Supplementary Material for

“Gaussian Mixture Models with Concave Penalized Fusion”

Yiwei Fan[1] and Guosheng Yin[2,3]

[1] *School of Mathematics and Statistics, Beijing Institute of Technology*

[2] *Department of Statistics and Actuarial Science, The University of Hong Kong*

[3] *Department of Mathematics, Imperial College London*

1. The ADMM algorithm for Gaussian mixture models

Algorithm S1 : The ADMM algorithm for Gaussian mixture models

```
Set initial values as $\boldsymbol{\beta}^{(0)}$, $\boldsymbol{\theta}^{(0)}_{[\cdot 1]}$, and $\boldsymbol{\theta}^{(0)}_{[\cdot 2]}$. Let $\Delta^{(0)}_{ijm} = \theta^{(0)}_{im} - \theta^{(0)}_{jm}$ and $\nu^{(0)}_{ijm} = 0$ for
any $1 \leq i < j \leq n$ and $m = 1, 2$.
Compute and store $\boldsymbol{E}^\top \boldsymbol{E}$.
$t \leftarrow 0$.
while $\|\boldsymbol{R}_p(\boldsymbol{\Theta}^{(t)}, \boldsymbol{\Delta}^{(t)})\|_F > \kappa^{\text{pri}}$ or $\|\boldsymbol{R}^{(t)}_d\|_F > \kappa^{\text{dual}}$ do
    Update $\boldsymbol{\theta}^{(t+1)}_{[\cdot 1]}$ by (3.2).
    Update $\boldsymbol{\beta}^{(t+1)}$ by (3.3).
    Update $\boldsymbol{\theta}^{(t+1)}_{[\cdot 2]}$ by repeatedly applying (3.4) in a cyclical manner until the
    relative distance of parameters between two cycles is smaller than a tolerance
    (e.g., $10^{-3}$).
    Update $\boldsymbol{\Delta}^{(t+1)}$ by (3.7) for the hard penalty or (3.8) for the
    SCAD penalty.
    Update $\boldsymbol{\nu}^{(t+1)}$ by (3.9).
    $t \leftarrow t + 1$.
end while
Output the final estimates $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}^{(t)}$, $\widehat{\boldsymbol{\theta}}_{[\cdot 1]} = \boldsymbol{\theta}^{(t)}_{[\cdot 1]}$, and $\widehat{\boldsymbol{\theta}}_{[\cdot 2]} = \boldsymbol{\theta}^{(t)}_{[\cdot 2]}$.
```

2. Proofs of conclusions in Section 3

2.1 Derivations of (3.2) – (3.4)

For the Gaussian mixture model, the log-likelihood function can be written as

$$L(\boldsymbol{\beta}, \boldsymbol{\Theta}) = \sum_{i=1}^{n} \frac{\log \theta_{i2}}{2} - \frac{\theta_{i2}(y_i - \boldsymbol{\beta}^\top \boldsymbol{x}_i - \theta_{i1})^2}{2} + C,$$

where C is a generic constant. By the first order condition of optimality, setting the derivative $\partial H(\boldsymbol{\beta}, \boldsymbol{\Theta}, \boldsymbol{\Delta}, \boldsymbol{\nu})/\partial \boldsymbol{\beta} = -\partial L(\boldsymbol{\beta}, \boldsymbol{\Theta})/\partial \boldsymbol{\beta}$ to zero, we have

$$\boldsymbol{\beta} = (\boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{W} (\boldsymbol{y} - \boldsymbol{\theta}_{[\cdot 1]}), \tag{S2.1}$$

where $\boldsymbol{W}$ is the diagonal matrix of $\boldsymbol{\theta}_{[\cdot 2]}$. Taking the partial derivative of $L(\boldsymbol{\beta}, \boldsymbol{\Theta})$ with respect to $\boldsymbol{\theta}_{[\cdot 1]}$, we have

$$\frac{\partial L(\boldsymbol{\beta}, \boldsymbol{\Theta})}{\partial \boldsymbol{\theta}_{[\cdot 1]}} = \left(\theta_{12}(y_1 - \boldsymbol{\beta}^\top \boldsymbol{x}_1 - \theta_{11}), \ldots, \theta_{n2}(y_n - \boldsymbol{\beta}^\top \boldsymbol{x}_n - \theta_{n1})\right)^\top .$$

One can verify that

$$\begin{aligned}
\frac{\partial H(\boldsymbol{\beta}, \boldsymbol{\Theta}, \boldsymbol{\Delta}, \boldsymbol{\nu})}{\partial \boldsymbol{\theta}_{[\cdot 1]}} =& - \left(\theta_{12}(y_1 - \boldsymbol{\beta}^\top \boldsymbol{x}_1 - \theta_{11}), \ldots, \theta_{n2}(y_n - \boldsymbol{\beta}^\top \boldsymbol{x}_n - \theta_{n1})\right)^\top \\
& + \rho \boldsymbol{E}^\top \left\{ \boldsymbol{E} \boldsymbol{\theta}_{[\cdot 1]} - \left(\boldsymbol{\Delta}_{[\cdot 1]} - \rho^{-1} \boldsymbol{\nu}_{[\cdot 1]} \right) \right\} \\
=& (\rho \boldsymbol{E}^\top \boldsymbol{E} + \boldsymbol{W}) \boldsymbol{\theta}_{[\cdot 1]} - \boldsymbol{W}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) - \rho \boldsymbol{E}^\top \left(\boldsymbol{\Delta}_{[\cdot 1]} - \rho^{-1} \boldsymbol{\nu}_{[\cdot 1]} \right) \\
=& (\rho \boldsymbol{E}^\top \boldsymbol{E} + \boldsymbol{A}) \boldsymbol{\theta}_{[\cdot 1]} - \boldsymbol{A}\boldsymbol{y} - \rho \boldsymbol{E}^\top \left(\boldsymbol{\Delta}_{[\cdot 1]} - \rho^{-1} \boldsymbol{\nu}_{[\cdot 1]} \right),
\end{aligned}$$

where the last equality is derived based on (S2.1) with $\boldsymbol{A} = \boldsymbol{W}(\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{W})$. By setting $\partial H(\boldsymbol{\beta}, \boldsymbol{\Theta}, \boldsymbol{\Delta}, \boldsymbol{\nu})/\partial \boldsymbol{\theta}_{[\cdot 1]} = \boldsymbol{0}$, we have

$$\boldsymbol{\theta}_{[\cdot 1]} = \left(\rho \boldsymbol{E}^\top \boldsymbol{E} + \boldsymbol{A}\right)^{-1} \left\{\boldsymbol{A}\boldsymbol{y} + \rho \boldsymbol{E}^\top \left(\boldsymbol{\Delta}_{[\cdot 1]} - \rho^{-1}\boldsymbol{\nu}_{[\cdot 1]}\right)\right\}.$$

Then (3.2) and (3.3) follow immediately.

Similarly, we take the partial derivative of $L(\boldsymbol{\beta}, \boldsymbol{\Theta})$ with respect to $\boldsymbol{\theta}_{[\cdot 2]}$,

$$\frac{\partial L(\boldsymbol{\beta}, \boldsymbol{\Theta})}{\partial \boldsymbol{\theta}_{[\cdot 2]}} = \left(\frac{1}{2\theta_{12}} - (y_1 - \boldsymbol{\beta}^\top \boldsymbol{x}_1 - \theta_{11})^2/2, \dots, \frac{1}{2\theta_{n2}} - (y_n - \boldsymbol{\beta}^\top \boldsymbol{x}_n - \theta_{n1})^2/2\right)^\top.$$

Thus, one can compute that

$$\begin{aligned}\frac{\partial H(\boldsymbol{\beta}, \boldsymbol{\Theta}, \boldsymbol{\Delta}, \boldsymbol{\nu})}{\partial \theta_{i2}} &= -\frac{1}{2\theta_{i2}} + (y_i - \boldsymbol{\beta}^\top \boldsymbol{x}_i - \theta_{i1})^2/2 + \sum_{j>i} \nu_{ij2} - \sum_{j<i} \nu_{ji2} \\ &\quad + \rho\left\{\sum_{j>i}(\theta_{i2} - \theta_{j2} - \Delta_{ij2}) - \sum_{j<i}(\theta_{j2} - \theta_{i2} - \Delta_{ji2})\right\} \\ &= -\frac{1}{2\theta_{i2}} + (y_i - \boldsymbol{\beta}^\top \boldsymbol{x}_i - \theta_{i1})^2/2 + \sum_{j>i} \nu_{ij2} - \sum_{j<i} \nu_{ji2} \\ &\quad + \rho\left\{\sum_{j>i}\theta_{i2} - \sum_{j>i}(\theta_{j2} + \Delta_{ij2}) + \sum_{j<i}\theta_{i2} - \sum_{j<i}(\theta_{j2} - \Delta_{ji2})\right\} \\ &= \rho(n-1)\theta_{i2} + b_i - \frac{1}{2\theta_{i2}},\end{aligned}$$

with

$$b_i = (y_i - \boldsymbol{\beta}^\top \boldsymbol{x}_i - \theta_{i1})^2/2 + \sum_{j>i} \nu_{ij2} - \sum_{j<i} \nu_{ji2} - \rho\left\{\sum_{j>i}(\theta_{j2} + \Delta_{ij2}) + \sum_{j<i}(\theta_{j2} - \Delta_{ji2})\right\}.$$

The solution of $\partial H(\boldsymbol{\beta}, \boldsymbol{\Theta}, \boldsymbol{\Delta}, \boldsymbol{\nu})/\partial \theta_{i2} = 0$ is the root of a quadratic equation $\rho(n-1)\theta_{i2}^2 + b_i\theta_{i2} - 1/2 = 0$. By $b_i^2 + 2\rho(n-1) > b_i^2$, there exists only one positive root,

$$\theta_{i2} = \{2\rho(n-1)\}^{-1}\left(-b_i + \sqrt{b_i^2 + 2\rho(n-1)}\right).$$

2.2 Derivations of (3.7) and (3.8)

Minimizing $H(\boldsymbol{\beta}, \boldsymbol{\Theta}, \boldsymbol{\Delta}, \boldsymbol{\nu})$ with respect to Δ_{ijm} is equivalent to minimizing

$$2^{-1}(r_{ijm} - \Delta_{ijm})^2 + \rho^{-1}p(|\Delta_{ijm}|, \lambda_m, \gamma_m).$$

The hard penalty (2.2) can be written as

$$p(|\Delta_{ijm}|, \lambda_m, \gamma_m) = \begin{cases} -\Delta_{ijm}^2/2 + \lambda_m|\Delta_{ijm}|, & \text{if } |\Delta_{ijm}| < \lambda_m, \\ \lambda_m^2/2, & \text{if } |\Delta_{ijm}| \geq \lambda_m. \end{cases}$$

For $|\Delta_{ijm}| < \lambda_m$, by the first order condition of optimality, any optimal solution Δ_{ijm} must satisfy

$$-(r_{ijm} - \Delta_{ijm}) + \rho^{-1}(-\Delta_{ijm} + \lambda_m g_{ijm}) = 0,$$

where $g_{ijm} \in \mathbb{R}$ belongs to the subdifferentials of the L_1 norm at Δ_{ijm}. In particular, we have $g_{ijm} = \text{sign}(\Delta_{ijm})$ if $\Delta_{ijm} \neq 0$; otherwise $g_{ijm} \in [-1, 1]$. By the KKT condition, we have $\Delta_{ijm} = 0$ if and only if $|r_{ijm}| \leq \rho^{-1}\lambda_m$. When $\Delta_{ijm} \neq 0$, it holds that $\Delta_{ijm} = (r_{ijm} - \rho^{-1}\lambda_m g_{ijm})/(1-\rho^{-1})$. Hence, we have $\Delta_{ijm} = \mathcal{S}(r_{ijm}, \rho^{-1}\lambda_m)/(1-\rho^{-1})$ with $\mathcal{S}(u, c) = \text{sign}(u)(|u| - c)_+$ being the soft thresholding function. For $|\Delta_{ijm}| \geq$

λ_m, the minimum value is obtained at $\Delta_{ijm} = r_{ijm}$. Hence, we have

$$\Delta_{ijm} = \begin{cases} \mathcal{S}(r_{ijm}, \rho^{-1}\lambda_m)/(1-\rho^{-1}), & \text{if } |r_{ijm}| < \lambda_m, \\ r_{ijm}, & \text{if } |r_{ijm}| \geq \lambda_m. \end{cases}$$

The SCAD penalty function can also be written as

$$p(|\Delta_{ijm}|, \lambda_m, \gamma_m) = \begin{cases} \lambda_m|\Delta_{ijm}|, & \text{if } |\Delta_{ijm}| \leq \lambda_m, \\ -\dfrac{(\Delta_{ijm})^2 - 2\gamma_m\lambda_m|\Delta_{ijm}| + \lambda_m^2}{2(\gamma_m - 1)}, & \text{if } \lambda_m < |\Delta_{ijm}| \leq \gamma_m\lambda_m, \\ \dfrac{(\gamma_m+1)\lambda_m^2}{2}, & \text{if } |\Delta_{ijm}| > \gamma_m\lambda_m. \end{cases}$$

For $|\Delta_{ijm}| \leq \lambda_m$, which is the same as Lasso, by the first order condition of optimality, any optimal solution Δ_{ijm} must satisfy $-(r_{ijm} - \Delta_{ijm}) + \rho^{-1}\lambda_m g_{ijm} = 0$. Thus, one can verify that $\Delta_{ijm} = \mathcal{S}(r_{ijm}, \rho^{-1}\lambda_m)$. For $\lambda_m < |\Delta_{ijm}| \leq \gamma_m\lambda_m$, by the first order condition of optimality, any optimal solution Δ_{ijm} must satisfy

$$-(r_{ijm} - \Delta_{ijm}) - \rho^{-1}\frac{\Delta_{ijm} - \gamma_m\lambda_m g_{ijm}}{\gamma_m - 1} = 0.$$

By the KKT condition, we have $\Delta_{ijm} = 0$ if and only if $|r_{ijm}| \leq \gamma_m\rho^{-1}\lambda_m/(\gamma_m - 1)$. When $\Delta_{ijm} \neq 0$, it holds that $\Delta_{ijm} = \{r_{ijm} - \gamma_m\rho^{-1}\lambda_m g_{ijm}/(\gamma_m - 1)\}/\{1 - \rho^{-1}/(\gamma_m - 1)\}$. Hence, we have

$$\Delta_{ijm} = \frac{\mathcal{S}(r_{ijm}, \gamma_m\rho^{-1}\lambda_m/(\gamma_m - 1))}{1 - \rho^{-1}/(\gamma_m - 1)}.$$

Moreover, by the first order condition of optimality, one can verify that for $|\Delta_{ijm}| > \gamma_m\lambda_m$, the SCAD yields $\Delta_{ijm} = r_{ijm}$. Combining the three cases, we update Δ_{ijm} as follows,

$$\Delta_{ijm} = \begin{cases} \mathcal{S}(r_{ijm}, \rho^{-1}\lambda_m), & \text{if } |r_{ijm}| \leq \lambda_m(1+\rho^{-1}), \\ \dfrac{\mathcal{S}(r_{ijm}, \gamma_m\rho^{-1}\lambda_m/(\gamma_m-1))}{1-\rho^{-1}/(\gamma_m-1)}, & \text{if } (1+\rho^{-1})\lambda_m < |r_{ijm}| \leq \gamma_m\lambda_m, \\ r_{ijm}, & \text{if } |r_{ijm}| > \gamma_m\lambda_m. \end{cases}$$

2.3 Proof of Lemma 1

Note that the objective function $Q(\boldsymbol{\beta}, \boldsymbol{\Theta}, \boldsymbol{\Delta})$ is coercive over the feasible set, that is, $Q(\boldsymbol{\beta}, \boldsymbol{\Theta}, \boldsymbol{\Delta}) \to \infty$ if $\boldsymbol{E\Theta} - \boldsymbol{\Delta} = \mathbf{0}$ and $\|(\boldsymbol{\beta}^\top, \boldsymbol{\theta}_{[\cdot 1]}^\top, \boldsymbol{\theta}_{[\cdot 2]}^\top, \boldsymbol{\Delta}_{[\cdot 1]}^\top, \boldsymbol{\Delta}_{[\cdot 2]}^\top)\|_2 \to \infty$. Moreover, as $\mathrm{Im}(\boldsymbol{E}) \subseteq \mathrm{Im}(\boldsymbol{I})$ with $\mathrm{Im}(\cdot)$ being the image of a matrix, there exists $\boldsymbol{\Delta}'$ such that $\boldsymbol{E\Theta}^{(t)} - \boldsymbol{\Delta}' = \mathbf{0}$. Therefore, we have

$$Q(\boldsymbol{\beta}^{(t)}, \boldsymbol{\Theta}^{(t)}, \boldsymbol{\Delta}') \geq \min_{\boldsymbol{\beta},\boldsymbol{\Theta},\boldsymbol{\Delta}} \{Q(\boldsymbol{\beta}, \boldsymbol{\Theta}, \boldsymbol{\Delta}) : \boldsymbol{E\Theta} - \boldsymbol{\Delta} = \mathbf{0}\} > -\infty. \tag{S2.2}$$

By the first order condition of optimality, it holds that for $m = 1, 2$,

$$\begin{aligned} &\frac{\partial H(\boldsymbol{\beta}^{(t)}, \boldsymbol{\Theta}^{(t)}, \boldsymbol{\Delta}^{(t)}, \boldsymbol{\nu}^{(t-1)})}{\partial \boldsymbol{\Delta}_{[\cdot m]}} \\ =&\frac{\partial \sum_{i<j} p(|\Delta_{ijm}^{(t)}|, \lambda_m, \gamma_m)}{\partial \boldsymbol{\Delta}_{[\cdot m]}} - \boldsymbol{\nu}_{[\cdot m]}^{(t-1)} - \rho(\boldsymbol{E}\boldsymbol{\theta}_{[\cdot m]}^{(t)} - \boldsymbol{\Delta}_{[\cdot m]}^{(t)}) \\ =&\frac{\partial \sum_{i<j} p(|\Delta_{ijm}^{(t)}|, \lambda_m, \gamma_m)}{\partial \boldsymbol{\Delta}_{[\cdot m]}} - \boldsymbol{\nu}_{[\cdot m]}^{(t)} = \mathbf{0}, \end{aligned} \tag{S2.3}$$

where $\partial \sum_{i<j} p(|\Delta_{ijm}^{(t)}|, \lambda_m, \gamma_m)/\partial \boldsymbol{\Delta}_{[\cdot m]}$ belongs to the subdifferentials of the penalty

function and the last equality is derived from (3.9). Therefore, we have

$$\boldsymbol{\nu}_{[\cdot m]}^{(t)} = \frac{\partial \sum_{i<j} p(|\Delta_{ijm}^{(t)}|, \lambda_m, \gamma_m)}{\partial \boldsymbol{\Delta}_{[\cdot m]}}, \quad m = 1, 2.$$

Furthermore, it can be verified that for $m = 1, 2$,

$$\begin{aligned}
&\sum_{i<j} p(|\Delta_{ijm}^{(t)}|, \lambda_m, \gamma_m) + (\boldsymbol{\nu}_{[\cdot m]}^{(t)})^\top (\boldsymbol{E}\boldsymbol{\theta}_{[\cdot m]}^{(t)} - \boldsymbol{\Delta}_{[\cdot m]}^{(t)}) \\
=&\sum_{i<j} p(|\Delta_{ijm}^{(t)}|, \lambda_m, \gamma_m) + \left\{ \frac{\partial \sum_{i<j} p(|\Delta_{ijm}^{(t)}|, \lambda_m, \gamma_m)}{\partial \boldsymbol{\Delta}_{[\cdot m]}} \right\}^\top (\boldsymbol{\Delta}'_{[\cdot m]} - \boldsymbol{\Delta}_{[\cdot m]}^{(t)}) \\
\geq&\sum_{i<j} p(|\Delta'_{ijm}|, \lambda_m, \gamma_m) + I_m - \frac{C_p}{2} \|\boldsymbol{\Delta}'_{[\cdot m]} - \boldsymbol{\Delta}_{[\cdot m]}^{(t)}\|_2^2,
\end{aligned}$$

with the inequality derived from the weakly convexity of the penalty function and

$$I_m = \left\{ \frac{\partial \sum_{i<j} p(|\Delta'_{ijm}|, \lambda_m, \gamma_m)}{\partial \boldsymbol{\Delta}_{[\cdot m]}} - \frac{\partial \sum_{i<j} p(|\Delta_{ijm}^{(t)}|, \lambda_m, \gamma_m)}{\partial \boldsymbol{\Delta}_{[\cdot m]}} \right\}^\top (\boldsymbol{\Delta}_{[\cdot m]}^{(t)} - \boldsymbol{\Delta}'_{[\cdot m]}).$$

