Appendix

A provides the detailed derivation of the updates for Algorithm 1.

B provides the proofs of theorems stated in Section 3.

C provides details on the simulated data in Section 4.

A Derivation of the Nodewise Tensor Lasso Estimator

A.1 Off-Diagonal updates

For $1 \le i_k < j_k \le m_k$, $T_{i_k j_k}(\boldsymbol{\Psi}_k^{\text{off}})$ can be computed in closed form:

$$(T_{i_k j_k}(\boldsymbol{\Psi}_k))^{\text{off}}_{i_k j_k} = \frac{S_{\frac{\lambda_k}{N}}\Big(F_{\boldsymbol{\mathcal{X}},\{\boldsymbol{\Psi}_k\}_{k=1}^K}\Big)}{(\frac{1}{N}\boldsymbol{\mathcal{X}}_{(k)}\boldsymbol{\mathcal{X}}_{(k)}^T)_{i_k i_k} + (\frac{1}{N}\boldsymbol{\mathcal{X}}_{(k)}\boldsymbol{\mathcal{X}}_{(k)}^T)_{j_k j_k}}, \tag{9}$$

where

$$\begin{aligned} F_{\boldsymbol{\mathcal{X}},\{\boldsymbol{\Psi}_k\}_{k=1}^K} = -\frac{1}{N}\Bigg(&\Big((\boldsymbol{\mathcal{W}}_{(k)} \circ \boldsymbol{\mathcal{X}}_{(k)})\boldsymbol{\mathcal{X}}_{(k)}^T\Big)_{i_k j_k} + \Big((\boldsymbol{\mathcal{W}}_{(k)} \circ \boldsymbol{\mathcal{X}}_{(k)})\boldsymbol{\mathcal{X}}_{(k)}^T\Big)_{j_k i_k} \\ &+ \Big(\boldsymbol{\mathcal{X}}_{(k)}(\boldsymbol{\mathcal{X}} \times_k \boldsymbol{\Psi}_k^{\text{off},i_k j_k})_{(k)}^T\Big)_{j_k i_k} + \Big(\boldsymbol{\mathcal{X}}_{(k)}(\boldsymbol{\mathcal{X}} \times_k \boldsymbol{\Psi}_k^{\text{off},i_k j_k})_{(k)}^T\Big)_{i_k j_k} \\ &+ \sum_{l \neq k}\Big(\boldsymbol{\mathcal{X}}_{(k)}(\boldsymbol{\mathcal{X}} \times_l \boldsymbol{\Psi}_l^{\text{off}})_{(k)}^T\Big)_{i_k j_k} + \sum_{l \neq k}\Big(\boldsymbol{\mathcal{X}}_{(k)}(\boldsymbol{\mathcal{X}} \times_l \boldsymbol{\Psi}_l^{\text{off}})_{(k)}^T\Big)_{j_k i_k}\Bigg). \end{aligned}$$

Here the $\circ$ operator denotes the Hadamard product between matrices; $\boldsymbol{\Psi}_k^{\text{off},i_k j_k}$ is $\boldsymbol{\Psi}_k^{\text{off}}$ with the (i_k, j_k) entry being zero; and $S_\lambda(x) := \text{sign}(x)(|x| - \lambda)_+$ is the soft-thresholding operator.

A.2 Diagonal updates

For $\boldsymbol{\mathcal{W}}$,

$$(T(\boldsymbol{\mathcal{W}}))_{i_{[1:K]}} = \frac{-\Big(\boldsymbol{\mathcal{X}}_{(N)}^T\boldsymbol{\mathcal{Y}}_{(N)}\Big)_{i_{[1:K]}} + \sqrt{\Big(\boldsymbol{\mathcal{X}}_{(N)}^T\boldsymbol{\mathcal{Y}}_{(N)}\Big)^2_{i_{[1:K]}} + 4\Big(\boldsymbol{\mathcal{X}}_{(N)}\boldsymbol{\mathcal{X}}_{(N)}^T\Big)_{i_{[1:K]}}}}{2\Big(\boldsymbol{\mathcal{X}}_{(N)}\boldsymbol{\mathcal{X}}_{(N)}^T\Big)_{i_{[1:K]}}}. \tag{10}$$

Here we define $\boldsymbol{\mathcal{Y}} := \sum_{k=1}^K \Big(\boldsymbol{\mathcal{X}} \times_k \boldsymbol{\Psi}_k^{\text{off}}\Big)$. Equations (9) and (10) give necessary ingredients for designing a coordinate descent approach to minimizing the objective function in (4). The optimization procedure is summarized in Algorithm 1.

A.3 Derivation of updates

Note that for $1 \le i_k < j_k \le m_k$, $1 \le k \le K$,

$$\begin{aligned} &Q_N(\{\boldsymbol{\Psi}_k\}_{k=1}^K) \\ &= (N/2)\Big(\sum_{i_{[1:k-1,k+1:K]}} ({\boldsymbol{\mathcal{X}}^{i_k}_{i_{[1:K]}}}^2 + {\boldsymbol{\mathcal{X}}^{j_k}_{i_{[1:K]}}}^2)\Big)\Big((\boldsymbol{\Psi}_k)_{i_k j_k}\Big)^2 \\ &+ NF_{\boldsymbol{\mathcal{X}},\{\boldsymbol{\Psi}\}_{k=1}^K}(\boldsymbol{\Psi}_k)_{i_k j_k} + \lambda_k|(\boldsymbol{\Psi}_k)_{i_k j_k}| \\ &+ \text{terms independent of } (\boldsymbol{\Psi}_k)_{i_k j_k}, \end{aligned}$$

where

$$
\begin{aligned}
F_{\boldsymbol{\mathcal{X}},\{\Psi\}_{k=1}^K} = - \sum_{i_{[1:k-1,k+1:K]}} \Big(& \boldsymbol{\mathcal{W}}^{i_k}_{i_{[1:K]}} \boldsymbol{\mathcal{X}}^{i_k}_{i_{[1:K]}} \boldsymbol{\mathcal{X}}^{j_k}_{i_{[1:K]}} + \boldsymbol{\mathcal{W}}^{j_k}_{i_{[1:K]}} \boldsymbol{\mathcal{X}}^{j_k}_{i_{[1:K]}} \boldsymbol{\mathcal{X}}^{i_k}_{i_{[1:K]}} \\
& + (\boldsymbol{\Psi}_k)^T_{i_k,\backslash\{i_k,j_k\}} \boldsymbol{\mathcal{X}}^{\backslash\{i_k,j_k\}}_{i_{[1:K]}} \boldsymbol{\mathcal{X}}^{j_k}_{i_{[1:K]}} \\
& + (\boldsymbol{\Psi}_k)^T_{j_k,\backslash\{i_k,j_k\}} \boldsymbol{\mathcal{X}}^{\backslash\{i_k,j_k\}}_{i_{[1:K]}} \boldsymbol{\mathcal{X}}^{i_k}_{i_{[1:K]}} \\
& + \sum_{l\in[1:k-1,k+1:K]} (\boldsymbol{\Psi}_l)^T_{i_l,\backslash i_l} \boldsymbol{\mathcal{X}}^{i_k,\backslash i_l}_{i_{[1:K]}} \boldsymbol{\mathcal{X}}^{j_k}_{i_{[1:K]}} \\
& + \sum_{l\in[1:k-1,k+1:K]} (\boldsymbol{\Psi}_l)^T_{i_l,\backslash i_l} \boldsymbol{\mathcal{X}}^{j_k,\backslash i_l}_{i_{[1:K]}} \boldsymbol{\mathcal{X}}^{i_k}_{i_{[1:K]}} \Big).
\end{aligned}
$$

