Partial Correlation Estimation by Joint Sparse Regression Models — Supplemental Material

Part I

In this section, we list properties of the loss function:

$$L(\theta,\sigma,Y)=\frac{1}{2}\sum_{i=1}^{p}w_i(y_i-\sum_{j\neq i}\sqrt{\sigma^{jj}/\sigma ii}\rho^{ij}y_j)^2=\frac{1}{2}\sum_{i=1}^{p}\tilde{w}_i(\tilde{y}_i-\sum_{j\neq i}\rho^{ij}\tilde{y}_j)^2, \quad \text{(S-1)}$$

where $Y=(y_1,\cdots,y_p)^T$ and $\tilde{y}_i=\sqrt{\sigma^{ii}}y_i,\tilde{w}_i=w_i/\sigma^{ii}$. These properties are used for the proof of the main results. Note: throughout the supplementary material, when evaluation is taken place at $\sigma=\bar{\sigma}$, sometimes we omit the argument σ in the notation for simplicity. Also we use $Y=(y_1,\cdots,y_p)^T$ to denote a generic sample and use $\mathbf{Y}$ to denote the $p\times n$ data matrix consisting of n i.i.d. such samples: $\mathbf{Y^1},\cdots,\mathbf{Y^n}$, and define

$$L_n(\theta,\sigma,\mathbf{Y}):=\frac{1}{n}\sum_{k=1}^{n}L(\theta,\sigma,\mathbf{Y}^k). \quad \text{(S-2)}$$

A1: for all θ,σ and $Y\in\mathcal{R}^p$, $L(\theta,\sigma,Y)\geq 0$.

A2: for any $Y\in\mathcal{R}^p$ and any $\sigma>0$, $L(\cdot,\sigma,Y)$ is convex in θ; and with probability one, $L(\cdot,\sigma,Y)$ is strictly convex.

A3: for $1 \le i < j \le p$

$$\overline{L}'_{ij}(\bar\theta, \bar\sigma) := E_{(\bar\theta,\bar\sigma)} \left(\frac{\partial L(\theta, \sigma, Y)}{\partial \rho^{ij}} \Big|_{\theta=\bar\theta, \sigma=\bar\sigma} \right) = 0.$$

A4: for $1 \le i < j \le p$ and $1 \le k < l \le p$,

$$\overline{L}''_{ij,kl}(\theta, \sigma) := E_{(\theta,\sigma)} \left(\frac{\partial^2 L(\theta, \sigma, Y)}{\partial \rho^{ij} \rho^{kl}} \right) = \frac{\partial}{\partial \rho^{kl}} \left[E_{(\theta,\sigma)} \left(\frac{\partial L(\theta, \sigma, Y)}{\partial \rho^{ij}} \right) \right],$$

and $\overline{L}''(\bar\theta, \bar\sigma)$ is positive semi-definite.

If assuming C0-C1, then we have

B0 : There exist constants $0 < \bar\sigma_0 \le \bar\sigma_\infty < \infty$ such that: $0 < \bar\sigma_0 \le \min\{\bar\sigma^{ii} : 1 \le i \le p\} \le \max\{\bar\sigma^{ii} : 1 \le i \le p\} \le \bar\sigma_\infty$.

B1 : There exist constants $0 < \Lambda^L_{\min}(\bar\theta) \le \Lambda^L_{\max}(\bar\theta) < \infty$, such that

$$0 < \Lambda^L_{\min}(\bar\theta) \le \lambda_{min}(\overline{L}''(\bar\theta)) \le \lambda_{\max}(\overline{L}''(\bar\theta)) \le \Lambda^L_{\max}(\bar\theta) < \infty$$

B1.1 : There exists a constant $K(\bar\theta) < \infty$, such that for all $1 \le i < j \le p$, $\overline{L}''_{ij,ij}(\bar\theta) \le K(\bar\theta)$.

B1.2 : There exist constants $M_1(\bar\theta), M_2(\bar\theta) < \infty$, such that for any $1 \le i < j \le p$

$$\mathrm{Var}_{(\bar\theta,\bar\sigma)}(L'_{ij}(\bar\theta, \bar\sigma, Y)) \le M_1(\bar\theta), \ \mathrm{Var}_{(\bar\theta,\bar\sigma)}(L''_{ij,ij}(\bar\theta, \bar\sigma, Y)) \le M_2(\bar\theta).$$

B1.3 : There exists a constant $0 < g(\bar\theta) < \infty$, such that for all $(i, j) \in \mathcal{A}$

$$\overline{L}''_{ij,ij}(\bar\theta, \bar\sigma) - \overline{L}''_{ij,\mathcal{A}_{ij}}(\bar\theta, \bar\sigma) \left[\overline{L}''_{\mathcal{A}_{ij},\mathcal{A}_{ij}}(\bar\theta, \bar\sigma) \right]^{-1} \overline{L}''_{\mathcal{A}_{ij},ij}(\bar\theta, \bar\sigma) \ge g(\bar\theta),$$

where $\mathcal{A}_{ij} = \mathcal{A}/\{(i,j)\}$.

B1.4 : There exists a constant $M(\bar{\theta}) < \infty$, such that for any $(i,j) \in \mathcal{A}^c$

$$||\overline{L}''_{ij,\mathcal{A}}(\bar{\theta})[\overline{L}''_{\mathcal{A}\mathcal{A}}(\bar{\theta})]^{-1}||_2 \leq M(\bar{\theta}).$$

B2 There exists a constant $K_1(\bar{\theta}) < \infty$, such that for any $1 \leq i \leq j \leq p$, $||E_{\bar{\theta}}(\tilde{y}_i \tilde{y}_j \tilde{y} \tilde{y}^T)|| \leq K_1(\bar{\theta})$, where $\tilde{y} = (\tilde{y}_1, \cdots, \tilde{y}_p)^T$.

B3 If we further assume that condition D holds for $\widehat{\sigma}$ and $q_n \sim o(\frac{n}{\log n})$, we have: for any $\eta > 0$, there exist constants $C_{1,\eta}, C_{2,\eta} > 0$, such that for sufficiently large n

$$\max_{1 \leq i < k \leq p} \left| L'_{n,ik}(\bar{\theta}, \bar{\sigma}, \mathbf{Y}) - L'_{n,ik}(\bar{\theta}, \widehat{\sigma}, \mathbf{Y}) \right| \leq C_{1,\eta}(\sqrt{\frac{\log n}{n}}),$$

$$\max_{1 \leq i < k \leq p, 1 \leq t < s \leq p} \left| L''_{n,ik,ts}(\bar{\theta}, \bar{\sigma}, \mathbf{Y}) - L''_{n,ik,ts}(\bar{\theta}, \widehat{\sigma}, \mathbf{Y}) \right| \leq C_{2,\eta}(\sqrt{\frac{\log n}{n}}),$$

hold with probability at least $1 - O(n^{-\eta})$.

B0 follows from C1 immediately. B1.1–B1.4 are direct consequences of B1. B2 follows from B1 and Gaussianity. B3 follows from conditions C0-C1 and D.

proof of A1: obvious.

proof of A2: obvious.

proof of A3: denote the residual for the ith term by

$$e_i(\theta, \sigma) = \tilde{y}_i - \sum_{j \neq i} \rho^{ij} \tilde{y}_j.$$

Then evaluated at the true parameter values $(\bar{\theta}, \bar{\sigma})$, we have $e_i(\bar{\theta}, \bar{\sigma})$ uncorrelated with $\tilde{y}_{(-i)}$ and $E_{(\bar{\theta},\bar{\sigma})}(e_i(\bar{\theta}, \bar{\sigma})) = 0$. It is easy to show

$$\frac{\partial L(\theta, \sigma, Y)}{\partial \rho^{ij}} = -\tilde{w}_i e_i(\theta, \sigma)\tilde{y}_j - \tilde{w}_j e_j(\theta, \sigma)\tilde{y}_i.$$

This proves A3.

proof of A4: see the proof of B1.