Then it holds that

$$
\begin{aligned}
&H(\boldsymbol{\beta}^{(t)}, \boldsymbol{\Theta}^{(t)}, \boldsymbol{\Delta}^{(t)}, \boldsymbol{\nu}^{(t)}) \\
=&Q(\boldsymbol{\beta}^{(t)}, \boldsymbol{\Theta}^{(t)}, \boldsymbol{\Delta}^{(t)}) + \sum_{m=1}^{2} (\boldsymbol{\nu}_{[\cdot m]}^{(t)})^\top (\boldsymbol{E}\boldsymbol{\theta}_{[\cdot m]}^{(t)} - \boldsymbol{\Delta}_{[\cdot m]}^{(t)}) + \frac{\rho}{2} \sum_{m=1}^{2} \|\boldsymbol{E}\boldsymbol{\theta}_{[\cdot m]}^{(t)} - \boldsymbol{\Delta}_{[\cdot m]}^{(t)}\|_2^2 \\
=&-L(\boldsymbol{\beta}^{(t)}, \boldsymbol{\Theta}^{(t)}) + \sum_{m=1}^{2} \sum_{i<j} p(|\Delta_{ijm}^{(t)}|, \lambda_m, \gamma_m) + \sum_{m=1}^{2} (\boldsymbol{\nu}_{[\cdot m]}^{(t)})^\top (\boldsymbol{E}\boldsymbol{\theta}_{[\cdot m]}^{(t)} - \boldsymbol{\Delta}_{[\cdot m]}^{(t)}) \\
&+ \frac{\rho}{2} \sum_{m=1}^{2} \|\boldsymbol{E}\boldsymbol{\theta}_{[\cdot m]}^{(t)} - \boldsymbol{\Delta}_{[\cdot m]}^{(t)}\|_2^2 \\
\geq&-L(\boldsymbol{\beta}^{(t)}, \boldsymbol{\Theta}^{(t)}) + \sum_{m=1}^{2} \sum_{i<j} p(|\Delta'_{ijm}|, \lambda_m, \gamma_m) + \sum_{m=1}^{2} \left(I_m + \frac{\rho - C_p}{2} \|\boldsymbol{\Delta}'_{[\cdot m]} - \boldsymbol{\Delta}_{[\cdot m]}^{(t)}\|_2^2 \right) \\
\geq&Q(\boldsymbol{\beta}^{(t)}, \boldsymbol{\Theta}^{(t)}, \boldsymbol{\Delta}') + \sum_{m=1}^{2} \left(-2n(n-1)C_s^2 \|\boldsymbol{\Delta}'_{[\cdot m]} - \boldsymbol{\Delta}_{[\cdot m]}^{(t)}\|_2 + \frac{\rho - C_p}{2} \|\boldsymbol{\Delta}'_{[\cdot m]} - \boldsymbol{\Delta}_{[\cdot m]}^{(t)}\|_2^2 \right).
\end{aligned}
$$

As $\rho - C_p > 0$, combing with (S2.2), we have $H(\boldsymbol{\beta}^{(t)}, \boldsymbol{\Theta}^{(t)}, \boldsymbol{\Delta}^{(t)}, \boldsymbol{\nu}^{(t)}) > -\infty$. This completes the proof. □

2.4 Proof of Lemma 2

To bound $H(\boldsymbol{\beta}^{(t)}, \boldsymbol{\Theta}^{(t)}, \boldsymbol{\Delta}^{(t)}, \boldsymbol{\nu}^{(t)}) - H(\boldsymbol{\beta}^{(t-1)}, \boldsymbol{\Theta}^{(t-1)}, \boldsymbol{\Delta}^{(t-1)}, \boldsymbol{\nu}^{(t-1)})$, we consider the following four terms,

(i) $H(\boldsymbol{\beta}^{(t)}, \boldsymbol{\Theta}^{(t)}, \boldsymbol{\Delta}^{(t)}, \boldsymbol{\nu}^{(t)}) - H(\boldsymbol{\beta}^{(t)}, \boldsymbol{\Theta}^{(t)}, \boldsymbol{\Delta}^{(t)}, \boldsymbol{\nu}^{(t-1)})$;

(ii) $H(\boldsymbol{\beta}^{(t)}, \boldsymbol{\Theta}^{(t)}, \boldsymbol{\Delta}^{(t)}, \boldsymbol{\nu}^{(t-1)}) - H(\boldsymbol{\beta}^{(t)}, \boldsymbol{\Theta}^{(t)}, \boldsymbol{\Delta}^{(t-1)}, \boldsymbol{\nu}^{(t-1)})$;

(iii) $H(\boldsymbol{\beta}^{(t)}, \boldsymbol{\Theta}^{(t)}, \boldsymbol{\Delta}^{(t-1)}, \boldsymbol{\nu}^{(t-1)}) - H(\boldsymbol{\beta}^{(t)}, (\boldsymbol{\theta}_{[\cdot 1]}^{(t)}, \boldsymbol{\theta}_{[\cdot 2]}^{(t-1)}), \boldsymbol{\Delta}^{(t-1)}, \boldsymbol{\nu}^{(t-1)})$;

(iv) $H(\boldsymbol{\beta}^{(t)}, (\boldsymbol{\theta}_{[\cdot 1]}^{(t)}, \boldsymbol{\theta}_{[\cdot 2]}^{(t-1)}), \boldsymbol{\Delta}^{(t-1)}, \boldsymbol{\nu}^{(t-1)}) - H(\boldsymbol{\beta}^{(t-1)}, \boldsymbol{\Theta}^{(t-1)}, \boldsymbol{\Delta}^{(t-1)}, \boldsymbol{\nu}^{(t-1)})$.

For (i), we have

$$
\begin{aligned}
&H(\boldsymbol{\beta}^{(t)}, \boldsymbol{\Theta}^{(t)}, \boldsymbol{\Delta}^{(t)}, \boldsymbol{\nu}^{(t)}) - H(\boldsymbol{\beta}^{(t)}, \boldsymbol{\Theta}^{(t)}, \boldsymbol{\Delta}^{(t)}, \boldsymbol{\nu}^{(t-1)}) \\
=&\sum_{m=1}^{2} (\boldsymbol{\nu}_{[\cdot m]}^{(t)} - \boldsymbol{\nu}_{[\cdot m]}^{(t-1)})^\top (\boldsymbol{E}\boldsymbol{\theta}_{[\cdot m]}^{(t)} - \boldsymbol{\Delta}_{[\cdot m]}^{(t)}) = \rho^{-1} \sum_{m=1}^{2} \|\boldsymbol{\nu}_{[\cdot m]}^{(t)} - \boldsymbol{\nu}_{[\cdot m]}^{(t-1)}\|_2^2 \leq 4\rho^{-1} n(n-1) C_s^2.
\end{aligned}
\tag{S2.4}
$$

By (S2.3), we have

$$
\boldsymbol{\nu}_{[\cdot m]}^{(t-1)} = \frac{\partial \sum_{i<j} p(|\Delta_{ijm}^{(t)}|, \lambda_m, \gamma_m)}{\partial \boldsymbol{\Delta}_{[\cdot m]}} - \rho(\boldsymbol{E}\boldsymbol{\theta}_{[\cdot m]}^{(t)} - \boldsymbol{\Delta}_{[\cdot m]}^{(t)}), \quad m = 1, 2. \tag{S2.5}
$$

Then it holds that

$$
\begin{aligned}
&H(\boldsymbol{\beta}^{(t)}, \boldsymbol{\Theta}^{(t)}, \boldsymbol{\Delta}^{(t)}, \boldsymbol{\nu}^{(t-1)}) - H(\boldsymbol{\beta}^{(t)}, \boldsymbol{\Theta}^{(t)}, \boldsymbol{\Delta}^{(t-1)}, \boldsymbol{\nu}^{(t-1)}) \\
=&\sum_{m=1}^{2} \sum_{i<j} p(|\Delta_{ijm}^{(t)}|, \lambda_m, \gamma_m) - \sum_{m=1}^{2} \sum_{i<j} p(|\Delta_{ijm}^{(t-1)}|, \lambda_m, \gamma_m) + \sum_{m=1}^{2} (\boldsymbol{\nu}_{[\cdot m]}^{(t-1)})^\top (\boldsymbol{\Delta}_{[\cdot m]}^{(t-1)} - \boldsymbol{\Delta}_{[\cdot m]}^{(t)}) \\
&+ \frac{\rho}{2} \sum_{m=1}^{2} \|\boldsymbol{E}\boldsymbol{\theta}_{[\cdot m]}^{(t)} - \boldsymbol{\Delta}_{[\cdot m]}^{(t)}\|_2^2 - \frac{\rho}{2} \sum_{m=1}^{2} \|\boldsymbol{E}\boldsymbol{\theta}_{[\cdot m]}^{(t)} - \boldsymbol{\Delta}_{[\cdot m]}^{(t-1)}\|_2^2 \\
=&\sum_{m=1}^{2} \sum_{i<j} p(|\Delta_{ijm}^{(t)}|, \lambda_m, \gamma_m) - \sum_{m=1}^{2} \sum_{i<j} p(|\Delta_{ijm}^{(t-1)}|, \lambda_m, \gamma_m) \\
&+ \sum_{m=1}^{2} \left(\frac{\partial \sum_{i<j} p(|\Delta_{ijm}^{(t)}|, \lambda_m, \gamma_m)}{\partial \boldsymbol{\Delta}_{[\cdot m]}} \right)^\top (\boldsymbol{\Delta}_{[\cdot m]}^{(t-1)} - \boldsymbol{\Delta}_{[\cdot m]}^{(t)}) \\
&+ \frac{\rho}{2} \sum_{m=1}^{2} \left\{ \|\boldsymbol{E}\boldsymbol{\theta}_{[\cdot m]}^{(t)} - \boldsymbol{\Delta}_{[\cdot m]}^{(t)}\|_2^2 - \|\boldsymbol{E}\boldsymbol{\theta}_{[\cdot m]}^{(t)} - \boldsymbol{\Delta}_{[\cdot m]}^{(t-1)}\|_2^2 - 2 \left(\boldsymbol{E}\boldsymbol{\theta}_{[\cdot m]}^{(t)} - \boldsymbol{\Delta}_{[\cdot m]}^{(t)} \right)^\top (\boldsymbol{\Delta}_{[\cdot m]}^{(t-1)} - \boldsymbol{\Delta}_{[\cdot m]}^{(t)}) \right\} \\
\leq&\sum_{m=1}^{2} \left(\frac{C_p}{2} \|\boldsymbol{\Delta}_{[\cdot m]}^{(t)} - \boldsymbol{\Delta}_{[\cdot m]}^{(t-1)}\|_2^2 - \frac{\rho}{2} \|\boldsymbol{\Delta}_{[\cdot m]}^{(t)} - \boldsymbol{\Delta}_{[\cdot m]}^{(t-1)}\|_2^2 \right) \\
=&- \frac{\rho - C_p}{2} \|\boldsymbol{\Delta}^{(t)} - \boldsymbol{\Delta}^{(t-1)}\|_F^2,
\end{aligned}
$$

where the inequality is derived from the weakly convexity of the penalty function and the strongly convexity of the function $\|u\|^2$.

Similarly, by the convexity of the negative log-likelihood function and the strongly convexity of the function $\|u\|^2$, it can be verified that

$$\begin{aligned}&H(\boldsymbol{\beta}^{(t)}, \boldsymbol{\Theta}^{(t)}, \boldsymbol{\Delta}^{(t-1)}, \boldsymbol{\nu}^{(t-1)}) - H(\boldsymbol{\beta}^{(t)}, (\boldsymbol{\theta}_{[\cdot 1]}^{(t)}, \boldsymbol{\theta}_{[\cdot 2]}^{(t-1)}), \boldsymbol{\Delta}^{(t-1)}, \boldsymbol{\nu}^{(t-1)})\\ \leq & -\frac{\rho}{2}\|\boldsymbol{E}\boldsymbol{\theta}_{[\cdot 2]}^{(t)} - \boldsymbol{E}\boldsymbol{\theta}_{[\cdot 2]}^{(t-1)}\|_2^2,\end{aligned}$$

and

$$\begin{aligned}&H(\boldsymbol{\beta}^{(t)}, (\boldsymbol{\theta}_{[\cdot 1]}^{(t)}, \boldsymbol{\theta}_{[\cdot 2]}^{(t-1)}), \boldsymbol{\Delta}^{(t-1)}, \boldsymbol{\nu}^{(t-1)}) - H(\boldsymbol{\beta}^{(t-1)}, \boldsymbol{\Theta}^{(t-1)}, \boldsymbol{\Delta}^{(t-1)}, \boldsymbol{\nu}^{(t-1)})\\ \leq & -\frac{\rho}{2}\|\boldsymbol{E}\boldsymbol{\theta}_{[\cdot 1]}^{(t)} - \boldsymbol{E}\boldsymbol{\theta}_{[\cdot 1]}^{(t-1)}\|_2^2.\end{aligned}$$

Therefore, we have

$$\begin{aligned}&H(\boldsymbol{\beta}^{(t)}, \boldsymbol{\Theta}^{(t)}, \boldsymbol{\Delta}^{(t)}, \boldsymbol{\nu}^{(t)}) - H(\boldsymbol{\beta}^{(t-1)}, \boldsymbol{\Theta}^{(t-1)}, \boldsymbol{\Delta}^{(t-1)}, \boldsymbol{\nu}^{(t-1)})\\ \leq & 4\rho^{-1}n(n-1)C_s^2 - \frac{\rho}{2}\|\boldsymbol{E}\boldsymbol{\Theta}^{(t)} - \boldsymbol{E}\boldsymbol{\Theta}^{(t-1)}\|_F^2 - \frac{\rho - C_p}{2}\|\boldsymbol{\Delta}^{(t)} - \boldsymbol{\Delta}^{(t-1)}\|_F^2.\end{aligned}$$

This completes the proof. □

2.5 Proof of Theorem 1

Proof. (i) From Lemma 1 and Lemma 2, $H(\boldsymbol{\beta}^{(t)}, \boldsymbol{\Theta}^{(t)}, \boldsymbol{\Delta}^{(t)}, \boldsymbol{\nu}^{(t)})$ is upper bounded and so are $Q(\boldsymbol{\beta}^{(t)}, \boldsymbol{\Theta}^{(t)}, \boldsymbol{\Delta}')$ and $\|\boldsymbol{E}\boldsymbol{\Theta}^{(t)} - \boldsymbol{\Delta}^{(t)}\|_F^2$. As the objective function $Q(\boldsymbol{\beta}, \boldsymbol{\Theta}, \boldsymbol{\Delta})$ is coercive over the feasible set, $\{\boldsymbol{\beta}^{(t)}, \boldsymbol{\Theta}^{(t)}\}$ is bounded and, therefore, $\{\boldsymbol{\Delta}^{(t)}\}$ is bounded. By the assumption that the subdifferential of the penalty function is bounded and (S2.3), $\{\boldsymbol{\nu}^{(t)}\}$ is also bounded.

By Lemma 2, we have

$$\begin{aligned} & H(\boldsymbol{\beta}^{(t)}, \boldsymbol{\Theta}^{(t)}, \boldsymbol{\Delta}^{(t)}, \boldsymbol{\nu}^{(t)}) - H(\boldsymbol{\beta}^{(0)}, \boldsymbol{\Theta}^{(0)}, \boldsymbol{\Delta}^{(0)}, \boldsymbol{\nu}^{(0)}) \\ \leq & \sum_{l=1}^{t} \left\{ 4\rho^{-1} n(n-1) C_s^2 - \frac{\rho}{2} \|\boldsymbol{E}\boldsymbol{\Theta}^{(l)} - \boldsymbol{E}\boldsymbol{\Theta}^{(l-1)}\|_F^2 - \frac{\rho - C_p}{2} \|\boldsymbol{\Delta}^{(l)} - \boldsymbol{\Delta}^{(l-1)}\|_F^2 \right\}. \end{aligned}$$

As $H(\boldsymbol{\beta}^{(t)}, \boldsymbol{\Theta}^{(t)}, \boldsymbol{\Delta}^{(t)}, \boldsymbol{\nu}^{(t)}) > -\infty$ by Lemma 1 and $\rho > 0$, $\sum_{l=1}^{t} \|\boldsymbol{E}\boldsymbol{\Theta}^{(l)} - \boldsymbol{E}\boldsymbol{\Theta}^{(l-1)}\|_F^2$ is upper bounded. Based on the fact that $\sum_{l=1}^{t} \|\boldsymbol{E}\boldsymbol{\Theta}^{(l)} - \boldsymbol{E}\boldsymbol{\Theta}^{(l-1)}\|_F^2 \leq \sum_{l=1}^{t+1} \|\boldsymbol{E}\boldsymbol{\Theta}^{(l)} - \boldsymbol{E}\boldsymbol{\Theta}^{(l-1)}\|_F^2$, $\sum_{l=1}^{t} \|\boldsymbol{E}\boldsymbol{\Theta}^{(l)} - \boldsymbol{E}\boldsymbol{\Theta}^{(l-1)}\|_F^2$ converges to a non-negative number as $t \to \infty$. As a result, $\lim_{t\to\infty} \|\boldsymbol{E}\boldsymbol{\Theta}^{(t)} - \boldsymbol{E}\boldsymbol{\Theta}^{(t-1)}\|_F^2 = 0$. Similarly, we have $\lim_{t\to\infty} \|\boldsymbol{\Delta}^{(t)} - \boldsymbol{\Delta}^{(t-1)}\|_F^2 = 0$, and thus,

$$\lim_{t\to\infty} \|\boldsymbol{R}_d^{(t)}\|_F = \lim_{t\to\infty} \|\rho \boldsymbol{E}^\top (\boldsymbol{\Delta}^{(t+1)} - \boldsymbol{\Delta}^{(t)})\|_F = 0.$$

As $\|\boldsymbol{\Theta}^{(t)} - \boldsymbol{\Theta}^{(t-1)}\|_F^2 \leq \lambda_{++}^{-1}(\boldsymbol{E}\boldsymbol{E}^\top) \|\boldsymbol{E}\boldsymbol{\Theta}^{(t)} - \boldsymbol{E}\boldsymbol{\Theta}^{(t-1)}\|_F^2$, where $\lambda_{++}(\boldsymbol{E}\boldsymbol{E}^\top)$ is the smallest strictly-positive eigenvalue of $\boldsymbol{E}\boldsymbol{E}^\top$, we have $\lim_{t\to\infty} \|\boldsymbol{\Theta}^{(t)} - \boldsymbol{\Theta}^{(t-1)}\|_F^2 = 0$. Moreover, by (3.3), it holds that $\lim_{t\to\infty} \|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^{(t-1)}\|_2^2 = 0$.