Here $\boldsymbol{\mathcal{X}}^{i_k}_{i_{[1:K]}}$ denotes the element of $\boldsymbol{\mathcal{X}}$ indexed by $i_{[1:K]}$ except that the kth index is replaced by i_k and $\boldsymbol{\mathcal{X}}^{i_k,j_l}_{i_{[1:K]}}$ denotes the element of $\boldsymbol{\mathcal{X}}$ indexed by $i_{[1:K]}$ except that the k,lth indices are replaced by i_k, j_l. Note the following equivalence:

$$
\begin{aligned}
\sum_{i_{[1:k-1,k+1:K]}} \boldsymbol{\mathcal{W}}^{i_k}_{i_{[1:K]}} \boldsymbol{\mathcal{X}}^{i_k}_{i_{[1:K]}} \boldsymbol{\mathcal{X}}^{j_k}_{i_{[1:K]}} &= \Big((\boldsymbol{\mathcal{W}}_{(k)} \circ \boldsymbol{\mathcal{X}}_{(k)}) \boldsymbol{\mathcal{X}}^T_{(k)} \Big)_{i_k j_k} \\
\sum_{i_{[1:k-1,k+1:K]}} \boldsymbol{\mathcal{X}}^{i_k}_{i_{[1:K]}} \boldsymbol{\mathcal{X}}^{j_k}_{i_{[1:K]}} &= (\boldsymbol{\mathcal{X}}_{(k)} \boldsymbol{\mathcal{X}}^T_{(k)})_{i_k j_k} \\
\sum_{i_{[1:k-1,k+1:K]}} (\boldsymbol{\Psi}_l)^T_{i_l,\cdot} \boldsymbol{\mathcal{X}}^{i_k,\cdot}_{i_{[1:K]}} \boldsymbol{\mathcal{X}}^{j_k}_{i_{[1:K]}} &= \Big(\boldsymbol{\mathcal{X}}_{(k)} (\boldsymbol{\mathcal{X}} \times_l \boldsymbol{\Psi}_l)^T_{(k)} \Big)_{j_k i_k},
\end{aligned}
$$

where $\boldsymbol{\mathcal{W}}$ is a tensor of the same dimensions of $\boldsymbol{\mathcal{X}}$, formed by tensorize values in $\boldsymbol{\mathcal{W}}$, and in the case of $N > 1$ the last mode of $\boldsymbol{\mathcal{W}}$ is the observation mode similarly to $\boldsymbol{\mathcal{X}}$ but with exact replicates. Using the tensor notation and standard sub-differential method, Equation (9) then follows.

For $\boldsymbol{\mathcal{W}}_{i_{[1:K]}}$, using similar tensor operations,

$$
\begin{aligned}
& \frac{\partial}{\partial \boldsymbol{\mathcal{W}}_{i_{[1:K]}}} Q_N(\boldsymbol{\mathcal{W}}, \{\boldsymbol{\Psi}^{\text{off}}_k\}_{k=1}^K) = 0 \\
\iff & -\frac{1}{\boldsymbol{\mathcal{W}}_{i_{[1:K]}}} + \boldsymbol{\mathcal{W}}^2_{i_{[1:K]}} \boldsymbol{\mathcal{X}}^2_{i_{[1:K]}} + \boldsymbol{\mathcal{W}}_{i_{[1:K]}} \Big(\boldsymbol{\mathcal{X}}_{i_{[1:K]}} \sum_{k=1}^K (\boldsymbol{\mathcal{X}} \times_k \boldsymbol{\Psi}^{\text{off}}_k)_{i_{[1:K]}}) \Big) = 0 \\
\iff & \boldsymbol{\mathcal{W}}^2_{i_{[1:K]}} \Big(\boldsymbol{\mathcal{X}}^T_{(N)} \boldsymbol{\mathcal{X}}_{(N)} \Big)_{i_{[1:K]}} + \boldsymbol{\mathcal{W}}_{i_{[1:K]}} \Big(\boldsymbol{\mathcal{X}}^T_{(N)} \sum_{k=1}^K (\boldsymbol{\mathcal{X}} \times_k \boldsymbol{\Psi}^{\text{off}}_k) \Big)_{i_{[1:K]}} - 1 = 0
\end{aligned}
$$

which is a quadratic equation in $\boldsymbol{\mathcal{W}}_{i_{[1:K]}}$ and since $\boldsymbol{\mathcal{W}}_{i_{[1:K]}} > 0$, so the positive root has been retained as the solution. Note that the estimation for one entry of $\boldsymbol{\mathcal{W}}$ is independent of the other entries. So during the estimation process we update all the entries at once by noting that $\operatorname{diag}\Big(\boldsymbol{\mathcal{X}}^T_{(N)} \boldsymbol{\mathcal{X}}_{(N)}\Big) = \Big(\Big(\boldsymbol{\mathcal{X}}^T_{(N)} \boldsymbol{\mathcal{X}}_{(N)}\Big)_{i_{[1:K]}}, \forall i_{[1:K]} \Big)$.

B Proofs of Main Theorems

We first list some properties of the loss function.

Lemma B.1. The following is true for the loss function:

(i) There exist constants $0 < \Lambda^L_{\min} \le \Lambda^L_{\max} < \infty$ such that for $\mathcal{S}_k := \{(i_k, j_k) : 1 \le i_k < j_k \le m_k\}, k = 1, \ldots, K$,

$$
\Lambda^L_{\min} \le \lambda_{\min}(\bar{L}''_{\mathcal{S}_k,\mathcal{S}_k}(\bar{\boldsymbol{\beta}})) \le \lambda_{\max}(\bar{L}''_{\mathcal{S}_k,\mathcal{S}_k}(\bar{\boldsymbol{\beta}})) \le \Lambda^L_{\max}
$$

(ii) There exists a constant $K(\bar{\boldsymbol{\beta}}) < \infty$ such that for all $1 \le i_k < j_k \le m_k$, $\bar{L}''_{i_k j_k, i_k j_k}(\bar{\boldsymbol{\beta}}) \le K(\bar{\boldsymbol{\beta}})$

(iii) There exist constant $M_1(\bar{\boldsymbol{\beta}}), M_2(\bar{\boldsymbol{\beta}}) < \infty$, such that for any $1 \le i_k < j_k \le m_k$

$$
\operatorname{Var}_{\bar{\boldsymbol{\mathcal{W}}},\bar{\boldsymbol{\beta}}}(L'_{i_k j_k}(\bar{\boldsymbol{\mathcal{W}}}, \bar{\boldsymbol{\beta}}, \boldsymbol{\mathcal{X}})) \le M_1(\bar{\boldsymbol{\beta}}),\ \operatorname{Var}_{\bar{\boldsymbol{\mathcal{W}}},\bar{\boldsymbol{\beta}}}(L''_{i_k j_k, i_k j_k}(\bar{\boldsymbol{\mathcal{W}}}, \bar{\boldsymbol{\beta}}, \boldsymbol{\mathcal{X}})) \le M_2(\bar{\boldsymbol{\beta}})
$$