proof of B1: Denote $\tilde{y} = (\tilde{y}_1, \cdots, \tilde{y}_p)^T$, and $\tilde{x} = (\tilde{x}_{(1,2)}, \tilde{x}_{(1,3)}, \cdots, \tilde{x}_{(p-1,p)})$ with $\tilde{x}_{(i,j)} = (0, \cdots, 0, \tilde{y}_j, \cdots, \tilde{y}_i, 0, \cdots, 0)^T$. Then the loss function (S-1) can be written as $L(\theta, \sigma, Y) = \frac{1}{2}||\tilde{w}(\tilde{y} - \tilde{x}\theta)||_2^2$, with $\tilde{w} = diag(\sqrt{\tilde{w}_1}, \cdots, \sqrt{\tilde{w}_p})$. Thus $\overline{L}''(\theta, \sigma) = E_{(\theta,\sigma)}\left[\tilde{x}^T\tilde{w}^2\tilde{x}\right]$ (this proves A4). Let $d = p(p-1)/2$, then $\tilde{x}$ is a p by d matrix. Denote its ith row by x_i^T $(1 \le i \le p)$. Then for any $a \in \mathcal{R}^d$, with $||a||_2 = 1$, we have

$$a^T\overline{L}''(\bar{\theta})a = E_{\bar{\theta}}(a^T\tilde{x}^T\tilde{w}^2\tilde{x}a) = E_{\bar{\theta}}\left(\sum_{i=1}^{p}\tilde{w}_i(x_i^Ta)^2\right).$$

Index the elements of a by $a = (a_{(1,2)}, a_{(1,3)}, \cdots, a_{(p-1,p)})^T$, and for each $1 \le i \le p$, define $a_i \in \mathcal{R}^p$ by $a_i = (a_{(1,i)}, \cdots, a_{(i-1,i)}, 0, a_{(i,i+1)}, \cdots, a_{(i,p)})^T$. Then by definition $x_i^Ta = \tilde{y}^Ta_i$. Also note that $\sum_{i=1}^p ||a_i||_2^2 = 2||a||_2^2 = 2$. This is because, for $i \ne j$, the jth entry of a_i appears exactly twice in a. Therefore

$$a^T\overline{L}''(\bar{\theta})a = \sum_{i=1}^{p}\tilde{w}_i E_{\bar{\theta}}\left(a_i^T\tilde{y}\tilde{y}^Ta_i\right) = \sum_{i=1}^{p}\tilde{w}_i a_i^T\tilde{\mathbf{\Sigma}}a_i \ge \sum_{i=1}^{p}\tilde{w}_i\lambda_{\min}(\tilde{\mathbf{\Sigma}})||a_i||_2^2 \ge 2\tilde{w}_0\lambda_{\min}(\tilde{\mathbf{\Sigma}}),$$

where $\tilde{\mathbf{\Sigma}} = \mathrm{Var}(\tilde{y})$ and $\tilde{w}_0 = w_0/\bar{\sigma}_\infty$. Similarly $a^T\overline{L}''(\bar{\theta})a \le 2\tilde{w}_\infty\lambda_{\max}(\tilde{\mathbf{\Sigma}})$, with $\tilde{w}_\infty = w_\infty/\bar{\sigma}_0$. By C1, $\tilde{\mathbf{\Sigma}}$ has bounded eigenvalues, thus B1 is proved.

proof of B1.1: obvious.

proof of B1.2: note that $\mathrm{Var}_{(\bar{\theta},\bar{\sigma})}(e_i(\bar{\theta},\bar{\sigma})) = 1/\bar{\sigma}^{ii}$ and $\mathrm{Var}_{(\bar{\theta},\bar{\sigma})}(\tilde{y}_i) = \bar{\sigma}^{ii}$. Then for any $1 \le i < j \le p$, by Cauchy-Schwartz

$$\begin{aligned}
\mathrm{Var}_{(\bar{\theta},\bar{\sigma})}(L'_{n,ij}(\bar{\theta},\bar{\sigma},Y)) &= \mathrm{Var}_{(\bar{\theta},\bar{\sigma})}(-\tilde{w}_i e_i(\bar{\theta},\bar{\sigma})\tilde{y}_j - \tilde{w}_j e_j(\bar{\theta},\bar{\sigma})\tilde{y}_i) \\
&\le E_{(\bar{\theta},\bar{\sigma})}(\tilde{w}_i^2 e_i^2(\bar{\theta},\bar{\sigma})\tilde{y}_j^2) + E_{(\bar{\theta},\bar{\sigma})}(\tilde{w}_j^2 e_j^2(\bar{\theta},\bar{\sigma})\tilde{y}_i^2) \\
&+ 2\sqrt{\tilde{w}_i^2 \tilde{w}_j^2 E_{(\bar{\theta},\bar{\sigma})}(e_i^2(\bar{\theta},\bar{\sigma})\tilde{y}_j^2) E_{(\bar{\theta},\bar{\sigma})}(e_j^2(\bar{\theta},\bar{\sigma})\tilde{y}_i^2)} \\
&= \frac{w_i^2 \bar{\sigma}^{jj}}{(\bar{\sigma}^{ii})^3} + \frac{w_j^2 \bar{\sigma}^{ii}}{(\bar{\sigma}^{jj})^3} + 2\frac{w_i w_j}{\bar{\sigma}^{ii}\bar{\sigma}^{jj}}.
\end{aligned}$$

The right hand side is bounded because of C0 and B0.

proof of B1.3: for $(i,j) \in \mathcal{A}$, denote

$$D := \overline{L}''_{ij,ij}(\bar{\theta},\bar{\sigma}) - \overline{L}''_{ij,\mathcal{A}_{ij}}(\bar{\theta},\bar{\sigma}) \left[\overline{L}''_{\mathcal{A}_{ij},\mathcal{A}_{ij}}(\bar{\theta},\bar{\sigma})\right]^{-1} \overline{L}''_{\mathcal{A}_{ij},ij}(\bar{\theta},\bar{\sigma}).$$

Then D^{-1} is the (ij,ij) entry in $\left[\overline{L}''_{\mathcal{A},\mathcal{A}}(\bar{\theta})\right]^{-1}$. Thus by B1, D^{-1} is positive and bounded from above, so D is bounded away from zero.

proof of B1.4: note that $||\overline{L}''_{ij,\mathcal{A}}(\bar{\theta})[\overline{L}''_{\mathcal{A}\mathcal{A}}(\bar{\theta})]^{-1}||_2^2 \le ||\overline{L}''_{ij,\mathcal{A}}(\bar{\theta})||_2^2 \lambda_{\max}([\overline{L}''_{\mathcal{A}\mathcal{A}}(\bar{\theta})]^{-2})$. By B1, $\lambda_{\max}([\overline{L}''_{\mathcal{A}\mathcal{A}}(\bar{\theta})]^{-2})$ is bounded from above, thus it suffices to show that $||\overline{L}''_{ij,\mathcal{A}}(\bar{\theta})||_2^2$ is bounded. Since $(i,j) \in \mathcal{A}^c$, define $\mathcal{A}^+ := (i,j) \cup \mathcal{A}$. Then $\overline{L}''_{ij,ij}(\bar{\theta}) - \overline{L}''_{ij,\mathcal{A}}(\bar{\theta})[\overline{L}''_{\mathcal{A}\mathcal{A}}(\bar{\theta})]^{-1}\overline{L}''_{\mathcal{A},ij}(\bar{\theta})$ is the inverse of the $(1,1)$ entry of $\overline{L}''_{\mathcal{A}^+,\mathcal{A}^+}(\bar{\theta})$. Thus by B1, it is bounded away from zero. Therefore by B1.1, $\overline{L}''_{ij,\mathcal{A}}(\bar{\theta})[\overline{L}''_{\mathcal{A}\mathcal{A}}(\bar{\theta})]^{-1}\overline{L}''_{\mathcal{A},ij}(\bar{\theta})$ is bounded from above. Since $\overline{L}''_{ij,\mathcal{A}}(\bar{\theta})[\overline{L}''_{\mathcal{A}\mathcal{A}}(\bar{\theta})]^{-1}\overline{L}''_{\mathcal{A},ij}(\bar{\theta}) \ge ||\overline{L}''_{ij,\mathcal{A}}(\bar{\theta})||_2^2 \lambda_{\min}([\overline{L}''_{\mathcal{A}\mathcal{A}}(\bar{\theta})]^{-1})$, and by B1, $\lambda_{\min}([\overline{L}''_{\mathcal{A}\mathcal{A}}(\bar{\theta})]^{-1})$ is bounded away from zero, we have $||\overline{L}''_{ij,\mathcal{A}}(\bar{\theta})||_2^2$ bounded from above.

proof of B2: the (k,l)-th entry of the matrix $\tilde{y}_i\tilde{y}_j\tilde{y}\tilde{y}^T$ is $\tilde{y}_i\tilde{y}_j\tilde{y}_k\tilde{y}_l$, for $1 \le k < l \le p$. Thus, the (k,l)-th entry of the matrix $\mathbb{E}[\tilde{y}_i\tilde{y}_j\tilde{y}\tilde{y}^T]$ is $\mathbb{E}[\tilde{y}_i\tilde{y}_j\tilde{y}_k\tilde{y}_l] = \tilde{\sigma}_{ij}\tilde{\sigma}_{kl} + \tilde{\sigma}_{ik}\tilde{\sigma}_{jl} + \tilde{\sigma}_{il}\tilde{\sigma}_{jk}$.