As $(\boldsymbol{\beta}^{(t)}, \boldsymbol{\theta}_{[\cdot 1]}^{(t)})$ is a minimizer of $H(\boldsymbol{\beta}, (\boldsymbol{\theta}_{[\cdot 1]}, \boldsymbol{\theta}_{[\cdot 2]}^{(t-1)}), \boldsymbol{\Delta}^{(t-1)}, \boldsymbol{\nu}^{(t-1)})$, by the first order condition of optimality, we have

$$\begin{aligned}
&\frac{\partial H(\boldsymbol{\beta}^{(t)},(\boldsymbol{\theta}_{[\cdot 1]}^{(t)},\boldsymbol{\theta}_{[\cdot 2]}^{(t-1)}),\boldsymbol{\Delta}^{(t-1)},\boldsymbol{\nu}^{(t-1)})}{\partial\boldsymbol{\theta}_{[\cdot 1]}}\\
=&\frac{-\partial L(\boldsymbol{\beta}^{(t)},(\boldsymbol{\theta}_{[\cdot 1]}^{(t)},\boldsymbol{\theta}_{[\cdot 2]}^{(t-1)}))}{\partial\boldsymbol{\theta}_{[\cdot 1]}}-\boldsymbol{E}^\top\boldsymbol{\nu}_{[\cdot 1]}^{(t-1)}-\rho\boldsymbol{E}^\top(\boldsymbol{E}\boldsymbol{\theta}_{[\cdot 1]}^{(t)}-\boldsymbol{\Delta}_{[\cdot 1]}^{(t-1)})\\
=&\frac{-\partial L(\boldsymbol{\beta}^{(t)},(\boldsymbol{\theta}_{[\cdot 1]}^{(t)},\boldsymbol{\theta}_{[\cdot 2]}^{(t-1)}))}{\partial\boldsymbol{\theta}_{[\cdot 1]}}-\boldsymbol{E}^\top\boldsymbol{\nu}_{[\cdot 1]}^{(t)}-\rho\boldsymbol{E}^\top(\boldsymbol{\Delta}_{[\cdot 1]}^{(t)}-\boldsymbol{\Delta}_{[\cdot 1]}^{(t-1)})=\mathbf{0}.
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
\boldsymbol{E}^\top(\boldsymbol{\nu}_{[\cdot 1]}^{(t+1)}-\boldsymbol{\nu}_{[\cdot 1]}^{(t)})=&\frac{\partial L(\boldsymbol{\beta}^{(t)},(\boldsymbol{\theta}_{[\cdot 1]}^{(t)},\boldsymbol{\theta}_{[\cdot 2]}^{(t-1)}))}{\partial\boldsymbol{\theta}_{[\cdot 1]}}-\frac{\partial L(\boldsymbol{\beta}^{(t+1)},(\boldsymbol{\theta}_{[\cdot 1]}^{(t+1)},\boldsymbol{\theta}_{[\cdot 2]}^{(t)}))}{\partial\boldsymbol{\theta}_{[\cdot 1]}}\\
&-\rho\boldsymbol{E}^\top(\boldsymbol{\Delta}_{[\cdot 1]}^{(t+1)}-\boldsymbol{\Delta}_{[\cdot 1]}^{(t)})+\rho\boldsymbol{E}^\top(\boldsymbol{\Delta}_{[\cdot 1]}^{(t)}-\boldsymbol{\Delta}_{[\cdot 1]}^{(t-1)}).
\end{aligned}$$

It can be verified that

$$\begin{aligned}
&\frac{\partial L(\boldsymbol{\beta}^{(t)},(\boldsymbol{\theta}_{[\cdot 1]}^{(t)},\boldsymbol{\theta}_{[\cdot 2]}^{(t-1)}))}{\partial\boldsymbol{\theta}_{[\cdot 1]}}-\frac{\partial L(\boldsymbol{\beta}^{(t+1)},(\boldsymbol{\theta}_{[\cdot 1]}^{(t+1)},\boldsymbol{\theta}_{[\cdot 2]}^{(t)}))}{\partial\boldsymbol{\theta}_{[\cdot 1]}}\\
=&\left\{\theta_{i2}^{(t-1)}(y_i-(\boldsymbol{\beta}^{(t)})^\top\boldsymbol{x}_i-\theta_{i1}^{(t)})-\theta_{i2}^{(t)}(y_i-(\boldsymbol{\beta}^{(t+1)})^\top\boldsymbol{x}_i-\theta_{i1}^{(t+1)})\right\}_{i=1}^n\\
=&\left\{\theta_{i2}^{(t-1)}((\boldsymbol{\beta}^{(t+1)}-\boldsymbol{\beta}^{(t)})^\top\boldsymbol{x}_i+\theta_{i1}^{(t+1)}-\theta_{i1}^{(t)})+(\theta_{i2}^{(t-1)}-\theta_{i2}^{(t)})(y_i-(\boldsymbol{\beta}^{(t+1)})^\top\boldsymbol{x}_i-\theta_{i1}^{(t+1)})\right\}_{i=1}^n.
\end{aligned}$$

Therefore, as $\lim_{t\to\infty}\|\boldsymbol{\Theta}^{(t)}-\boldsymbol{\Theta}^{(t-1)}\|_F^2=0$ and $\lim_{t\to\infty}\|\boldsymbol{\beta}^{(t)}-\boldsymbol{\beta}^{(t-1)}\|_2^2=0$, it holds that

$$\lim_{t\to\infty}\left\|\frac{\partial L(\boldsymbol{\beta}^{(t)},(\boldsymbol{\theta}_{[\cdot 1]}^{(t)},\boldsymbol{\theta}_{[\cdot 2]}^{(t-1)}))}{\partial\boldsymbol{\theta}_{[\cdot 1]}}-\frac{\partial L(\boldsymbol{\beta}^{(t+1)},(\boldsymbol{\theta}_{[\cdot 1]}^{(t+1)},\boldsymbol{\theta}_{[\cdot 2]}^{(t)}))}{\partial\boldsymbol{\theta}_{[\cdot 1]}}\right\|_2^2=0,$$

and $\lim_{t\to\infty}\|\boldsymbol{E}^\top(\boldsymbol{\nu}_{[\cdot 1]}^{(t+1)}-\boldsymbol{\nu}_{[\cdot 1]}^{(t)})\|_2^2=0$. By $\|\boldsymbol{\nu}_{[\cdot 1]}^{(t+1)}-\boldsymbol{\nu}_{[\cdot 1]}^{(t)}\|_2^2\le\lambda_{++}^{-1}(\boldsymbol{E}^\top\boldsymbol{E})\|\boldsymbol{E}^\top(\boldsymbol{\nu}_{[\cdot 1]}^{(t+1)}-\boldsymbol{\nu}_{[\cdot 1]}^{(t)})\|_2^2$, we have $\lim_{t\to\infty}\|\boldsymbol{\nu}_{[\cdot 1]}^{(t+1)}-\boldsymbol{\nu}_{[\cdot 1]}^{(t)}\|_2^2=0$.

As $\boldsymbol{\theta}_{[\cdot 2]}^{(t)}$ is a minimizer of $H(\boldsymbol{\beta}^{(t)},(\boldsymbol{\theta}_{[\cdot 1]}^{(t)},\boldsymbol{\theta}_{[\cdot 2]}),\boldsymbol{\Delta}^{(t-1)},\boldsymbol{\nu}^{(t-1)})$, following similar steps, it can be verified that $\lim_{t\to\infty}\|\boldsymbol{\nu}_{[\cdot 2]}^{(t+1)}-\boldsymbol{\nu}_{[\cdot 2]}^{(t)}\|_2^2=0$. Therefore, we have

$$\lim_{t\to\infty}\|\boldsymbol{R}_p(\boldsymbol{\Theta}^{(t)},\boldsymbol{\Delta}^{(t)})\|_F=\lim_{t\to\infty}\|\boldsymbol{E}\boldsymbol{\Theta}^{(t)}-\boldsymbol{\Delta}^{(t)}\|_F=\lim_{t\to\infty}\|\boldsymbol{\nu}^{(t)}-\boldsymbol{\nu}^{(t-1)}\|_F=0.$$

(ii) Next we prove that $\|\partial H(\boldsymbol{\beta}^{(t)},\boldsymbol{\Theta}^{(t)},\boldsymbol{\Delta}^{(t)},\boldsymbol{\nu}^{(t)})\|_2\to 0$ as $t\to\infty$. Note that

$$\begin{aligned}&\frac{\partial H(\boldsymbol{\beta}^{(t)},\boldsymbol{\Theta}^{(t)},\boldsymbol{\Delta}^{(t)},\boldsymbol{\nu}^{(t)})}{\partial\boldsymbol{\beta}}=\frac{-\partial L(\boldsymbol{\beta}^{(t)},\boldsymbol{\Theta}^{(t)})}{\partial\boldsymbol{\beta}}\\ =&\frac{-\partial L(\boldsymbol{\beta}^{(t)},(\boldsymbol{\theta}_{[\cdot 1]}^{(t)},\boldsymbol{\theta}_{[\cdot 2]}^{(t-1)}))}{\partial\boldsymbol{\beta}}+\frac{-\partial L(\boldsymbol{\beta}^{(t)},\boldsymbol{\Theta}^{(t)})}{\partial\boldsymbol{\beta}}-\frac{-\partial L(\boldsymbol{\beta}^{(t)},(\boldsymbol{\theta}_{[\cdot 1]}^{(t)},\boldsymbol{\theta}_{[\cdot 2]}^{(t-1)}))}{\partial\boldsymbol{\beta}}\\ =&\sum_{i=1}^n(\theta_{i2}^{(t-1)}-\theta_{i2}^{(t)})(y_i-(\boldsymbol{\beta}^{(t)})^\top\boldsymbol{x}_i-\theta_{i1}^{(t)})\boldsymbol{x}_i.\end{aligned}$$

Based on the fact that $\|(y_i-(\boldsymbol{\beta}^{(t)})^\top\boldsymbol{x}_i-\theta_{i1}^{(t)})\boldsymbol{x}_i\|_2$ is bounded and $\lim_{t\to\infty}\|\boldsymbol{\theta}_{[\cdot 2]}^{(t)}-\boldsymbol{\theta}_{[\cdot 2]}^{(t-1)}\|_2=0$, we have $\lim_{t\to\infty}\|\partial H/\partial\boldsymbol{\beta}\|_2=0$. Similarly, it can be verified that $\lim_{t\to\infty}\|\partial H/\partial\boldsymbol{\theta}_{[\cdot 1]}\|_2=0$ and $\lim_{t\to\infty}\|\partial H/\partial\boldsymbol{\theta}_{[\cdot 2]}\|_2=0$.

Furthermore, as

$$\lim_{t\to\infty}\|\partial H/\partial\boldsymbol{\nu}_{[\cdot m]}\|_2=\lim_{t\to\infty}\|\boldsymbol{E}\boldsymbol{\theta}_{[\cdot m]}^{(t)}-\boldsymbol{\Delta}_{[\cdot m]}^{(t)}\|_2=0,\quad m=1,2,$$

and

$$\lim_{t\to\infty}\|\partial H/\partial\boldsymbol{\Delta}_{[\cdot m]}\|_2=\lim_{t\to\infty}\|\boldsymbol{\nu}_{[\cdot m]}^{(t-1)}-\boldsymbol{\nu}_{[\cdot m]}^{(t)}\|_2=0,\quad m=1,2,$$

we have $\|\partial H(\boldsymbol{\beta}^{(t)}, \boldsymbol{\Theta}^{(t)}, \boldsymbol{\Delta}^{(t)}, \boldsymbol{\nu}^{(t)})\|_2 \to 0$ as $t \to \infty$.

As $\{\boldsymbol{\beta}^{(t)}, \boldsymbol{\Theta}^{(t)}, \boldsymbol{\Delta}^{(t)}, \boldsymbol{\nu}^{(t)}\}$ is bounded, there exists a convergent subsequence and a limit point, denoted by $\{\boldsymbol{\beta}^{(t_s)}, \boldsymbol{\Theta}^{(t_s)}, \boldsymbol{\Delta}^{(t_s)}, \boldsymbol{\nu}^{(t_s)}\} \to (\boldsymbol{\beta}^*, \boldsymbol{\Theta}^*, \boldsymbol{\Delta}^*, \boldsymbol{\nu}^*)$ as $s \to \infty$. And moreover, we have $\|\partial H(\boldsymbol{\beta}^{(t_s)}, \boldsymbol{\Theta}^{(t_s)}, \boldsymbol{\Delta}^{(t_s)}, \boldsymbol{\nu}^{(t_s)})\|_2 \to 0$ as $s \to \infty$. Since H is a continuous function, it holds that $H(\boldsymbol{\beta}^*, \boldsymbol{\Theta}^*, \boldsymbol{\Delta}^*, \boldsymbol{\nu}^*) = \lim_{s\to\infty} H(\boldsymbol{\beta}^{(t_s)}, \boldsymbol{\Theta}^{(t_s)}, \boldsymbol{\Delta}^{(t_s)}, \boldsymbol{\nu}^{(t_s)})$. By the definition of general subdifferential, we have $0 \in \partial H(\boldsymbol{\beta}^*, \boldsymbol{\Theta}^*, \boldsymbol{\Delta}^*, \boldsymbol{\nu}^*)$.

Thus, the sequence $\{\boldsymbol{\beta}^{(t)}, \boldsymbol{\Theta}^{(t)}, \boldsymbol{\Delta}^{(t)}, \boldsymbol{\nu}^{(t)}\}$ has at least a limit point $\{\boldsymbol{\beta}^*, \boldsymbol{\Theta}^*, \boldsymbol{\Delta}^*, \boldsymbol{\nu}^*\}$, and any limit point is a stationary point. Similar conclusions and steps are found in Wang et al. (2019) (see Proposition 2 in their paper). This completes the proof. □

3. Proofs of theorems in Section 4

3.1 Proof of Theorem 2

Proof. For $a, b \in \mathbb{R}^+$, we denote $a \lesssim b$ if $a \le cb$ for some generic constant $0 < c < +\infty$. As $((\widehat{\boldsymbol{\theta}}^{\text{or}}_{[\cdot 1]})^\top, (\widehat{\boldsymbol{\theta}}^{\text{or}}_{[\cdot 2]})^\top) = ((\boldsymbol{Z}^{(1)}\widehat{\boldsymbol{\mu}}^{\text{or}})^\top, (\boldsymbol{Z}^{(2)}\widehat{\boldsymbol{\tau}}^{\text{or}})^\top)$, to prove Theorem 2, it suffices to show

$$\begin{aligned}
&\|((\widehat{\boldsymbol{\beta}}^{\text{or}})^\top, (\widehat{\boldsymbol{\mu}}^{\text{or}})^\top, (\widehat{\boldsymbol{\tau}}^{\text{or}})^\top) - ((\boldsymbol{\beta}^0)^\top, (\boldsymbol{\mu}^0)^\top, (\boldsymbol{\tau}^0)^\top)\|_2 \\
&\lesssim \max(K_1, K_2)\left(\frac{(p+K_1)K_2^2\log n}{n}\right)^{1/2} + \max(K_1, K_2)\left(\frac{K_2\log n}{n}\right)^{1/2},
\end{aligned} \tag{S3.6}$$

with probability converging to 1 as n goes to infinity.

Denote $\boldsymbol{\eta} = (\boldsymbol{\beta}^\top, \boldsymbol{\mu}^\top)^\top$, $\boldsymbol{\alpha} = (\boldsymbol{\eta}^\top, \boldsymbol{\tau}^\top)^\top$, and $\widetilde{\boldsymbol{x}}_i = (\boldsymbol{x}_i^\top, (\boldsymbol{z}^{(1)}_{[i\cdot]})^\top)^\top$. Define

$$S_n(\boldsymbol{\alpha}) = n^{-1}\sum_{i=1}^{n}\sum_{k'=1}^{K_2} z^{(2)}_{ik'}\left\{\log\tau_{k'} - \tau_{k'}(y_i - \boldsymbol{\eta}^\top\widetilde{\boldsymbol{x}}_i)^2\right\}.$$

It can be seen that $((\widehat{\boldsymbol{\beta}}^{\text{or}})^\top, (\widehat{\boldsymbol{\mu}}^{\text{or}})^\top, (\widehat{\boldsymbol{\tau}}^{\text{or}})^\top)^\top = \operatorname{argmax} S_n(\boldsymbol{\alpha})$. Furthermore, define the function evaluating the error between $\boldsymbol{\alpha}$ and the true parameter $\boldsymbol{\alpha}^0 = ((\boldsymbol{\beta}^0)^\top, (\boldsymbol{\mu}^0)^\top, (\boldsymbol{\tau}^0)^\top)^\top$ as

$$g(\Delta\boldsymbol{\alpha}) = S_n(\boldsymbol{\alpha}^0 + \Delta\boldsymbol{\alpha}) - S_n(\boldsymbol{\alpha}^0).$$

As $\widehat{\boldsymbol{\alpha}}^{\text{or}} = ((\widehat{\boldsymbol{\beta}}^{\text{or}})^\top, (\widehat{\boldsymbol{\mu}}^{\text{or}})^\top, (\widehat{\boldsymbol{\tau}}^{\text{or}})^\top)^\top$ is the maximum point of $S_n(\boldsymbol{\alpha})$, we have $g(\widetilde{\Delta\boldsymbol{\alpha}}) \geq 0$ with $\widetilde{\Delta\boldsymbol{\alpha}} = \boldsymbol{\alpha}^{\text{or}} - \boldsymbol{\alpha}^0$. We prove (S3.6) by contradiction. For any $\delta_1, \delta_2 > 0$, let

$$\begin{aligned}
\iota_1 =& (p + K_1)^{1/2} K_2 \left(\frac{2\max\{M, 1\} \log(2(p + K_1)K_2/\delta_1)}{\tau_{\min} n} \right)^{1/2}, \\
\iota_2 =& (K_2)^{1/2} \max\{(2n^{-1}\log(K_2/\delta_2))^{1/2}, 2n^{-1}\log(K_2/\delta_2)\},
\end{aligned}$$

and

$$\begin{aligned}
\xi_1 &= \tilde{c}_1^{-1} \max(K_1, K_2)(2(1 + c_2)\tau_{\max}\iota_1 + 16\tau_{\min}^{-1}\iota_2), \\
\xi_2 &= 2\tilde{c}_2^{-1} c_2 \tau_{\max} \max\{K_1, K_2\}\iota_1, \\
\xi_3 &= 4\tilde{c}_3^{-1} c_2 \tau_{\max} \max(K_1, K_2)\iota_1,
\end{aligned}$$

where $\tilde{c}_1, \tilde{c}_2, c_2$ and $\tilde{c}_3$ are finite constants defined below. Let $\xi = \max\{2\xi_1, 2\xi_2, 2\xi_3\}$. If $\|\widetilde{\Delta\boldsymbol{\alpha}}\|_2 \geq \xi$, there exists some $0 < t^* \leq 1$ such that $\|t^*\widetilde{\Delta\boldsymbol{\alpha}}\|_2 = \xi/2$. Denote $\boldsymbol{\alpha}^* = t^*\boldsymbol{\alpha}^{\text{or}} + (1 - t^*)\boldsymbol{\alpha}^0$. We then prove the following two results:

(i) $g(t^*\widetilde{\Delta\boldsymbol{\alpha}}) < 0$ with probability at least $1 - 2(\delta_1 + \delta_2)$.

(ii) $g(t^*\widetilde{\Delta\boldsymbol{\alpha}}) \geq g(\widetilde{\Delta\boldsymbol{\alpha}}) \geq 0$ with probability at least $1 - (4\delta_1 + 2\delta_2)$.