(iv) There exists a constant $0 < g(\bar{\boldsymbol{\beta}}) < \infty$, such that for all $(i,j) \in \mathcal{A}_k$

$$\bar{L}''_{ij,ij}(\bar{\boldsymbol{\mathcal{W}}}, \bar{\boldsymbol{\beta}}) - \bar{L}''_{ij,\mathcal{A}_k^{ij}}(\bar{\boldsymbol{\mathcal{W}}}, \bar{\boldsymbol{\beta}})[\bar{L}''_{\mathcal{A}_k^{ij},\mathcal{A}_k^{ij}}(\bar{\boldsymbol{\mathcal{W}}}, \bar{\boldsymbol{\beta}})]^{-1}\bar{L}''_{\mathcal{A}_k^{ij},ij}(\bar{\boldsymbol{\mathcal{W}}}, \bar{\boldsymbol{\beta}}) \geq g(\bar{\boldsymbol{\beta}}),$$

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

(v) There exists a constant $M(\bar{\boldsymbol{\beta}}) < \infty$, such that for any $(i,j) \in \mathcal{A}_k^c$

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

proof of Lemma B.1. We prove (i). $(ii-v)$ are then direct consequences, and the proofs follow from the proofs of B1.1-B1.4 in Peng et al. (2009), with the modifications being that the indexing is now with respect to each k for $1 \leq k \leq K$.

Consider the loss function in matrix form as in (5). Then $\bar{L}''_{\mathcal{S}_k,\mathcal{S}_k}(\bar{\boldsymbol{\beta}})$ is equivalent to $\frac{\partial^2}{\partial \boldsymbol{\Psi}_k^{\text{off}} \partial \boldsymbol{\Psi}_k^{\text{off}}} L(\boldsymbol{\mathcal{W}}, \{\boldsymbol{\Psi}_k^{\text{off}}\}_{k=1}^K)$, which is

$$\begin{aligned}
&\frac{\partial^2}{\partial \boldsymbol{\Psi}_k^{\text{off}} \partial \boldsymbol{\Psi}_k^{\text{off}}} \left(\text{tr}(\boldsymbol{\Psi}_k^T \mathbf{S} \boldsymbol{\Psi}_k) + \text{first order terms in } \boldsymbol{\Psi}_k + \text{terms independent of } \boldsymbol{\Psi}_k \right) \\
&= \frac{\partial^2}{\partial \boldsymbol{\Psi}_k^{\text{off}} \partial \boldsymbol{\Psi}_k^{\text{off}}} \Bigg(\text{tr}((\boldsymbol{\Psi}_k^{\text{off}} + \text{diag}(\boldsymbol{\Psi}_k))^T \mathbf{S} (\boldsymbol{\Psi}_k^{\text{off}} + \text{diag}(\boldsymbol{\Psi}_k))) + \text{first order terms in } \boldsymbol{\Psi}_k^{\text{off}} \\
&\qquad + \text{terms independent of } \boldsymbol{\Psi}_k^{\text{off}} \Bigg) \\
&= \frac{\partial^2}{\partial \boldsymbol{\Psi}_k^{\text{off}} \partial \boldsymbol{\Psi}_k^{\text{off}}} \left(\text{tr}((\boldsymbol{\Psi}_k^{\text{off}})^T \mathbf{S} \boldsymbol{\Psi}_k^{\text{off}}) + \text{first order terms in } \boldsymbol{\Psi}_k^{\text{off}} + \text{terms independent of } \boldsymbol{\Psi}_k^{\text{off}} \right) \\
&= \mathbf{S} = \frac{1}{N} \text{vec}(\boldsymbol{\mathcal{X}})^T \text{vec}(\boldsymbol{\mathcal{X}}).
\end{aligned}$$

Thus $\bar{L}''_{\mathcal{S}_k,\mathcal{S}_k}(\boldsymbol{\beta}) = E_{\boldsymbol{\mathcal{W}},\boldsymbol{\beta}}(\mathbf{S})$. Then for any non-zero $\mathbf{a} \in \mathbb{R}^p$, we have

$$\mathbf{a}^T \bar{L}''_{\mathcal{S}_k,\mathcal{S}_k}(\bar{\boldsymbol{\beta}})\mathbf{a} = \mathbf{a}^T \bar{\boldsymbol{\Sigma}} \mathbf{a} \geq \|\mathbf{a}\|_2^2 \lambda_{\min}(\bar{\boldsymbol{\Sigma}}).$$

Similarly, $\mathbf{a}^T \bar{L}''_{\mathcal{S}_k,\mathcal{S}_k}(\bar{\boldsymbol{\beta}})\mathbf{a} \leq \|\mathbf{a}\|_2^2 \lambda_{\max}(\bar{\boldsymbol{\Sigma}})$. By (A2), $\bar{\boldsymbol{\Sigma}}$ has bounded eigenvalues, thus the lemma is proved.

□

Lemma B.2. Suppose conditions (A1-A2) hold, then for any $\eta > 0$, there exist constant $c_{0,\eta}, c_{1,\eta}, c_{2,\eta}, c_{3,\eta}$, such that for any $u \in \mathbb{R}^{q_k}$ the following events hold with probability at least $1 - O(\exp(-\eta \log p))$ for sufficiently large N:

(i) $\|L'_{N,\mathcal{A}_k}(\bar{\boldsymbol{\mathcal{W}}}, \bar{\boldsymbol{\beta}}, \boldsymbol{\mathcal{X}})\|_2 \leq c_{0,\eta}\sqrt{q_k \frac{\log p}{N}}$

(ii) $|u^T L'_{N,\mathcal{A}_k}(\bar{\boldsymbol{\mathcal{W}}}, \bar{\boldsymbol{\beta}}, \boldsymbol{\mathcal{X}})| \leq c_{1,\eta}\|u\|_2\sqrt{q_k \frac{\log p}{N}}$

(iii) $|u^T L''_{N,\mathcal{A}_k\mathcal{A}_k}(\bar{\boldsymbol{\mathcal{W}}}, \bar{\boldsymbol{\beta}}, \boldsymbol{\mathcal{X}})u - u^T \bar{L}''_{\mathcal{A}_k\mathcal{A}_k}(\bar{\boldsymbol{\beta}})u| \leq c_{2,\eta}\|u\|_2^2 q_k \sqrt{\frac{\log p}{N}}$

(iv) $|L''_{N,\mathcal{A}_k\mathcal{A}_k}(\bar{\boldsymbol{\mathcal{W}}}, \bar{\boldsymbol{\beta}}, \boldsymbol{\mathcal{X}})u - \bar{L}''_{\mathcal{A}_k\mathcal{A}_k}(\bar{\boldsymbol{\beta}})u| \leq c_{3,\eta}\|u\|_2^2 q_k \sqrt{\frac{\log p}{N}}$

proof of Lemma B.2. (i) By Cauchy-Schwartz inequality,

$$\|L'_{N,\mathcal{A}_k}(\bar{\boldsymbol{\mathcal{W}}}, \bar{\boldsymbol{\beta}}, \boldsymbol{\mathcal{X}})\|_2 \leq \sqrt{q_k} \max_{i \in \mathcal{A}_k} |L'_{N,i}(\bar{\boldsymbol{\mathcal{W}}}, \bar{\boldsymbol{\beta}}, \boldsymbol{\mathcal{X}})|.$$