Thus, we can write

$$\mathbb{E}[\tilde{y}_i\tilde{y}_j\tilde{y}\tilde{y}^T] = \tilde{\sigma}_{ij}\tilde{\boldsymbol{\Sigma}} + \tilde{\sigma}_{i\cdot}\tilde{\sigma}_{j\cdot}^T + \tilde{\sigma}_{j\cdot}\tilde{\sigma}_{i\cdot}^T, \tag{S-3}$$

where $\tilde{\sigma}_{i\cdot}$ is the $p \times 1$ vector $(\tilde{\sigma}_{ik})_{k=1}^p$. From (S-3), we have

$$\| \mathbb{E}[\tilde{y}_i\tilde{y}_j\tilde{y}\tilde{y}^T] \| \leq |\tilde{\sigma}_{ij}| \| \tilde{\boldsymbol{\Sigma}} \| +2 \| \tilde{\sigma}_{i\cdot} \|_2 \| \tilde{\sigma}_{j\cdot} \|_2, \tag{S-4}$$

where $|| \cdot ||$ is the operator norm. By C0-C1, the first term on the right hand side is uniformly bounded. Now, we also have,

$$\tilde{\sigma}_{ii} - \tilde{\sigma}_{i\cdot}^T \tilde{\boldsymbol{\Sigma}}_{(-i)}^{-1} \tilde{\sigma}_{i\cdot} > 0 \tag{S-5}$$

where $\tilde{\boldsymbol{\Sigma}}_{(-i)}$ is the submatrix of $\tilde{\boldsymbol{\Sigma}}$ removing i-th row and column. From this, it follows that

$$\begin{aligned} \| \tilde{\sigma}_{i\cdot} \|_2 &= \| \tilde{\boldsymbol{\Sigma}}_{(-i)}^{1/2} \tilde{\boldsymbol{\Sigma}}_{(-i)}^{-1/2} \tilde{\sigma}_{i\cdot} \|_2 \\ &\leq \| \tilde{\boldsymbol{\Sigma}}_{(-i)}^{1/2} \| \quad \| \tilde{\boldsymbol{\Sigma}}_{(-i)}^{-1/2} \tilde{\sigma}_{i\cdot} \|_2 \\ &\leq \sqrt{\| \tilde{\boldsymbol{\Sigma}} \|} \sqrt{\tilde{\sigma}_{ii}}, \end{aligned} \tag{S-6}$$

where the last inequality follows from (S-5), and the fact that $\tilde{\boldsymbol{\Sigma}}_{(-i)}$ is a principal submatrix of $\tilde{\boldsymbol{\Sigma}}$. Thus the result follows by applying (S-6) to bound the last term in (S-4).

proof of B3:

$$\begin{aligned} L'_{n,ik}(\bar{\theta}, \sigma, \mathbf{Y}) &= \frac{1}{n} \sum_{l=1}^{n} -w_i \left(y_i^l - \sum_{j \neq i} \sqrt{\frac{\sigma^{jj}}{\sigma^{ii}}} \rho^{ij} y_j^l \right) \sqrt{\frac{\sigma^{kk}}{\sigma^{ii}}} y_k^l \\ & \quad -w_k \left(y_k^l - \sum_{j \neq k} \sqrt{\frac{\sigma^{jj}}{\sigma^{kk}}} \rho^{kj} y_j^l \right) \sqrt{\frac{\sigma^{ii}}{\sigma^{kk}}} y_i^l. \end{aligned}$$

Thus,

$$\begin{aligned}
& L'_{n,ik}(\bar{\theta}, \bar{\sigma}, \mathbf{Y}) - L'_{n,ik}(\bar{\theta}, \widehat{\sigma}, \mathbf{Y}) \\
= \quad & -w_i \left[\overline{y_i y_k} \left(\sqrt{\frac{\sigma^{kk}}{\sigma^{ii}}} - \sqrt{\frac{\widehat{\sigma}^{kk}}{\widehat{\sigma}^{ii}}} \right) - \sum_{j \neq i} \overline{y_j y_k} \rho^{ij} \left(\frac{\sqrt{\sigma^{jj}\sigma^{kk}}}{\sigma^{ii}} - \frac{\sqrt{\widehat{\sigma}^{jj}\widehat{\sigma}^{kk}}}{\widehat{\sigma}^{ii}} \right) \right] \\
& -w_k \left[\overline{y_i y_k} \left(\sqrt{\frac{\sigma^{ii}}{\sigma^{kk}}} - \sqrt{\frac{\widehat{\sigma}^{ii}}{\widehat{\sigma}^{kk}}} \right) - \sum_{j \neq k} \overline{y_j y_i} \rho^{kj} \left(\frac{\sqrt{\sigma^{jj}\sigma^{ii}}}{\sigma^{kk}} - \frac{\sqrt{\widehat{\sigma}^{jj}\widehat{\sigma}^{ii}}}{\widehat{\sigma}^{kk}} \right) \right],
\end{aligned}$$

where for $1 \leq i, j \leq p$, $\overline{y_i y_j} := \frac{1}{n} \sum_{l=1}^{n} y_i^l y_j^l$. Let σ_{ij} denote the (i,j)-th element of the true covariance matrix $\overline{\Sigma}$. By C1, $\{\sigma_{ij} : 1 \leq i, j \leq p\}$ are bounded from below and above, thus

$$\max_{1 \leq i,j \leq p} |\overline{y_i y_j} - \sigma_{ij}| = O_p(\sqrt{\frac{\log n}{n}}).$$

(Throughout the proof, $O_p(\cdot)$ means that for any $\eta > 0$, for sufficiently large n, the left hand side is bounded by the order within $O_p(\cdot)$ with probability at least $1 - O(n^{-\eta})$.) Therefore

$$\sum_{j \neq i} |\overline{y_j y_k} - \sigma_{jk}||\rho^{ij}| \leq (\sum_{j \neq i} |\rho^{ij}|) \max_{1 \leq i,j \leq p} |\overline{y_i y_j} - \sigma_{ij}| \leq (\sqrt{q_n \sum_{j \neq i} (\rho^{ij})^2}) \max_{1 \leq i,j \leq p} |\overline{y_i y_j} - \sigma_{ij}| = o(1),$$

where the last inequality is by Cauchy-Schwartz and the fact that, for fixed i, there are at most q_n non-zero ρ^{ij}. The last equality is due to the assumption $q_n \sim o(\frac{n}{\log n})$, and the fact that $\sum_{j \neq i} (\rho^{ij})^2$ is bounded which is in turn implied by condition C1. Therefore,

$$\begin{aligned}
& |L'_{n,ik}(\bar{\theta}, \bar{\sigma}, \mathbf{Y}) - L'_{n,ik}(\bar{\theta}, \widehat{\sigma}, \mathbf{Y})| \\
\leq \quad & (w_i|\sigma_{ik}| + w_k|\sigma_{ik}|) \max_{i,k} \left| \sqrt{\frac{\sigma^{kk}}{\sigma^{ii}}} - \sqrt{\frac{\widehat{\sigma}^{kk}}{\widehat{\sigma}^{ii}}} \right| + (w_i \tau_{ki} + w_k \tau_{ik}) \max_{i,j,k} \left| \frac{\sqrt{\sigma^{jj}\sigma^{kk}}}{\sigma^{ii}} - \frac{\sqrt{\widehat{\sigma}^{jj}\widehat{\sigma}^{kk}}}{\widehat{\sigma}^{ii}} \right| + R_n,
\end{aligned}$$

where $\tau_{ki} := \sum_{j \neq i} |\sigma_{jk}\rho^{ij}|$, and the reminder term R_n is of smaller order of the leading terms. Since C1 implies B0, thus together with condition D, we have

$$\max_{1\leq i,k\leq p} \left| \sqrt{\frac{\sigma^{ii}}{\sigma^{kk}}} - \sqrt{\frac{\widehat{\sigma}^{ii}}{\widehat{\sigma}^{kk}}} \right| = O_p(\sqrt{\frac{\log n}{n}}),$$

$$\max_{1\leq i,j,k\leq p} \left| \frac{\sqrt{\sigma^{jj}\sigma^{ii}}}{\sigma^{kk}} - \frac{\sqrt{\widehat{\sigma}^{jj}\widehat{\sigma}^{ii}}}{\widehat{\sigma}^{kk}} \right| = O_p(\sqrt{\frac{\log n}{n}}).$$

