Supplementary Materials for "EFFECTS-NESTED MULTI-LEVEL SUPERVISED HETEROGENEITY ANALYSIS"

Ruiyue Wang[1,2], Sanguo Zhang[1], and Shuangge Ma[2]

[1] *University of Chinese Academy of Sciences*

[2] *Yale University*

S1 Proofs

S1.1 Conditions

Condition 1. For some positive constants κ_1 and κ_2, $0 < \kappa_1 < \sigma_k^{u*} < \kappa_2$ for $k = 1, 2, \cdots, K_u$ and $0 < \kappa_1 < \sigma_k^{k*} < \kappa_2$ for $k = 1, 2, \cdots, K_l$. Further, $\|\boldsymbol{\beta}^*\|_\infty = \max_{k=1,\ldots,K_u} \|\boldsymbol{\beta}_k^*\|_\infty$, $\|\boldsymbol{\alpha}^*\|_\infty = \max_{k=1,\ldots,K_l} \|\boldsymbol{\alpha}_k^*\|_\infty$ and $\|\boldsymbol{\gamma}^*\|_\infty = \max_{k=1,\ldots,K_l} \|\boldsymbol{\gamma}_k^*\|_\infty$ are bounded.

Similar conditions have been commonly assumed in heterogeneity analysis and high-dimensional regression analysis.

Condition 2. Design matrix $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ has bounded elements. Let $\boldsymbol{X}_{\mathcal{S}}$ be the sub-matrix of $\boldsymbol{X}$ with the support of nonzero coefficient set S, and $\boldsymbol{X}_{S^c}$ is the corresponding complement. Define $\mathbb{E}\left(L_k\left(y; \boldsymbol{x}, \boldsymbol{\Omega}\right)\right) = \int L_k\left(y; \boldsymbol{x}, \boldsymbol{\Omega}\right) f\left(y; \boldsymbol{x}, \boldsymbol{\Omega}\right) dy$, and $\boldsymbol{G}_k = \operatorname{diag}\left(\mathbb{E}\left(L_k\left(y; \boldsymbol{x}, \boldsymbol{\Omega}\right)\right)\right)$ is a $n \times n$ diagonal matrix with $\mathbb{E}\left(L_k\left(y_i; \boldsymbol{x}_i, \boldsymbol{\Omega}\right)\right)$'s as its elements, where $\boldsymbol{\Omega} \in \{\boldsymbol{\Omega} \mid \|\boldsymbol{\Omega} - \boldsymbol{\Omega}^*\|_2 \leqslant \alpha\}$. $\boldsymbol{G}_k^u$ and $\boldsymbol{G}_k^l$ are defined depending on $\boldsymbol{\Omega}^u$ and $\{\boldsymbol{\Omega}^l, \boldsymbol{\gamma}\}$. For a positive constant C_0 and any k,

$$\lambda_{\min}\left(\boldsymbol{X}_{\mathcal{S}}^\top \boldsymbol{G}_k^l \boldsymbol{X}_{\mathcal{S}}/n\right) \geq C_0, \lambda_{\min}\left(\boldsymbol{X}_{\mathcal{S}_1}^\top \boldsymbol{G}_k^u \boldsymbol{X}_{\mathcal{S}_1}/n\right) \geq C_0.$$

This condition controls the correlations between the variables and has multiple counterparts in literature. If the underlying heterogeneity structure were known, this condition would simplify to Condition 4 of Fan and Lv (2011). These matrices can be viewed as the weighted covariance matrices across the subgroups. This condition ensures that the design matrices are not too ill-conditioned and facilitate the EM iterations.

Condition 3. The K_l subgroups are sufficiently separable such that, for any $\boldsymbol{\Omega}$ in a α_0-neighborhood of $\boldsymbol{\Omega}^*$($\boldsymbol{\Omega} \in B_{\alpha_0}\left(\boldsymbol{\Omega}^*\right) := \{\boldsymbol{\Omega} : \|\boldsymbol{\Omega} - \boldsymbol{\Omega}^*\|_2 \leqslant \alpha_0\}$), and each pair $\{(k, k'), 1 \leqslant k \neq k' \leqslant K_0\}$,

$$\operatorname{pr}\left(y \in A_k \mid \boldsymbol{\Omega}\right) \cdot \operatorname{pr}\left(y \in A_{k'} \mid \boldsymbol{\Omega}\right) \leqslant \frac{\varrho}{24\left(K_l - 1\right)\sqrt{K_l \max\left\{W, W', W''\right\}}},$$

where A_{l_1} is the l_1-th subgroup, $\varrho = c \cdot \min\left\{C_0\kappa_1, 0.5\left(\kappa_2 + \alpha_0\right)^{-2}\right\}$ for a constant c, and the definitions of W, W', W'' are as in Lemma 2. Additionally, the K_u groups satisfy similar conditions, for which, we also refer to Condition C6 of Li et al. (2023).

This sufficiently separable condition is common in the heterogeneity analysis literature. It is noted that as K_l grows, the problem gets more challenging, and a stronger condition is needed. Relevant discussions can be found in Hao et al. (2018) and others.

Condition 4. $\min_{k=1,\ldots,K_l} \pi_k^{l*} = O\left(\max_{k=1,\ldots,K_l} \pi_k^{l*}\right)$, $K_l^2 = o\left(p(\log n)^{-1}\right)$. Also, $\min_{k=1,\ldots,K_u} \pi_k^{u*} = O\left(\max_{k=1,\ldots,K_u} \pi_k^{u*}\right)$, $K_u^2 = o\left(p(\log n)^{-1}\right)$.

This condition assumes that the subgroups (groups) are not too imbalanced. Furthermore, the number of subgroups (groups) is allowed to grow with sample size n and dimension p, at a rate slower than $p(\log n)^{-1}$. Intuitively, a larger n allows more subgroups (groups), but it needs to be indirectly limited by p. It can also be fixed, which leads to the optimal rate discussed in the main article and is comparable to those in the literature.

Condition 5. $\rho(t) = \lambda^{-1}p(t, \lambda)$ is concave in $t \in [0, \infty)$ with a continuous derivative $\rho'(t)$ satisfying $\rho(0+) = 1$, and $\rho'(0+)$ is independent of λ.