Then note that

$$
\begin{aligned}
&L'_{N,i}(\boldsymbol{\mathcal{W}},\boldsymbol{\beta},\boldsymbol{\mathcal{X}}) \\
&= \sum_{i_{[1:k-1,k+1:K]}} (e_{i_{[1:k-1]},p,i_{[k+1:K]}}(\boldsymbol{\mathcal{W}},\boldsymbol{\beta})\boldsymbol{\mathcal{X}}_{i_{[1:k-1]},q,i_{[k+1:K]}} + e_{i_{[1:k-1]},q,i_{[k+1:K]}}(\boldsymbol{\mathcal{W}},\boldsymbol{\beta})\boldsymbol{\mathcal{X}}_{i_{[1:k-1]},p,i_{[k+1:K]}}),
\end{aligned}
$$

where $e_{i_{[1:k-1]},p,i_{[k+1:K]}}\boldsymbol{\mathcal{X}}_{i_{[1:k-1]},q,i_{[k+1:K]}}(\boldsymbol{\mathcal{W}},\boldsymbol{\beta})$ is defined by

$$
w_{i_{[1:k-1]},p,i_{[k+1:K]}}\boldsymbol{\mathcal{X}}_{i_{[1:k-1]},p,i_{[k+1:K]}} + \sum_{j_k\neq p}(\boldsymbol{\Psi}_k)_{p,j_k}\boldsymbol{\mathcal{X}}_{i_{[1:k-1]},j_k,i_{[k+1:K]}} + \sum_{l\neq k}\sum_{j_l\neq i_l}(\boldsymbol{\Psi}_l)_{i_l,j_l}\boldsymbol{\mathcal{X}}_{i_{[1:k-1]},p,i_{[k+1:K]}}.
$$

Then evaluated at the true parameter values $(\bar{\boldsymbol{\mathcal{W}}},\bar{\boldsymbol{\beta}})$, we have $e_{i_{[1:k-1]},p,i_{[k+1:K]}}(\bar{\boldsymbol{\mathcal{W}}},\bar{\boldsymbol{\beta}})$ uncorrelated with $\boldsymbol{\mathcal{X}}_{i_{[1:k-1]},\backslash p,i_{[k+1:K]}}$ and $E_{(\bar{\boldsymbol{\mathcal{W}}},\bar{\boldsymbol{\beta}})}(e_{i_{[1:k-1]},p,i_{[k+1:K]}}(\bar{\boldsymbol{\mathcal{W}}},\bar{\boldsymbol{\beta}})) = 0$. Also, since $\boldsymbol{\mathcal{X}}$ is subgaussian and $\mathrm{Var}(L'_{N,i}(\bar{\boldsymbol{\mathcal{W}}},\bar{\boldsymbol{\beta}},\boldsymbol{\mathcal{X}}))$ is bounded by Lemma C.1. $\forall i$, $L'_{N,i}(\bar{\boldsymbol{\mathcal{W}}},\bar{\boldsymbol{\beta}},\boldsymbol{\mathcal{X}})$ has subexponential tails. Thus, by Bernstein inequality,

$$
\begin{aligned}
&P(\|L'_{N,\mathcal{A}_k}(\bar{\boldsymbol{\mathcal{W}}},\bar{\boldsymbol{\beta}},\boldsymbol{\mathcal{X}})\|_2 \leq c_{0,\eta}\sqrt{q_k\frac{\log p}{N}}) \\
&\geq P(\sqrt{q_k}\max_{i\in\mathcal{A}_k}|L'_{N,i}(\bar{\boldsymbol{\mathcal{W}}},\bar{\boldsymbol{\beta}},\boldsymbol{\mathcal{X}})| \leq c_{0,\eta}\sqrt{q_k\frac{\log p}{N}}) \geq 1 - O(\exp(-\eta\log p)).
\end{aligned}
$$

(iii) By Cauchy-Schwartz,

$$
\begin{aligned}
&|u^T L''_{N,\mathcal{A}_k\mathcal{A}_k}(\bar{\boldsymbol{\mathcal{W}}},\bar{\boldsymbol{\beta}},\boldsymbol{\mathcal{X}})u - u^T\bar{L}''_{\mathcal{A}_k\mathcal{A}_k}(\bar{\boldsymbol{\beta}})u| \\
&\leq \|u\|_2\|u^T L''_{N,\mathcal{A}_k\mathcal{A}_k}(\bar{\boldsymbol{\mathcal{W}}},\bar{\boldsymbol{\beta}},\boldsymbol{\mathcal{X}}) - u^T\bar{L}''_{\mathcal{A}_k\mathcal{A}_k}(\bar{\boldsymbol{\beta}})\|_2 \\
&\leq \|u\|_2\sqrt{q_k}\max_i|u^T L''_{N,\mathcal{A}_k,i}(\bar{\boldsymbol{\mathcal{W}}},\bar{\boldsymbol{\beta}},\boldsymbol{\mathcal{X}}) - u^T\bar{L}''_{\mathcal{A}_k,i}(\bar{\boldsymbol{\beta}})| \\
&= \|u\|_2\sqrt{q_k}|u^T L''_{N,\mathcal{A}_k,i_{\max}}(\bar{\boldsymbol{\mathcal{W}}},\bar{\boldsymbol{\beta}},\boldsymbol{\mathcal{X}}) - u^T\bar{L}''_{\mathcal{A}_k,i_{\max}}(\bar{\boldsymbol{\beta}})| \\
&= \|u\|_2\sqrt{q_k}|\sum_{j=1}^{q_k}(u_j L''_{N,j,i_{\max}}(\bar{\boldsymbol{\mathcal{W}}},\bar{\boldsymbol{\beta}},\boldsymbol{\mathcal{X}}) - u_j\bar{L}''_{j,i_{\max}}(\bar{\boldsymbol{\beta}}))| \\
&\leq \|u\|_2 q_k|u_{j_{\max}}||L''_{N,j_{\max},i_{\max}}(\bar{\boldsymbol{\mathcal{W}}},\bar{\boldsymbol{\beta}},\boldsymbol{\mathcal{X}}) - \bar{L}''_{j_{\max},i_{\max}}(\bar{\boldsymbol{\beta}}))| \\
&\leq \|u\|_2^2 q_k|L''_{N,j_{\max},i_{\max}}(\bar{\boldsymbol{\mathcal{W}}},\bar{\boldsymbol{\beta}},\boldsymbol{\mathcal{X}}) - \bar{L}''_{j_{\max},i_{\max}}(\bar{\boldsymbol{\beta}}))|.
\end{aligned}
$$

Then by Bernstein inequality,

$$
\begin{aligned}
&P(|u^T L''_{N,\mathcal{A}_k\mathcal{A}_k}(\bar{\boldsymbol{\mathcal{W}}},\bar{\boldsymbol{\beta}},\boldsymbol{\mathcal{X}})u - u^T\bar{L}''_{\mathcal{A}_k\mathcal{A}_k}(\bar{\boldsymbol{\beta}})u| \leq c_{2,\eta}\|u\|_2^2 q_k\sqrt{\frac{\log p}{N}}) \\
&\geq P(\|u\|_2^2 q_k|L''_{N,j_{\max},i_{\max}}(\bar{\boldsymbol{\mathcal{W}}},\bar{\boldsymbol{\beta}},\boldsymbol{\mathcal{X}}) - \bar{L}''_{j_{\max},i_{\max}}(\bar{\boldsymbol{\beta}}))| \leq c_{2,\eta}\|u\|_2^2 q_k\sqrt{\frac{\log p}{N}}) \\
&\geq 1 - O(\exp(-\eta\log p)).
\end{aligned}
$$

(ii) and (iv) can be proved using similar arguments. □

Lemma C.3. and C.4. are used later to prove Theorem 1.