By the results in (i) and (ii), $\|\widetilde{\Delta\boldsymbol{\alpha}}\|_2 \geq \xi$ leads to a contradiction and thus, we have $\|\widetilde{\Delta\boldsymbol{\alpha}}\|_2 < \xi$ with probability at least $1-(6\delta_1+4\delta_2)$. Under the condition $(p+K_1)K_2 = o(n)$, setting $\delta_1 = (p+K_1)K_2/n$ and $\delta_2 = K_2/n$, it holds that

$$\|\boldsymbol{\alpha}^{\text{or}} - \boldsymbol{\alpha}^0\|_2 = O_p\left(\max(K_1,K_2)\left(\frac{(p+K_1)K_2^2\log n}{n}\right)^{1/2} + \max(K_1,K_2)\left(\frac{K_2\log n}{n}\right)^{1/2}\right).$$

Step 1: We prove the result in (i). Note that for any $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}'$,

$$S_n(\boldsymbol{\alpha}) - S_n(\boldsymbol{\alpha}^{'}) = I_1(\boldsymbol{\alpha},\boldsymbol{\alpha}') + I_2(\boldsymbol{\alpha},\boldsymbol{\alpha}'),$$

where

$$I_1(\boldsymbol{\alpha},\boldsymbol{\alpha}') = n^{-1}\sum_{i=1}^{n}\sum_{k'=1}^{K_2} z_{ik'}^{(2)}\left(-\tau_{k'}(y_i - \boldsymbol{\eta}^\top\widetilde{\boldsymbol{x}}_i)^2 + \tau_{k'}(y_i - (\boldsymbol{\eta}^{'})^\top\widetilde{\boldsymbol{x}}_i)^2\right),$$
$$I_2(\boldsymbol{\alpha},\boldsymbol{\alpha}') = n^{-1}\sum_{i=1}^{n}\sum_{k'=1}^{K_2} z_{ik'}^{(2)}\left(-\tau_{k'}(y_i - (\boldsymbol{\eta}')^\top\widetilde{\boldsymbol{x}}_i)^2 + \tau_{k'}^{'}(y_i - (\boldsymbol{\eta}')^\top\widetilde{\boldsymbol{x}}_i)^2 + \log\tau_{k'} - \log\tau_{k'}^{'}\right).$$

On the other hand, we have

$$\langle\nabla_{\boldsymbol{\alpha}}S_n(\boldsymbol{\alpha}^{'}), \boldsymbol{\alpha}-\boldsymbol{\alpha}^{'}\rangle = \langle\nabla_{\boldsymbol{\eta}}S_n(\boldsymbol{\alpha}^{'}), \boldsymbol{\eta}-\boldsymbol{\eta}^{'}\rangle + \langle\nabla_{\boldsymbol{\tau}}S_n(\boldsymbol{\alpha}^{'}), \boldsymbol{\tau}-\boldsymbol{\tau}^{'}\rangle.$$

Then one can verify that

$$
\begin{aligned}
&I_1(\boldsymbol{\alpha},\boldsymbol{\alpha}') - \langle \nabla_{\boldsymbol{\eta}} S_n(\boldsymbol{\alpha}^{'}), \boldsymbol{\eta}-\boldsymbol{\eta}^{'}\rangle \\
=&I_1(\boldsymbol{\alpha},\boldsymbol{\alpha}') - n^{-1}\sum_{i=1}^{n}\sum_{k'=1}^{K_2} 2z_{ik'}^{(2)}\tau_{k'}'(y_i-(\boldsymbol{\eta}^{'})^{\top}\widetilde{\boldsymbol{x}}_i)\widetilde{\boldsymbol{x}}_i^{\top}(\boldsymbol{\eta}-\boldsymbol{\eta}^{'}) \\
=&I_1(\boldsymbol{\alpha},\boldsymbol{\alpha}') - n^{-1}\sum_{i=1}^{n}\sum_{k'=1}^{K_2} 2z_{ik'}^{(2)}\tau_{k'}(y_i-(\boldsymbol{\eta}^{'})^{\top}\widetilde{\boldsymbol{x}}_i)\widetilde{\boldsymbol{x}}_i^{\top}(\boldsymbol{\eta}-\boldsymbol{\eta}^{'}) \\
&+n^{-1}\sum_{i=1}^{n}\sum_{k'=1}^{K_2} 2z_{ik'}^{(2)}(\tau_{k'}-\tau_{k'}')(y_i-(\boldsymbol{\eta}^{'})^{\top}\widetilde{\boldsymbol{x}}_i)\widetilde{\boldsymbol{x}}_i^{\top}(\boldsymbol{\eta}-\boldsymbol{\eta}^{'}) \\
=&-n^{-1}\sum_{i=1}^{n}\sum_{k'=1}^{K_2} z_{ik'}^{(2)}\tau_{k'}\|(\boldsymbol{\eta}-\boldsymbol{\eta}^{'})^{\top}\widetilde{\boldsymbol{x}}_i\|_2^2 \\
&+n^{-1}\sum_{i=1}^{n}\sum_{k'=1}^{K_2} 2z_{ik'}^{(2)}(\tau_{k'}-\tau_{k'}')(y_i-(\boldsymbol{\eta}^{'})^{\top}\widetilde{\boldsymbol{x}}_i)\widetilde{\boldsymbol{x}}_i^{\top}(\boldsymbol{\eta}-\boldsymbol{\eta}^{'}).
\end{aligned}
\tag{S3.7}
$$

We then consider $I_2(\boldsymbol{\alpha},\boldsymbol{\alpha}') - \langle \nabla_{\boldsymbol{\tau}} S_n(\boldsymbol{\alpha}^{'}), \boldsymbol{\tau}-\boldsymbol{\tau}^{'}\rangle$. Define

$$
\psi_n(\boldsymbol{\tau}) = n^{-1}\sum_{i=1}^{n}\sum_{k'=1}^{K_2} z_{ik'}^{(2)}\left\{-\tau_{k'}(y_i-(\boldsymbol{\eta}')^{\top}\widetilde{\boldsymbol{x}}_i)^2+\log\tau_{k'}\right\}.
$$

Note that

$$
\begin{aligned}
I_2(\boldsymbol{\alpha},\boldsymbol{\alpha}') - \langle \nabla_{\boldsymbol{\tau}} S_n(\boldsymbol{\alpha}^{'}), \boldsymbol{\tau}-\boldsymbol{\tau}^{'}\rangle =&\psi_n(\boldsymbol{\tau})-\psi_n(\boldsymbol{\tau}^{'})-\langle \nabla\psi_n(\boldsymbol{\tau}^{'}),\boldsymbol{\tau}-\boldsymbol{\tau}^{'}\rangle \\
=&2^{-1}(\boldsymbol{\tau}-\boldsymbol{\tau}^{'})^{\top}\nabla^2\psi_n(\boldsymbol{\tau}^{''})(\boldsymbol{\tau}-\boldsymbol{\tau}^{'}),
\end{aligned}
$$

with $\boldsymbol{\tau}^{''} = t\boldsymbol{\tau} + (1-t)\boldsymbol{\tau}^{'}$ for some $t \in [0,1]$. It can be verified that

$$\begin{cases} \dfrac{\partial^2 \psi_n(\boldsymbol{\tau}^{''})}{\partial \tau_{k'} \partial \tau_{k''}} =0, \quad \text{for } k' \neq k^{''}, \\ \dfrac{\partial^2 \psi_n(\boldsymbol{\tau}^{''})}{\partial \tau_{k'}^2} = -\, n^{-1} \displaystyle\sum_{i=1}^n z_{ik'}^{(2)} (\tau_{k'}^{''})^{-2}. \end{cases}$$

Thus, we have

$$I_2(\boldsymbol{\alpha}, \boldsymbol{\alpha}') - \langle \nabla_{\boldsymbol{\tau}} S_n(\boldsymbol{\alpha}^{'}), \boldsymbol{\tau} - \boldsymbol{\tau}^{'} \rangle \leq -2^{-1} \min_{k'} \left(n^{-1} \sum_{i=1}^n z_{ik'}^{(2)} \right) \min_{k'} \{ (\tau_{k'}^{''})^{-2} \} \| \boldsymbol{\tau} - \boldsymbol{\tau}^{'} \|_2^2. \tag{S3.8}$$

For simplicity, let $\widetilde{\epsilon}_i = y_i - (\boldsymbol{\eta}^0)^\top \widetilde{\boldsymbol{x}}_i$. Obviously, $\widetilde{\epsilon}_i$ is from a normal distribution with mean zero and precision $\tau_{k'}^0$ if $z_{ik'}^{(2)} = 1$ for some k'. We now plug $\boldsymbol{\alpha}^*$ and $\boldsymbol{\alpha}^0$ in (S3.7). By the Central Limit Theory, there exists some constant $0 < c_2 < 1/2$ such that $(1-c_2)\tau_{k'}^0 \leq \tau_{k'}^* \leq (1+c_2)\tau_{k'}^0$ for each k' when n is sufficiently large. One can verify that

$$\begin{aligned} & I_1(\boldsymbol{\alpha}^*, \boldsymbol{\alpha}^0) \\ = & -n^{-1} \sum_{i=1}^n \sum_{k'=1}^{K_2} z_{ik'}^{(2)} \tau_{k'}^* \| (\boldsymbol{\eta}^* - \boldsymbol{\eta}^0)^\top \widetilde{\boldsymbol{x}}_i \|_2^2 \\ & + n^{-1} \sum_{i=1}^n \sum_{k'=1}^{K_2} 2 z_{ik'}^{(2)} (\tau_{k'}^* - \tau_{k'}^0)(y_i - (\boldsymbol{\eta}^0)^\top \widetilde{\boldsymbol{x}}_i) \widetilde{\boldsymbol{x}}_i^\top (\boldsymbol{\eta}^* - \boldsymbol{\eta}^0) + \langle \nabla_{\boldsymbol{\eta}} S_n(\boldsymbol{\alpha}^0), \boldsymbol{\eta}^* - \boldsymbol{\eta}^0 \rangle \\ = & -n^{-1} \sum_{i=1}^n \sum_{k'=1}^{K_2} z_{ik'}^{(2)} \tau_{k'}^* \| (\boldsymbol{\eta}^* - \boldsymbol{\eta}^0)^\top \widetilde{\boldsymbol{x}}_i \|_2^2 + n^{-1} \sum_{i=1}^n \sum_{k'=1}^{K_2} 2 z_{ik'}^{(2)} \tau_{k'}^* (y_i - (\boldsymbol{\eta}^0)^\top \widetilde{\boldsymbol{x}}_i) \widetilde{\boldsymbol{x}}_i^\top (\boldsymbol{\eta}^* - \boldsymbol{\eta}^0) \\ \leq & -n^{-1} \min_{k'} \{ \tau_{k'}^* \} (\boldsymbol{\eta}^* - \boldsymbol{\eta}^0)^\top \widetilde{\boldsymbol{X}}^\top \widetilde{\boldsymbol{X}} (\boldsymbol{\eta}^* - \boldsymbol{\eta}^0) + \sum_{k'=1}^{K_2} 2\tau_{k'}^* \left(n^{-1} \sum_{i=1}^n z_{ik'}^{(2)} \widetilde{\epsilon}_i \widetilde{\boldsymbol{x}}_i^\top (\boldsymbol{\eta}^* - \boldsymbol{\eta}^0) \right). \end{aligned}$$

Note that

$$
\begin{aligned}
&-n^{-1}\min_{k'}\{\tau_{k'}^*\}(\boldsymbol{\eta}^*-\boldsymbol{\eta}^0)^\top\widetilde{\boldsymbol{X}}^\top\widetilde{\boldsymbol{X}}(\boldsymbol{\eta}^*-\boldsymbol{\eta}^0)\le -n^{-1}c_1\min_{k'}\{\tau_{k'}^*\}|\mathcal{G}_{\min}^{(1)}|\|\boldsymbol{\eta}^*-\boldsymbol{\eta}^0\|_2^2\\
=&-c_1\min_{k'}\{\tau_{k'}^*\}\min_{k}\left(n^{-1}\sum_{i=1}^n z_{ik}^{(1)}\right)\|\boldsymbol{\eta}^*-\boldsymbol{\eta}^0\|_2^2.
\end{aligned}
$$

We derive a concentration bound for the random term $\{\min_k(n^{-1}\sum_{i=1}^n z_{ik}^{(1)})\}$. For simplicity, denote $k^*=\operatorname{argmin}_k n^{-1}\sum_{i=1}^n z_{ik}^{(1)}$. Applying Hoeffding's inequality, we have

$$
P\left(\left|n^{-1}\sum_{i=1}^n z_{ik}^{(1)}-\mathbb{E}(z_{ik}^{(1)})\right|\le u\right)\ge 1-2\exp(-2nu^2),\quad k=1,\dots,K_1,
$$

which implies

$$
\left|n^{-1}\sum_{i=1}^n z_{ik}^{(1)}-\mathbb{E}(z_{ik}^{(1)})\right|\le ((2n)^{-1}\log(2K_1K_2/\delta_1))^{1/2},
$$

with probability at least $1-\delta_1/(K_1K_2)$. Therefore, it holds that

$$
P\left(\left|n^{-1}\sum_{i=1}^n z_{ik^*}^{(1)}-\mathbb{E}(z_{ik^*}^{(1)})\right|\le ((2n)^{-1}\log(2K_1K_2/\delta_1))^{1/2}\right)\ge 1-\delta_1/K_2.
$$

Note that for $\delta_1=(p+K_1)K_2/n$, $((2n)^{-1}\log(2K_1K_2/\delta_1))^{1/2}\to 0$, and by Condition (C3), $\mathbb{E}(z_{ik^*}^{(1)})=\sum_{k'=1}^{K_2}\pi_{k^*k'}=O(K_1^{-1})$. Moreover, as $\max\{K_1,K_2\}\sqrt{p+K_1}K_2=o(\sqrt{n(\log n)^{-1}})$, there exists a constant $c_3>0$ such that $n^{-1}\sum_{i=1}^n z_{ik^*}^{(1)}\ge c_3/K_1$ with probability at least $1-\delta_1/K_2$ for sufficiently large n. Furthermore, we have

$\min_{k'}\{\tau^*_{k'}\}\min_k\left(n^{-1}\sum_{i=1}^n z^{(1)}_{ik}\right) \geq (1-c_2)c_3\tau_{\min}K_1^{-1}$ with probability at least $1-\delta_1$.

Moreover, we have

$$\sum_{k'=1}^{K_2} 2\tau^*_{k'}\left(n^{-1}\sum_{i=1}^n z^{(2)}_{ik'}\widetilde{\epsilon}_i\widetilde{\boldsymbol{x}}_i^\top(\boldsymbol{\eta}^*-\boldsymbol{\eta}^0)\right) \leq \sum_{k'=1}^{K_2} 2\tau^*_{k'}\left\|n^{-1}\sum_{i=1}^n z^{(2)}_{ik'}\widetilde{\epsilon}_i\widetilde{\boldsymbol{x}}_i\right\|_2 \left\|\boldsymbol{\eta}^*-\boldsymbol{\eta}^0)\right\|_2.$$

Since $\widetilde{\epsilon}_i$ is sub-Gaussian with parameter $\tau_{\min}^{-1}$, by Condition (C1), $z^{(2)}_{ik'}\widetilde{\epsilon}_i\widetilde{x}_{ij}$ is also sub-Gaussian with parameter $\max\{M,1\}\tau_{\min}^{-1}$ for $j=1,\ldots,p+K_1$ and $k'=1,\ldots,K_2$. Applying Hoeffding's inequality, we have

$$P\left(\left|n^{-1}\sum_{i=1}^n z^{(2)}_{ik'}\widetilde{\epsilon}_i\widetilde{x}_{ij}\right| \leq u\right) \geq 1-2\exp\left(-\frac{\tau_{\min}nu^2}{2\max\{M,1\}}\right), \quad j=1,\ldots,p+K_1,$$

which implies

$$\left|n^{-1}\sum_{i=1}^n z^{(2)}_{ik'}\widetilde{\epsilon}_i\widetilde{x}_{ij}\right| \leq \left(\frac{2\max\{M,1\}\log(2(p+K_1)K_2/\delta_1)}{\tau_{\min}n}\right)^{1/2}, \quad j=1,\ldots,p+K_1,$$

with probability at least $1-\delta_1/((p+K_1)K_2)$. Therefore, it holds that

$$\left\|n^{-1}\sum_{i=1}^n z^{(2)}_{ik'}\widetilde{\epsilon}_i\widetilde{\boldsymbol{x}}_i\right\|_2 \leq (p+K_1)^{1/2}\left(\frac{2\max\{M,1\}\log(2(p+K_1)K_2/\delta_1)}{\tau_{\min}n}\right)^{1/2},$$

with probability at least $1-\delta_1/K_2$. And thus, we have

$$\sum_{k'=1}^{K_2} 2\tau^*_{k'}\left\|n^{-1}\sum_{i=1}^n z^{(2)}_{ik'}\widetilde{\epsilon}_i\widetilde{\boldsymbol{x}}_i\right\|_2 \leq 2(1+c_2)\tau_{\max}\iota_1,$$

with probability at least $1-\delta_1$ and

$$\iota_1 = (p+K_1)^{1/2} K_2 \left(\frac{2\max\{M,1\}\log(2(p+K_1)K_2/\delta_1)}{\tau_{\min} n} \right)^{1/2}.$$

As a result, it holds that

$$I_1(\boldsymbol{\alpha}^*, \boldsymbol{\alpha}^0) \leq -c_1(1-c_2)c_3\tau_{\min}K_1^{-1}\|\boldsymbol{\eta}^* - \boldsymbol{\eta}^0\|_2^2 + 2(1+c_2)\tau_{\max}\iota_1\|\boldsymbol{\eta}^* - \boldsymbol{\eta}^0\|_2, \quad \text{(S3.9)}$$

with probability at least $1-2\delta_1$.

On the other hand, we plug $\boldsymbol{\alpha}^*$ and $\boldsymbol{\alpha}^0$ in (S3.8). As $\boldsymbol{\tau}'' = t\boldsymbol{\tau}^* + (1-t)\boldsymbol{\tau}^0$, similarly, we have $(\tau_{k'}^{''})^{-2} \geq \{(1+c_2)\tau_{k'}^0\}^{-2}$ when n is large enough. Thus, we have

$$\begin{aligned}
I_2(\boldsymbol{\alpha}^*, \boldsymbol{\alpha}^0) &\leq -2^{-1} \min_{k'} \left\{ (\tau_{k'}^{''})^{-2} n^{-1} \sum_{i=1}^n z_{ik'}^{(2)} \right\} \|\boldsymbol{\tau}^* - \boldsymbol{\tau}^0\|_2^2 + \langle \nabla_{\boldsymbol{\tau}} S_n(\boldsymbol{\alpha}^0), \boldsymbol{\tau}^* - \boldsymbol{\tau}^0 \rangle \\
\leq &- 2^{-1} \min_{k'} \left\{ (\tau_{k'}^{''})^{-2} n^{-1} \sum_{i=1}^n z_{ik'}^{(2)} \right\} \|\boldsymbol{\tau}^* - \boldsymbol{\tau}^0\|_2^2 \\
&+ n^{-1} \sum_{i=1}^n \sum_{k'=1}^{K_2} z_{ik'}^{(2)} \{1/\tau_{k'}^0 - (y_i - (\boldsymbol{\eta}^0)^\top \widetilde{\boldsymbol{x}}_i)^2\}(\tau_{k'}^* - \tau_{k'}^0) \\
\leq &- 2^{-1} \min_{k'} \left\{ (\tau_{k'}^{''})^{-2} n^{-1} \sum_{i=1}^n z_{ik'}^{(2)} \right\} \|\boldsymbol{\tau}^* - \boldsymbol{\tau}^0\|_2^2 + n^{-1} \sum_{i=1}^n \sum_{k'=1}^{K_2} z_{ik'}^{(2)} \{\mathbb{E}(\widetilde{\epsilon}_i^2) - \widetilde{\epsilon}_i^2\}(\tau_{k'}^* - \tau_{k'}^0) \\
\leq &- 2^{-1} \min_{k'} \left\{ (\tau_{k'}^{''})^{-2} n^{-1} \sum_{i=1}^n z_{ik'}^{(2)} \right\} \|\boldsymbol{\tau}^* - \boldsymbol{\tau}^0\|_2^2 + \left\| n^{-1} \sum_{i=1}^n \{\mathbb{E}(\widetilde{\epsilon}_i^2) - \widetilde{\epsilon}_i^2\} \boldsymbol{z}_{[i\cdot]}^{(2)} \right\|_2 \|\boldsymbol{\tau}^* - \boldsymbol{\tau}^0\|_2.
\end{aligned}$$

Similarly, we derive a concentration bound for $\min_{k'}\{(\tau_{k'}^{''})^{-2} n^{-1} \sum_{i=1}^n z_{ik'}^{(2)}\}$, i.e., $\min_{k'}\{(\tau_{k'}^{''})^{-2} n^{-1} \sum_{i=1}^n z_{ik'}^{(2)}\} \geq \{(1+c_2)\tau_{\max}\}^{-2} c_3/K_2$ with probability at least $1-\delta_2$ for sufficiently large n as $((2n)^{-1}\log(2K_2/\delta_2))^{1/2} \to 0$ with $\delta_2 = K_2/n$ and Condition (C3) $\mathbb{E}(z_{ik^*}^{(2)}) = \sum_{k=1}^{K_1} \pi_{kk^*} = O(K_2^{-1})$.

We now consider $n^{-1}\sum_{i=1}^{n}\{\mathbb{E}(\widetilde{\epsilon}_i^2)-\widetilde{\epsilon}_i^2\}z_{ik'}^{(2)}$ for $k'=1,\ldots,K_2$. Since $\widetilde{\epsilon}_i$ is sub-Gaussian with parameter $\tau_{\min}^{-1}$, $\{\mathbb{E}(\widetilde{\epsilon}_i^2)-\widetilde{\epsilon}_i^2\}z_{ik'}^{(2)}$ is sub-exponential with parameter $16\tau_{\min}^{-1}$. By Bernstein's inequality, it holds that

$$P\left(\left|n^{-1}\sum_{i=1}^{n}\{\mathbb{E}(\widetilde{\epsilon}_i^2)-\widetilde{\epsilon}_i^2\}z_{ik'}^{(2)}\right|\leq u\right)\geq 1-\exp\left\{-\frac{n}{2}\min\left(\frac{u^2}{(16\tau_{\min}^{-1})^2},\frac{u}{16\tau_{\min}^{-1}}\right)\right\},$$

which implies

$$\left|n^{-1}\sum_{i=1}^{n}\{\mathbb{E}(\widetilde{\epsilon}_i^2)-\widetilde{\epsilon}_i^2\}z_{ik'}^{(2)}\right|\leq 16\tau_{\min}^{-1}\max\{(2n^{-1}\log(K_2/\delta_2))^{1/2},2n^{-1}\log(K_2/\delta_2)\},$$

with probability at least $1-\delta_2/K_2$. Therefore, we have

$$\left\|n^{-1}\sum_{i=1}^{n}\{\mathbb{E}(\widetilde{\epsilon}_i^2)-\widetilde{\epsilon}_i^2\}\boldsymbol{z}_{[i\cdot]}^{(2)}\right\|_2\leq 16\tau_{\min}^{-1}\iota_2,$$

with probability at least $1-\delta_2$ and $\iota_2=(K_2)^{1/2}\max\{(2n^{-1}\log(K_2/\delta_2))^{1/2},2n^{-1}\log(K_2/\delta_2)\}$.

