{\Omega}$ and $\mathbf{\Gamma}$ are hub graphs.

E.2 Estimation of $\widehat{A}_p$

In Figure 2, we plot the ratio $\widehat{A}_p/A_p$ for $n = 20$ (left figure) and $n = 100$ (right figure) as $p = q$ increases from 50 to 400. Due to space constraints, we only show the case when both $\mathbf{\Omega}$ and $\mathbf{\Gamma}$ are generated from hub graphs (the plots

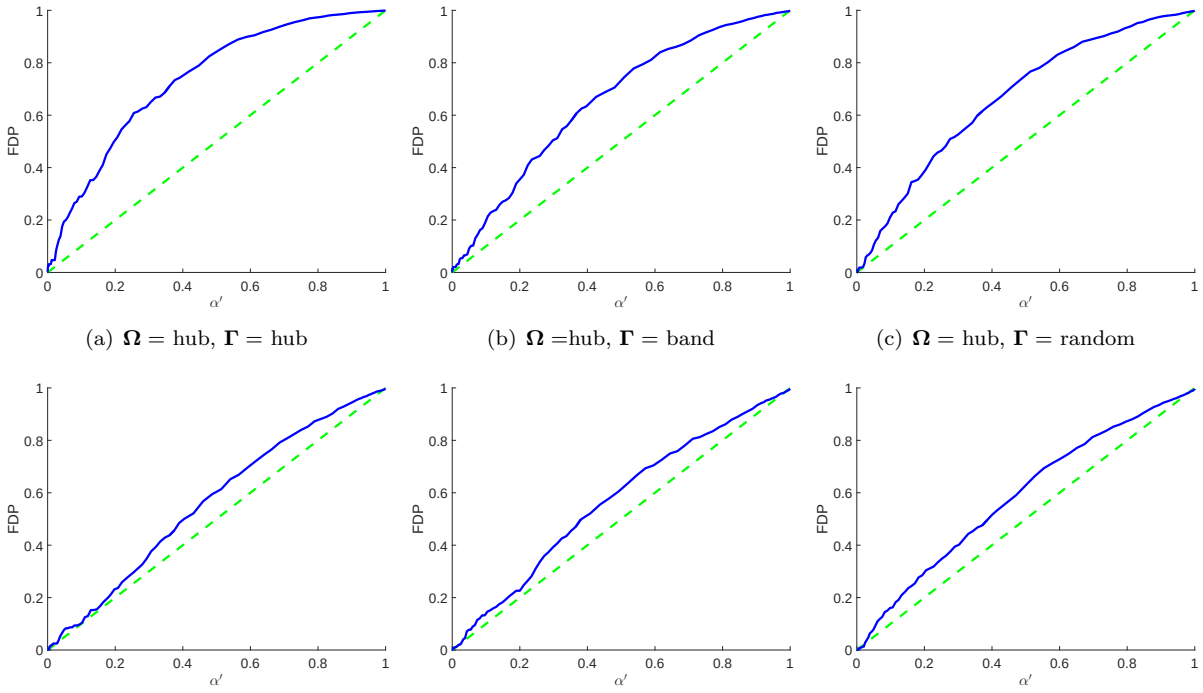
Table 1: Averaged empirical FDP and power using the penalized likelihood method with the SCAD penalty.

p	q	$\mathbf{\Omega}$	$\mathbf{\Gamma}$	$n = 20$		$n = 100$	
				FDP	Power	FDP	Power
100	100	hub	hub	0.037	0.347	0.038	0.346
		hub	band	0.455	0.410	0.348	0.381
		hub	random	0.722	0.926	0.738	0.910
		band	band	0.341	0.206	0.332	0.167
		band	random	0.339	0.964	0.246	0.971
		random	random	0.703	0.962	0.622	0.943
200	200	hub	hub	0.027	0.322	0.059	0.370
		hub	band	0.357	0.314	0.333	0.306
		hub	random	0.629	0.903	0.654	0.888
		band	band	0.333	0.187	0.333	0.167
		band	random	0.246	0.966	0.151	0.982
		random	random	0.481	0.828	0.326	0.863
200	50	hub	hub	0.060	0.381	0.054	0.385
		hub	band	0.362	0.330	0.371	0.517
		hub	random	0.754	0.927	0.704	0.909
		band	band	0.340	0.205	0.334	0.174
		band	random	0.588	0.833	0.551	0.855
		random	random	0.784	0.961	0.607	0.947
400	400	hub	hub	0.946	0.978	0.952	1.000
		hub	band	0.982	1.000	0.962	0.992
		hub	random	0.949	0.999	0.910	1.000
		band	band	0.338	0.179	0.336	0.175
		band	random	0.233	0.848	0.237	0.864
		random	random	0.141	0.610	0.099	0.644