Lemma B.3. Assuming conditions of Theorem 1. Then there exists a constant $C_1(\bar{\boldsymbol{\beta}}) > 0$ such that for any $\eta > 0$, there exists a global minimizer of the restricted problem (8) within the disc:

$$
\{\boldsymbol{\beta} : \|\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}\|_2 \leq C_1(\bar{\boldsymbol{\beta}})\sqrt{K}\max_k\sqrt{q_k}\lambda_{N,k}\}
$$

with probability at least $1 - O(\exp(-\eta\log p))$ for sufficiently large N.

proof of Lemma B.3. Let $\alpha_N = \max_k \sqrt{q_k}\lambda_{N,k}$. Further for $1 \le k \le K$ let $C_k > 0$ and $u^k \in \mathbb{R}^{m_k(m_k-1)/2}$ such that $u^k_{\mathcal{A}_k^c} = 0$, $\|u^k\|_2 = C_k$, and $u = (u_1, \dots, u_K)$ with $\sqrt{K}\min_k C_k \le \|u\|_2 \le \sqrt{K}\max_k C_k$.

Then by Cauchy-Schwartz and triangle inequality, we have

$$\|\bar{\boldsymbol{\beta}}^k + \alpha_N u^k - \alpha_N u^k\|_1 \le \|\bar{\boldsymbol{\beta}}^k + \alpha_N u^k\|_1 + \alpha_N \|u^k\|_1,$$

and

$$\|\bar{\boldsymbol{\beta}}^k\|_1 - \|\bar{\boldsymbol{\beta}}^k + \alpha_N u^k\|_1 \le \alpha_N \|u^k\|_1 \le \alpha_N \sqrt{q_k}\|u^k\|_2 = C_k \alpha_N \sqrt{q_k}.$$

Thus,

$$\begin{aligned}
&Q_N(\bar{\boldsymbol{\beta}} + \alpha_N u, \boldsymbol{\mathcal{X}}, \{\lambda_{N,k}\}_{k=1}^K) - Q_N(\bar{\boldsymbol{\beta}}, \boldsymbol{\mathcal{X}}, \{\lambda_{N,k}\}_{k=1}^K) \\
&= L_N(\bar{\boldsymbol{\beta}} + \alpha_N u, \boldsymbol{\mathcal{X}}) - L_N(\bar{\boldsymbol{\beta}}, \boldsymbol{\mathcal{X}}) - \sum_{k=1}^K \lambda_{N,k}\big(\|\bar{\boldsymbol{\beta}}^k\|_1 - \|\bar{\boldsymbol{\beta}}^k + \alpha_N u^k\|_1\big) \\
&\ge L_N(\bar{\boldsymbol{\beta}} + \alpha_N u, \boldsymbol{\mathcal{X}}) - L_N(\bar{\boldsymbol{\beta}}, \boldsymbol{\mathcal{X}}) - \sum_{k=1}^K \lambda_{N,k} C_k \alpha_N \sqrt{q_k} \\
&\ge L_N(\bar{\boldsymbol{\beta}} + \alpha_N u, \boldsymbol{\mathcal{X}}) - L_N(\bar{\boldsymbol{\beta}}, \boldsymbol{\mathcal{X}}) - \alpha_N K \max_k C_k \sqrt{q_k}\lambda_{N,k} \\
&\ge L_N(\bar{\boldsymbol{\beta}} + \alpha_N u, \boldsymbol{\mathcal{X}}) - L_N(\bar{\boldsymbol{\beta}}, \boldsymbol{\mathcal{X}}) - K\alpha_N^2 \max_k C_k.
\end{aligned}$$

Next,

$$\begin{aligned}
&L_N(\bar{\boldsymbol{\beta}} + \alpha_N u, \boldsymbol{\mathcal{X}}) - L_N(\bar{\boldsymbol{\beta}}, \boldsymbol{\mathcal{X}}) = \alpha_N u_{\mathcal{A}}^T L'_{N,\mathcal{A}}(\bar{\boldsymbol{\beta}}, \boldsymbol{\mathcal{X}}) + \frac{1}{2}\alpha_N^2 u_{\mathcal{A}}^T L''_{N,\mathcal{A}\mathcal{A}}(\bar{\boldsymbol{\beta}}, \boldsymbol{\mathcal{X}}) u_{\mathcal{A}} \\
&= \alpha_N \sum_{k=1}^K (u^k_{\mathcal{A}_k})^T L'_{N,\mathcal{A}_k}(\bar{\boldsymbol{\beta}}, \boldsymbol{\mathcal{X}}) + \frac{1}{2}\alpha_N^2 \sum_{k=1}^K (u^k_{\mathcal{A}_k})^T L''_{N,\mathcal{A}_k\mathcal{A}_k}(\bar{\boldsymbol{\beta}}, \boldsymbol{\mathcal{X}}) u^k_{\mathcal{A}_k} \\
&= \alpha_N \sum_{k=1}^K (u^k_{\mathcal{A}_k})^T L'_{N,\mathcal{A}_k}(\bar{\boldsymbol{\beta}}, \boldsymbol{\mathcal{X}}) + \frac{1}{2}\alpha_N^2 \sum_{k=1}^K (u^k_{\mathcal{A}_k})^T (L''_{N,\mathcal{A}_k\mathcal{A}_k}(\bar{\boldsymbol{\beta}}, \boldsymbol{\mathcal{X}}) - \bar{L}''_{N,\mathcal{A}_k\mathcal{A}_k}(\bar{\boldsymbol{\beta}}, \boldsymbol{\mathcal{X}})) u^k_{\mathcal{A}_k} \\
&+ \frac{1}{2}\alpha_N^2 \sum_{k=1}^K (u^k_{\mathcal{A}_k})^T \bar{L}''_{N,\mathcal{A}_k\mathcal{A}_k}(\bar{\boldsymbol{\beta}}, \boldsymbol{\mathcal{X}}) u^k_{\mathcal{A}_k} \\
&\ge \frac{1}{2}\alpha_N^2 \sum_{k=1}^K (u^k_{\mathcal{A}_k})^T \bar{L}''_{N,\mathcal{A}_k\mathcal{A}_k}(\bar{\boldsymbol{\beta}}, \boldsymbol{\mathcal{X}}) u^k_{\mathcal{A}_k} - \alpha_N K(\max_k c_{1,\eta}\|u^k_{\mathcal{A}_k}\|_2 \sqrt{q_k \frac{\log p}{N}}) \\
&- \frac{1}{2}\alpha_N^2 K(\max_k c_{2,\eta}\|u^k_{\mathcal{A}_k}\|_2^2 q_k \sqrt{\frac{\log p}{N}}).
\end{aligned}$$