Moreover, by Cauchy-Schwartz

$$\tau_{ki} \leq \sqrt{\sum_j (\rho^{ij})^2} \sqrt{\sum_j (\sigma_{jk})^2},$$

and the right hand side is uniformly bounded (over (i,k)) due to condition C1. Thus by C0,C1 and D, we have showed

$$\max_{i,k} |L'_{n,ik}(\bar{\theta}, \bar{\sigma}, \mathbf{Y}) - L'_{n,ik}(\bar{\theta}, \widehat{\sigma}, \mathbf{Y})| = O_p(\sqrt{\frac{\log n}{n}}).$$

Observe that, for $1 \leq i < k \leq p, 1 \leq t < s \leq p$

$$L''_{n,ik,ts} = \begin{cases} \frac{1}{n}\sum_{l=1}^n w_i \frac{\sigma^{kk}}{\sigma^{ii}} y_k^l + w_k \frac{\sigma^{ii}}{\sigma^{kk}} y_i^l & , if \quad (i,k) = (t,s) \\ \frac{1}{n}\sum_{l=1}^n w_i \frac{\sqrt{\sigma^{kk}\sigma^{ss}}}{\sigma^{ii}} y_s^l y_k^l, & if \quad i = t, k \neq s \\ \frac{1}{n}\sum_{l=1}^n w_k \frac{\sqrt{\sigma^{tt}\sigma^{ii}}}{\sigma^{kk}} y_t^l y_i^l, & if \quad i \neq t, k = s \\ 0 & if \quad otherwise. \end{cases}$$

Thus by similar arguments as in the above, it is easy to proof the claim.

Part II

In this section, we proof the main results (Theorems 1–3). We first give a few lemmas.

Lemma S-1 *(Karush-Kuhn-Tucker condition)* $\widehat{\theta}$ *is a solution of the optimization problem*

$$\arg\min_{\theta:\theta_{\mathcal{S}^c}=0} L_n(\theta,\widehat{\sigma},\mathbf{Y}) + \lambda_n||\theta||_1,$$

where $\mathcal{S}$ *is a subset of* $\mathcal{T} := \{(i,j) : 1 \leq i < j \leq p\}$, *if and only if*

$$\begin{array}{rcll} L'_{n,ij}(\widehat{\theta},\widehat{\sigma},\mathbf{Y}) & = & \lambda_n \text{sign}(\widehat{\theta}_{ij}), & \text{if} \quad \widehat{\theta}_{ij} \neq 0 \\ |L'_{n,ij}(\widehat{\theta},\widehat{\sigma},\mathbf{Y})| & \leq & \lambda_n, & \text{if} \quad \widehat{\theta}_{ij} = 0, \end{array}$$

for $(i,j) \in \mathcal{S}$. *Moreover, if the solution is not unique,* $|L'_{n,ij}(\tilde{\theta},\widehat{\sigma},\mathbf{Y})| < \lambda_n$ *for some specific solution* $\tilde{\theta}$ *and* $L'_{n,ij}(\theta,\widehat{\sigma},\mathbf{Y})$ *being continuous in* θ *imply that* $\widehat{\theta}_{ij} = 0$ *for all solutions* $\widehat{\theta}$. *(Note that optimization problem (9) corresponds to* $\mathcal{S} = \mathcal{T}$ *and the restricted optimization problem (11) corresponds to* $\mathcal{S} = \mathcal{A}$.*)*

Lemma S-2 *For the loss function defined by (S-2), if conditions C0-C1 hold and condition D holds for* $\widehat{\sigma}$ *and if* $q_n \sim o(\frac{n}{\log n})$, *then for any* $\eta > 0$, *there exist constants* $c_{0,\eta}, c_{1,\eta}, c_{2,\eta}, c_{3,\eta} > 0$, *such that for any* $u \in R^{q_n}$ *the following hold with probability as least* $1 - O(n^{-\eta})$ *for sufficiently large* n:

$$||L'_{n,\mathcal{A}}(\overline{\theta},\widehat{\sigma},\mathbf{Y})||_2 \leq c_{0,\eta}\sqrt{\frac{q_n \log n}{n}}$$

$$|u^T L'_{n,\mathcal{A}}(\overline{\theta},\widehat{\sigma},\mathbf{Y})| \leq c_{1,\eta}||u||_2(\sqrt{\frac{q_n \log n}{n}})$$

$$|u^T L''_{n,\mathcal{A}\mathcal{A}}(\overline{\theta},\widehat{\sigma},\mathbf{Y})u - u^T \overline{L}''_{\mathcal{A}\mathcal{A}}(\overline{\theta})u| \leq c_{2,\eta}||u||_2^2(q_n\sqrt{\frac{\log n}{n}})$$

$$||L''_{n,\mathcal{A}\mathcal{A}}(\overline{\theta},\widehat{\sigma},\mathbf{Y})u - \overline{L}''_{\mathcal{A}\mathcal{A}}(\overline{\theta})u||_2 \leq c_{3,\eta}||u||_2(q_n\sqrt{\frac{\log n}{n}})$$

proof of Lemma S-2: If we replace $\widehat{\sigma}$ by $\bar{\sigma}$ on the left hand side, then the above results follow easily from Cauchy-Schwartz and Bernstein's inequalities by using B1.2.

Further observe that,

$$||L'_{n,\mathcal{A}}(\overline{\theta}, \widehat{\sigma}, \mathbf{Y})||_2 \leq ||L'_{n,\mathcal{A}}(\overline{\theta}, \bar{\sigma}, \mathbf{Y})||_2 + ||L'_{n,\mathcal{A}}(\overline{\theta}, \bar{\sigma}, \mathbf{Y}) - L'_{n,\mathcal{A}}(\overline{\theta}, \widehat{\sigma}, \mathbf{Y})||_2,$$

and the second term on the right hand side has order $\sqrt{\frac{q_n \log n}{n}}$, since there are q_n terms and by B3, they are uniformly bounded by $\sqrt{\frac{\log n}{n}}$. The rest of the lemma can be proved by similar arguments.

The following two lemmas are used for proving Theorem 1.

Lemma S-3 *Assuming the same conditions of Theorem 1. Then there exists a constant $C_1(\overline{\theta}) > 0$, such that for any $\eta > 0$, the probability that there exists a local minima of the restricted problem (11) within the disc:*

$$\{\theta : ||\theta - \overline{\theta}||_2 \leq C_1(\overline{\theta})\sqrt{q_n}\lambda_n\}.$$

is at least $1 - O(n^{-\eta})$ for sufficiently large n.

proof of Lemma S-3: Let $\alpha_n = \sqrt{q_n}\lambda_n$, and $Q_n(\theta, \widehat{\sigma}, \mathbf{Y}, \lambda_n) = L_n(\theta, \widehat{\sigma}, \mathbf{Y}) + \lambda_n||\theta||_1$. Then for any given constant $C > 0$ and any vector $u \in R^p$ such that $u_{\mathcal{A}^c} = 0$ and $||u||_2 = C$, by the triangle inequality and Cauchy-Schwartz inequality, we have

$$||\overline{\theta}||_1 - ||\overline{\theta} + \alpha_n u||_1 \leq \alpha_n||u||_1 \leq C\alpha_n\sqrt{q_n}.$$

Thus

$$\begin{aligned} & Q_n(\overline{\theta} + \alpha_n u, \widehat{\sigma}, \mathbf{Y}, \lambda_n) - Q_n(\overline{\theta}, \widehat{\sigma}, \mathbf{Y}, \lambda_n) \\ = \; & \{L_n(\overline{\theta} + \alpha_n u, \widehat{\sigma}, \mathbf{Y}) - L_n(\overline{\theta}, \widehat{\sigma}, \mathbf{Y})\} - \lambda_n\{||\overline{\theta}||_1 - ||\overline{\theta} + \alpha_n u||_1\} \\ \geq \; & \{L_n(\overline{\theta} + \alpha_n u, \widehat{\sigma}, \mathbf{Y}) - L_n(\overline{\theta}, \widehat{\sigma}, \mathbf{Y})\} - C\alpha_n\sqrt{q_n}\lambda_n \\ = \; & \{L_n(\overline{\theta} + \alpha_n u, \widehat{\sigma}, \mathbf{Y}) - L_n(\overline{\theta}, \widehat{\sigma}, \mathbf{Y})\} - C\alpha_n^2. \end{aligned}$$