when $\mathbf{\Omega}$ and $\mathbf{\Gamma}$ are generated from other graphs structures are similar). As one can see from Figure 2, when either n is fixed and $p = q$ increases or $p = q$ is fixed and n increases from 20 to 100, the mean ratio becomes closer to one and the standard deviation of the ratios decreases. This study empirically verifies Proposition 3, which claims that the ratio $\widehat{A}_p/A_p$ converges to 1 in probability as $nq \to \infty$.

E.3 Comparison to the penalized likelihood approach

We compare our procedure with the penalized likelihood approach in Leng and Tang (2012). We adopt the same (regularization) parameter-tuning procedure as Leng and Tang (2012), i.e., we generate an extra random test dataset with the sample size equal to the training set and choose the parameter that maximizes the log-likelihood on the test dataset. Due to space constraints, we only report the result using the SCAD penalty (Fan and Li, 2001) rather than the L1 penalty since the SCAD penalty leads to slightly better performance (also observed in Leng and Tang (2012)). The averaged empirical FDPs and powers for different settings of $n, p, q, \mathbf{\Omega}, \mathbf{\Gamma}$ are shown in Table 1. As one can see from Table 1, each setting has either a large FDP or a small power. In fact, for those settings with small averaged FDPs (e.g., $n = 100, p = 200, q = 50$ and $\mathbf{\Omega}$ and $\mathbf{\Gamma}$ generated from hub graphs with the averaged FDP 0.054), the corresponding powers are also small (e.g., 0.385 for the aforementioned case), which indicates that the estimated $\widehat{\mathbf{\Omega}}$ or $\widehat{\mathbf{\Gamma}}$ is too sparse. On the other hand, for those settings with large averaged powers (e.g., $n = 100, p = q = 400$ and $\mathbf{\Omega}$ from hub and $\mathbf{\Gamma}$ from random with the averaged power equal to 1), the corresponding FDPs are also large (e.g., 0.910 for the aforementioned case), which indicates that the estimated $\widehat{\mathbf{\Omega}}$ or $\widehat{\mathbf{\Gamma}}$ is too dense. We also note that when



(a) $\mathbf{\Omega}$ = hub, $\mathbf{\Gamma}$ = hub (b) $\mathbf{\Omega}$ =hub, $\mathbf{\Gamma}$ = band (c) $\mathbf{\Omega}$ = hub, $\mathbf{\Gamma}$ = random

(d) $\mathbf{\Omega}$ = band, $\mathbf{\Gamma}$ = band (e) $\mathbf{\Omega}$ = band, $\mathbf{\Gamma}$ = random (f) $\mathbf{\Omega}$ = random, $\mathbf{\Gamma}$ = random

Figure 3: Averaged empirical FDPs (y-axis) against different estimated FDP α's (x-axis) for the de-correlation approach. For the purpose of controlling FDP, the blue line should be close to or below the dashed green line, which represents FDP=α'.

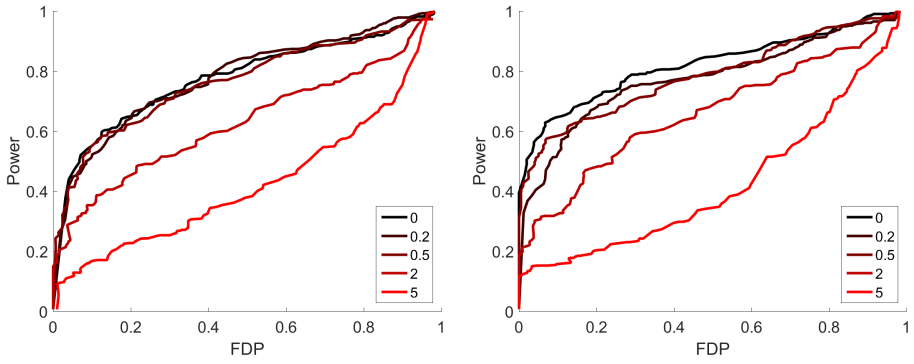
p, q are small as compared to n, the penalized likelihood approach still achieves good support recovery performance (e.g., the case $n = 100, p = q = 20$ as reported in Leng and Tang (2012)). When p, q are comparable to or lager than n, our testing based method achieves better support recovery performance.