This condition is satisfied by Lasso, SCAD, MCP, and others. For some penalties, there exists a constant $0 < a < \infty$ such that $\rho(t)$ is constant for all $t \geqslant a\lambda$.

Condition 6.

$$\begin{gathered}\left\|\left[\frac{1}{n}\boldsymbol{X}_{\mathcal{S}^c}^\top \boldsymbol{G}_k^l \boldsymbol{X}_{\mathcal{S}}\right]\left[\frac{1}{n}\boldsymbol{X}_{\mathcal{S}}^\top \boldsymbol{G}_k^l \boldsymbol{X}_{\mathcal{S}}\right]^{-1}\right\|_\infty \leqslant O(h_l), \\ \left\|\left[\frac{1}{n}\boldsymbol{X}_{\mathcal{S}_1^c}^\top \boldsymbol{G}_k^u \boldsymbol{X}_{\mathcal{S}_1}\right]\left[\frac{1}{n}\boldsymbol{X}_{\mathcal{S}_1}^\top \boldsymbol{G}_k^u \boldsymbol{X}_{\mathcal{S}_1}\right]^{-1}\right\|_\infty \leqslant O(h_u),\end{gathered} \tag{S1.1}$$

where $h_1 \leqslant \min(\sqrt{\frac{ns_2}{s^2 K_l^4 \log p}}, \sqrt{K_l^2 s_2})$ and $h_2 \leqslant \min(\sqrt{\frac{n}{sK_u^4 \log p}}, \sqrt{K_u^2 s})$;

$$\left\| \left[\frac{1}{n}\boldsymbol{X}_{\mathcal{S}_1^c}^\top \boldsymbol{G}_k^l \boldsymbol{X}_{\mathcal{S}}\right] \left[\frac{1}{n}\boldsymbol{X}_{\mathcal{S}}^\top \boldsymbol{G}_k^l \boldsymbol{X}_{\mathcal{S}}\right]^{-1} \right\|_\infty \leqslant O(h),$$

where $h \leqslant \min(\sqrt{\frac{n}{sK_l^4 \log p}}, \sqrt{K_l^2 s})$.

This is the irrepresentable condition and regulates the correlations between the important and unimportant variables within each subgroup. This is a common assumption in the high-dimensional regression literature. Analogous to Condition C3 in Li et al. (2023), we adopt a sample-weighted version for both levels. Furthermore, we restrict the correlations between the heterogeneous and homogeneous variables, ensuring that the heterogeneous and homogeneous parts can be effectively identified.

S1.2 Proof of Theorem 1

Denote $\boldsymbol{\Omega} = (\boldsymbol{\Omega}^{u\top}, \boldsymbol{\Omega}^{l\top}, \boldsymbol{\gamma}^\top)^\top$, $\boldsymbol{\Omega}^u = \left(\boldsymbol{\Omega}_1^{u\top}, \cdots, \boldsymbol{\Omega}_{K_u}^{u\top}\right)^\top$, $\boldsymbol{\Omega}_k^u = \mathrm{vec}\,(\boldsymbol{\beta}_k, \sigma_k^u) = \left(\beta_{k1}, \beta_{k2}, \cdots, \beta_{kp}^\top, \sigma_k^u\right)$, $\boldsymbol{\Omega}^l = \left(\boldsymbol{\Omega}_1^{l\top}, \cdots, \boldsymbol{\Omega}_{K_l}^{l\top}\right)^\top$, where $\boldsymbol{\Omega}_k^l = \mathrm{vec}\,(\boldsymbol{\alpha}_k, \sigma_k^u) = \left(\alpha_{k1}, \alpha_{k2}, \cdots, \alpha_{kp}^\top, \sigma_k^l\right)$ and $\boldsymbol{\gamma} = \{\boldsymbol{\gamma}_1^\top, \boldsymbol{\gamma}_2^\top, \cdots, \boldsymbol{\gamma}_{K_l}^\top\}^\top$. Accordingly, $\boldsymbol{\Omega}^* = (\boldsymbol{\Omega}^{u*\top}, \boldsymbol{\Omega}^{l*\top}, \boldsymbol{\gamma}^{*\top})^\top$. Consider the objective function:

$$\begin{aligned}\mathcal{Q}(\boldsymbol{\Omega}, \boldsymbol{\pi}^u, \boldsymbol{\pi}^l) =& \frac{1}{n}\sum_{i=1}^n \log\left(\sum_{k=1}^{K_u} \pi_k^u f_k\left(y_i; \boldsymbol{x}_i, \boldsymbol{\Omega}_k^u\right)\right) \\ &+ \frac{1}{n}\sum_{i=1}^n \log\left(\sum_{k=1}^{K_l} \pi_k^l f_k\left(y_i; \boldsymbol{x}_i, \boldsymbol{\Omega}_k^l, \boldsymbol{\gamma}_k\right)\right) - \mathcal{P}_1(\boldsymbol{\Omega}),\end{aligned} \tag{S1.2}$$

where $\boldsymbol{\pi}^u = \left(\pi_1^u, \pi_2^u, \ldots, \pi_{K_u}^u\right)^\top$ and $\boldsymbol{\pi}^l = \left(\pi_1^l, \pi_2^l, \ldots, \pi_{K_l}^l\right)^\top$, and the penalty

$$\mathcal{P}_1(\boldsymbol{\Omega}) = \sum_{j=1}^{p} P\left(\sqrt{\sum_{k=1}^{K_u}(\beta_{kj}^u)^2 + \sum_{k=1}^{K_l}(\alpha_{kj}^l)^2}, \lambda_1\right) + \sum_{j=1}^{p} P\left(\sqrt{\sum_{k=1}^{K_l}(\gamma_{kj}^l)^2}, \lambda_2\right). \tag{S1.3}$$