Thus for any $\eta > 0$, there exists $c_{1,\eta}, c_{2,\eta} > 0$, such that, with probability at least $1 - O(n^{-\eta})$

$$
\begin{aligned}
& L_n(\overline{\theta} + \alpha_n u, \widehat{\sigma}, \mathbf{Y}) - L_n(\overline{\theta}, \widehat{\sigma}, \mathbf{Y}) = \alpha_n u_{\mathcal{A}}^T L'_{n,\mathcal{A}}(\overline{\theta}, \widehat{\sigma}, \mathbf{Y}) + \frac{1}{2}\alpha_n^2 u_{\mathcal{A}}^T L''_{n,\mathcal{A}\mathcal{A}}(\overline{\theta}, \widehat{\sigma}, \mathbf{Y}) u_{\mathcal{A}} \\
= & \frac{1}{2}\alpha_n^2 u_{\mathcal{A}}^T \overline{L}''_{\mathcal{A}\mathcal{A}}(\overline{\theta}) u_{\mathcal{A}} + \alpha_n u_{\mathcal{A}}^T L'_{n,\mathcal{A}}(\overline{\theta}, \widehat{\sigma}, \mathbf{Y}) + \frac{1}{2}\alpha_n^2 u_{\mathcal{A}}^T \left(L''_{n,\mathcal{A}\mathcal{A}}(\overline{\theta}, \widehat{\sigma}, \mathbf{Y}) - \overline{L}''_{\mathcal{A}\mathcal{A}}(\overline{\theta}) \right) u_{\mathcal{A}} \\
\geq & \frac{1}{2}\alpha_n^2 u_{\mathcal{A}}^T \overline{L}''_{\mathcal{A}\mathcal{A}}(\overline{\theta}) u_{\mathcal{A}} - c_{1,\eta}(\alpha_n q_n^{1/2} n^{-1/2} \sqrt{\log n}) - c_{2,\eta}(\alpha_n^2 q_n n^{-1/2} \sqrt{\log n}).
\end{aligned}
$$

In the above, the first equation is because the loss function $L(\theta, \sigma, Y)$ is quadratic in θ and $u_{\mathcal{A}^c} = 0$. The inequality is due to Lemma S-2 and the union bound. By the **assumption** $\lambda_n \sqrt{\frac{n}{\log n}} \to \infty$, we have $\alpha_n q_n^{1/2} n^{-1/2} \sqrt{\log n} = o(\alpha_n \sqrt{q_n} \lambda_n) = o(\alpha_n^2)$. Also by the **assumption that** $q_n \sim o(\sqrt{n / \log n})$, we have $\alpha_n^2 q_n n^{-1/2} \sqrt{\log n} = o(\alpha_n^2)$. Thus, with n sufficiently large

$$
Q_n(\overline{\theta} + \alpha_n u, \widehat{\sigma}, \mathbf{Y}, \lambda_n) - Q_n(\overline{\theta}, \widehat{\sigma}, \mathbf{Y}, \lambda_n) \geq \frac{1}{4}\alpha_n^2 u_{\mathcal{A}}^T \overline{L}''_{\mathcal{A}\mathcal{A}}(\overline{\theta}) u_{\mathcal{A}} - C\alpha_n^2
$$

with probability at least $1 - O(n^{-\eta})$. By B1, $u_{\mathcal{A}}^T \overline{L}''_{\mathcal{A}\mathcal{A}}(\overline{\theta}) u_{\mathcal{A}} \geq \Lambda^L_{\min}(\overline{\theta}) ||u_{\mathcal{A}}||_2^2 = \Lambda^L_{\min}(\overline{\theta}) C^2$. Thus, if we choose $C = 4/\Lambda^L_{\min}(\overline{\theta}) + \epsilon$, then for any $\eta > 0$, for sufficiently large n, the following holds

$$
\inf_{u: u_{\mathcal{A}^c} = 0, ||u||_2 = C} Q_n(\overline{\theta} + \alpha_n u, \widehat{\sigma}, \mathbf{Y}, \lambda_n) > Q_n(\overline{\theta}, \widehat{\sigma}, \mathbf{Y}, \lambda_n),
$$

with probability at least $1 - O(n^{-\eta})$. This means that a local minima exists within the disc $\{\theta : ||\theta - \overline{\theta}||_2 \leq C\alpha_n = C\sqrt{q_n}\lambda_n\}$ with probability at least $1 - O(n^{-\eta})$.

Lemma S-4 *Assuming the same conditions of Theorem 1. Then there exists a constant $C_2(\overline{\theta}) > 0$, such that for any $\eta > 0$, for sufficiently large n, the following holds with probability at least $1 - O(n^{-\eta})$: for any θ belongs to the set $S = \{\theta : ||\theta - \overline{\theta}||_2 \geq C_2(\overline{\theta})\sqrt{q_n}\lambda_n, \theta_{\mathcal{A}^c} = 0\}$, it has $||L'_{n,\mathcal{A}}(\theta, \widehat{\sigma}, \mathbf{Y})||_2 > \sqrt{q_n}\lambda_n$.*

<u>*proof of Lemma S-4*</u>: Let $\alpha_n = \sqrt{q_n}\lambda_n$. Any θ belongs to S can be written as: $\theta = \overline{\theta} + \alpha_n u$, with $u_{\mathcal{A}^c} = 0$ and $||u||_2 \geq C_2(\bar{\theta})$. Note that

$$
\begin{aligned}
L'_{n,\mathcal{A}}(\theta, \widehat{\sigma}, \mathbf{Y}) &= L'_{n,\mathcal{A}}(\overline{\theta}, \widehat{\sigma}, \mathbf{Y}) + \alpha_n L''_{n,\mathcal{A}\mathcal{A}}(\overline{\theta}, \widehat{\sigma}, \mathbf{Y})u \\
&= L'_{n,\mathcal{A}}(\overline{\theta}, \widehat{\sigma}, \mathbf{Y}) + \alpha_n (L''_{n,\mathcal{A}\mathcal{A}}(\overline{\theta}, \widehat{\sigma}, \mathbf{Y}) - \overline{L}''_{\mathcal{A}\mathcal{A}}(\overline{\theta}))u + \alpha_n \overline{L}''_{\mathcal{A}\mathcal{A}}(\overline{\theta}))u.
\end{aligned}
$$

By the triangle inequality and Lemma S-2, for any $\eta > 0$, there exists constants $c_{0,\eta}, c_{3,\eta} > 0$, such that

$$
||L'_{n,\mathcal{A}}(\theta, \widehat{\sigma}, \mathbf{Y})||_2 \geq \alpha_n ||\overline{L}''_{\mathcal{A}\mathcal{A}}(\overline{\theta}))u||_2 - c_{0,\eta}(q_n^{1/2} n^{-1/2}\sqrt{\log n}) - c_{3,\eta}||u||_2(\alpha_n q_n n^{-1/2}\sqrt{\log n})
$$

with probability at least $1 - O(n^{-\eta})$. Thus, similar as in Lemma S-3, for n sufficiently large, $||L'_{n,\mathcal{A}}(\theta, \widehat{\sigma}, \mathbf{Y})||_2 \geq \frac{1}{2}\alpha_n ||\overline{L}''_{\mathcal{A}\mathcal{A}}(\overline{\theta}))u||_2$ with probability at least $1 - O(n^{-\eta})$. By B1, $||\overline{L}''_{\mathcal{A}\mathcal{A}}(\overline{\theta}))u||_2 \geq \Lambda^L_{\min}(\bar{\theta})||u||_2$. Therefore $C_2(\overline{\theta})$ can be taken as $2/\Lambda^L_{\min}(\bar{\theta}) + \epsilon$.

The following lemma is used in proving Theorem 2.

Lemma S-5 *Assuming conditions C0-C1. Let* $D_{\mathcal{A}\mathcal{A}}(\bar{\theta}, Y) = L''_{1,\mathcal{A}\mathcal{A}}(\bar{\theta}, Y) - \overline{L}''_{\mathcal{A}\mathcal{A}}(\bar{\theta})$. *Then there exists a constant* $K_2(\bar{\theta}) < \infty$, *such that for any* $(k,l) \in \mathcal{A}$, $\lambda_{\max}(\mathrm{Var}_{\bar{\theta}}(D_{\mathcal{A},kl}(\bar{\theta}, Y))) \leq K_2(\bar{\theta})$.