E.4 De-correlation method

In Remark 1, we illustrate why the de-correlation approach is not applicable in our problem setup from a theoretical perspective. Here, we provide some empirical evidences. Due to space constraints, we only report the comparison when $n = 20$, $p = q = 100$, and, in fact, the performance becomes even worse when p, q gets larger. Ideally, the empirical FDP should be close to (or below) the FDP estimate α' in (3.21). However, as one can see from Figure 3, the empirical FDP is much larger than the corresponding α' in many cases. Moreover, by setting the FDR level for individual $\mathbf{\Gamma}$ and $\mathbf{\Omega}$ to be $\alpha = 0.1$, we present the corresponding empirical FDP and α' in Table 2, where the FDP can be twice as large as α' in some cases. The experimental results from Table 2 and Figure 3 empirically verify that the de-correlation approach does not control FDP well.

Table 2: Averaged empirical FDP and the estimated FDP α' for the de-correlation approach.

$\mathbf{\Omega}$	$\mathbf{\Gamma}$	FDP	(α')
hub	hub	0.376	(0.146)
hub	band	0.272	(0.152)
hub	random	0.323	(0.154)
band	band	0.176	(0.161)
band	random	0.199	(0.164)
random	random	0.250	(0.164)



(a) $\mathbf{\Omega}$ = band, $\mathbf{\Gamma}$ =random (b) $\mathbf{\Omega}$ = hub, $\mathbf{\Gamma}$ = random

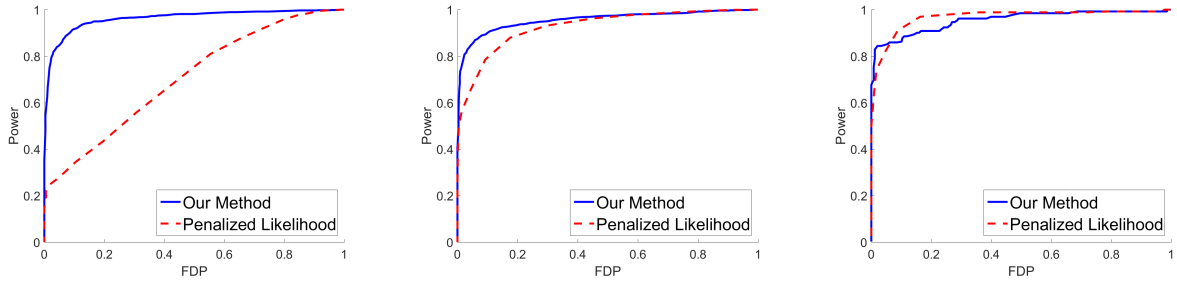
Figure 4: ROC curves for different perturbation levels of $\nu = 0, 0.2, 0.5, 2$ and 5. The larger the ν is, the more "red" color of the ROC curve is (from black to red).