Denote $\boldsymbol{\omega}^u = (\omega_{ik}^u)_{n\times K_u}$ and $\boldsymbol{\omega}^l = (\omega_{ik}^u)_{n\times K_l}$. If $\boldsymbol{\omega}^u$ and $\boldsymbol{\omega}^l$ were available, the penalized log-likelihood function for the complete data could be written as:

$$\begin{aligned}\mathcal{Q}\left(\boldsymbol{\Omega} \mid \boldsymbol{X}, \boldsymbol{\omega}\right) :=& \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K_u}\omega_{ik}^u\left[\log \pi_k^u + \log f_k\left(y_i; \boldsymbol{x}_i, \boldsymbol{\Omega}_k^u\right)\right] \\ &+ \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K_l}\omega_{ik}^l\left[\log \pi_k^l + \log f_k\left(y_i; \boldsymbol{x}_i, \boldsymbol{\Omega}_k^l, \boldsymbol{\gamma}_k\right)\right] - \mathcal{P}_1(\boldsymbol{\Omega}).\end{aligned} \tag{S1.4}$$

Here, ω_{ik}^l is a latent Bernoulli variable with expectation $\mathbb{E}\left(\omega_{ik}^l \mid y, \boldsymbol{x}, \boldsymbol{\Omega}^l, \boldsymbol{\gamma}\right) = \mathbb{P}\left(\omega_{ik} = 1 \mid y, \boldsymbol{x}, \boldsymbol{\Omega}^l, \boldsymbol{\gamma}\right)$. Note that it can be computed as $L_k^l\left(\boldsymbol{x}_i, \boldsymbol{\Omega}^l, \boldsymbol{\gamma}\right) = \frac{\pi_k^l f_k\left(y_i; \boldsymbol{x}_i, \boldsymbol{\Omega}_k^l, \boldsymbol{\gamma}_k\right)}{\sum_{k=1}^{K_l}\pi_k^l f_k\left(y_i; \boldsymbol{x}_i, \boldsymbol{\Omega}_k^l, \boldsymbol{\gamma}_k\right)}$. ω_{ik}^u can be defined accordingly. In the t-th step of the EM algorithm, the conditional expectation is computed as:

$$\mathbb{E}_{\boldsymbol{\omega}\mid\boldsymbol{X},\boldsymbol{\Omega}^{(t-1)}}\left[\mathcal{Q}\left(\boldsymbol{\Omega} \mid \boldsymbol{X}, \boldsymbol{\omega}\right)\right] = \mathcal{H}_n(\boldsymbol{\Omega} \mid \boldsymbol{\Omega}^{(t-1)}) - \mathcal{P}_1(\boldsymbol{\Omega}), \tag{S1.5}$$

where

$$\mathcal{H}_n(\boldsymbol{\Omega} \mid \boldsymbol{\Omega}^{(t-1)}) = \mathcal{H}_n^u(\boldsymbol{\Omega}^u \mid \boldsymbol{\Omega}^{u(t-1)}) + \mathcal{H}_n^l(\boldsymbol{\Omega}^l, \boldsymbol{\gamma} \mid \boldsymbol{\Omega}^{l(t-1)}, \boldsymbol{\gamma}^{(t-1)}), \tag{S1.6}$$

$$\mathcal{H}_n^u(\boldsymbol{\Omega}^u \mid \boldsymbol{\Omega}^{u(t-1)}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K_u}\omega_{ik}^u(\boldsymbol{\Omega}^{u(t-1)})\left[\log \pi_k^u + \log f_k\left(y_i; \boldsymbol{x}_i, \boldsymbol{\Omega}_k^u\right)\right],$$
$$\mathcal{H}_n^l(\boldsymbol{\Omega}^l, \boldsymbol{\gamma} \mid \boldsymbol{\Omega}^{l(t-1)}, \boldsymbol{\gamma}^{(t-1)}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K_l}\omega_{ik}^l(\boldsymbol{\Omega}^{l(t-1)}, \boldsymbol{\gamma}^{(t-1)})$$
$$\left[\log \pi_k^l + \log f_k\left(y_i; \boldsymbol{x}_i, \boldsymbol{\Omega}_k^l, \boldsymbol{\gamma}_k\right)\right], \tag{S1.7}$$

and here $\omega_{ik}^l(\boldsymbol{\Omega}^{l(t-1)}, \boldsymbol{\gamma}^{(t-1)})$ can be computed as $L_k^l\left(y_i, \boldsymbol{x}_i, \boldsymbol{\Omega}^{l(t-1)}, \boldsymbol{\gamma}^{(t-1)}\right) = \frac{\pi_k^{l(t-1)} f_k\left(y_i;\boldsymbol{x}_i,\boldsymbol{\Omega}_k^{l(t-1)},\boldsymbol{\gamma}_k^{(t-1)}\right)}{\sum_{k=1}^{K_l}\pi_k^{l(t-1)} f_k\left(y_i;\boldsymbol{x}_i,\boldsymbol{\Omega}_k^{l(t-1)},\boldsymbol{\gamma}_k^{(t-1)}\right)}$, which depends on $\pi_k^{l(t-1)}, \boldsymbol{\Omega}^{l(t-1)}$, and $\boldsymbol{\gamma}^{(t-1)}$ obtained in the previous iteration. $\omega_{ik}^u(\boldsymbol{\Omega}^{u(t-1)})$ can be similarly defined. Note that the update of these weights is independent of the updates of the other parameters. To establish Theorem 1, a corresponding population version of $\mathcal{H}_n$ needs to be defined as:

$$\mathcal{H}^u(\boldsymbol{\Omega}^u \mid \boldsymbol{\Omega}^{u(t-1)}) = \mathbb{E}\left[\sum_{k=1}^{K_u}\omega_{ik}^u(\boldsymbol{\Omega}^{u(t-1)})\left[\log \pi_k^u + \log f_k\left(y_i; \boldsymbol{x}_i, \boldsymbol{\Omega}_k^u\right)\right]\right];$$
$$\mathcal{H}^l(\boldsymbol{\Omega}^l, \boldsymbol{\gamma} \mid \boldsymbol{\Omega}^{l(t-1)}, \boldsymbol{\gamma}^{(t-1)}) = \mathbb{E}\left[\sum_{k=1}^{K_l}\omega_{ik}^l(\boldsymbol{\Omega}^{l(t-1)}, \boldsymbol{\gamma}^{(t-1)})\right.$$
$$\left.\left[\log \pi_k^l + \log f_k\left(y_i; \boldsymbol{x}_i, \boldsymbol{\Omega}_k^l, \boldsymbol{\gamma}_k\right)\right]\right]. \tag{S1.8}$$