<u>*proof of Lemma S-5*</u>: $\mathrm{Var}_{\bar{\theta}}(D_{\mathcal{A},kl}(\bar{\theta}, Y)) = E_{\bar{\theta}}(L''_{1,\mathcal{A},kl}(\bar{\theta}, Y)L''_{1,\mathcal{A},kl}(\bar{\theta}, Y)^T) - \overline{L}''_{\mathcal{A},kl}(\bar{\theta})\overline{L}''_{\mathcal{A},kl}(\bar{\theta})^T$. Thus it suffices to show that, there exists a constant $K_2(\bar{\theta}) > 0$, such that for all (k,l)

$$
\lambda_{\max}(E_{\bar{\theta}}(L''_{1,\mathcal{A},kl}(\bar{\theta}, Y)L''_{1,\mathcal{A},kl}(\bar{\theta}, Y)^T)) \leq K_2(\bar{\theta}).
$$

Use the same notations as in the proof of B1. Note that $L''_{1,\mathcal{A},kl}(\overline{\theta}, Y) = \tilde{x}^T \tilde{w}^2 \tilde{x}_{(k,l)} = \tilde{w}_k \tilde{y}_l x_k + \tilde{w}_l \tilde{y}_k x_l$. Thus

$$
E_{\bar{\theta}}(L''_{1,\mathcal{A},kl}(\bar{\theta}, Y)L''_{1,\mathcal{A},kl}(\bar{\theta}, Y)^T) = \tilde{w}_k^2 \mathbb{E}[\tilde{y}_l^2 x_k x_k^T] + \tilde{w}_l^2 \mathbb{E}[\tilde{y}_k^2 x_l x_l^T] + \tilde{w}_k \tilde{w}_l \mathbb{E}[\tilde{y}_k \tilde{y}_l (x_k x_l^T + x_l x_k^T)],
$$

and for $a \in \mathcal{R}^{p(p-1)/2}$

$$
\begin{aligned}
& a^T E_{\bar{\theta}}(L''_{1,\mathcal{A},kl}(\bar{\theta}, Y) L''_{1,\mathcal{A},kl}(\bar{\theta}, Y)^T) a \\
= \; & \tilde{w}_k^2 a_k^T \mathbb{E}[\tilde{y}_l^2 \tilde{y}\tilde{y}^T] a_k + \tilde{w}_l^2 a_l^T \mathbb{E}[\tilde{y}_k^2 \tilde{y}\tilde{y}^T] a_l + 2\tilde{w}_k \tilde{w}_l a_k^T \mathbb{E}[\tilde{y}_k \tilde{y}_l \tilde{y}\tilde{y}^T] a_l.
\end{aligned}
$$

Since $\sum_{k=1}^{p} ||a_k||_2^2 = 2||a||_2^2$, and by B2: $\lambda_{\max}(\mathbb{E}[\tilde{y}_i \tilde{y}_j \tilde{y}\tilde{y}^T]) \leq K_1(\bar{\theta})$ for any $1 \leq i \leq j \leq p$, the conclusion follows.

proof of Theorem 1: The existence of a solution of (11) follows from Lemma S-3. By the Karush-Kuhn-Tucker condition (Lemma S-1), for any solution $\widehat{\theta}$ of (11), it has $||L'_{n,\mathcal{A}}(\widehat{\theta}, \widehat{\sigma}, \mathbf{Y})||_\infty \leq \lambda_n$. Thus $||L'_{n,\mathcal{A}}(\widehat{\theta}, \widehat{\sigma}, \mathbf{Y})||_2 \leq \sqrt{q_n} ||L'_{n,\mathcal{A}}(\widehat{\theta}, \widehat{\sigma}, \mathbf{Y})||_\infty \leq \sqrt{q_n}\lambda_n$. Thus by Lemma S-4, for any $\eta > 0$, for n sufficiently large with probability at least $1 - O(n^{-\eta})$, all solutions of (11) are inside the disc $\{\theta : ||\theta - \overline{\theta}||_2 \leq C_2(\overline{\theta})\sqrt{q_n}\lambda_n\}$. Since $\frac{s_n}{\sqrt{q_n}\lambda_n} \to \infty$, for sufficiently large n and $(i,j) \in \mathcal{A}$: $\overline{\theta}_{ij} \geq s_n > 2C_2(\overline{\theta})\sqrt{q_n}\lambda_n$. Thus

$$
\begin{aligned}
1 - O(n^{-\eta}) \; & \leq \; P_{\overline{\theta}}\left(||\widehat{\theta}^{\mathcal{A},\lambda_n} - \overline{\theta}_{\mathcal{A}}||_2 \leq C_2(\overline{\theta})\sqrt{q_n}\lambda_n, \overline{\theta}_{ij} > 2C_2(\overline{\theta})\sqrt{q_n}\lambda_n, \text{ for all}(i,j) \in \mathcal{A} \right) \\
& \leq \; P_{\overline{\theta}}\left(\text{sign}(\widehat{\theta}_{ij}^{\mathcal{A},\lambda_n}) = \text{sign}(\overline{\theta}_{ij}), \text{ for all}(i,j) \in \mathcal{A} \right).
\end{aligned}
$$

proof of Theorem 2: For any given $\eta > 0$, let $\eta' = \eta + \kappa$. Let $\mathcal{E}_n = \{\text{sign}(\widehat{\theta}^{\mathcal{A},\lambda_n}) = \text{sign}(\bar{\theta})\}$. Then by Theorem 1, $P_{\bar{\theta}}(\mathcal{E}_n) \geq 1 - O(n^{-\eta'})$ for sufficiently large n. On $\mathcal{E}_n$, by the Karush-Kuhn-Tucker condition and the expansion of $L'_{n,\mathcal{A}}(\widehat{\theta}^{\mathcal{A},\lambda_n}, \widehat{\sigma}, \mathbf{Y})$ at $\bar{\theta}$

$$
\begin{aligned}
-\lambda_n \text{sign}(\bar{\theta}_{\mathcal{A}}) \; & = \; L'_{n,\mathcal{A}}(\widehat{\theta}^{\mathcal{A},\lambda_n}, \widehat{\sigma}, \mathbf{Y}) = L'_{n,\mathcal{A}}(\bar{\theta}, \widehat{\sigma}, \mathbf{Y}) + L''_{n,\mathcal{A}\mathcal{A}}(\bar{\theta}, \widehat{\sigma}, \mathbf{Y})\nu_n \\
& = \; \overline{L}''_{\mathcal{A}\mathcal{A}}(\bar{\theta})\nu_n + L'_{n,\mathcal{A}}(\bar{\theta}, \widehat{\sigma}, \mathbf{Y}) + \left(L''_{n,\mathcal{A}\mathcal{A}}(\bar{\theta}, \widehat{\sigma}, \mathbf{Y}) - \overline{L}''_{\mathcal{A}\mathcal{A}}(\bar{\theta}) \right)\nu_n,
\end{aligned}
$$

where $\nu_n := \widehat{\theta}^{\mathcal{A},\lambda_n}_{\mathcal{A}} - \bar{\theta}_{\mathcal{A}}$. By the above expression

$$\nu_n = -\lambda_n[\overline{L}''_{\mathcal{A}\mathcal{A}}(\bar{\theta})]^{-1}\text{sign}(\bar{\theta}_{\mathcal{A}}) - [\overline{L}''_{\mathcal{A}\mathcal{A}}(\bar{\theta})]^{-1}[L'_{n,\mathcal{A}}(\bar{\theta},\widehat{\sigma},\mathbf{Y}) + D_{n,\mathcal{A}\mathcal{A}}(\bar{\theta},\widehat{\sigma},\mathbf{Y})\nu_n], \quad \text{(S-7)}$$

where $D_{n,\mathcal{A}\mathcal{A}}(\bar{\theta},\widehat{\sigma},\mathbf{Y}) = L''_{n,\mathcal{A}\mathcal{A}}(\bar{\theta},\widehat{\sigma},\mathbf{Y}) - \overline{L}''_{\mathcal{A}\mathcal{A}}(\bar{\theta})$. Next, fix $(i,j) \in \mathcal{A}^c$, and consider the expansion of $L'_{n,ij}(\widehat{\theta}^{\mathcal{A},\lambda_n},\widehat{\sigma},\mathbf{Y})$ around $\bar{\theta}$:

$$L'_{n,ij}(\widehat{\theta}^{\mathcal{A},\lambda_n},\widehat{\sigma},\mathbf{Y}) \quad = \quad L'_{n,ij}(\bar{\theta},\widehat{\sigma},\mathbf{Y}) + L''_{n,ij,\mathcal{A}}(\bar{\theta},\widehat{\sigma},\mathbf{Y})\nu_n. \quad \text{(S-8)}$$

Then plug in (S-7) into (S-8), we get

$$\begin{aligned} L'_{n,ij}(\widehat{\theta}^{\mathcal{A},\lambda_n},\widehat{\sigma},\mathbf{Y}) &= -\lambda_n\overline{L}''_{ij,\mathcal{A}}(\bar{\theta})[\overline{L}''_{\mathcal{A}\mathcal{A}}(\bar{\theta})]^{-1}\text{sign}(\bar{\theta}_{\mathcal{A}}) - \overline{L}''_{ij,\mathcal{A}}(\bar{\theta})[\overline{L}''_{\mathcal{A}\mathcal{A}}(\bar{\theta})]^{-1}L'_{n,\mathcal{A}}(\bar{\theta},\widehat{\sigma},\mathbf{Y}) \\ + \quad L'_{n,ij}(\bar{\theta},\widehat{\sigma},\mathbf{Y}) &+ \left[D_{n,ij,\mathcal{A}}(\bar{\theta},\widehat{\sigma},\mathbf{Y}) - \overline{L}''_{ij,\mathcal{A}}(\bar{\theta})[\overline{L}''_{\mathcal{A}\mathcal{A}}(\bar{\theta})]^{-1}D_{n,\mathcal{A}\mathcal{A}}(\bar{\theta},\widehat{\sigma},\mathbf{Y})\right]\nu_n. \quad \text{(S-9)} \end{aligned}$$