E.5 Simulation study when the covariance is not a Kronecker product

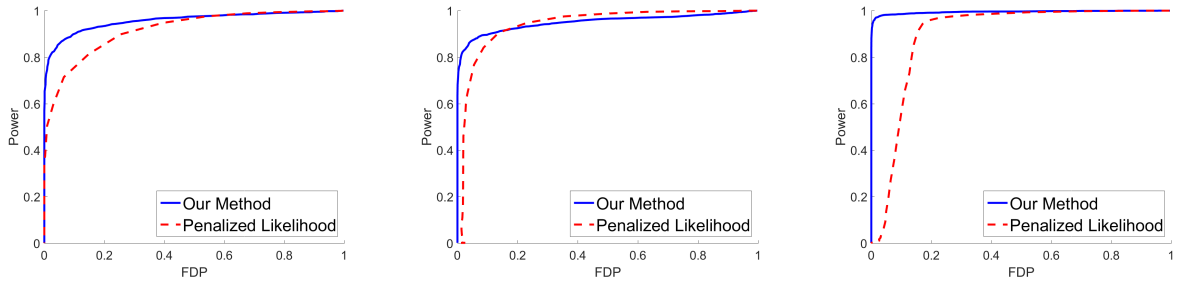
In this section, we present simulation study when the covariance matrix does not follow the form of a Kronecker product. More precisely, we generate the covariance matrix in the form of $\mathbf{\Sigma} \otimes \mathbf{\Psi} + \nu \mathbf{I}$, where $\mathbf{I}$ is the $pq \times pq$ identity matrix and ν is the level of perturbation. Due to space constraints, we only present the case when $n = 20$, $p = q = 30$, $\mathbf{\Omega}$ is either a band or a hub graph, $\mathbf{\Gamma}$ is a random graph. The observation is similar for other settings. Figure 4 plots the ROC curves for different perturbation parameters $\nu = 0, 0.2, 0.5, 2$, and 5. As one can see, when the perturbation level ν is small, the ROC curve is almost identical to the case when the covariance is a Kronecker product (i.e., $\nu = 0$). However, when ν becomes larger, the support recovery performance becomes inferior.

E.6 Additional ROC curve comparisons

In Figure 5, we fix the factor $f = 3$ and consider different types of $\mathbf{\Omega}$ and $\mathbf{\Gamma}$. For most cases, our method achieves better performance. The only exception is that, for hub/random and band/random graphs, the power of the penalized likelihood approach outperforms our method when FDP is large. However, for support recovery in high-dimensional settings, one is more interested in the scenario when FDP is very small. In such a case, our method consistently leads to a larger power than the penalized likelihood approach.



(a) $\mathbf{\Omega}$ = hub, $\mathbf{\Gamma}$ = hub (b) $\mathbf{\Omega}$ =hub, $\mathbf{\Gamma}$ = band (c) $\mathbf{\Omega}$ = hub, $\mathbf{\Gamma}$ = random



(d) $\mathbf{\Omega}$ = band, $\mathbf{\Gamma}$ = band (e) $\mathbf{\Omega}$ = band, $\mathbf{\Gamma}$ = random (f) $\mathbf{\Omega}$ = random, $\mathbf{\Gamma}$ = random

Figure 5: ROC curves for different types of the construction of $\mathbf{\Omega}$ and $\mathbf{\Gamma}$ when $f = 3$.

Table 3: No. of edges for the export data. For 13 regions, there are 78 possible edges in total. For 36 products, there are 630 possible edges in total.

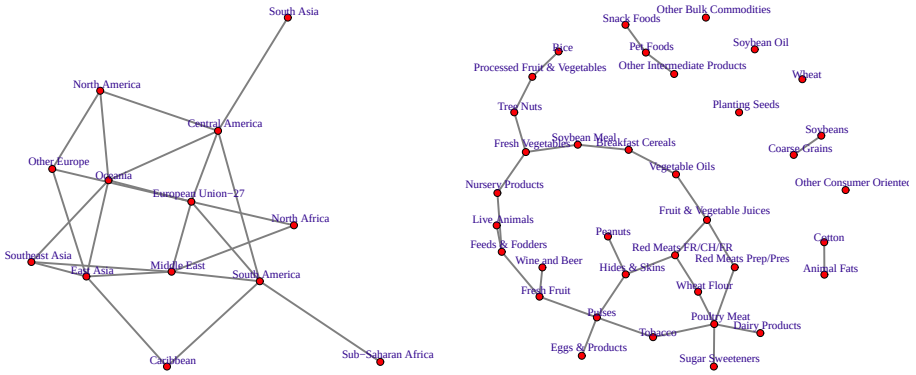
	Region			Product		
	$\alpha = 0.1$	$\alpha = 0.2$	$\alpha = 0.3$	$\alpha = 0.1$	$\alpha = 0.2$	$\alpha = 0.3$
No. of Edges	2	23	31	19	30	37
Density of the Graph	2.56%	29.49%	39.74%	3.01%	4.76%	5.87%

E.7 Real data analysis

In this section, we investigate the performance of the proposed method on two real datasets, the U.S. agricultural export data from Leng and Tang (2012) and the climatological data from Lozano et al. (2009).