Define the function evaluating the error between the iterative estimator $\boldsymbol{\Omega}$ and true parameter $\boldsymbol{\Omega}^*$ as:

$$
\begin{aligned}
\boldsymbol{q}(\boldsymbol{v}) =& \mathcal{H}_n(\boldsymbol{\Omega}^* + \boldsymbol{v} \mid \boldsymbol{\Omega}^{(t-1)}) - \mathcal{H}_n(\boldsymbol{\Omega}^* \mid \boldsymbol{\Omega}^{(t-1)}) - [\mathcal{P}_1(\boldsymbol{\Omega}^* + \boldsymbol{v}) - \mathcal{P}_1(\boldsymbol{\Omega}^*)] \\
=& \left[\mathcal{H}_n^u(\boldsymbol{\Omega}^{u*} + \boldsymbol{v}_1 \mid \boldsymbol{\Omega}^{u(t-1)}) - \mathcal{H}_n^u(\boldsymbol{\Omega}^{u*} \mid \boldsymbol{\Omega}^{u(t-1)})\right] \\
&+ \left[\mathcal{H}_n^l(\boldsymbol{\Omega}^{l*} + \boldsymbol{v}_2, \boldsymbol{\gamma}^* + \boldsymbol{v}_\gamma \mid \boldsymbol{\Omega}^{l(t-1)}, \boldsymbol{\gamma}^{(t-1)}) - \mathcal{H}_n^l(\boldsymbol{\Omega}^{l*}, \boldsymbol{\gamma}^* \mid \boldsymbol{\Omega}^{l(t-1)}, \boldsymbol{\gamma}^{(t-1)})\right] \\
&- [\mathcal{P}_1(\boldsymbol{\Omega}^* + \boldsymbol{v}) - \mathcal{P}_1(\boldsymbol{\Omega}^*)].
\end{aligned}
$$

Note that we set $\boldsymbol{v} = (\boldsymbol{v}_1^\top, \boldsymbol{v}_2^\top, \boldsymbol{v}_\gamma^\top)^\top$ corresponding to $\boldsymbol{\Omega}_0 = (\boldsymbol{\Omega}_0^{u\top}, \boldsymbol{\Omega}_0^{l\top}, \boldsymbol{\gamma}_0^\top)^\top$, where $\boldsymbol{\Omega}_0^u = \left(\boldsymbol{\Omega}_{01}^{u\top}, \cdots, \boldsymbol{\Omega}_{0K_u}^{u\top}\right)^\top$, $\boldsymbol{\Omega}_{0k}^u = \mathrm{vec}\left(\boldsymbol{\beta}_{k\mathcal{S}_1}, \sigma_k^u\right)$ with $\boldsymbol{\beta}_{k\mathcal{S}_1} = \{\boldsymbol{\beta}_{kj} : j \in \mathcal{S}_1\}$ containing the important parameters. $\boldsymbol{\Omega}_0^l$, $\boldsymbol{\gamma}_0$ and $\boldsymbol{\Omega}_0^*$ are defined similarly.

Step 1:

To prove estimation consistency, we first show that, if the estimate obtained in the $(t-1)$-th iteration of the EM algorithm $\boldsymbol{\Omega}_0^{(t-1)} \in \mathcal{B}_\alpha(\boldsymbol{\Omega}_0^*)$, where $\alpha = O\left(K_l^3 s_1 \sqrt{\log p/n}\right)$, then there is a conditional local maximizer $\boldsymbol{\Omega}_0$ in $\{\boldsymbol{\Omega}_0^u : \|\boldsymbol{\Omega}_0^u - \boldsymbol{\Omega}_0^{u*}\|_2 \leqslant \chi_1\}$, $\left\{\boldsymbol{\Omega}_0^l : \left\|\boldsymbol{\Omega}_0^l - \boldsymbol{\Omega}_0^{l*}\right\|_2 \leqslant \chi_2\right\}$, and $\{\boldsymbol{\gamma}_0 : \|\boldsymbol{\gamma}_0 - \boldsymbol{\gamma}_0^*\|_2 \leqslant \chi_\gamma\}$, where $\chi_1 = \frac{4\epsilon_1}{\varrho} + \iota\left\|\boldsymbol{\Omega}_0^{(t-1)} - \boldsymbol{\Omega}_0^*\right\|_2$, $\chi_2 = \frac{4\epsilon_2}{\varrho} + \iota\left\|\boldsymbol{\Omega}_0^{(t-1)} - \boldsymbol{\Omega}_0^*\right\|_2$, $\chi_\gamma = \frac{4\epsilon_\gamma}{\varrho} + \iota\sqrt{\frac{s_2}{s_1}}\left\|\boldsymbol{\Omega}_0^{(t-1)} - \boldsymbol{\Omega}_0^*\right\|_2$, $\epsilon_1 \simeq \epsilon_2 = O(\sqrt{\frac{K_l^3 s \log p}{n}})$, $\epsilon_\gamma = O(\sqrt{\frac{K_l^3 s_2 \log p}{n}})$ and $1/3 \leqslant \iota < 1$ is a positive constant. It is noted that $\alpha \gg \epsilon$, so the selection of α is reasonable. It suffices to show that:

$$
\lim_{n\to\infty} P\left(\sup_{\|\boldsymbol{v}_1\|=\chi_1, \|\boldsymbol{v}_2\|=\chi_2, \|\boldsymbol{v}_\gamma\|=\chi_\gamma} \boldsymbol{q}(\boldsymbol{v}) < 0)\right) = 1. \tag{S1.9}
$$

Next, we establish an upper bound for $\boldsymbol{q}(\boldsymbol{v})$ over the set $C(\chi) := \{\boldsymbol{v} : \|\boldsymbol{v}_1\| = \chi_1, \|\boldsymbol{v}_2\| = \chi_2, \|\boldsymbol{v}_\gamma\| = \chi_\gamma\}$. Note that for a sufficiently large n, we can obtain that $\chi_1, \chi_2, \chi_\gamma \leqslant \alpha$.