By condition C2, for any $(i,j) \in \mathcal{A}^c$: $|\overline{L}''_{ij,\mathcal{A}}(\bar{\theta})[\overline{L}''_{\mathcal{A}\mathcal{A}}(\bar{\theta})]^{-1}\text{sign}(\bar{\theta}_{\mathcal{A}})| \leq \delta < 1$. Thus it suffices to prove that the remaining terms in (S-9) are all $o(\lambda_n)$ with probability at least $1 - O(n^{-\eta'})$ (uniformly for all $(i,j) \in \mathcal{A}^c$). Then since $|\mathcal{A}^c| \leq p \sim O(n^\kappa)$, by the union bound, the event $\max_{(i,j)\in\mathcal{A}^c} |L'_{n,ij}(\widehat{\theta}^{\mathcal{A},\lambda_n},\widehat{\sigma},\mathbf{Y})| < \lambda_n$ holds with probability at least $1 - O(n^{\kappa-\eta'}) = 1 - O(n^{-\eta})$, when n is sufficiently large.

By B1.4, for any $(i,j) \in \mathcal{A}^c$: $||\overline{L}''_{ij,\mathcal{A}}(\bar{\theta})[\overline{L}''_{\mathcal{A}\mathcal{A}}(\bar{\theta})]^{-1}||_2 \leq M(\bar{\theta})$. Therefore by Lemma S-2, for any $\eta > 0$, there exists a constant $C_{1,\eta} > 0$, such that

$$\max_{(i,j)\in\mathcal{A}^c} |\overline{L}''_{ij,\mathcal{A}}(\bar{\theta})[\overline{L}''_{\mathcal{A}\mathcal{A}}(\bar{\theta})]^{-1}L'_{n,\mathcal{A}}(\bar{\theta},\widehat{\sigma},\mathbf{Y})| \leq C_{1,\eta}(\sqrt{\frac{q_n \log n}{n}}) = (o(\lambda_n))$$

with probability at least $1-O(n^{-\eta})$. The claim follows by the **assumption** $\sqrt{\frac{q_n \log n}{n}} \sim o(\lambda_n)$.

By B1.2, $||\text{Var}_{\bar{\theta}}(L'_{ij}(\bar{\theta},\bar{\sigma},\mathbf{Y}))||_2 \leq M_1(\bar{\theta})$. Then similarly as in Lemma S-2, for any $\eta > 0$, there exists a constant $C_{2,\eta} > 0$, such that $\max_{i,j} |L'_{n,ij}(\bar{\theta}, \widehat{\sigma}, \mathbf{Y})| \leq C_{2,\eta}(\sqrt{\frac{\log n}{n}}) = (o(\lambda_n))$, with probability at least $1 - O(n^{-\eta})$. The claims follows by the **assumption that** $\lambda_n \sqrt{\frac{n}{\log n}} \to \infty$**.**

Note that by Theorem 1, for any $\eta > 0$, $||\nu_n||_2 \leq C(\bar{\theta})\sqrt{q_n}\lambda_n$ with probability at least $1 - O(n^{-\eta})$ for large enough n. Thus, similarly as in Lemma S-2, for any $\eta > 0$, there exists a constant $C_{3,\eta}$, such $|D_{n,ij,\mathcal{A}}(\bar{\theta}, \widehat{\sigma}, \mathbf{Y})\nu_n| \leq C_{3,\eta}(\sqrt{\frac{q_n \log n}{n}}\sqrt{q_n}\lambda_n)(= o(\lambda_n))$, with probability at least $1 - O(n^{-\eta})$. The claims follows from **the assumption** $q_n \sim o(\sqrt{\frac{n}{\log n}})$.

Finally, let $b^T = |\overline{L}''_{ij,\mathcal{A}}(\bar{\theta})[\overline{L}''_{\mathcal{A}\mathcal{A}}(\bar{\theta})]^{-1}$. By Cauchy-Schwartz inequality

$$|b^T D_{n,\mathcal{A}\mathcal{A}}(\bar{\theta}, \bar{\sigma}, \mathbf{Y})\nu_n| \leq ||b^T D_{n,\mathcal{A}\mathcal{A}}(\bar{\theta}, \bar{\sigma}, \mathbf{Y})||_2||\nu_n||_2 \leq q_n\lambda_n \max_{(k,l)\in\mathcal{A}} |b^T D_{n,\mathcal{A},kl}(\bar{\theta}, \bar{\sigma}, \mathbf{Y})|.$$

In order to show the right hand side is $o(\lambda_n)$ with probability at least $1 - O(n^{-\eta})$, it suffices to show $\max_{(k,l)\in\mathcal{A}} |b^T D_{n,\mathcal{A},kl}(\bar{\theta}, \bar{\sigma}, \mathbf{Y})| = O(\sqrt{\frac{\log n}{n}})$ with probability at least $1 - O(n^{-\eta})$, because of the **the assumption** $q_n \sim o(\sqrt{\frac{n}{\log n}})$. This is implied by

$$E_{\bar{\theta}}(|b^T D_{\mathcal{A},kl}(\bar{\theta}, \bar{\sigma}, Y)|^2) \leq ||b||_2^2 \lambda_{\max}(\mathrm{Var}_{\bar{\theta}}(D_{\mathcal{A},kl}(\bar{\theta}, \bar{\sigma}, Y)))$$

being bounded, which follows immediately from B1.4 and Lemma S-5. Finally, similarly as in Lemma S-2,

$$|b^T D_{n,\mathcal{A}\mathcal{A}}(\bar{\theta}, \widehat{\sigma}, \mathbf{Y})\nu_n| \leq |b^T D_{n,\mathcal{A}\mathcal{A}}(\bar{\theta}, \bar{\sigma}, \mathbf{Y})\nu_n| + |b^T(D_{n,\mathcal{A}\mathcal{A}}(\bar{\theta}, \bar{\sigma}, \mathbf{Y}) - D_{n,\mathcal{A}\mathcal{A}}(\bar{\theta}, \widehat{\sigma}, \mathbf{Y}))\nu_n|,$$

where by B3, the second term on the right hand side is bounded by $O_p(\sqrt{\frac{\log n}{n}})||b||_2||\nu_n||_2$. Note that $||b||_2 \sim \sqrt{q_n}$, thus the second term is also of order $o(\lambda_n)$ by **the assumption** $q_n \sim o(\sqrt{\frac{n}{\log n}})$. This completes the proof.

<u>*proof of Theorem 3*</u>: By Theorems 1 and 2 and the Karush-Kuhn-Tucker condition,

for any $\eta > 0$, with probability at least $1 - O(n^{-\eta})$, a solution of the restricted problem is also a solution of the original problem. On the other hand, by Theorem 2 and the Karush-Kuhn-Tucker condition, with high probability, any solution of the original problem is a solution of the restricted problem. Therefore, by Theorem 1, the conclusion follows.

Part III

In this section, we provide details for the implementation of `space` which takes advantage of the sparse structure of $\mathcal{X}$. Denote the target loss function as

$$f(\theta) = \frac{1}{2} \|\mathcal{Y} - \mathcal{X}\theta\|^2 + \lambda_1 \sum_{i<j} |\rho^{ij}|. \tag{S-10}$$

Our goal is to find $\widehat{\theta} = \text{argmin}_\theta f(\theta)$ for a given λ_1. We will employ `active-shooting` algorithm (Section 2.3) to solve this optimization problem.

Without loss of generality, we assume $\text{mean}(\mathbf{Y}_i) = 1/n \sum_{k=1}^n y_i^k = 0$ for $i = 1, \ldots, p$. Denote $\xi_i = \mathbf{Y}_i^T \mathbf{Y}_i$. We have

$$\mathcal{X}_{(i,j)}^T \mathcal{X}_{(i,j)} = \xi_j \frac{\sigma^{jj}}{\sigma^{ii}} + \xi_i \frac{\sigma^{ii}}{\sigma^{jj}};$$

$$\mathcal{Y}^T \mathcal{X}_{(i,j)} = \sqrt{\frac{\sigma^{jj}}{\sigma^{ii}}} \mathbf{Y}_i^T \mathbf{Y}_j + \sqrt{\frac{\sigma^{ii}}{\sigma^{jj}}} \mathbf{Y}_j^T \mathbf{Y}_i.$$

Denote $\rho^{ij} = \rho_{(i,j)}$. We now present details of the initialization step and the updating steps in the `active-shooting` algorithm.