Here the first equality is due to the second order expansion of the loss function and the inequality is due to Lemma B.2. For sufficiently large N, by assumption that $\lambda_{N,k}\sqrt{N/\log p} \to \infty$ if $m_k \to \infty$ and $\sqrt{\log p/N} = o(1)$, the second term in the last line above is $o(\alpha_N \sqrt{q_k}\lambda_{N,k}) = o(\alpha_N^2)$; the last term is $o(\alpha_N^2)$. Therefore, for sufficiently large N

$$\begin{aligned}
Q_N(\bar{\boldsymbol{\beta}} + \alpha_N u, \boldsymbol{\mathcal{X}}, \{\lambda_{N,k}\}_{k=1}^K) - Q_N(\bar{\boldsymbol{\beta}}, \boldsymbol{\mathcal{X}}, \{\lambda_{N,k}\}_{k=1}^K) &\ge \frac{1}{2}\alpha_N^2 \sum_{k=1}^K (u^k_{\mathcal{A}_k})^T \bar{L}''_{N,\mathcal{A}_k\mathcal{A}_k}(\bar{\boldsymbol{\beta}}, \boldsymbol{\mathcal{X}}) u^k_{\mathcal{A}_k} \\
&- K\alpha_N^2 \max_k C_k \\
&\ge \frac{1}{2}\alpha_N^2 K \min_k \big((u^k_{\mathcal{A}_k})^T \bar{L}''_{N,\mathcal{A}_k\mathcal{A}_k}(\bar{\boldsymbol{\beta}}, \boldsymbol{\mathcal{X}}) u^k_{\mathcal{A}_k}\big) \\
&- K\alpha_N^2 \max_k C_k,
\end{aligned}$$

with probability at least $1 - O(N^{-\eta})$. By Lemma B.1., for each k, $(u^k_{\mathcal{A}_k})^T \bar{L}''_{N,\mathcal{A}_k\mathcal{A}_k}(\bar{\boldsymbol{\beta}}, \boldsymbol{\mathcal{X}}) u^k_{\mathcal{A}_k} \ge \Lambda^L_{\min}\|u^k_{\mathcal{A}_k}\|_2^2 = \Lambda^L_{\min}(C_k)^2$. So, if we choose $\min_k C_k$ and $\max_k C_k$ such that the upper bound is minimized, then for N sufficiently large, the following holds

$$\inf_{u:u_{(\mathcal{A}_k)^c}=0,\|u^k\|_2=C_k,k=1,\dots,K} Q_N(\bar{\boldsymbol{\beta}}+\alpha_N u,\boldsymbol{\mathcal{X}},\{\lambda_{N,k}\}_{k=1}^K) > Q_N(\bar{\boldsymbol{\beta}},\boldsymbol{\mathcal{X}},\{\lambda_{N,k}\}_{k=1}^K),$$

with probability at least $1-O(\exp(-\eta\log p))$, which means any solution to the problem defined in (8) is within the disc $\{\boldsymbol{\beta}: \|\boldsymbol{\beta}-\bar{\boldsymbol{\beta}}\|_2 \le \alpha_N\|u\|_2 \le \alpha_N\sqrt{K}\max_k C_k\}$ with probability at least $1-O(\exp(-\eta\log p))$.

□

Lemma B.4. Assuming conditions of Theorems 1. Then there exists a constant $C_2(\bar{\boldsymbol{\beta}})>0$, such that for any $\eta>0$, for sufficiently large N, the following event holds with probability at least $1-O(\exp(-\eta\log p))$: if for any $\boldsymbol{\beta}\in S=\{\boldsymbol{\beta}: \|\boldsymbol{\beta}-\bar{\boldsymbol{\beta}}\|_2 \ge C_2(\bar{\boldsymbol{\beta}})\sqrt{K}\max_k\sqrt{q_k}\lambda_{N,k}, \boldsymbol{\beta}_{\mathcal{A}_N^c}=0\}$, then $\|L'_{N,\mathcal{A}_N}(\bar{\boldsymbol{\mathcal{W}}},\bar{\boldsymbol{\beta}},\boldsymbol{\mathcal{X}})\|_2 > \sqrt{K}\max_k\sqrt{q_k}\lambda_{N,k}$.

proof of Lemma B.4. Let $\alpha_N=\max_k\sqrt{q_k}\lambda_{N,k}$. For $\boldsymbol{\beta}\in S$, we have $\boldsymbol{\beta}=\bar{\boldsymbol{\beta}}+\alpha_N u$, with $u_{(\mathcal{A})^c}$ and $\|u\|_2\ge C_2(\bar{\boldsymbol{\beta}})$. Note that by Taylor expansion of $L'_{N,\mathcal{A}}(\bar{\boldsymbol{\mathcal{W}}},\boldsymbol{\beta},\boldsymbol{\mathcal{X}})$ at $\bar{\boldsymbol{\beta}}$

$$\begin{aligned}
L'_{N,\mathcal{A}}(\bar{\boldsymbol{\mathcal{W}}},\boldsymbol{\beta},\boldsymbol{\mathcal{X}}) &= L'_{N,\mathcal{A}}(\bar{\boldsymbol{\mathcal{W}}},\boldsymbol{\beta},\boldsymbol{\mathcal{X}}) + \alpha_N L''_{N,\mathcal{A}\mathcal{A}}(\bar{\boldsymbol{\mathcal{W}}},\boldsymbol{\beta},\boldsymbol{\mathcal{X}})u_{\mathcal{A}}\\
&= L'_{N,\mathcal{A}}(\bar{\boldsymbol{\mathcal{W}}},\boldsymbol{\beta},\boldsymbol{\mathcal{X}}) + \alpha_N\big(L''_{N,\mathcal{A}\mathcal{A}}(\bar{\boldsymbol{\mathcal{W}}},\boldsymbol{\beta},\boldsymbol{\mathcal{X}}) - \bar{L}''_{N,\mathcal{A}\mathcal{A}}(\bar{\boldsymbol{\beta}})\big)u_{\mathcal{A}}\\
&\quad + \alpha_N\bar{L}''_{N,\mathcal{A}\mathcal{A}}(\bar{\boldsymbol{\beta}})u_{\mathcal{A}}.
\end{aligned}$$