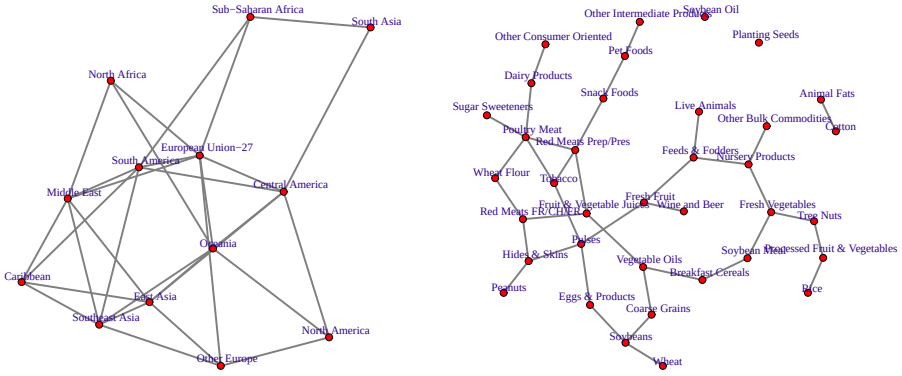
U.S. agricultural export data

We first apply our method to the U.S. agricultural export data studied in Leng and Tang (2012). The dataset contains annual U.S. agriculture export data for 40 years, from 1970 to 2009. Each annual dataset contains the amount (in thousands U.S. dollars) of exports for 36 products (e.g., pet foods, snack foods, breakfast cereals, soybean meal, meats, eggs, dairy products, etc.) in 13 different regions (e.g., North America, Central America, South America, South Asia, etc.). Thus, the dataset can be organized into 40 matrix-variate observations, where each observation is a $(p = 13) \times (q = 36)$ matrix. We adopt the method proposed in Leng and Tang (2012) to remove the dependence in this matrix-variate time series data. In particular, we take the logarithm of the original data plus one and then take the lag-one difference for each matrix observation so that the number of observations becomes $n = 39$. Please refer



(a) Graph for Regions ($\alpha = 0.2$)

(b) Graph for Products ($\alpha = 0.2$)



(c) Graph for Regions ($\alpha = 0.3$)

(d) Graph for Products ($\alpha = 0.3$)

Figure 6: Estimated graphs for the export data

to Leng and Tang (2012) for more details on the pre-processing of the data.

We apply the proposed FDR control procedure to estimate the support of the precision matrices for regions and products under different $\alpha \in \{0.1, 0.2, 0.3\}$. In Table 3, we report the number of edges/discoveries for different α's. We observe that for the product graphs, the number of discoveries is relatively small as compared to the number of hypotheses, which indicates that many pairs of products are conditionally independent. In Figure 6, we plot the graphs corresponding to the estimated supports of $\mathbf{\Omega}$ (corresponding to Regions) and $\mathbf{\Gamma}$ (corresponding to Products) for $\alpha = 0.2$ and $\alpha = 0.3$. Figures 6(a) and 6(c) show the estimated graphs for $p = 13$ regions. As we can see, the regions in the following sets, {East Asia, Southeast Asia}, {European Union, Other Europe, Oceania} and {Central America, North America, South America}, are always connected. Such an observation should be expected since regions in the aforementioned sets are close geographically. This observation is consistent with the result

Table 4: No. of edges for the climate data. For $p = 17$ meteorological factors, there are 136 edges in total. For $q = 125$ locations, there are 7,750 possible edges in total.