1. Initialization

Let

$$\begin{aligned}\rho^{(0)}_{(i,j)} &= \frac{\left(|\mathcal{Y}^T\mathcal{X}_{(i,j)}|-\lambda_1\right)_+ \cdot \mathrm{sign}(\mathcal{Y}^T\mathcal{X}_{(i,j)})}{\mathcal{X}^T_{(i,j)}\mathcal{X}_{(i,j)}} \\ &= \frac{\left(\left|\sqrt{\frac{\sigma^{jj}}{\sigma^{ii}}}\mathbf{Y}_i^T\mathbf{Y}_j+\sqrt{\frac{\sigma^{ii}}{\sigma^{jj}}}\mathbf{Y}_j^T\mathbf{Y}_i\right|-\lambda_1\right)_+ \cdot \mathrm{sign}(\mathbf{Y}_i^T\mathbf{Y}_j)}{\xi_j\frac{\sigma^{jj}}{\sigma^{ii}}+\xi_i\frac{\sigma^{ii}}{\sigma^{jj}}}.\end{aligned} \tag{S-11}$$

For $j = 1, \ldots, p$, compute

$$\widehat{\mathbf{Y}}^{(0)}_j = \left(\sqrt{\frac{\sigma^{11}}{\sigma^{jj}}}\mathbf{Y}_1, ..., \sqrt{\frac{\sigma^{pp}}{\sigma^{jj}}}\mathbf{Y}_p\right) \cdot \begin{pmatrix} \rho^{(0)}_{(1,j)} \\ \vdots \\ \rho^{(0)}_{(p,j)} \end{pmatrix}, \tag{S-12}$$

and

$$E^{(0)} = \mathcal{Y} - \widehat{\mathcal{Y}}^{(0)} = \left((E^{(0)}_1)^T, ..., (E^{(0)}_p)^T\right), \tag{S-13}$$

where $E^{(0)}_j = \mathbf{Y}_j - \widehat{\mathbf{Y}}^{(0)}_j$, for $1 \le j \le p$.

2. Update $\rho^{(0)}_{(i,j)} \longrightarrow \rho^{(1)}_{(i,j)}$

Let

$$A_{(i,j)} = (E^{(0)}_j)^T \cdot \sqrt{\frac{\sigma^{ii}}{\sigma^{jj}}}\mathbf{Y}_i, \tag{S-14}$$

$$A_{(j,i)} = (E^{(0)}_i)^T \cdot \sqrt{\frac{\sigma^{jj}}{\sigma^{ii}}}\mathbf{Y}_j. \tag{S-15}$$

We have

$$\begin{aligned}(E^{(0)})^T\mathcal{X}_{(i,j)} &= (E^{(0)}_i)^T \cdot \sqrt{\tfrac{\sigma^{jj}}{\sigma^{ii}}}\mathbf{Y}_j + (E^{(0)}_j)^T \cdot \sqrt{\tfrac{\sigma^{ii}}{\sigma^{jj}}}\mathbf{Y}_i \\ &= A_{(j,i)} + A_{(i,j)}.\end{aligned} \tag{S-16}$$

It follows

$$\begin{aligned}\rho^{(1)}_{(i,j)} &= \mathrm{sign}\left(\frac{(E^{(0)})^T\mathcal{X}_{(i,j)}}{\mathcal{X}^T_{(i,j)}\mathcal{X}_{(i,j)}} + \rho^{(0)}_{(i,j)}\right)\left(\left|\frac{(E^{(0)})^T\mathcal{X}_{(i,j)}}{\mathcal{X}^T_{(i,j)}\mathcal{X}_{(i,j)}} + \rho^{(0)}_{(i,j)}\right| - \frac{\lambda_1}{\mathcal{X}^T_{(i,j)}\mathcal{X}_{(i,j)}}\right)_+ \\ &= \mathrm{sign}\left(\frac{A_{(j,i)}+A_{(i,j)}}{\xi_j\frac{\sigma^{jj}}{\sigma^{ii}}+\xi_i\frac{\sigma^{ii}}{\sigma^{jj}}} + \rho^{(0)}_{(i,j)}\right)\left(\left|\frac{A_{(j,i)}+A_{(i,j)}}{\xi_j\frac{\sigma^{jj}}{\sigma^{ii}}+\xi_i\frac{\sigma^{ii}}{\sigma^{jj}}} + \rho^{(0)}_{(i,j)}\right| - \frac{\lambda_1}{\xi_j\frac{\sigma^{jj}}{\sigma^{ii}}+\xi_i\frac{\sigma^{ii}}{\sigma^{jj}}}\right)_+.\end{aligned} \tag{S-17}$$

3. Update $\rho^{(t)} \longrightarrow \rho^{(t+1)}$

From the previous iteration, we have

- $E^{(t-1)}$: residual in the previous iteration ($np \times 1$ vector).
- (i_0, j_0): index of coefficient that is updated in the previous iteration.
- $\rho^{(t)}_{(i,j)} = \begin{cases} \rho^{(t-1)}_{(i,j)} & \text{if } (i,j) \neq (i_0, j_0), \text{ nor } (j_0, i_0) \\ \rho^{(t-1)}_{(i,j)} - \Delta & \text{if } (i,j) = (i_0, j_0), \text{ or } (j_0, i_0) \end{cases}$

Then,

$$\begin{aligned} E_k^{(t)} &= E_k^{(t-1)} \text{ for } k \neq i_0, j_0; \\ E_{j_0}^{(t)} &= E_{j_0}^{(t-1)} + \widehat{\mathbf{Y}}_{j_0}^{(t-1)} - \widehat{\mathbf{Y}}_{j_0}^{(t)} \\ &= E_{j_0}^{(t-1)} + \sum_{i=1}^{p} \sqrt{\tfrac{\sigma^{ii}}{\sigma^{j_0 j_0}}} \mathbf{Y}_i (\rho^{(t-1)}_{(i,j_0)} - \rho^{(t)}_{(i,j_0)}) \\ &= E_{j_0}^{(t-1)} + \sqrt{\tfrac{\sigma^{i_0 i_0}}{\sigma^{j_0 j_0}}} \mathbf{Y}_{i_0} \cdot \Delta; \\ E_{i_0}^{(t)} &= E_{i_0}^{(t-1)} + \sqrt{\tfrac{\sigma^{j_0 j_0}}{\sigma^{i_0 i_0}}} \mathbf{Y}_{j_0} \cdot \Delta. \end{aligned} \tag{S-18}$$

Suppose the index of the coefficient we would like to update in this iteration is (i_1, j_1), then let

$$A_{(i_1,j_1)} = (E_{j_1}^{(t)})^T \cdot \sqrt{\frac{\sigma^{i_1 i_1}}{\sigma^{j_1 j_1}}} \mathbf{Y}_{i_1},$$

$$A_{(j_1,i_1)} = (E_{i_1}^{(t)})^T \cdot \sqrt{\frac{\sigma^{j_1 j_1}}{\sigma^{i_1 i_1}}} \mathbf{Y}_{j_1}.$$

We have

$$\begin{aligned} \rho^{(t+1)}_{(i,j)} = \ & \text{sign}\left(\frac{A_{(j_1,i_1)} + A_{(i_1,j_1)}}{\xi_j \frac{\sigma^{j_1 j_1}}{\sigma^{i_1 i_1}} + \xi_{i_1} \frac{\sigma^{i_1 i_1}}{\sigma^{j_1 j_1}}} + \rho^{(t)}_{(i_1,j_1)} \right) \\ & \times \left(\left| \frac{A_{(j_1,i_1)} + A_{(i_1,j_1)}}{\xi_j \frac{\sigma^{j_1 j_1}}{\sigma^{i_1 i_1}} + \xi_{i_1} \frac{\sigma^{i_1 i_1}}{\sigma^{j_1 j_1}}} + \rho^{(t)}_{(i_1,j_1)} \right| - \frac{\lambda_1}{\xi_j \frac{\sigma^{jj}}{\sigma^{ii}} + \xi_i \frac{\sigma^{ii}}{\sigma^{jj}}} \right)_+ . \end{aligned} \tag{S-19}$$

Using the above steps 1–3, we have implemented the `active-shooting` algorithm in `c`, and the corresponding `R` package `space` to fit the `space` model is available on `cran`.