By triangle inequality and similar proof strategies as in Lemma B.3., for sufficiently large N

$$\begin{aligned}
\|L'_{N,\mathcal{A}}(\bar{\boldsymbol{\mathcal{W}}},\boldsymbol{\beta},\boldsymbol{\mathcal{X}})\|_2 &\ge \|L'_{N,\mathcal{A}}(\bar{\boldsymbol{\mathcal{W}}},\boldsymbol{\beta},\boldsymbol{\mathcal{X}})\|_2 + \alpha_N\|L''_{N,\mathcal{A}\mathcal{A}}(\bar{\boldsymbol{\mathcal{W}}},\boldsymbol{\beta},\boldsymbol{\mathcal{X}})u_{\mathcal{A}} - \bar{L}''_{N,\mathcal{A}\mathcal{A}}(\bar{\boldsymbol{\beta}})u_{\mathcal{A}}\|_2\\
&\quad + \alpha_N\|\bar{L}''_{N,\mathcal{A}\mathcal{A}}(\bar{\boldsymbol{\beta}})u_{\mathcal{A}}\|_2\\
&\ge \alpha_N\|\bar{L}''_{N,\mathcal{A}\mathcal{A}}(\bar{\boldsymbol{\beta}})u_{\mathcal{A}}\|_2 + o(\alpha_N)
\end{aligned}$$

with probability at least $1-O(\exp(-\eta\log p))$. By Lemma B.1., $\|\bar{L}''_{N,\mathcal{A}\mathcal{A}}(\bar{\boldsymbol{\beta}})u_{\mathcal{A}}\|_2 \ge \Lambda^L_{\min}(\bar{\boldsymbol{\beta}})\|u_{\mathcal{A}}\|_2$. Therefore, taking $C_2(\bar{\boldsymbol{\beta}})$ to be $1/\Lambda^L_{\min}(\bar{\boldsymbol{\beta}})+\epsilon$ completes the proof. □

proof of Theorem 1. By the Karush-Kuhn-Tucker condition, for any solution $\hat{\boldsymbol{\beta}}$ of (8), it satisfies $\|L'_{N,\mathcal{A}_k}(\boldsymbol{\mathcal{W}},\hat{\boldsymbol{\beta}},\boldsymbol{\mathcal{X}})\|_\infty \le \lambda_{N,k}$. Thus,

$$\begin{aligned}
\|L'_{N,\mathcal{A}_N}(\boldsymbol{\mathcal{W}},\hat{\boldsymbol{\beta}},\boldsymbol{\mathcal{X}})\|_2 &\le \sqrt{K}\max_k\|L'_{N,\mathcal{A}_k}(\boldsymbol{\mathcal{W}},\hat{\boldsymbol{\beta}},\boldsymbol{\mathcal{X}})\|_2\\
&\le \sqrt{K}\max_k\sqrt{q_k}\|L'_{N,\mathcal{A}_k}(\boldsymbol{\mathcal{W}},\hat{\boldsymbol{\beta}},\boldsymbol{\mathcal{X}})\|_\infty\\
&\le \sqrt{K}\max_k\sqrt{q_k}\lambda_{N,k}.
\end{aligned}$$

Then by Lemmas B.4., for any $\eta>0$, for N sufficiently large, all solutions of (8) are inside the disc $\{\boldsymbol{\beta}: \|\boldsymbol{\beta}-\bar{\boldsymbol{\beta}}\|_2 \le C_2(\bar{\boldsymbol{\beta}})\max_k\sqrt{q_k}\lambda_{N,k}, \boldsymbol{\beta}_{\mathcal{A}_N^c}=0\}$ with probability at least $1-O(\exp(-\eta\log p))$. If we further assume that $\min_{(i,j)\in\mathcal{A}_k}|\bar{\boldsymbol{\beta}}_{i,j}| \ge 2C(\bar{\boldsymbol{\beta}})\max_k\sqrt{q_k}\lambda_{N,k}$ for each k, then

$$\begin{aligned}
&1-O(\exp(-\eta\log p))\\
&\qquad\le P_{\bar{\boldsymbol{\mathcal{W}}},\bar{\boldsymbol{\beta}}}(\|\hat{\boldsymbol{\beta}}^{\mathcal{A}}-\bar{\boldsymbol{\beta}}^{\mathcal{A}}\|_2 \le C_2(\bar{\boldsymbol{\beta}})\max_k\sqrt{q_k}\lambda_{N,k}, \min_{(i,j)\in\mathcal{A}_k}|\bar{\boldsymbol{\beta}}_{i,j}| \ge 2C(\bar{\boldsymbol{\beta}})\max_k\sqrt{q_k}\lambda_{N,k}, \forall k)\\
&\qquad\le P_{\bar{\boldsymbol{\mathcal{W}}},\bar{\boldsymbol{\beta}}}(\mathrm{sign}(\hat{\boldsymbol{\beta}}^{\mathcal{A}_k}_{i_kj_k}) = \mathrm{sign}(\bar{\boldsymbol{\beta}}^{\mathcal{A}_k}_{i_kj_k}), \forall(i_k,j_k)\in\mathcal{A}_k, \forall k).
\end{aligned}$$

□

proof of Theorem 2. Let $\mathcal{E}_{N,k}=\{\mathrm{sign}(\hat{\boldsymbol{\beta}}^{\mathcal{A}_k}_{i_kj_k})=\mathrm{sign}(\bar{\boldsymbol{\beta}}^{\mathcal{A}_k}_{i_kj_k})\}$. Then by Theorem 1, $P_{\bar{\boldsymbol{\mathcal{W}}},\bar{\boldsymbol{\beta}}}(\mathcal{E}_{N,k}) \ge 1-O(\exp(-\eta\log p))$ for large N. On $\mathcal{E}_{N,k}$, By the KKT condition and the expansion of $L'_{N,\mathcal{A}_k}(\bar{\boldsymbol{\mathcal{W}}},\hat{\boldsymbol{\beta}}^{\mathcal{A}_k},\boldsymbol{\mathcal{X}})$ at $\bar{\boldsymbol{\beta}}^{\mathcal{A}_k}$

$$\begin{aligned}
&-\lambda_{N,k}\mathrm{sign}(\bar{\boldsymbol{\beta}}^{\mathcal{A}_k})\\
&\quad= L'_{N,\mathcal{A}_k}(\bar{\boldsymbol{\mathcal{W}}},\hat{\boldsymbol{\beta}}^{\mathcal{A}_k},\boldsymbol{\mathcal{X}})\\
&\quad= L'_{N,\mathcal{A}_k}(\bar{\boldsymbol{\mathcal{W}}},\bar{\boldsymbol{\beta}}^{\mathcal{A}_k},\boldsymbol{\mathcal{X}}) + L''_{N,\mathcal{A}_k\mathcal{A}_k}(\bar{\boldsymbol{\mathcal{W}}},\bar{\boldsymbol{\beta}},\boldsymbol{\mathcal{X}})v_{N,k}\\
&\quad= \bar{L}''_{\mathcal{A}_k\mathcal{A}_k}v_{N,k} + L'_{N,\mathcal{A}_k}(\bar{\boldsymbol{\mathcal{W}}},\bar{\boldsymbol{\beta}}^{\mathcal{A}_k},\boldsymbol{\mathcal{X}}) + (L''_{N,\mathcal{A}_k\mathcal{A}_k}(\bar{\boldsymbol{\mathcal{W}}},\bar{\boldsymbol{\beta}},\boldsymbol{\mathcal{X}}) - \bar{L}''_{\mathcal{A}_k\mathcal{A}_k})v_{N,k},
\end{aligned}$$

where $v_{N,k} = \hat{\boldsymbol{\beta}}^{\mathcal{A}_k} - \bar{\boldsymbol{\beta}}^{\mathcal{A}_k}$. By rearranging the terms