We introduce the strong concavity mentioned in Balakrishnan et al. (2017). First consider the lower level. For any $\boldsymbol{\Omega}_0^l \in \mathcal{B}_\alpha(\boldsymbol{\Omega}_0^{l*})$, $\boldsymbol{\gamma}_0 \in \mathcal{B}_\alpha(\boldsymbol{\gamma}_0^*)$, with probability at least $1 - 1/p$ and a sufficiently large n, each $\boldsymbol{\Omega}_0^{l\prime} \in \left\{\boldsymbol{\Omega}_0^{l\prime} \mid \|\boldsymbol{\Omega}_0^{l\prime} - \boldsymbol{\Omega}_0^{l*}\|_2 \leqslant \alpha\right\}$ and $\boldsymbol{\gamma}_0' \in \{\boldsymbol{\gamma}_0' \mid \|\boldsymbol{\gamma}_0' - \boldsymbol{\gamma}_0^*\|_2 \leqslant \alpha\}$ satisfies:

$$
\begin{aligned}
&\mathcal{H}_n^l(\boldsymbol{\Omega}_0^{l\prime}, \boldsymbol{\gamma}_0' \mid \boldsymbol{\Omega}_0^l, \boldsymbol{\gamma}_0) - \mathcal{H}_n^l(\boldsymbol{\Omega}_0^{l*}, \boldsymbol{\gamma}_0^* \mid \boldsymbol{\Omega}_0^l, \boldsymbol{\gamma}_0) \\
&\quad \leqslant \left\langle \nabla_{\boldsymbol{\Omega}_0^{l*}} \mathcal{H}_n\left(\boldsymbol{\Omega}_0^{l*}, \boldsymbol{\gamma}_0^* \mid \boldsymbol{\Omega}_0^l, \boldsymbol{\gamma}_0\right), \boldsymbol{\Omega}_0^{l\prime} - \boldsymbol{\Omega}_0^{l*} \right\rangle - \frac{\varrho}{2}\|\boldsymbol{\Omega}_0^{l\prime} - \boldsymbol{\Omega}_0^{l*}\|_2^2 \\
&\quad + \left\langle \nabla_{\boldsymbol{\gamma}_0^*} \mathcal{H}_n\left(\boldsymbol{\Omega}_0^{l*}, \boldsymbol{\gamma}_0^* \mid \boldsymbol{\Omega}_0^l, \boldsymbol{\gamma}_0\right), \boldsymbol{\gamma}_0' - \boldsymbol{\gamma}_0^* \right\rangle - \frac{\varrho}{2}\|\boldsymbol{\gamma}_0' - \boldsymbol{\gamma}_0^*\|_2^2,
\end{aligned}
\tag{S1.10}
$$

where ϱ is a positive constant. Then we consider $k = 1, \cdots, K_l$ individually:

$$
\begin{aligned}
&\nabla_{\boldsymbol{\alpha}_{k\mathcal{S}_1}^*} \mathcal{H}_n\left(\boldsymbol{\Omega}_{0k}^{l*}, \boldsymbol{\gamma}_{0k}^* \mid \boldsymbol{\Omega}_0^l, \boldsymbol{\gamma}_0\right) \\
&\qquad = \frac{\sigma_k^{l*}}{n} \sum_{i=1}^n \left[L_k^l\left(y_i, \boldsymbol{x}_i, \boldsymbol{\Omega}_0^l, \boldsymbol{\gamma}_0\right)\left(y_i - \boldsymbol{x}_{i\mathcal{S}_1}^\top \boldsymbol{\alpha}_{k\mathcal{S}_1}^* - \boldsymbol{x}_{i\mathcal{S}_2}^\top \boldsymbol{\gamma}_{k\mathcal{S}_2}^*\right)\boldsymbol{x}_{i\mathcal{S}_1}\right]
\end{aligned}
\tag{S1.11}
$$

$$
\begin{aligned}
&\nabla_{\boldsymbol{\gamma}_{k\mathcal{S}_2}^*} \mathcal{H}_n\left(\boldsymbol{\Omega}_{0k}^{l*}, \boldsymbol{\gamma}_{0k}^* \mid \boldsymbol{\Omega}_0^l, \boldsymbol{\gamma}_0\right) \\
&\qquad = \frac{\sigma_k^{l*}}{n} \sum_{i=1}^n \left[L_k^l\left(y_i, \boldsymbol{x}_i, \boldsymbol{\Omega}_0^l, \boldsymbol{\gamma}_0\right)\left(y_i - \boldsymbol{x}_{i\mathcal{S}_1}^\top \boldsymbol{\alpha}_{k\mathcal{S}_1}^* - \boldsymbol{x}_{i\mathcal{S}_2}^\top \boldsymbol{\gamma}_{k\mathcal{S}_2}^*\right)\boldsymbol{x}_{i\mathcal{S}_2}\right]
\end{aligned}
\tag{S1.12}
$$

$$
\begin{aligned}
\nabla_{\boldsymbol{\sigma}_k^{l*}} \mathcal{H}_n\left(\boldsymbol{\Omega}_{0k}^{l*}, \boldsymbol{\gamma}_{0k}^* \mid \boldsymbol{\Omega}_0^l, \boldsymbol{\gamma}_0\right) &= \frac{1}{2n\sigma_k^{l*}} \sum_{i=1}^n \left[L_k^l\left(y_i, \boldsymbol{x}_i, \boldsymbol{\Omega}_0^l, \boldsymbol{\gamma}_0\right)\right] \\
&- \frac{1}{2n} \sum_{i=1}^n \left[L_k^l\left(y_i, \boldsymbol{x}_i, \boldsymbol{\Omega}_0^l, \boldsymbol{\gamma}_0\right)\left(y_i - \boldsymbol{x}_{i\mathcal{S}_1}^\top \boldsymbol{\alpha}_{k\mathcal{S}_1}^* - \boldsymbol{x}_{i\mathcal{S}_2}^\top \boldsymbol{\gamma}_{k\mathcal{S}_2}^*\right)^2\right],
\end{aligned}
\tag{S1.13}
$$

where it is defined, for example, $\boldsymbol{x}_{i\mathcal{S}_2} = \{x_{ij} : j \in \mathcal{S}_2\}$. And we can also show that:

$$\begin{aligned}
&\mathcal{H}_n^l(\boldsymbol{\Omega}_{0k}^{l\prime}, \boldsymbol{\gamma}_{0k}' \mid \boldsymbol{\Omega}_0^l, \boldsymbol{\gamma}_0) - \mathcal{H}_n^l(\boldsymbol{\Omega}_{0k}^{l*}, \boldsymbol{\gamma}_{0k}^* \mid \boldsymbol{\Omega}_0^l, \boldsymbol{\gamma}_0) \\
&\quad = \frac{1}{n}\sum_{i=1}^n \Bigg[L_k^l\left(y_i, \boldsymbol{x}_i, \boldsymbol{\Omega}_0^l, \boldsymbol{\gamma}_0\right)\Bigg\{\frac{1}{2}\log(\sigma_k^{l\prime}) - \frac{1}{2}\log(\sigma_k^{l*}) \\
&\quad + \frac{\sigma_k^{l*}}{2}\left(y_i - \boldsymbol{x}_{i\mathcal{S}_1}^\top \boldsymbol{\alpha}_{k\mathcal{S}_1}^* - \boldsymbol{x}_{i\mathcal{S}_2}^\top \boldsymbol{\gamma}_{k\mathcal{S}_2}^*\right)^2 - \frac{\sigma_k^{l\prime}}{2}\left(y_i - \boldsymbol{x}_{i\mathcal{S}_1}^\top \boldsymbol{\alpha}_{k\mathcal{S}_1}' - \boldsymbol{x}_{i\mathcal{S}_2}^\top \boldsymbol{\gamma}_{k\mathcal{S}_2}'\right)^2\Bigg\}\Bigg].
\end{aligned} \tag{S1.14}$$

Then,

$$\begin{aligned}
&\mathcal{H}_n^l(\boldsymbol{\Omega}_{0k}^{l\prime}, \boldsymbol{\gamma}_{0k}' \mid \boldsymbol{\Omega}_0^l, \boldsymbol{\gamma}_0) - \mathcal{H}_n^l(\boldsymbol{\Omega}_{0k}^{l*}, \boldsymbol{\gamma}_{0k}^* \mid \boldsymbol{\Omega}_0^l, \boldsymbol{\gamma}_0) \\
&\qquad - \left\langle \nabla_{\boldsymbol{\Omega}_{0k}^{l*}} \mathcal{H}_n\left(\boldsymbol{\Omega}_{0k}^{l*}, \boldsymbol{\gamma}_{0k}^* \mid \boldsymbol{\Omega}_0^l, \boldsymbol{\gamma}_0\right), \boldsymbol{\Omega}_{0k}^{l\prime} - \boldsymbol{\Omega}_{0k}^{l*}\right\rangle \\
&\qquad - \left\langle \nabla_{\boldsymbol{\gamma}_{0k}^*} \mathcal{H}_n\left(\boldsymbol{\Omega}_{0k}^{l*}, \boldsymbol{\gamma}_{0k}^* \mid \boldsymbol{\Omega}_0^l, \boldsymbol{\gamma}_0\right), \boldsymbol{\gamma}_{0k}' - \boldsymbol{\gamma}_{0k}^*\right\rangle = I + II + III.
\end{aligned} \tag{S1.15}$$

By Taylor expansion, and recalling that $\boldsymbol{\gamma}_{0k} = \boldsymbol{\gamma}_{k\mathcal{S}_2}$, we have:

$$\begin{aligned}
I =& \frac{1}{n}\sum_{i=1}^n \Bigg[L_k^l\left(y_i, \boldsymbol{x}_i, \boldsymbol{\Omega}_0^l, \boldsymbol{\gamma}_0\right)\Bigg\{\frac{\sigma_k^{l*}}{2}\left(y_i - \boldsymbol{x}_{i\mathcal{S}_1}^\top \boldsymbol{\alpha}_{k\mathcal{S}_1}^* - \boldsymbol{x}_{i\mathcal{S}_2}^\top \boldsymbol{\gamma}_{k\mathcal{S}_2}^*\right)^2 \\
&- \frac{\sigma_k^{l*}}{2}\left(y_i - \boldsymbol{x}_{i\mathcal{S}_1}^\top \boldsymbol{\alpha}_{k\mathcal{S}_1}' - \boldsymbol{x}_{i\mathcal{S}_2}^\top \boldsymbol{\gamma}_{k\mathcal{S}_2}'\right)^2\Bigg\}\Bigg] - \left\langle \nabla_{\boldsymbol{\gamma}_{0k}^*} \mathcal{H}_n\left(\boldsymbol{\Omega}_{0k}^{l*}, \boldsymbol{\gamma}_{0k}^* \mid \boldsymbol{\Omega}_0^l, \boldsymbol{\gamma}_0\right), \boldsymbol{\gamma}_{0k}' - \boldsymbol{\gamma}_{0k}^*\right\rangle \\
&- \left\langle \nabla_{\boldsymbol{\alpha}_{k\mathcal{S}_1}^*} \mathcal{H}_n\left(\boldsymbol{\Omega}_{0k}^{l*}, \boldsymbol{\gamma}_{0k}^* \mid \boldsymbol{\Omega}_0^l, \boldsymbol{\gamma}_0\right), \boldsymbol{\alpha}_{k\mathcal{S}_1}' - \boldsymbol{\alpha}_{k\mathcal{S}_1}^*\right\rangle \\
&= -\sigma_k^{l*}\,\mathrm{vec}(\boldsymbol{\alpha}_{k\mathcal{S}_1}' - \boldsymbol{\alpha}_{k\mathcal{S}_1}^*, \boldsymbol{\gamma}_{0k}' - \boldsymbol{\gamma}_{0k}^*)^\top \left[\frac{1}{n}\sum_{i=1}^n L_k^l\left(y_i, \boldsymbol{x}_i, \boldsymbol{\Omega}_0^l, \boldsymbol{\gamma}_0\right)\widetilde{\boldsymbol{x}}_{i\mathcal{S}}\widetilde{\boldsymbol{x}}_{i\mathcal{S}}^\top\right] \\
&\qquad \mathrm{vec}(\boldsymbol{\alpha}_{k\mathcal{S}_1}' - \boldsymbol{\alpha}_{k\mathcal{S}_1}^*, \boldsymbol{\gamma}_{0k}' - \boldsymbol{\gamma}_{0k}^*),
\end{aligned} \tag{S1.16}$$

where $\widetilde{\boldsymbol{x}}_{i\mathcal{S}}$ is an rearrangement of $\boldsymbol{x}_{i\mathcal{S}}$.