As a result, it holds that

$$I_2(\boldsymbol{\alpha}^*,\boldsymbol{\alpha}^0)\leq -2^{-1}\{(1+c_2)\tau_{\max}\}^{-2}c_3K_2^{-1}\|\boldsymbol{\tau}^*-\boldsymbol{\tau}^0\|_2^2+16\tau_{\min}^{-1}\iota_2\|\boldsymbol{\tau}^*-\boldsymbol{\tau}^0\|_2, \quad \text{(S3.10)}$$

with probability at least $1-2\delta_2$.

By the definition of $\boldsymbol{\alpha}^*$, one can verify that $\|\boldsymbol{\alpha}^*-\boldsymbol{\alpha}^0\|_2=\xi/2$. Combining (S3.9) and (S3.10), we have

$$
\begin{aligned}
&I_1(\boldsymbol{\alpha}^*, \boldsymbol{\alpha}^0) + I_2(\boldsymbol{\alpha}^*, \boldsymbol{\alpha}^0) \\
\leq& - c_1(1-c_2)c_3\tau_{\min}K_1^{-1}\|\boldsymbol{\eta}^* - \boldsymbol{\eta}^0\|_2^2 + 2(1+c_2)\tau_{\max}\iota_1\|\boldsymbol{\eta}^* - \boldsymbol{\eta}^0\|_2 \\
&- 2^{-1}\{(1+c_2)\tau_{\max}\}^{-2}c_3K_2^{-1}\|\boldsymbol{\tau}^* - \boldsymbol{\tau}^0\|_2^2 + 16\tau_{\min}^{-1}\iota_2\|\boldsymbol{\tau}^* - \boldsymbol{\tau}^0\|_2, \\
=& - \tilde{c}_1 \min(K_1^{-1}, K_2^{-1})(\xi/2)^2 + (2(1+c_2)\tau_{\max}\iota_1 + 16\tau_{\min}^{-1}\iota_2)(\xi/2),
\end{aligned}
$$

with probability at least $1 - 2(\delta_1 + \delta_2)$, where $\tilde{c}_1 = \min\{c_1(1-c_2)c_3\tau_{\min}, 2^{-1}\{(1+c_2)\tau_{\max}\}^{-2}c_3\}$. Based on the definition of ξ, we have $\xi/2 \geq \xi_1 = \tilde{c}_1^{-1}\max(K_1, K_2)(2(1+c_2)\tau_{\max}\iota_1+16\tau_{\min}^{-1}\iota_2)$. Therefore, it holds that $g(t^*\widetilde{\Delta\boldsymbol{\alpha}}) = S_n(\boldsymbol{\alpha}^*)-S_n(\boldsymbol{\alpha}^0) = I_1(\boldsymbol{\alpha}^*, \boldsymbol{\alpha}^0)+I_2(\boldsymbol{\alpha}^*, \boldsymbol{\alpha}^0) \leq 0$ with probability at least $1 - 2(\delta_1 + \delta_2)$.

Step 2: We prove the result in (ii). We first prove

$$
S_n(\boldsymbol{\alpha}^0) - S_n(\boldsymbol{\alpha}^*) \leq \langle \nabla S_n(\boldsymbol{\alpha}^*), \boldsymbol{\alpha}^0 - \boldsymbol{\alpha}^* \rangle = \langle \nabla S_n(\boldsymbol{\alpha}^*), -t^*(\boldsymbol{\alpha}^{\mathrm{or}} - \boldsymbol{\alpha}^0) \rangle, \qquad \text{(S3.11)}
$$

with probability at least $1 - (2\delta_1 + \delta_2)$.

Note that $(1-c_2)\tau_{k'}^0 \leq \tau_{k'}^* \leq (1+c_2)\tau_{k'}^0$ for sufficiently large n. Plugging $\boldsymbol{\alpha}^0$ and

$\boldsymbol{\alpha}^*$ in (S3.7), we have

$$\begin{aligned}
& I_1(\boldsymbol{\alpha}^0, \boldsymbol{\alpha}^*) - \langle \nabla_{\boldsymbol{\eta}} S_n(\boldsymbol{\alpha}^*), \boldsymbol{\eta}^0 - \boldsymbol{\eta}^* \rangle \\
\leq & - n^{-1} \sum_{i=1}^{n} \sum_{k'=1}^{K_2} z_{ik'}^{(2)} \tau_{k'}^0 \|(\boldsymbol{\eta}^0 - \boldsymbol{\eta}^*)^\top \widetilde{\boldsymbol{x}}_i\|_2^2 \\
& + n^{-1} \sum_{i=1}^{n} \sum_{k'=1}^{K_2} 2 z_{ik'}^{(2)} (\tau_{k'}^0 - \tau_{k'}^*)(y_i - (\boldsymbol{\eta}^*)^\top \widetilde{\boldsymbol{x}}_i) \widetilde{\boldsymbol{x}}_i^\top (\boldsymbol{\eta}^0 - \boldsymbol{\eta}^*) \\
= & - n^{-1} \sum_{i=1}^{n} \sum_{k'=1}^{K_2} z_{ik'}^{(2)} \tau_{k'}^0 \|(\boldsymbol{\eta}^0 - \boldsymbol{\eta}^*)^\top \widetilde{\boldsymbol{x}}_i\|_2^2 \\
& + n^{-1} \sum_{i=1}^{n} \sum_{k'=1}^{K_2} 2 z_{ik'}^{(2)} (\tau_{k'}^0 - \tau_{k'}^*)(y_i - (\boldsymbol{\eta}^0)^\top \widetilde{\boldsymbol{x}}_i + (\boldsymbol{\eta}^0 - \boldsymbol{\eta}^*)^\top \widetilde{\boldsymbol{x}}_i) \widetilde{\boldsymbol{x}}_i^\top (\boldsymbol{\eta}^0 - \boldsymbol{\eta}^*) \\
= & - n^{-1} \sum_{i=1}^{n} \sum_{k'=1}^{K_2} z_{ik'}^{(2)} (2\tau_{k'}^* - \tau_{k'}^0) \|(\boldsymbol{\eta}^0 - \boldsymbol{\eta}^*)^\top \widetilde{\boldsymbol{x}}_i\|_2^2 \\
& + n^{-1} \sum_{i=1}^{n} \sum_{k'=1}^{K_2} 2 z_{ik'}^{(2)} (\tau_{k'}^0 - \tau_{k'}^*)(y_i - (\boldsymbol{\eta}^0)^\top \widetilde{\boldsymbol{x}}_i) \widetilde{\boldsymbol{x}}_i^\top (\boldsymbol{\eta}^0 - \boldsymbol{\eta}^*) \\
\leq & - c_1 \min_{k'} \{2\tau_{k'}^* - \tau_{k'}^0\} \min_{k} \left(n^{-1} \sum_{i=1}^{n} z_{ik}^{(1)} \right) \|\boldsymbol{\eta}^0 - \boldsymbol{\eta}^*\|_2^2 \\
& + \sum_{k'=1}^{K_2} 2 |\tau_{k'}^0 - \tau_{k'}^*| \left\| n^{-1} \sum_{i=1}^{n} z_{ik'}^{(2)} \widetilde{\epsilon}_i \widetilde{\boldsymbol{x}}_i \right\|_2 \|\boldsymbol{\eta}^0 - \boldsymbol{\eta}^*\|_2.
\end{aligned}$$

Since $\min_{k'}\{2\tau_{k'}^* - \tau_{k'}^0\} n^{-1} \sum_{i=1}^n z_{ik^*}^{(1)} \geq (1 - 2c_2) c_3 \tau_{\min}/K_1$ with probability at least $1 - \delta_1$ for sufficiently large n and

$$\sum_{k'=1}^{K_2} 2 |\tau_{k'}^0 - \tau_{k'}^*| \left\| n^{-1} \sum_{i=1}^{n} z_{ik'}^{(2)} \widetilde{\epsilon}_i \widetilde{\boldsymbol{x}}_i \right\|_2 \leq 2 c_2 \tau_{\max} \iota_1,$$

with probability at least $1 - \delta_1$, it holds that

$$I_1(\boldsymbol{\alpha}^0, \boldsymbol{\alpha}^*) - \langle \nabla_{\boldsymbol{\eta}} S_n(\boldsymbol{\alpha}^*), \boldsymbol{\eta}^0 - \boldsymbol{\eta}^* \rangle \leq -c_1 (1 - 2c_2) c_3 K_1^{-1} \tau_{\min} \|\boldsymbol{\eta}^0 - \boldsymbol{\eta}^*\|_2^2 + 2 c_2 \tau_{\max} \iota_1 \|\boldsymbol{\eta}^0 - \boldsymbol{\eta}^*\|_2,$$

with probability at least $1-2\delta_1$. On the other hand, by (S3.8), it holds that

$$I_2(\boldsymbol{\alpha}^0,\boldsymbol{\alpha}^*)-\langle\nabla_{\boldsymbol{\tau}}S_n(\boldsymbol{\alpha}^*),\boldsymbol{\tau}^0-\boldsymbol{\tau}^*\rangle\le -2^{-1}\{(1+c_2)\tau_{\max}\}^{-2}c_3K_2^{-1}\|\boldsymbol{\tau}^0-\boldsymbol{\tau}^*\|_2^2,$$

with probability at least $1-\delta_2$. Therefore, we have

$$\begin{aligned}
&S_n(\boldsymbol{\alpha}^0)-S_n(\boldsymbol{\alpha}^*)-\langle\nabla S_n(\boldsymbol{\alpha}^*),\boldsymbol{\alpha}^0-\boldsymbol{\alpha}^*\rangle\\
\le&-c_1(1-2c_2)c_3K_1^{-1}\tau_{\min}\|\boldsymbol{\eta}^0-\boldsymbol{\eta}^*\|_2^2+2c_2\tau_{\max}\iota_1\|\boldsymbol{\eta}^0-\boldsymbol{\eta}^*\|_2\\
&-2^{-1}\{(1+c_2)\tau_{\max}\}^{-2}c_3K_2^{-1}\|\boldsymbol{\tau}^0-\boldsymbol{\tau}^*\|_2^2\\
\le&-\tilde{c}_2\min(K_1^{-1},K_2^{-1})(\xi/2)^2+2c_2\tau_{\max}\iota_1(\xi/2),
\end{aligned}$$

with probability at least $1-(2\delta_1+\delta_2)$, where $\tilde{c}_2=\min\{c_1(1-2c_2)c_3\tau_{\min},2^{-1}\{(1+c_2)\tau_{\max}\}^{-2}c_3\}$. Therefore, for $\xi/2\ge\xi_2=2\tilde{c}_2^{-1}c_2\tau_{\max}\max\{K_1,K_2\}\iota_1$, it holds that $S_n(\boldsymbol{\alpha}^0)-S_n(\boldsymbol{\alpha}^*)\le\langle\nabla S_n(\boldsymbol{\alpha}^*),\boldsymbol{\alpha}^0-\boldsymbol{\alpha}^*\rangle$ with probability at least $1-(2\delta_1+\delta_2)$.

We next prove

$$S_n(\boldsymbol{\alpha}^{\mathrm{or}})-S_n(\boldsymbol{\alpha}^*)\le\langle\nabla S_n(\boldsymbol{\alpha}^*),\boldsymbol{\alpha}^{\mathrm{or}}-\boldsymbol{\alpha}^*\rangle=\langle\nabla S_n(\boldsymbol{\alpha}^*),(1-t^*)(\boldsymbol{\alpha}^{\mathrm{or}}-\boldsymbol{\alpha}^0)\rangle,\quad\text{(S3.12)}$$

with probability at least $1-(2\delta_1+\delta_2)$.

Plugging $\boldsymbol{\alpha}^{\text{or}}$ and $\boldsymbol{\alpha}^*$ in (S3.7), we have

$$
\begin{aligned}
& I_1(\boldsymbol{\alpha}^{\text{or}}, \boldsymbol{\alpha}^*) - \langle \nabla_{\boldsymbol{\eta}} S_n(\boldsymbol{\alpha}^*), \boldsymbol{\eta}^{\text{or}} - \boldsymbol{\eta}^* \rangle \\
\leq & - n^{-1} \sum_{i=1}^{n} \sum_{k'=1}^{K_2} z_{ik'}^{(2)} \tau_{k'}^{\text{or}} \| (\boldsymbol{\eta}^{\text{or}} - \boldsymbol{\eta}^*)^\top \widetilde{\boldsymbol{x}}_i \|_2^2 \\
& + n^{-1} \sum_{i=1}^{n} \sum_{k'=1}^{K_2} 2 z_{ik'}^{(2)} (\tau_{k'}^{\text{or}} - \tau_{k'}^*)(y_i - (\boldsymbol{\eta}^0)^\top \widetilde{\boldsymbol{x}}_i + (\boldsymbol{\eta}^0 - \boldsymbol{\eta}^{\text{or}})^\top \widetilde{\boldsymbol{x}}_i + (\boldsymbol{\eta}^{\text{or}} - \boldsymbol{\eta}^*)^\top \widetilde{\boldsymbol{x}}_i) \widetilde{\boldsymbol{x}}_i^\top (\boldsymbol{\eta}^{\text{or}} - \boldsymbol{\eta}^*) \\
= & - n^{-1} \sum_{i=1}^{n} \sum_{k'=1}^{K_2} z_{ik'}^{(2)} (2\tau_{k'}^* - \tau_{k'}^{\text{or}}) \| (\boldsymbol{\eta}^{\text{or}} - \boldsymbol{\eta}^*)^\top \widetilde{\boldsymbol{x}}_i \|_2^2 \\
& + n^{-1} \sum_{i=1}^{n} \sum_{k'=1}^{K_2} 2 z_{ik'}^{(2)} (\tau_{k'}^{\text{or}} - \tau_{k'}^*)(\boldsymbol{\eta}^0 - \boldsymbol{\eta}^{\text{or}})^\top \widetilde{\boldsymbol{x}}_i \widetilde{\boldsymbol{x}}_i^\top (\boldsymbol{\eta}^{\text{or}} - \boldsymbol{\eta}^*) \\
& + n^{-1} \sum_{i=1}^{n} \sum_{k'=1}^{K_2} 2 z_{ik'}^{(2)} (\tau_{k'}^{\text{or}} - \tau_{k'}^*)(y_i - (\boldsymbol{\eta}^0)^\top \widetilde{\boldsymbol{x}}_i) \widetilde{\boldsymbol{x}}_i^\top (\boldsymbol{\eta}^{\text{or}} - \boldsymbol{\eta}^*) \\
\leq & - c_1 \min_{k'} \{2\tau_{k'}^* - \tau_{k'}^{\text{or}}\} \min_{k} \left(n^{-1} \sum_{i=1}^{n} z_{ik}^{(1)} \right) \| \boldsymbol{\eta}^{\text{or}} - \boldsymbol{\eta}^* \|_2^2 \\
& + \sum_{k'=1}^{K_2} 2 |\tau_{k'}^{\text{or}} - \tau_{k'}^*| \left\| n^{-1} \sum_{i=1}^{n} z_{ik'}^{(2)} (\boldsymbol{\eta}^0 - \boldsymbol{\eta}^{\text{or}})^\top \widetilde{\boldsymbol{x}}_i \widetilde{\boldsymbol{x}}_i^\top \right\|_2 \| \boldsymbol{\eta}^{\text{or}} - \boldsymbol{\eta}^* \|_2 \\
& + \sum_{k'=1}^{K_2} 2 |\tau_{k'}^{\text{or}} - \tau_{k'}^*| \left\| n^{-1} \sum_{i=1}^{n} z_{ik'}^{(2)} \widetilde{\epsilon}_i \widetilde{\boldsymbol{x}}_i \right\|_2 \| \boldsymbol{\eta}^{\text{or}} - \boldsymbol{\eta}^* \|_2 .
\end{aligned}
$$

Similarly, it holds that

$$
\begin{aligned}
& I_1(\boldsymbol{\alpha}^{\text{or}}, \boldsymbol{\alpha}^*) - \langle \nabla_{\boldsymbol{\eta}} S_n(\boldsymbol{\alpha}^*), \boldsymbol{\eta}^{\text{or}} - \boldsymbol{\eta}^* \rangle \\
\leq & - c_1 (1 - 3c_2) c_3 \tau_{\min} K_1^{-1} \| \boldsymbol{\eta}^{\text{or}} - \boldsymbol{\eta}^* \|_2^2 + o(1) + 4 c_2 \tau_{\max} \iota_1 \| \boldsymbol{\eta}^{\text{or}} - \boldsymbol{\eta}^* \|_2,
\end{aligned}
$$

with probability at least $1 - 2\delta_1$. By (S3.8), we have

$$
I_2(\boldsymbol{\alpha}^{\text{or}}, \boldsymbol{\alpha}^*) - \langle \nabla_{\boldsymbol{\tau}} S_n(\boldsymbol{\alpha}^*), \boldsymbol{\tau}^{\text{or}} - \boldsymbol{\tau}^* \rangle \leq -2^{-1} \{(1 + c_2)\tau_{\max}\}^{-2} c_3 K_2^{-1} \| \boldsymbol{\tau}^{\text{or}} - \boldsymbol{\tau}^* \|_2^2,
$$

with probability at least $1-\delta_2$. Thus, it holds that

$$\begin{aligned}
&S_n(\boldsymbol{\alpha}^{\mathrm{or}})-S_n(\boldsymbol{\alpha}^*)-\langle\nabla S_n(\boldsymbol{\alpha}^*),\boldsymbol{\alpha}^{\mathrm{or}}-\boldsymbol{\alpha}^*\rangle\\
\leq&-c_1(1-3c_2)c_3\tau_{\min}K_1^{-1}\|\boldsymbol{\eta}^{\mathrm{or}}-\boldsymbol{\eta}^*\|_2^2+4c_2\tau_{\max}\iota_1\|\boldsymbol{\eta}^{\mathrm{or}}-\boldsymbol{\eta}^*\|_2\\
&-2^{-1}\{(1+c_2)\tau_{\max}\}^{-2}c_3K_2^{-1}\|\boldsymbol{\tau}^{\mathrm{or}}-\boldsymbol{\tau}^*\|_2^2\\
\leq&-\tilde{c}_3\min(K_1^{-1},K_2^{-1})\|\boldsymbol{\alpha}^{\mathrm{or}}-\boldsymbol{\alpha}^*\|_2^2+4c_2\tau_{\max}\iota_1\|\boldsymbol{\alpha}^{\mathrm{or}}-\boldsymbol{\alpha}^*\|_2,
\end{aligned}$$

with probability at least $1-(2\delta_1+\delta_2)$, where $\tilde{c}_3=\min\{c_1(1-3c_2)c_3\tau_{\min},2^{-1}\{(1+c_2)\tau_{\max}\}^{-2}c_3\}$. Therefore, for $\xi/2\geq\xi_3=4\tilde{c}_3^{-1}c_2\tau_{\max}\max(K_1,K_2)\iota_1$, it holds that $S_n(\boldsymbol{\alpha}^{\mathrm{or}})-S_n(\boldsymbol{\alpha}^*)\leq\langle\nabla S_n(\boldsymbol{\alpha}^*),\boldsymbol{\alpha}^{\mathrm{or}}-\boldsymbol{\alpha}^*\rangle$ with probability at least $1-(2\delta_1+\delta_2)$.