$$\begin{aligned} &v_{N,k} = \\ &-\lambda_{N,k}[\bar{L}''_{\mathcal{A}_k\mathcal{A}_k}]^{-1}\text{sign}(\bar{\boldsymbol{\beta}}^{\mathcal{A}_k}) - [\bar{L}''_{\mathcal{A}_k\mathcal{A}_k}]^{-1}[L'_{N,\mathcal{A}_k}(\bar{\boldsymbol{\mathcal{W}}}, \bar{\boldsymbol{\beta}}^{\mathcal{A}_k}, \boldsymbol{\mathcal{X}}) + D_{N,\mathcal{A}_k\mathcal{A}_k}(\bar{\boldsymbol{\mathcal{W}}}, \bar{\boldsymbol{\beta}}^{\mathcal{A}_k})v_{N,k}], \end{aligned} \tag{11}$$

where $D_{N,\mathcal{A}_k\mathcal{A}_k} = L''_{N,\mathcal{A}_k\mathcal{A}_k}(\bar{\boldsymbol{\mathcal{W}}}, \bar{\boldsymbol{\beta}}, \boldsymbol{\mathcal{X}}) - \bar{L}''_{\mathcal{A}_k\mathcal{A}_k}$. Next, for fixed $(i,j) \in \mathcal{A}_k^c$, by expanding $L'_{N,\mathcal{A}_k}(\bar{\boldsymbol{\mathcal{W}}}, \hat{\boldsymbol{\beta}}^{\mathcal{A}_k}, \boldsymbol{\mathcal{X}})$ at $\bar{\boldsymbol{\beta}}^{\mathcal{A}_k}$

$$L'_{N,ij}(\bar{\boldsymbol{\mathcal{W}}}, \hat{\boldsymbol{\beta}}^{\mathcal{A}_k}, \boldsymbol{\mathcal{X}}) = L'_{N,ij}(\bar{\boldsymbol{\mathcal{W}}}, \bar{\boldsymbol{\beta}}^{\mathcal{A}_k}, \boldsymbol{\mathcal{X}}) + L''_{N,ij,\mathcal{A}_k}(\bar{\boldsymbol{\mathcal{W}}}, \bar{\boldsymbol{\beta}}^{\mathcal{A}_k}, \boldsymbol{\mathcal{X}})v_{N,k}. \tag{12}$$

Then combining (11) and (12) we get

$$\begin{aligned} &L'_{N,ij}(\bar{\boldsymbol{\mathcal{W}}}, \hat{\boldsymbol{\beta}}^{\mathcal{A}_k}, \boldsymbol{\mathcal{X}}) \\ &= -\lambda_{N,k}\bar{L}''_{ij,\mathcal{A}_k}(\bar{\boldsymbol{\beta}}^{\mathcal{A}_k})[\bar{L}''_{\mathcal{A}_k\mathcal{A}_k}]^{-1}\text{sign}(\bar{\boldsymbol{\beta}}^{\mathcal{A}_k}) - \bar{L}''_{ij,\mathcal{A}_k}(\bar{\boldsymbol{\beta}}^{\mathcal{A}_k})[\bar{L}''_{\mathcal{A}_k\mathcal{A}_k}]^{-1}L'_{N,\mathcal{A}_k}(\bar{\boldsymbol{\mathcal{W}}}, \bar{\boldsymbol{\beta}}^{\mathcal{A}_k}, \boldsymbol{\mathcal{X}}) \\ &+ [D_{N,ij,\mathcal{A}_k}(\bar{\boldsymbol{\mathcal{W}}}, \bar{\boldsymbol{\beta}}^{\mathcal{A}_k}) - \bar{L}''_{ij,\mathcal{A}_k}(\bar{\boldsymbol{\beta}}^{\mathcal{A}_k})[\bar{L}''_{\mathcal{A}_k\mathcal{A}_k}]^{-1}D_{N,\mathcal{A}_k\mathcal{A}_k}(\bar{\boldsymbol{\mathcal{W}}}, \bar{\boldsymbol{\beta}}^{\mathcal{A}_k})]v_{N,k} \\ &+ L'_{N,ij}(\bar{\boldsymbol{\mathcal{W}}}, \bar{\boldsymbol{\beta}}^{\mathcal{A}_k}, \boldsymbol{\mathcal{X}}). \end{aligned} \tag{13}$$

By the incoherence condition outlined in condition (A3), for any $(i,j) \in \mathcal{A}_k$,

$$|\bar{L}''_{ij,\mathcal{A}_k}(\bar{\boldsymbol{\mathcal{W}}}, \bar{\boldsymbol{\beta}})[\bar{L}''_{\mathcal{A}_k,\mathcal{A}_k}(\bar{\boldsymbol{\mathcal{W}}}, \bar{\boldsymbol{\beta}})]^{-1}\text{sign}(\bar{\boldsymbol{\beta}}_{\mathcal{A}_k})| \leq \delta < 1.$$

Thus, following straightforwardly (with the modification that we are considering each $\mathcal{A}_k$ instead of $\mathcal{A}$) from the proofs of Theorem 2 of Peng et al. (2009), the remaining terms in (13) can be shown to be all $o(\lambda_{N,k})$, and the event $\max_{(i,j)\in\mathcal{A}_k^c} |L'_{N,ij}(\bar{\boldsymbol{\mathcal{W}}}, \hat{\boldsymbol{\beta}}^{\mathcal{A}_k}, \boldsymbol{\mathcal{X}})| < \lambda_{N,k}$ with probability at least $1 - O(\exp(-\eta \log p))$ for sufficiently large N. Thus, it has been proved that for sufficiently large N, no wrong edge will be included for each true edge set $\mathcal{A}_k$ and hence, no wrong edge will be included in $\mathcal{A} = \cup_k \mathcal{A}_k$. □

proof of Theorem 3. By Theorem 1 and Theorem 2, with probability tending to 1, any solution of the restricted problem is also a solution of the original problem. On the other hand, by Theorem 2 and the KKT condition, with probability tending to 1, any solution of the original problem is also a solution of the restricted problem. Therefore, Theorem 3 follows. □

C Simulated Precision Matrix

1. **AR1(ρ)**: The covariance matrix of the form $\mathbf{A} = (\rho^{|i-j|})_{ij}$ for $\rho \in (0,1)$.
2. **Star-Block (SB):** A block-diagonal covariance matrix, where each block's precision matrix corresponds to a star-structured graph with $(\boldsymbol{\Psi}_k)_{ij} = 1$. Then, for $\rho \in (0,1)$, we have that $\mathbf{A}_{ij} = \rho$ if $(i,j) \in E$ and $\mathbf{A}_{ij} = \rho^2$ for $(i,j) \notin E$, where E is the corresponding edge set.
3. **Erdos-Renyi random graph (ER):** The precision matrix is initialized at $\mathbf{A} = 0.25\mathbf{I}$, and d edges are randomly selected. For the selected edge (i,j), we randomly choose $\psi \in [0.6, 0.8]$ and update $\mathbf{A}_{ij} = \mathbf{A}_{ji} \to \mathbf{A}_{ij} - \psi$ and $\mathbf{A}_{ii} \to \mathbf{A}_{ii} + \psi$, $\mathbf{A}_{jj} \to \mathbf{A}_{jj} + \psi$.