By taking Weyl's theorem, we can obtain that:

$$\begin{aligned}
&\lambda_{\min}\left(\frac{1}{n}\sum_{i=1}^{n} L_k^l\left(y_i, \boldsymbol{x}_i, \boldsymbol{\Omega}_0^l, \boldsymbol{\gamma}_0\right) \widetilde{\boldsymbol{x}}_{i\mathcal{S}} \widetilde{\boldsymbol{x}}_{i\mathcal{S}}^\top\right) \\
&\geqslant \lambda_{\min}\left(\frac{1}{n}\boldsymbol{X}_{\mathcal{S}}^\top \boldsymbol{G}_k^l \boldsymbol{X}_{\mathcal{S}}\right) \qquad \text{(S1.17)}\\
&\quad - \lambda_{\max}\left(\frac{1}{n}\boldsymbol{X}_{\mathcal{S}}^\top \boldsymbol{G}_k^l \boldsymbol{X}_{\mathcal{S}} - \frac{1}{n}\sum_{i=1}^{n} L_k^l\left(y_i, \boldsymbol{x}_i, \boldsymbol{\Omega}_0^l, \boldsymbol{\gamma}_0\right) \widetilde{\boldsymbol{x}}_{i\mathcal{S}} \widetilde{\boldsymbol{x}}_{i\mathcal{S}}^\top\right),
\end{aligned}$$

where $L_k^l\left(y_i, \boldsymbol{x}_i, \boldsymbol{\Omega}_0^l, \boldsymbol{\gamma}_0\right) \widetilde{\boldsymbol{x}}_{i\mathcal{S}} \widetilde{\boldsymbol{x}}_{i\mathcal{S}}^\top$ is bounded by some matrix A. Denote $M = \|A\|^2$, where $\|\cdot\|$ is the spectral norm. According to the matrix Hoeffding's inequality, we have:

$$P\left(\lambda_{\max}\left(\frac{1}{n}\boldsymbol{X}_{\mathcal{S}}^\top \boldsymbol{G}_k^l \boldsymbol{X}_{\mathcal{S}} - \frac{1}{n}\sum_{i=1}^{n} L_k^l\left(y_i, \boldsymbol{x}_i, \boldsymbol{\Omega}_0^l, \boldsymbol{\gamma}_0\right) \widetilde{\boldsymbol{x}}_{i\mathcal{S}} \widetilde{\boldsymbol{x}}_{i\mathcal{S}}^\top\right) \leqslant t\right) \geqslant 1 - s e^{\frac{-nt^2}{8M}}, \tag{S1.18}$$

where $s = |\mathcal{S}|$. Let $t = \sqrt{\frac{8M}{n}\log(2K_l sp)}$. With probability $1 - \frac{1}{2K_l p}$, we have:

$$\lambda_{\max}\left(\frac{1}{n}\boldsymbol{X}_{\mathcal{S}}^\top \boldsymbol{G}_k^l \boldsymbol{X}_{\mathcal{S}} - \frac{1}{n}\sum_{i=1}^{n} L_k^l\left(y_i, \boldsymbol{x}_i, \boldsymbol{\Omega}_0^l, \boldsymbol{\gamma}_0\right) \widetilde{\boldsymbol{x}}_{i\mathcal{S}} \widetilde{\boldsymbol{x}}_{i\mathcal{S}}^\top\right) \leqslant \sqrt{\frac{8M}{n}\log(2K_l sp)}. \tag{S1.19}$$

Here, note that $\sqrt{\frac{8M}{n}\log(2K_l sp)} = o(1)$. Therefore, $I \leqslant -\frac{C_0\kappa_1}{2}(\|\boldsymbol{\alpha}'_{k\mathcal{S}_1} - \boldsymbol{\alpha}^*_{k\mathcal{S}_1}\|_2^2 + \|\boldsymbol{\gamma}'_{0k} - \boldsymbol{\gamma}^*_{0k}\|_2^2)$.

Similar to I, by Taylor expansion, we have:

$$
\begin{aligned}
II =& \frac{1}{n}\sum_{i=1}^{n}\Bigg[L_k^l\left(y_i, \boldsymbol{x}_i, \boldsymbol{\Omega}_0^l, \boldsymbol{\gamma}_0\right)\Bigg\{\frac{1}{2}\log(\sigma_k^{l\prime}) - \frac{1}{2}\log(\sigma_k^{l*}) - \frac{1}{2\sigma_k^{l*}\sigma_k^{l\prime}}(\sigma_k^{l\prime} - \sigma_k^{l*}) \\
&+ \frac{\sigma_k^{l*}}{2}\left(y_i - \boldsymbol{x}_{i\mathcal{S}_1}^\top\boldsymbol{\alpha}_{k\mathcal{S}_1}^\prime - \boldsymbol{x}_{i\mathcal{S}_2}^\top\boldsymbol{\gamma}_{k\mathcal{S}_2}^\prime\right)^2 - \frac{\sigma_k^{l\prime}}{2}\left(y_i - \boldsymbol{x}_{i\mathcal{S}_1}^\top\boldsymbol{\alpha}_{k\mathcal{S}_1}^\prime - \boldsymbol{x}_{i\mathcal{S}_2}^\top\boldsymbol{\gamma}_{k\mathcal{S}_2}^\prime\right)^2\Bigg\}\Bigg] \\
&+ \frac{1}{2n}\sum_{i=1}^{n}\left[L_k^l\left(y_i, \boldsymbol{x}_i, \boldsymbol{\Omega}_0^l, \boldsymbol{\gamma}_0\right)\left(y_i - \boldsymbol{x}_{i\mathcal{S}_1}^\top\boldsymbol{\alpha}_{k\mathcal{S}_1}^\prime - \boldsymbol{x}_{i\mathcal{S}_2}^\top\boldsymbol{\gamma}_{k\mathcal{S}_2}^\prime\right)^2\right](\sigma_k^{l\prime} - \sigma_k^{l*}) \\
&= -\frac{1}{4n}\sum_{i=1}^{n}L_k^l\left(y_i, \boldsymbol{x}_i, \boldsymbol{\Omega}_0^l, \boldsymbol{\gamma}_0\right)\frac{(\sigma_k^{l\prime} - \sigma_k^{l*})^2}{(\sigma_k^{l*} + t(\sigma_k^{l\prime} - \sigma_k^{l*}))^2},
\end{aligned}
\tag{S1.20}
$$