As $0<t^*\leq 1$, by adding (S3.11) and (S3.12) together with proper scaling, we have

$$t^*S_n(\boldsymbol{\alpha}^{\mathrm{or}})+(1-t^*)S_n(\boldsymbol{\alpha}^0)\leq S_n(\boldsymbol{\alpha}^*),$$

with probability at least $1-(4\delta_1+2\delta_2)$. Therefore, it holds that

$$g(t^*\widetilde{\Delta\boldsymbol{\alpha}})=S_n(\boldsymbol{\alpha}^*)-S_n(\boldsymbol{\alpha}^0)\geq t^*\{S_n(\boldsymbol{\alpha}^{\mathrm{or}})-S_n(\boldsymbol{\alpha}^0)\}=t^*g(\widetilde{\Delta\boldsymbol{\alpha}})\geq 0,$$

with probability at least $1-(4\delta_1+2\delta_2)$. This completes the proof. □

3.2 Proof of Theorem 3

Proof. Note that

$$Q(\boldsymbol{\beta}, \boldsymbol{\Theta}) = -\,L(\boldsymbol{\beta}, \boldsymbol{\Theta}) + \sum_{m=1}^{2} \sum_{1\le i<j\le n} p(|\theta_{im} - \theta_{jm}|, \lambda_m, \gamma_m).$$

When the true group structures $\boldsymbol{Z}^{(1)}$ and $\boldsymbol{Z}^{(2)}$ are known, define $T^{(1)} : \mathbb{R}^{K_1} \to \mathbb{R}^n$ satisfying $T^{(1)}(\boldsymbol{\mu}) = \boldsymbol{Z}^{(1)}\boldsymbol{\mu}$, and $T^{(2)} : \mathbb{R}^{K_2} \to \mathbb{R}^n$ with $T^{(2)}(\boldsymbol{\tau}) = \boldsymbol{Z}^{(2)}\boldsymbol{\tau}$. Furthermore, define $\widetilde{T}^{(1)} : \mathbb{R}^n \to \mathbb{R}^{K_1}$ satisfying $\widetilde{T}^{(1)}(\boldsymbol{\theta}_{[\cdot 1]}) = \{(\boldsymbol{Z}^{(1)})^\top \boldsymbol{Z}^{(1)}\}^{-1}(\boldsymbol{Z}^{(1)})^\top \boldsymbol{\theta}_{[\cdot 1]}$, and $\widetilde{T}^{(2)} : \mathbb{R}^n \to \mathbb{R}^{K_2}$ with $\widetilde{T}^{(2)}(\boldsymbol{\theta}_{[\cdot 2]}) = \{(\boldsymbol{Z}^{(2)})^\top \boldsymbol{Z}^{(2)}\}^{-1}(\boldsymbol{Z}^{(2)})^\top \boldsymbol{\theta}_{[\cdot 2]}$.

By the results in Theorem 2, for any $\kappa > 0$, there exists a finite $M_\kappa > 0$ and a finite $N_\kappa > 0$ such that for any $n > N_\kappa$,

$$P\left(\|((\widehat{\boldsymbol{\beta}}^{\text{or}})^\top, (\widehat{\boldsymbol{\theta}}^{\text{or}}_{[\cdot 1]})^\top, (\widehat{\boldsymbol{\theta}}^{\text{or}}_{[\cdot 2]})^\top) - ((\boldsymbol{\beta}^0)^\top, (\boldsymbol{\theta}^0_{[\cdot 1]})^\top, (\boldsymbol{\theta}^0_{[\cdot 2]})^\top)\|_\infty > M_\kappa \psi_n\right) < \kappa,$$

with

$$\psi_n = \max(K_1, K_2)\sqrt{n^{-1}\log n}\left(\sqrt{(p+K_1)K_2^2} + \sqrt{K_2}\right).$$

Let $\phi_n = M_\kappa \psi_n$. We now consider the neighborhood of $((\boldsymbol{\beta}^0)^\top, (\boldsymbol{\theta}^0_{[\cdot 1]})^\top, (\boldsymbol{\theta}^0_{[\cdot 2]})^\top)$, which is defined as

$$\mathcal{A} = \left\{(\boldsymbol{\beta}^\top, \boldsymbol{\theta}^\top_{[\cdot 1]}, \boldsymbol{\theta}^\top_{[\cdot 2]}) : \left\|(\boldsymbol{\beta}^\top, \boldsymbol{\theta}^\top_{[\cdot 1]}, \boldsymbol{\theta}^\top_{[\cdot 2]}) - ((\boldsymbol{\beta}^0)^\top, (\boldsymbol{\theta}^0_{[\cdot 1]})^\top, (\boldsymbol{\theta}^0_{[\cdot 2]})^\top)\right\|_\infty \le \phi_n\right\}.$$

Denote the event $\{((\widehat{\boldsymbol{\beta}}^{\text{or}})^\top, (\widehat{\boldsymbol{\theta}}^{\text{or}}_{[\cdot 1]})^\top, (\widehat{\boldsymbol{\theta}}^{\text{or}}_{[\cdot 2]})^\top) \in \mathcal{A}\}$ by $\mathcal{F}_1$, which satisfies $P(\mathcal{F}_1^c) < \kappa$ for $n > N_\kappa$ with $\mathcal{F}_1^c$ being the complement of $\mathcal{F}_1$. For any $\boldsymbol{\theta}_{[\cdot m]} \in \mathbb{R}^n$, let $\boldsymbol{\theta}^*_{[\cdot m]} =$

$T^{(m)}(\widetilde{T}^{(m)}(\boldsymbol{\theta}_{[\cdot m]}))$ for $m = 1, 2$.

We show that $((\widehat{\boldsymbol{\beta}}^{\text{or}})^\top, (\widehat{\boldsymbol{\theta}}^{\text{or}}_{[\cdot 1]})^\top, (\widehat{\boldsymbol{\theta}}^{\text{or}}_{[\cdot 2]})^\top)$ is a strictly local minimizer of the objective function with probability approaching 1 as $n \to \infty$ through two steps:

(i) On the event $\mathcal{F}_1$, it holds that $Q(\boldsymbol{\beta}, \boldsymbol{\Theta}^*) > Q(\widehat{\boldsymbol{\beta}}^{\text{or}}, \widehat{\boldsymbol{\Theta}}^{\text{or}})$ for any $(\boldsymbol{\beta}^\top, \boldsymbol{\theta}^\top_{[\cdot 1]}, \boldsymbol{\theta}^\top_{[\cdot 2]}) \in \mathcal{A}$ and $(\boldsymbol{\beta}^\top, (\boldsymbol{\theta}^*_{[\cdot 1]})^\top, (\boldsymbol{\theta}^*_{[\cdot 2]})^\top) \neq ((\widehat{\boldsymbol{\beta}}^{\text{or}})^\top, (\widehat{\boldsymbol{\theta}}^{\text{or}}_{[\cdot 1]})^\top, (\widehat{\boldsymbol{\theta}}^{\text{or}}_{[\cdot 2]})^\top)$.

(ii) There is an event $\mathcal{F}_2$ such that $P(\mathcal{F}_2^c) \to 0$ as $n \to \infty$. On $\mathcal{F}_1 \cap \mathcal{F}_2$, there is a neighborhood of $((\widehat{\boldsymbol{\beta}}^{\text{or}})^\top, (\widehat{\boldsymbol{\theta}}^{\text{or}}_{[\cdot 1]})^\top, (\widehat{\boldsymbol{\theta}}^{\text{or}}_{[\cdot 2]})^\top)$, denoted by $\mathcal{A}_n$, such that $Q(\boldsymbol{\beta}, \boldsymbol{\Theta}) \geq Q(\boldsymbol{\beta}, \boldsymbol{\Theta}^*)$ for any $(\boldsymbol{\beta}^\top, \boldsymbol{\theta}^\top_{[\cdot 1]}, \boldsymbol{\theta}^\top_{[\cdot 2]}) \in \mathcal{A}_n \cap \mathcal{A}$ when n is sufficiently large.

By the results in (i) and (ii), we have $Q(\boldsymbol{\beta}, \boldsymbol{\Theta}) > Q(\widehat{\boldsymbol{\beta}}^{\text{or}}, \widehat{\boldsymbol{\Theta}}^{\text{or}})$ for $(\boldsymbol{\beta}^\top, \boldsymbol{\theta}^\top_{[\cdot 1]}, \boldsymbol{\theta}^\top_{[\cdot 2]}) \in \mathcal{A}_n \cap \mathcal{A}$ and $(\boldsymbol{\beta}^\top, (\boldsymbol{\theta}^*_{[\cdot 1]})^\top, (\boldsymbol{\theta}^*_{[\cdot 2]})^\top) \neq ((\widehat{\boldsymbol{\beta}}^{\text{or}})^\top, (\widehat{\boldsymbol{\theta}}^{\text{or}}_{[\cdot 1]})^\top, (\widehat{\boldsymbol{\theta}}^{\text{or}}_{[\cdot 2]})^\top)$, thus $((\widehat{\boldsymbol{\beta}}^{\text{or}})^\top, (\widehat{\boldsymbol{\theta}}^{\text{or}}_{[\cdot 1]})^\top, (\widehat{\boldsymbol{\theta}}^{\text{or}}_{[\cdot 2]})^\top)$ is a strictly local minimizer of $Q(\boldsymbol{\beta}, \boldsymbol{\Theta})$ on the event $\mathcal{F}_1 \cap \mathcal{F}_2$ with $P(\mathcal{F}_1 \cap \mathcal{F}_2) \to 1$ as $n \to \infty$.

First, we prove the result in (i). We begin with showing $\sum_{m=1}^{2} \sum_{1 \leq i < j \leq n} p(|\theta^*_{im} - \theta^*_{jm}|, \lambda_m, \gamma_m) = C_n$ for any $(\boldsymbol{\beta}^\top, \boldsymbol{\theta}^\top_{[\cdot 1]}, \boldsymbol{\theta}^\top_{[\cdot 2]}) \in \mathcal{A}$, where C_n is a constant that does not depend on $(\boldsymbol{\beta}^\top, \boldsymbol{\theta}^\top_{[\cdot 1]}, \boldsymbol{\theta}^\top_{[\cdot 2]})$. Note that for any $(\boldsymbol{\beta}^\top, \boldsymbol{\theta}^\top_{[\cdot 1]}, \boldsymbol{\theta}^\top_{[\cdot 2]}) \in \mathcal{A}$,

$$
\begin{aligned}
&|\theta^*_{im} - \theta^*_{jm}| + \|\boldsymbol{\theta}^*_{[\cdot m]} - \boldsymbol{\theta}^0_{[\cdot m]}\|_\infty \geq |\theta^*_{im} - \theta^*_{jm}| + |\theta^*_{jm} - \theta^0_{jm}| \\
\geq& |\theta^*_{im} - \theta^0_{jm}| = |\theta^*_{im} - \theta^0_{jm} + \theta^0_{im} - \theta^0_{im}| \geq |\theta^0_{im} - \theta^0_{jm}| - |\theta^*_{im} - \theta^0_{im}| \qquad \text{(S3.13)} \\
\geq& |\theta^0_{im} - \theta^0_{jm}| - \|\boldsymbol{\theta}^*_{[\cdot m]} - \boldsymbol{\theta}^0_{[\cdot m]}\|_\infty,
\end{aligned}
$$

and

$$\begin{aligned}\|\boldsymbol{\theta}^*_{[\cdot m]}-\boldsymbol{\theta}^0_{[\cdot m]}\|_\infty =&\sup_k\left|\sum_{i\in\mathcal{G}_k^{(m)}}(\theta_{im}-\theta^0_{im})/|\mathcal{G}_k|\right| \le \sup_k \sup_{i\in\mathcal{G}_k^{(m)}}|\theta_{im}-\theta^0_{im}| \\ =&\|\boldsymbol{\theta}_{[\cdot m]}-\boldsymbol{\theta}^0_{[\cdot m]}\|_\infty \le \phi_n.\end{aligned} \tag{S3.14}$$

By (S3.13) and (S3.14), when $i\in\mathcal{G}_k^{(m)}, j\in\mathcal{G}_{k'}^{(m)}$ for some $k\neq k'$, we have

$$|\theta^*_{im}-\theta^*_{jm}|\ge|\theta^0_{im}-\theta^0_{jm}|-2\|\boldsymbol{\theta}^*_{[\cdot m]}-\boldsymbol{\theta}^0_{[\cdot m]}\|_\infty\ge b_n-2\phi_n>a\lambda_m,$$

where the last inequality follows from the assumption that $b_n>a\lambda_m\gg\psi_n$. Hence, by Condition (C4), we have $\sum_{m=1}^2\sum_{1\le i<j\le n}p(|\theta^*_{im}-\theta^*_{jm}|,\lambda_m,\gamma_m)=C_n$, and $Q(\boldsymbol{\beta},\boldsymbol{\Theta})=-L(\boldsymbol{\beta},\boldsymbol{\Theta})+C_n$, for all $(\boldsymbol{\beta}^\top,\boldsymbol{\theta}^\top_{[\cdot 1]},\boldsymbol{\theta}^\top_{[\cdot 2]})\in\mathcal{A}$. Because $((\widehat{\boldsymbol{\beta}}^{\rm or})^\top,(\widehat{\boldsymbol{\theta}}^{\rm or}_{[\cdot 1]})^\top,(\widehat{\boldsymbol{\theta}}^{\rm or}_{[\cdot 2]})^\top)$ is a local minimizer of $-L(\boldsymbol{\beta},\boldsymbol{\Theta})$, we have $Q(\boldsymbol{\beta},\boldsymbol{\Theta}^*)>Q(\widehat{\boldsymbol{\beta}}^{\rm or},\widehat{\boldsymbol{\Theta}}^{\rm or})$ for any $(\boldsymbol{\beta}^\top,\boldsymbol{\theta}^\top_{[\cdot 1]},\boldsymbol{\theta}^\top_{[\cdot 2]})\in\mathcal{A}$ and $(\boldsymbol{\beta}^\top,(\boldsymbol{\theta}^*_{[\cdot 1]})^\top,(\boldsymbol{\theta}^*_{[\cdot 2]})^\top)\neq((\widehat{\boldsymbol{\beta}}^{\rm or})^\top,(\widehat{\boldsymbol{\theta}}^{\rm or}_{[\cdot 1]})^\top,(\widehat{\boldsymbol{\theta}}^{\rm or}_{[\cdot 2]})^\top)$.

Next, we prove the result in (ii). Based on the results in Theorem 2 and Condition (C2), we have $\widehat{\theta}^{\rm or}_{i2}>\tau_{\min}/2$ for sufficiently large n. In addition, there exists a positive sequence $t_n=o(1)$ such that $\theta_{i2}>0$ for any $(\boldsymbol{\theta}^\top_{[\cdot 1]},\boldsymbol{\theta}^\top_{[\cdot 2]})\in\mathcal{A}_n$ with

$$\mathcal{A}_n=\left\{(\boldsymbol{\theta}^\top_{[\cdot 1]},\boldsymbol{\theta}^\top_{[\cdot 2]}):\left\|(\boldsymbol{\theta}^\top_{[\cdot 1]},\boldsymbol{\theta}^\top_{[\cdot 2]})-((\widehat{\boldsymbol{\theta}}^{\rm or}_{[\cdot 1]})^\top,(\widehat{\boldsymbol{\theta}}^{\rm or}_{[\cdot 2]})^\top)\right\|_\infty\le t_n\right\}.$$

For $(\boldsymbol{\beta}^\top,\boldsymbol{\theta}^\top_{[\cdot 1]},\boldsymbol{\theta}^\top_{[\cdot 2]})\in\mathcal{A}_n\cap\mathcal{A}$, by Taylor's expansion, we have

$$Q(\boldsymbol{\beta},\boldsymbol{\Theta})-Q(\boldsymbol{\beta},\boldsymbol{\Theta}^*)=\frac{\partial Q(\boldsymbol{\beta},\widetilde{\boldsymbol{\Theta}})}{\partial(\boldsymbol{\theta}^\top_{[\cdot 1]},\boldsymbol{\theta}^\top_{[\cdot 2]})}((\boldsymbol{\theta}^\top_{[\cdot 1]},\boldsymbol{\theta}^\top_{[\cdot 2]})-((\boldsymbol{\theta}^*_{[\cdot 1]})^\top,(\boldsymbol{\theta}^*_{[\cdot 2]})^\top))^\top,$$

with $\widetilde{\boldsymbol{\theta}}_{[\cdot m]} = \zeta\boldsymbol{\theta}_{[\cdot m]} + (1-\zeta)\boldsymbol{\theta}^*_{[\cdot m]}$ for some $\zeta \in (0,1)$ and $m = 1, 2$. By (S3.14), it holds that

$$\|\widetilde{\boldsymbol{\theta}}_{[\cdot m]} - \boldsymbol{\theta}^0_{[\cdot m]}\|_\infty \leq \phi_n. \tag{S3.15}$$

Define

$$\begin{aligned}
\Gamma_1 =& \frac{\partial L(\boldsymbol{\beta}, \widetilde{\boldsymbol{\Theta}})}{\partial(\boldsymbol{\theta}^\top_{[\cdot 1]}, \boldsymbol{\theta}^\top_{[\cdot 2]})}((\boldsymbol{\theta}^\top_{[\cdot 1]}, \boldsymbol{\theta}^\top_{[\cdot 2]}) - ((\boldsymbol{\theta}^*_{[\cdot 1]})^\top, (\boldsymbol{\theta}^*_{[\cdot 2]})^\top))^\top, \\
\Gamma_2 =& \frac{\partial \sum_{m=1}^2 \sum_{1\leq i<j\leq n} p(|\widetilde{\theta}_{im} - \widetilde{\theta}_{jm}|, \lambda_m, \gamma_m)}{\partial(\boldsymbol{\theta}^\top_{[\cdot 1]}, \boldsymbol{\theta}^\top_{[\cdot 2]})}((\boldsymbol{\theta}^\top_{[\cdot 1]}, \boldsymbol{\theta}^\top_{[\cdot 2]}) - ((\boldsymbol{\theta}^*_{[\cdot 1]})^\top, (\boldsymbol{\theta}^*_{[\cdot 2]})^\top))^\top.
\end{aligned}$$

We first consider Γ_1, which can be computed as

$$\begin{aligned}
\Gamma_1 =& (\boldsymbol{w}_1^\top, \boldsymbol{w}_2^\top)(\boldsymbol{\theta}^\top_{[\cdot 1]} - (\boldsymbol{\theta}^*_{[\cdot 1]})^\top, \boldsymbol{\theta}^\top_{[\cdot 2]} - (\boldsymbol{\theta}^*_{[\cdot 2]})^\top)^\top \\
=& \sum_{k=1}^{K_1} \sum_{i\in\mathcal{G}_k^{(1)}} w_{1i}\left(\theta_{i1} - \frac{\sum_{j\in\mathcal{G}_k^{(1)}} \theta_{j1}}{|\mathcal{G}_k^{(1)}|}\right) + \sum_{k'=1}^{K_2} \sum_{i\in\mathcal{G}_{k'}^{(2)}} w_{2i}\left(\theta_{i2} - \frac{\sum_{j\in\mathcal{G}_{k'}^{(2)}} \theta_{j2}}{|\mathcal{G}_{k'}^{(2)}|}\right) \\
=& \sum_{k=1}^{K_1} \sum_{i,j\in\mathcal{G}_k^{(1)}} \frac{w_{1i}(\theta_{i1} - \theta_{j1})}{|\mathcal{G}_k^{(1)}|} + \sum_{k'=1}^{K_2} \sum_{i,j\in\mathcal{G}_{k'}^{(2)}} \frac{w_{2i}(\theta_{i2} - \theta_{j2})}{|\mathcal{G}_{k'}^{(2)}|} \\
=& \sum_{k=1}^{K_1} \sum_{i,j\in\mathcal{G}_k^{(1)}, i<j} \frac{(w_{1j} - w_{1i})(\theta_{j1} - \theta_{i1})}{|\mathcal{G}_k^{(1)}|} + \sum_{k'=1}^{K_2} \sum_{i,j\in\mathcal{G}_{k'}^{(2)}, i<j} \frac{(w_{2j} - w_{2i})(\theta_{j2} - \theta_{i2})}{|\mathcal{G}_{k'}^{(2)}|} \\
\leq& |\mathcal{G}_{\min}^{(1)}|^{-1} \sum_{k=1}^{K_1} \sum_{i,j\in\mathcal{G}_k^{(1)}, i<j} |w_{1j} - w_{1i}||\theta_{j1} - \theta_{i1}| + |\mathcal{G}_{\min}^{(2)}|^{-1} \sum_{k'=1}^{K_2} \sum_{i,j\in\mathcal{G}_{k'}^{(2)}, i<j} |w_{2j} - w_{2i}||\theta_{j2} - \theta_{i2}| \\
\leq& \max_{m=1,2}\left(|\mathcal{G}_{\min}^{(m)}|^{-1} \max_{i,j} |w_{mj} - w_{mi}|\right) \sum_{m=1}^2 \sum_{i<j, \theta^*_{jm}=\theta^*_{im}} |\theta_{jm} - \theta_{im}|,
\end{aligned}$$

with

$$\begin{aligned}
\boldsymbol{w}_1 &= \Big(\widetilde{\theta}_{12}(y_1 - \boldsymbol{\beta}^\top \boldsymbol{x}_1 - \widetilde{\theta}_{11}), \ldots, \widetilde{\theta}_{n2}(y_n - \boldsymbol{\beta}^\top \boldsymbol{x}_n - \widetilde{\theta}_{n1})\Big)^\top, \\
\boldsymbol{w}_2 &= \left(\frac{1}{2\widetilde{\theta}_{12}} - (y_1 - \boldsymbol{\beta}^\top \boldsymbol{x}_1 - \widetilde{\theta}_{11})^2/2, \ldots, \frac{1}{2\widetilde{\theta}_{n2}} - (y_n - \boldsymbol{\beta}^\top \boldsymbol{x}_n - \widetilde{\theta}_{n1})^2/2\right)^\top.
\end{aligned}$$