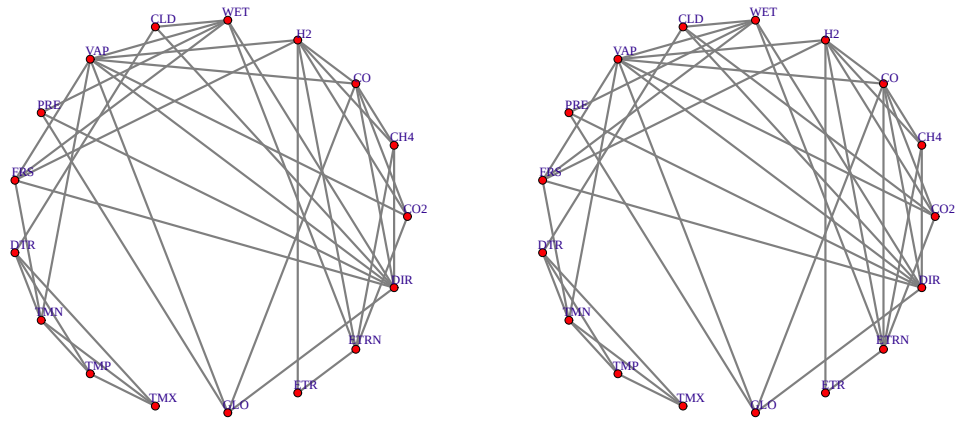
	Meteorological factors			Locations		
	$\alpha = 0.1$	$\alpha = 0.2$	$\alpha = 0.3$	$\alpha = 0.1$	$\alpha = 0.2$	$\alpha = 0.3$
No. of Edges	30	40	42	1059	1539	2065
Density of the Graph	22.05%	29.41%	30.88%	13.66%	19.85%	26.65%

obtained by penalized likelihood approach in Leng and Tang (2012), which claims that "the magnitude between Europe Union and Other Europe, and that between East Asia and Southeast Asia are the strongest." The regions South Asia, Sub-Saharan Africa, and North Africa connect to fewer regions. This observation is also consistent with the result in Leng and Tang (2012), noting that "interestingly, none of the 11 largest edges corresponds to either North Africa or Sub-Saharan Africa." The estimated graphs for products shown in Figures 6(b) and 6(d) are quite sparse, which indicates many pairs of products are conditionally independent given the information of the rest of the products. The product graphs also lead to many interesting observations. For example, the products in the following sets, {Pet foods, Snack foods, Other Intermediate Products},
{Dairy Products, Red Meats FR/CH/FR, Red Meats Prep/Pres, Poultry Meat, Wheat Flour}, are always connected (not necessarily directly). Such observations also make sense since different kinds of meats and dairy products are closely related products and thus should be highly correlated.

Climate data analysis

In this section, we study the climatological data from Lozano et al. (2009), which contains monthly data of $p = 17$ different meteorological factors during 144 months, from 1990 to 2002. The observations span $q = 125$ locations in the U.S. The 17 meteorological factors measured for each month include CO_2, CH_4, H_2, CO, average temperature (TMP), diurnal temperature range (DTR), minimum temperate (TMN), maximum temperature (TMX), precipitation (PRE), vapor (VAP), cloud cover (CLD), wet days (WET), frost days (FRS), global solar radiation (GLO), direct solar radiation (DIR), extraterrestrial radiation (ETR) and extraterrestrial normal radiation (ETRN). We note that we ignore the UV aerosol index factor in Lozano et al. (2009) since most measurements of this factor are missing. We adopt the same procedure as described in Section E.7 to reduce the level of dependence in this matrix-variate time series data.

We apply the proposed FDR control procedure to estimate the support of the precision matrices for meteorological factors and locations under different $\alpha \in \{0.1, 0.2, 0.3\}$. In Table 4, we report the number of edges/discoveries for different α's. From Table 4, the number of discoveries for meteorological factors is quite stable as α increases from 0.1 to 0.3. Moreover, the number of discoveries for locations is relatively large, which indicates many strong correlations among pairs of locations. We plot the graphs corresponding to the estimated supports of the precision matrices for meteorological factors in Figure 7 (the plots for locations are omitted since they are too dense to visualize). An interesting observation is that the factors TMX, TMP, TMN and DTR form a clique. This pattern is reasonable since the factors TMX, TMP, TMN and DTR are all related to temperature and thus should be highly correlated. Other sparsity patterns might also provide insight for understanding dependency relationships among meteorological factors.



(a) Graph for meteorological factors ($\alpha = 0.2$) (b) Graph for meteorological factors ($\alpha = 0.3$)

Figure 7: Estimated graphs for the climate data

Bibliography

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

Leng, C. and Tang, C. Y. (2012) Sparse matrix graphical models. *Journal of the American Statistical Association*, **107**, 1187–1200.

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

Lozano, A. C., Li, H., Niculescu-Mizil, A., Liu, Y., Perlich, C., Hosking, J. and Abe, N. (2009) Spatial-temporal causal modeling for climate change attribution. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.