Recall that $\widetilde{\boldsymbol{\epsilon}} = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^0 - \boldsymbol{\theta}^0_{[\cdot 1]}$. One can verify that for $t_n = o(1)$,

$$\begin{aligned}
&\max_{i,j} |w_{1j} - w_{1i}| \le 2\|\boldsymbol{w}_1\|_\infty \le 2\|\widetilde{\boldsymbol{\theta}}_{[\cdot 2]}\|_\infty \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \widetilde{\boldsymbol{\theta}}_{[\cdot 1]}\|_\infty \\
=&2\|\widetilde{\boldsymbol{\theta}}_{[\cdot 2]} - \boldsymbol{\theta}^0_{[\cdot 2]} + \boldsymbol{\theta}^0_{[\cdot 2]}\|_\infty \|\boldsymbol{y} - \boldsymbol{X}(\boldsymbol{\beta} - \boldsymbol{\beta}^0 + \boldsymbol{\beta}^0) - \widetilde{\boldsymbol{\theta}}_{[\cdot 1]} - \boldsymbol{\theta}^0_{[\cdot 1]} + \boldsymbol{\theta}^0_{[\cdot 1]}\|_\infty \\
\le&2\|\widetilde{\boldsymbol{\theta}}_{[\cdot 2]} - \boldsymbol{\theta}^0_{[\cdot 2]} + \boldsymbol{\theta}^0_{[\cdot 2]}\|_\infty \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^0 - \boldsymbol{\theta}^0_{[\cdot 1]} - \boldsymbol{X}(\boldsymbol{\beta} - \boldsymbol{\beta}^0) + \boldsymbol{\theta}^0_{[\cdot 1]} - \widetilde{\boldsymbol{\theta}}_{[\cdot 1]}\|_\infty \\
\le&2(\|\widetilde{\boldsymbol{\theta}}_{[\cdot 2]} - \boldsymbol{\theta}^0_{[\cdot 2]}\|_\infty + \|\boldsymbol{\theta}^0_{[\cdot 2]}\|_\infty)(\|\widetilde{\boldsymbol{\epsilon}} - \boldsymbol{X}(\boldsymbol{\beta} - \boldsymbol{\beta}^0) + \boldsymbol{\theta}^0_{[\cdot 1]} - \widetilde{\boldsymbol{\theta}}_{[\cdot 1]}\|_\infty) \\
\le&2(\|\widetilde{\boldsymbol{\theta}}_{[\cdot 2]} - \boldsymbol{\theta}^0_{[\cdot 2]}\|_\infty + \|\boldsymbol{\theta}^0_{[\cdot 2]}\|_\infty)(\|\widetilde{\boldsymbol{\epsilon}}\|_\infty + \max_{1\le i\le n}\sum_{j=1}^p |x_{ij}|\|\boldsymbol{\beta} - \boldsymbol{\beta}^0\|_\infty + \|\boldsymbol{\theta}^0_{[\cdot 1]} - \widetilde{\boldsymbol{\theta}}_{[\cdot 1]}\|_\infty) \\
\le&2(\phi_n + \tau_{\max})(\|\widetilde{\boldsymbol{\epsilon}}\|_\infty + Mp\phi_n + \phi_n),
\end{aligned}$$

where the last inequality is derived by (S3.15) and Condition (C1). Similarly, we have

$$\begin{aligned}
&\max_{i,j} |w_{2j} - w_{2i}| \le 2\|\boldsymbol{w}_2\|_\infty \le 2\max_{1\le i\le n}\left|\frac{1}{2\widetilde{\theta}_{i2}} - (y_i - \boldsymbol{\beta}^\top \boldsymbol{x}_i - \widetilde{\theta}_{i1})^2/2\right| \\
&\le (\min_{1\le i\le n}|\widetilde{\theta}_{i2}|)^{-1} \le (\min_{1\le i\le n}|\widetilde{\theta}_{i2} - \theta^0_{i2} + \theta^0_{i2}|)^{-1} \le (\min_{1\le i\le n}|\theta^0_{i2}| - \max_{1\le i\le n}|\widetilde{\theta}_{i2} - \theta^0_{i2}|)^{-1} \\
&\le (\tau_{\min} - \phi_n)^{-1},
\end{aligned}$$

where the third inequality is derived from the fact that $\widetilde{\theta}_{i2} > 0$. Noting that $\{\widetilde{\epsilon}_i, i =$

$1,\ldots,n\}$ is sub-Gaussian with $\tau_{\min}^{-1}$, we have

$$P\left(\|\boldsymbol{\epsilon}\|_\infty > (4\tau_{\min}^{-1}\log n)^{1/2}\right) \leq \sum_{i=1}^{n} P(|\epsilon_i| > (4\tau_{\min}^{-1}\log n)^{1/2}) \leq 2n^{-1}.$$

Thus, there exists an event $\mathcal{F}_2$ such that $P(\mathcal{F}_2^c) \leq 2n^{-1}$, and on the event $\mathcal{F}_2$,

$$\begin{aligned}\max_{i,j}|w_{1j}-w_{1i}| \leq& 2(\phi_n+\tau_{\max})((4\tau_{\min}^{-1}\log n)^{1/2}+Mp\phi_n+\phi_n),\\ \max_{i,j}|w_{2j}-w_{2i}| \leq& (\tau_{\min}-\phi_n)^{-1}.\end{aligned}$$

Next, we consider Γ_2. We have that for $m=1,2$,

$$\begin{aligned}&\frac{\partial\sum_{1\leq i<j\leq n} p(|\widetilde{\theta}_{im}-\widetilde{\theta}_{jm}|,\lambda_m,\gamma_m)}{\partial\theta_{lm}}\\ =&\sum_{j>l}\frac{\partial p(|\widetilde{\theta}_{lm}-\widetilde{\theta}_{jm}|,\lambda_m,\gamma_m)}{\partial\theta_{lm}}+\sum_{j<l}\frac{\partial p(|\widetilde{\theta}_{jm}-\widetilde{\theta}_{lm}|,\lambda_m,\gamma_m)}{\partial\theta_{lm}}\\ =&\sum_{j>l}p'(|\widetilde{\theta}_{lm}-\widetilde{\theta}_{jm}|,\lambda_m,\gamma_m)\text{sign}(\widetilde{\theta}_{lm}-\widetilde{\theta}_{jm})-\sum_{j<l}p'(|\widetilde{\theta}_{jm}-\widetilde{\theta}_{lm}|,\lambda_m,\gamma_m)\text{sign}(\widetilde{\theta}_{jm}-\widetilde{\theta}_{lm})\\ =&\sum_{j>l}p'(|\widetilde{\theta}_{lm}-\widetilde{\theta}_{jm}|,\lambda_m,\gamma_m)\text{sign}(\widetilde{\theta}_{lm}-\widetilde{\theta}_{jm})+\sum_{j<l}p'(|\widetilde{\theta}_{lm}-\widetilde{\theta}_{jm}|,\lambda_m,\gamma_m)\text{sign}(\widetilde{\theta}_{lm}-\widetilde{\theta}_{jm}).\end{aligned}$$

Therefore, it holds that

$$\begin{aligned}
&\frac{\partial \sum_{1\le i<j\le n} p(|\widetilde{\theta}_{im}-\widetilde{\theta}_{jm}|,\lambda_m,\gamma_m)}{\partial \theta_{lm}}(\theta_{lm}-\theta^*_{lm})\\
=&\sum_{j>l} p'(|\widetilde{\theta}_{lm}-\widetilde{\theta}_{jm}|,\lambda_m,\gamma_m)\mathrm{sign}(\widetilde{\theta}_{lm}-\widetilde{\theta}_{jm})(\theta_{lm}-\theta^*_{lm})\\
&+\sum_{j<l} p'(|\widetilde{\theta}_{lm}-\widetilde{\theta}_{jm}|,\lambda_m,\gamma_m)\mathrm{sign}(\widetilde{\theta}_{lm}-\widetilde{\theta}_{jm})(\theta_{lm}-\theta^*_{lm})\\
=&\sum_{j>l} p'(|\widetilde{\theta}_{lm}-\widetilde{\theta}_{jm}|,\lambda_m,\gamma_m)\mathrm{sign}(\widetilde{\theta}_{lm}-\widetilde{\theta}_{jm})(\theta_{lm}-\theta^*_{lm})\\
&-\sum_{l<j} p'(|\widetilde{\theta}_{lm}-\widetilde{\theta}_{jm}|,\lambda_m,\gamma_m)\mathrm{sign}(\widetilde{\theta}_{lm}-\widetilde{\theta}_{jm})(\theta_{jm}-\theta^*_{jm})\\
=&\sum_{j>l} p'(|\widetilde{\theta}_{lm}-\widetilde{\theta}_{jm}|,\lambda_m,\gamma_m)\mathrm{sign}(\widetilde{\theta}_{lm}-\widetilde{\theta}_{jm})\left\{\theta_{lm}-\theta^*_{lm}-(\theta_{jm}-\theta^*_{jm})\right\}.
\end{aligned}$$

Thus, we have

$$\Gamma_2=\sum_{m=1}^{2}\sum_{l=1}^{n}\sum_{j>l} p'(|\widetilde{\theta}_{lm}-\widetilde{\theta}_{jm}|,\lambda_m,\gamma_m)\mathrm{sign}(\widetilde{\theta}_{lm}-\widetilde{\theta}_{jm})\left\{\theta_{lm}-\theta^*_{lm}-(\theta_{jm}-\theta^*_{jm})\right\}.$$

For some m, if there exists $k\neq k'$ such that $l\in\mathcal{G}_k^{(m)}, j\in\mathcal{G}_{k'}^{(m)}$, by (S3.15) and the same line of (S3.13), we have

$$|\widetilde{\theta}_{lm}-\widetilde{\theta}_{jm}|\ge|\theta^0_{lm}-\theta^0_{jm}|-2\|\widetilde{\boldsymbol{\theta}}_{[\cdot m]}-\boldsymbol{\theta}^0_{[\cdot m]}\|_\infty\ge b_n-2\phi_n>a\lambda_m,$$

and thus $p'(|\widetilde{\theta}_{lm} - \widetilde{\theta}_{jm}|, \lambda_m, \gamma_m) = 0$. Therefore, it holds that

$$\begin{aligned}\Gamma_2 =& \sum_{m=1}^{2} \sum_{\{j>l, \theta^*_{lm}=\theta^*_{jm}\}} p'(|\widetilde{\theta}_{lm} - \widetilde{\theta}_{jm}|, \lambda_m, \gamma_m)\mathrm{sign}(\widetilde{\theta}_{lm} - \widetilde{\theta}_{jm})\,(\theta_{lm} - \theta_{jm})\\ =& \sum_{m=1}^{2} \sum_{\{j>l, \theta^*_{lm}=\theta^*_{jm}\}} p'(|\widetilde{\theta}_{lm} - \widetilde{\theta}_{jm}|, \lambda_m, \gamma_m)\,|\theta_{lm} - \theta_{jm}|\,,\end{aligned}$$

where the last equality is derived from the fact that $\widetilde{\theta}_{lm} - \widetilde{\theta}_{jm}$ has the same sign as $\theta_{lm} - \theta_{jm}$ for $j > l$ with $\theta^*_{lm} = \theta^*_{jm}$. Furthermore, by the same line of (S3.14), we have $\|\boldsymbol{\theta}^*_{[\cdot m]} - \widehat{\boldsymbol{\theta}}^{\mathrm{or}}_{[\cdot m]}\|_\infty \leq \|\boldsymbol{\theta}_{[\cdot m]} - \widehat{\boldsymbol{\theta}}^{\mathrm{or}}_{[\cdot m]}\|_\infty$. Then one can verify that for $j > l$ with $\theta^*_{lm} = \theta^*_{jm}$,

$$\begin{aligned}&|\widetilde{\theta}_{lm} - \widetilde{\theta}_{jm}| = |\widetilde{\theta}_{lm} - \widetilde{\theta}_{jm} + \theta^*_{jm} - \theta^*_{lm}| \leq |\widetilde{\theta}_{lm} - \theta^*_{lm}| + |\widetilde{\theta}_{jm} - \theta^*_{jm}| \leq 2\|\widetilde{\boldsymbol{\theta}}_{[\cdot m]} - \boldsymbol{\theta}^*_{[\cdot m]}\|_\infty\\ \leq& 2\|\boldsymbol{\theta}_{[\cdot m]} - \boldsymbol{\theta}^*_{[\cdot m]}\|_\infty \leq 2(\|\boldsymbol{\theta}_{[\cdot m]} - \widehat{\boldsymbol{\theta}}^{\mathrm{or}}_{[\cdot m]}\|_\infty + \|\boldsymbol{\theta}^*_{[\cdot m]} - \widehat{\boldsymbol{\theta}}^{\mathrm{or}}_{[\cdot m]}\|_\infty) \leq 4\|\boldsymbol{\theta}_{[\cdot m]} - \widehat{\boldsymbol{\theta}}^{\mathrm{or}}_{[\cdot m]}\|_\infty \leq 4t_n.\end{aligned} \tag{S3.16}$$

As a result, for $t_n = o(1)$, by the concavity of the penalty function and Condition (C4), we have

$$\Gamma_2 \geq \sum_{m=1}^{2} \sum_{\{j>l, \theta^*_{lm}=\theta^*_{jm}\}} \lambda_m |\theta_{lm} - \theta_{jm}|.$$

Note that by Condition (C3), we have $|\mathcal{G}^{(m)}_{\min}| = O(n/K_m)$. Combining with the assumption $\max\{K_1, K_2\}\sqrt{p + K_1}K_2 = o(\sqrt{n(\log n)^{-1}})$, it holds that $|\mathcal{G}^{(m)}_{\min}|^{-1}p = o(1)$. As $\lambda_m \gg \psi_n \gg |\mathcal{G}^{(m)}_{\min}|^{-1}p\phi_n$ and $\lambda \gg \psi_n \gg |\mathcal{G}^{(m)}_{\min}|^{-1}(\log n)^{1/2}$, on the event $\mathcal{F}_2$, we have for $(\boldsymbol{\beta}^\top, \boldsymbol{\theta}_{[\cdot 1]}^\top, \boldsymbol{\theta}_{[\cdot 2]}^\top) \in \mathcal{A}_n \cap \mathcal{A}$,

$$\begin{aligned}
&Q(\boldsymbol{\beta}, \boldsymbol{\Theta}) - Q(\boldsymbol{\beta}, \boldsymbol{\Theta}^*) = \Gamma_2 - \Gamma_1 \\
\geq &\sum_{m=1}^{2} \sum_{\{j>l, \theta_{lm}^* = \theta_{jm}^*\}} \left\{ \lambda_m - \max_{m=1,2} \left(|\mathcal{G}_{\min}^{(m)}|^{-1} \max_{i,j} |w_{mj} - w_{mi}| \right) \right\} |\theta_{lm} - \theta_{jm}| > 0,
\end{aligned}$$

for sufficiently large n. This completes the proof. □

3.3 Proof of Theorem 4

Proof. The proof of Conclusion (1) in Theorem 4 is the same as that of Theorem 2 by letting $K_1 = K_2 = 1$, which is omitted here. The proof of Conclusion (2), as detailed below, follows a similar procedure as that of Theorem 3.

Let $T^{(m)} : \mathbb{R} \to \mathcal{I}_m$ for $m = 1, 2$ be the mapping such that $T^{(1)}(\mu) = \mathbf{1}_n \mu$ and $T^{(2)}(\tau) = \mathbf{1}_n \tau$. Let $\widetilde{T}^{(m)} : \mathbb{R}^n \to \mathbb{R}$ for $m = 1, 2$ be the mapping such that $\widetilde{T}^{(1)}(\boldsymbol{\theta}_{[\cdot 1]}) = n^{-1} \sum_{i=1}^n \theta_{i1}$ and $\widetilde{T}^{(2)}(\boldsymbol{\theta}_{[\cdot 2]}) = n^{-1} \sum_{i=1}^n \theta_{i2}$.

Consider the neighborhood of $((\boldsymbol{\beta}^0)^\top, (\boldsymbol{\theta}_{[\cdot 1]}^0)^\top, (\boldsymbol{\theta}_{[\cdot 2]}^0)^\top)$, which is defined as

$$\mathcal{A} = \left\{ (\boldsymbol{\beta}^\top, \boldsymbol{\theta}_{[\cdot 1]}^\top, \boldsymbol{\theta}_{[\cdot 2]}^\top) : \left\| (\boldsymbol{\beta}^\top, \boldsymbol{\theta}_{[\cdot 1]}^\top, \boldsymbol{\theta}_{[\cdot 2]}^\top) - ((\boldsymbol{\beta}^0)^\top, (\boldsymbol{\theta}_{[\cdot 1]}^0)^\top, (\boldsymbol{\theta}_{[\cdot 2]}^0)^\top) \right\|_\infty \leq M_\kappa \psi_n \right\},$$

with $\psi_n = \sqrt{(p+1)n^{-1} \log n} + \sqrt{n^{-1} \log n}$. Let $\phi_n = M_\kappa \psi_n$. By Conclusion (1) of Theorem 4, the event $\mathcal{F}_1 = \{((\widehat{\boldsymbol{\beta}}^{\text{or}})^\top, (\widehat{\boldsymbol{\theta}}_{[\cdot 1]}^{\text{or}})^\top, (\widehat{\boldsymbol{\theta}}_{[\cdot 2]}^{\text{or}})^\top) \in \mathcal{A}\}$ satisfies $P(\mathcal{F}_1^c) < \kappa$ for $n > N_\kappa$. For any $\boldsymbol{\theta}_{[\cdot m]} \in \mathbb{R}^n$, let $\boldsymbol{\theta}_{[\cdot m]}^* = T^{(m)}(\widetilde{T}^{(m)}(\boldsymbol{\theta}_{[\cdot m]})), m = 1, 2$.

We prove the conclusion through the same two steps as those in Theorem 3. We first prove (i). Note that

$$\sum_{m=1}^{2}\sum_{1\leq i<j\leq n} p(|\theta_{im}^{*}-\theta_{jm}^{*}|,\lambda_m,\gamma_m)=\sum_{m=1}^{2}\sum_{1\leq i<j\leq n} p(|\widehat{\theta}_{im}^{\mathrm{or}}-\widehat{\theta}_{jm}^{\mathrm{or}}|,\lambda_m,\gamma_m)=0.$$

By the definition of $((\widehat{\boldsymbol{\beta}}^{\mathrm{or}})^{\top},(\widehat{\boldsymbol{\theta}}_{[\cdot 1]}^{\mathrm{or}})^{\top},(\widehat{\boldsymbol{\theta}}_{[\cdot 2]}^{\mathrm{or}})^{\top})$, we have $Q(\boldsymbol{\beta},\boldsymbol{\Theta}^{*})>Q(\boldsymbol{\beta},\widehat{\boldsymbol{\Theta}}^{\mathrm{or}})$ for any $(\boldsymbol{\beta}^{\top},\boldsymbol{\theta}_{[\cdot 1]}^{\top},\boldsymbol{\theta}_{[\cdot 2]}^{\top})\in\mathcal{A}$ and $(\boldsymbol{\beta}^{\top},(\boldsymbol{\theta}_{[\cdot 1]}^{*})^{\top},(\boldsymbol{\theta}_{[\cdot 2]}^{*})^{\top})\neq((\widehat{\boldsymbol{\beta}}^{\mathrm{or}})^{\top},(\widehat{\boldsymbol{\theta}}_{[\cdot 1]}^{\mathrm{or}})^{\top},(\widehat{\boldsymbol{\theta}}_{[\cdot 2]}^{\mathrm{or}})^{\top})$.

Next, we prove the result in (ii). Based on the results in Conclusion (1), we have $\widehat{\theta}_{i2}^{\mathrm{or}}>\tau^{0}/2$ for sufficiently large n. In addition, there exists a positive sequence $t_n=o(1)$ such that $\theta_{i2}>0$ for any $(\boldsymbol{\theta}_{[\cdot 1]}^{\top},\boldsymbol{\theta}_{[\cdot 2]}^{\top})\in\mathcal{A}_n$ with

$$\mathcal{A}_n=\left\{(\boldsymbol{\theta}_{[\cdot 1]}^{\top},\boldsymbol{\theta}_{[\cdot 2]}^{\top}):\left\|(\boldsymbol{\theta}_{[\cdot 1]}^{\top},\boldsymbol{\theta}_{[\cdot 2]}^{\top})-((\widehat{\boldsymbol{\theta}}_{[\cdot 1]}^{\mathrm{or}})^{\top},(\widehat{\boldsymbol{\theta}}_{[\cdot 2]}^{\mathrm{or}})^{\top})\right\|_{\infty}\leq t_n\right\}.$$

For $(\boldsymbol{\beta}^{\top},\boldsymbol{\theta}_{[\cdot 1]}^{\top},\boldsymbol{\theta}_{[\cdot 2]}^{\top})\in\mathcal{A}_n\cap\mathcal{A}$, by Taylor's expansion, we have

$$Q(\boldsymbol{\beta},\boldsymbol{\Theta})-Q(\boldsymbol{\beta},\boldsymbol{\Theta}^{*})=\Gamma_1+\Gamma_2,$$

with Γ_1 and Γ_2 defined in Theorem 3. Note that Γ_1 can be computed as

$$\begin{aligned}\Gamma_1=&(\boldsymbol{w}_1^{\top},\boldsymbol{w}_2^{\top})(\boldsymbol{\theta}_{[\cdot 1]}^{\top}-(\boldsymbol{\theta}_{[\cdot 1]}^{*})^{\top},\boldsymbol{\theta}_{[\cdot 2]}^{\top}-(\boldsymbol{\theta}_{[\cdot 2]}^{*})^{\top})^{\top}\\=&n^{-1}\left\{\sum_{i=1}^{n}\sum_{j=1}^{n}w_{1i}\left(\theta_{i1}-\theta_{j1}\right)+\sum_{i=1}^{n}\sum_{j=1}^{n}w_{2i}\left(\theta_{i2}-\theta_{j2}\right)\right\}\\\leq&n^{-1}\left(\sum_{i<j}|w_{1j}-w_{1i}||\theta_{j1}-\theta_{i1}|+\sum_{i<j}|w_{2j}-w_{2i}||\theta_{j2}-\theta_{i2}|\right).\end{aligned}$$

By the proof of Theorem 3, there exists an event $\mathcal{F}_2$ such that $P(\mathcal{F}_2^{c})\to 0$, and on the event $\mathcal{F}_2$, we have

$$\begin{aligned}\max_{i,j}|w_{1j}-w_{1i}| \leq & 2(\phi_n+\tau^0)\{(4(\tau^0)^{-1}\log n)^{1/2}+Mp\phi_n+\phi_n\},\\ \max_{i,j}|w_{2j}-w_{2i}| \leq & (\tau^0-\phi_n)^{-1}.\end{aligned}$$

Next, we consider Γ_2. It holds that for $t_n = o(1)$,

$$\begin{aligned}\Gamma_2 = & \sum_{m=1}^{2}\sum_{j>l} p'(|\widetilde{\theta}_{lm}-\widetilde{\theta}_{jm}|,\lambda_m,\gamma_m)\mathrm{sign}(\widetilde{\theta}_{lm}-\widetilde{\theta}_{jm})\left[\theta_{lm}-\theta^*_{lm}-(\theta_{jm}-\theta^*_{jm})\right]\\ = & \sum_{m=1}^{2}\sum_{j>l} p'(|\widetilde{\theta}_{lm}-\widetilde{\theta}_{jm}|,\lambda_m,\gamma_m)\,|\theta_{lm}-\theta_{jm}| \geq \sum_{m=1}^{2}\sum_{j>l}\lambda_m|\theta_{lm}-\theta_{jm}|.\end{aligned}$$

By the assumption $p = o(n(\log n)^{-1})$, it holds that $n^{-1}p = o(1)$. As $\lambda_m \gg \psi_n \gg n^{-1}p\phi_n$ and $\lambda \gg \psi_n \gg n^{-1}(\log n)^{1/2}$, on the event $\mathcal{F}_2$, when n is sufficiently large, it holds that for $(\boldsymbol{\beta}^\top, \boldsymbol{\theta}^\top_{[\cdot 1]}, \boldsymbol{\theta}^\top_{[\cdot 2]}) \in \mathcal{A}_n \cap \mathcal{A}$,

$$\begin{aligned}& Q(\boldsymbol{\beta},\boldsymbol{\Theta}) - Q(\boldsymbol{\beta},\boldsymbol{\Theta}^*) = \Gamma_2 - \Gamma_1\\ \geq & \sum_{m=1}^{2}\sum_{j>l}\left(\lambda_m - n^{-1}\max\left\{\max_{i,j}|w_{1j}-w_{1i}|, \max_{i,j}|w_{2j}-w_{2i}|\right\}\right)|\theta_{lm}-\theta_{jm}| > 0.\end{aligned}$$

This completes the proof. □

4. Additional numerical results

4.1 Results of clustering analysis

We use the Rand Index measure to evaluate the performance of clustering. A pair of instances is true positive if they are from the same component and are also assigned to the same cluster. A pair of instances is true negative if they are from different components and are also assigned to different clusters. The Rand Index (RI) is defined as

$$\mathrm{RI} = \frac{n_{\mathrm{TP}} + n_{\mathrm{TN}}}{n(n-1)/2},$$

where n_{TP} is the number of true positive pairs and n_{TN} is the number of true negative pairs. The Rand Index lies between 0 and 1, and a higher value is preferred. Table S1 reports the average value and standard deviation of the Rand Index for clustering means and precisions, respectively. In Scenario 1, Hard-GMM, SCAD-GMM, and SubAna can always correctly cluster means. The proposed methods outperform the EM-based methods FlexMix and MS-GMM in clustering precisions. In Scenario 2, our method also shows great advantages in clustering. We also note that the clustering accuracy of FlexMix or MS-GMM is larger in Scenario 2 than that in Scenario 1, while the former is a much more complicated case. As FlexMix and MS-GMM assume that the means and precisions share the same structure, the heterogeneity among means can improve the clustering ability of the two methods, which is the case in Scenario 2.

Table S1: The average value and standard deviation of the Rand Index for clustering means and precisions over 100 replications.

	Hard-GMM	SCAD-GMM	SubAna	FlexMix	MS-GMM
			Scenario 1		
Mean	1_0	1_0	1_0	$0.585_{0.165}$	$0.547_{0.171}$
Precision	$0.849_{0.091}$	$0.838_{0.086}$	–	$0.680_{0.116}$	$0.630_{0.109}$
			Scenario 2		
Mean	$0.910_{0.028}$	$0.909_{0.038}$	$0.895_{0.048}$	$0.878_{0.077}$	$0.870_{0.073}$
Precision	$0.848_{0.040}$	$0.846_{0.045}$	–	$0.772_{0.040}$	$0.792_{0.061}$

4.2 Plots of relative residuals

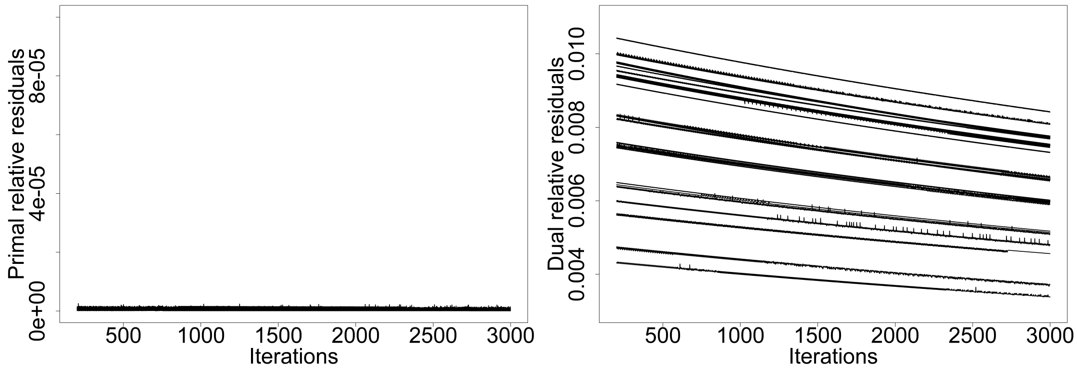

Figure S1: The primal and dual relative residuals against the number of iterations by Hard-GMM for 20 simulated datasets under Scenario 1 (after 200 iterations). Each curve represents one dataset.

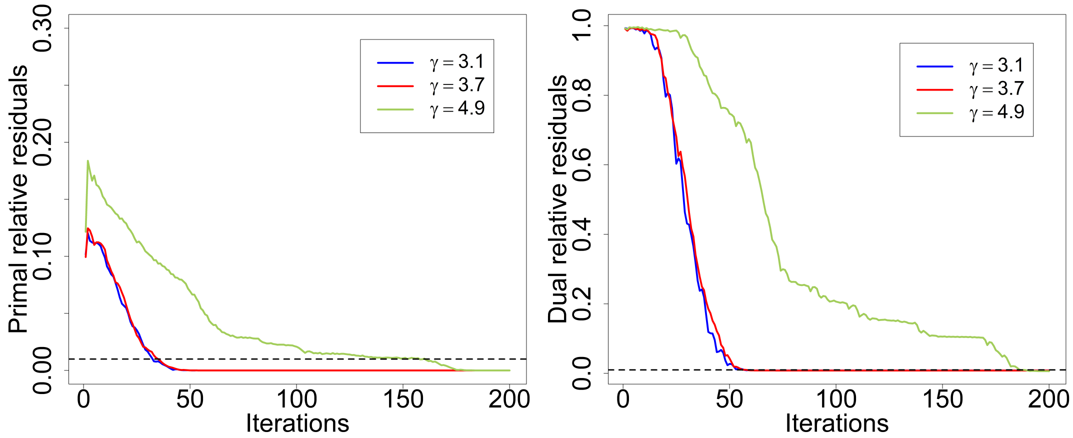

Figure S2: Average curves for the primal and dual relative residuals against the number of iterations with different values of γ by SCAD-GMM over 100 repetitions under Scenario 1.

4.3 Illustration on initialization

We use one simulated dataset under Scenario 1 to illustrate how the initialization procedure works. Figure S3 shows the responses, errors, initial points, and final estimates by Hard-GMM. As one can see, there is no clear pattern in the responses. The errors, which are unobserved, are centered around 0 with some points close to the centroid and others diversely distributed. In the initialization procedure, the ridge penalty shrinks the differences among means and precisions, but it does not lead to sparsity. We then cluster the observations into $\lfloor n^{1/2} \rfloor$ subgroups based on the ridge estimators. In Figure S3, the structures of initial points are clearer than those of the original observations. The initial points for means become more compact and those for precisions are separately distributed. After we apply Hard-GMM, the structures of means and precisions are identified as shown in the right panel in Figure S3. Figure S4 shows the results for one dataset under Scenario 2.

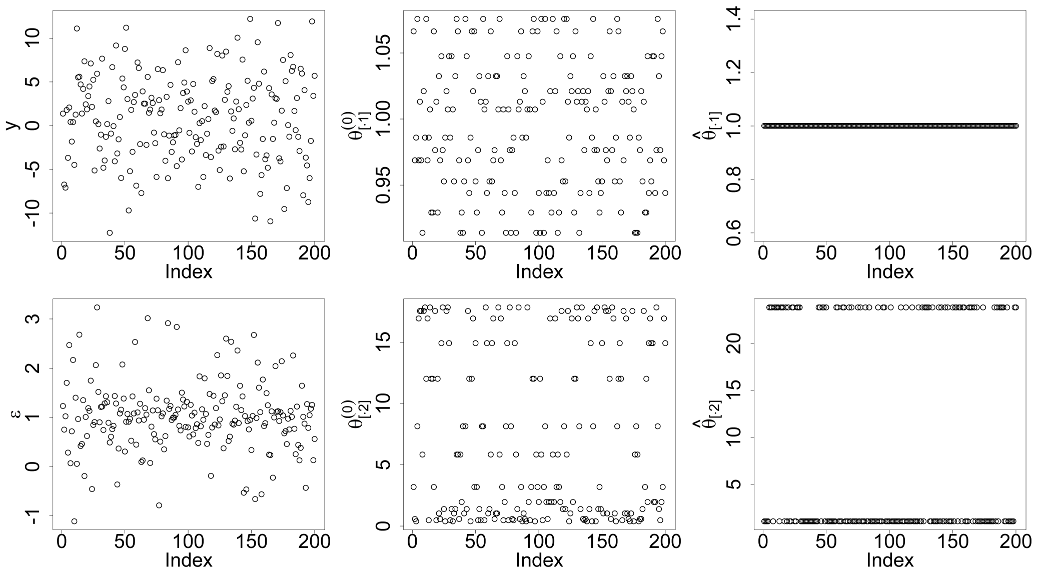

Figure S3: The responses, errors (unobserved), initial points, and estimates by Hard-GMM for one dataset under Scenario 1.

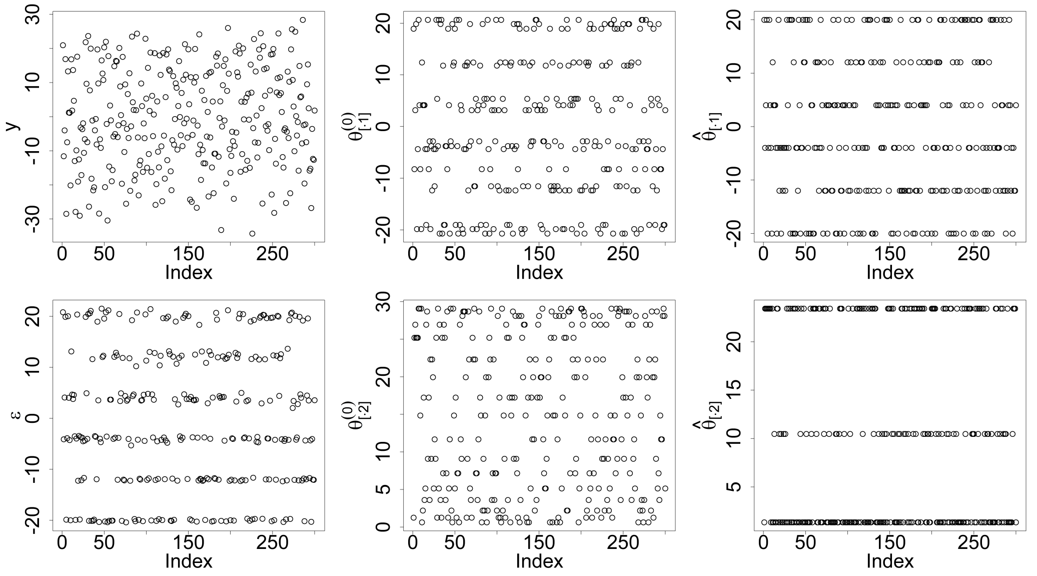

Figure S4: The responses, errors (unobserved), initial points, and estimates by SCAD-GMM for one dataset under Scenario 2.

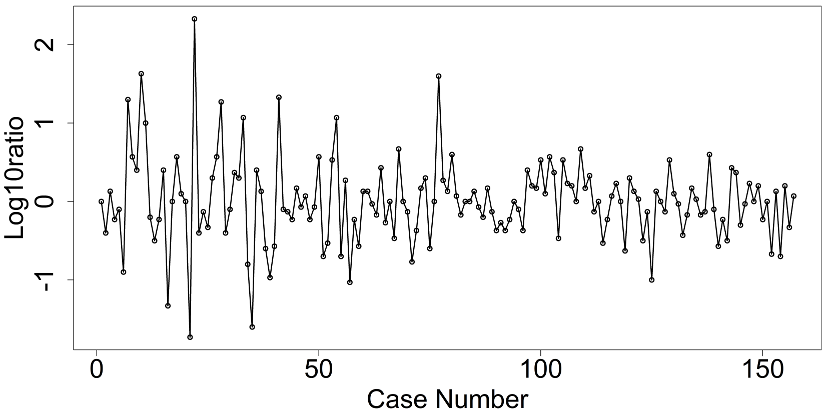

Figure S5: Sequence of the log of the ratio of samples in the Gold Mine Sampling data.

4.4 Application to the Gold Mine Sampling data

African gold miners extract samples from sections of ore on the basis of face sampling and submit them for chemical assay to check their gold concentrations. As a quality control measure, supervisors then randomly select some locations that have already been sampled by gold miners and cut fresh samples next to the spot, where the gold concentrations are also measured later. In this way, we have several pairs of samples and gold concentration, by the gold miners and supervisors respectively. To measure whether the operation of gold miners is effective, the log of the ratio of the gold concentration of the original sampler to that of the supervisor is calculated, which approximately follows a Gaussian mixture distribution. Rowland and Sichel (1961) provided several such datasets, and we focus on the one that is available in Jandhyala et al. (2002). In this dataset, there are a total of 157 observations, where the sequence of the log of the ratio of samples is shown in Figure S5.

We apply Hard-GMM and SCAD-GMM to this dataset by fitting a null regression model for the observations. The SubAna, FlexMix, and MS-GMM methods are also implemented for comparisons. Table S2 shows the estimated values of K_1, K_2, $\boldsymbol{\mu}$, and $(\boldsymbol{\tau})^{-1/2}$, where the latter two are the distinct values in means and standard deviations. The sizes of subgroups of means and precisions, denoted by $|\widehat{\mathcal{G}}^{(1)}|$ and $|\widehat{\mathcal{G}}^{(2)}|$, are also presented. By the Hard-GMM and SCAD-GMM methods, we obtain $\widehat{K}_1 = 1$ and $\widehat{K}_2 = 2$. The result corresponds to the sequence in Figure S5, where the first part possesses a relatively large volatility and the second part has a smaller variance. Such an increase in precision may indicate a highly desirable improvement in gold mining quality, perhaps caused by gaining knowledge and learning skills.

To compare the performances of various methods in clustering, we calculate the generalized Dunn (GD) index (Bezdek and Pal, 1998),

$$\mathrm{GD} = \min_{k \neq k'} o_{kk'} / (2 \max_{k} h_k),$$

where $o_{kk'}$ is the maximum distance between two samples from different clusters, and h_k is the average distance of all samples in the k-th cluster to its centroid. The GD index quantifies the ratio of between-clusters and within-groups distances, for which a larger value indicates better performance in clustering. The values of the GD index for Hard-GMM, SCAD-GMM, FlexMix, and MS-GMM are 1.738, 1.715, 1.279, and 1.212, respectively, indicating the proposed method performs much better than the other approaches.

Table S2: Estimated values of K_1, K_2, $\boldsymbol{\mu}$, and $(\boldsymbol{\tau})^{-1/2}$, and the sizes of subgroups in means and precisions, respectively denoted by $|\widehat{\mathcal{G}}^{(1)}|$ and $|\widehat{\mathcal{G}}^{(2)}|$, for the Gold Mine Sampling data.

	$\widehat{K}_1$	$\widehat{K}_2$	$\widehat{\boldsymbol{\mu}}$	$\lvert\widehat{\mathcal{G}}^{(1)}\rvert$	$(\widehat{\boldsymbol{\tau}})^{-1/2}$	$\lvert\widehat{\mathcal{G}}^{(2)}\rvert$
Hard-GMM	1	2	-0.014	157	(0.876, 0.244)	(58, 99)
SCAD-GMM	1	2	-0.015	157	(0.895, 0.206)	(53, 104)
SubAna	1	–	0.003	157	–	–
FlexMix	2	2	(0.102, -0.032)	(20, 137)	(0.819, 0.107)	(20, 137)
MS-GMM	2	2	(0.153, -0.072)	(16, 141)	(1.012, 0.323)	(16, 141)

References

Bezdek, J. C. and N. R. Pal (1998). Some new indexes of cluster validity. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 28*(3), 301–315.

Jandhyala, V. K., S. B. Fotopoulos, and D. M. Hawkins (2002). Detection and estimation of abrupt changes in the variability of a process. *Computational Statistics & Data Analysis 40*(1), 1–19.

Rowland, R. and H. Sichel (1961). Statistical quality control of routine underground sampling. *Journal of the South African Institute for Mining and Metallurgy 60*, 251–284.

Wang, Y., W. Yin, and J. Zeng (2019). Global convergence of ADMM in nonconvex nonsmooth optimization. *Journal of Scientific Computing 78*, 29–63.