where $t \in (0,1)$. So, $II \leqslant -\frac{1}{4n}\sum_{i=1}^{n}L_k^l\left(y_i, \boldsymbol{x}_i, \boldsymbol{\Omega}_0^l, \boldsymbol{\gamma}_0\right)\frac{(\sigma_k^{l\prime} - \sigma_k^{l*})^2}{(\kappa_2+\alpha)^2}$.

Since $0 \leqslant L_k^l\left(y_i, \boldsymbol{x}_i, \boldsymbol{\Omega}_0^l, \boldsymbol{\gamma}_0\right) \leqslant 1$, according to Hoeffding's inequality,

$$
\left|\frac{1}{n}\sum_{i=1}^{n}\left[L_k^l\left(y_i, \boldsymbol{x}_i, \boldsymbol{\Omega}_0^l, \boldsymbol{\gamma}_0\right) - \mathbb{E}\left(L_k^l\left(y_i, \boldsymbol{x}_i, \boldsymbol{\Omega}_0^l, \boldsymbol{\gamma}_0\right)\right)\right]\right| \leq \sqrt{\frac{1}{2n}\log(4K_0p)},
\tag{S1.21}
$$

with probability at least $1 - \frac{1}{2K_lp}$. Since $\sqrt{\frac{1}{2n}\log(2K_lp)} = o(1)$ and $0 \leq \mathbb{E}\left(L_k^l\left(y_i, \boldsymbol{x}_i, \boldsymbol{\Omega}_0^l, \boldsymbol{\gamma}_0\right)\right) \leq 1$, there exists some positive constant c such that $-\frac{1}{n}\sum_{i=1}^{n}L_k^l\left(y_i, \boldsymbol{x}_i, \boldsymbol{\Omega}_0^l, \boldsymbol{\gamma}_0\right) \leq -c$ when n is large enough.

For III, with Taylor expansion, we have:

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\left[L_k^l\left(y_i,\boldsymbol{x}_i,\boldsymbol{\Omega}_0^l,\boldsymbol{\gamma}_0\right)\left(y_i-\boldsymbol{x}_{i\mathcal{S}_1}^\top\boldsymbol{\alpha}_{k\mathcal{S}_1}^*-\boldsymbol{x}_{i\mathcal{S}_2}^\top\boldsymbol{\gamma}_{k\mathcal{S}_2}^*\right)\widetilde{\boldsymbol{x}}_{i\mathcal{S}}^\top\right]\\
&\qquad\qquad\operatorname{vec}(\boldsymbol{\alpha}_{k\mathcal{S}_1}'-\boldsymbol{\alpha}_{k\mathcal{S}_1}^*,\boldsymbol{\gamma}_{0k}'-\boldsymbol{\gamma}_{0k}^*)(\sigma_k^{l\prime}-\sigma_k^{l*})\\
&-\operatorname{vec}(\boldsymbol{\alpha}_{k\mathcal{S}_1}'-\boldsymbol{\alpha}_{k\mathcal{S}_1}^*,\boldsymbol{\gamma}_{0k}'-\boldsymbol{\gamma}_{0k}^*)^\top\left[\frac{\sigma_k^{l\prime}-\sigma_k^{l*}}{2n}\sum_{i=1}^{n}L_k^l\left(y_i,\boldsymbol{x}_i,\boldsymbol{\Omega}_0^l,\boldsymbol{\gamma}_0\right)\widetilde{\boldsymbol{x}}_{i\mathcal{S}}\widetilde{\boldsymbol{x}}_{i\mathcal{S}}^\top\right]\\
&\qquad\qquad\operatorname{vec}(\boldsymbol{\alpha}_{k\mathcal{S}_1}'-\boldsymbol{\alpha}_{k\mathcal{S}_1}^*,\boldsymbol{\gamma}_{0k}'-\boldsymbol{\gamma}_{0k}^*),
\end{aligned}
\tag{S1.22}
$$

where the second term is dominated by I. Note that the first term is small enough with the decomposition $L_k^l\left(y_i,\boldsymbol{x}_i,\boldsymbol{\Omega}_0^l,\boldsymbol{\gamma}_0\right)=\mathbb{E}L_k^l\left(y_i,\boldsymbol{x}_i,\boldsymbol{\Omega}_0^l,\boldsymbol{\gamma}_0\right)-(\mathbb{E}L_k^l\left(y_i,\boldsymbol{x}_i,\boldsymbol{\Omega}_0^l,\boldsymbol{\gamma}_0\right)-L_k^l\left(y_i,\boldsymbol{x}_i,\boldsymbol{\Omega}_0^l,\boldsymbol{\gamma}_0\right))$, where the first part of the first term can be small enough by the subgaussian tail, and the second part can be small enough similar to (S1.21). Then, III can be negative with a large enough n. So, we can get:

$$
\begin{aligned}
&\mathcal{H}_n^l(\boldsymbol{\Omega}_{0k}^{l\prime},\boldsymbol{\gamma}_{0k}'\mid\boldsymbol{\Omega}_0^l,\boldsymbol{\gamma}_0)-\mathcal{H}_n^l(\boldsymbol{\Omega}_{0k}^{l*},\boldsymbol{\gamma}_{0k}^*\mid\boldsymbol{\Omega}_0^l,\boldsymbol{\gamma}_0)\\
&-\left\langle\nabla_{\boldsymbol{\Omega}_{0k}^{l*}}\mathcal{H}_n\left(\boldsymbol{\Omega}_{0k}^{l*},\boldsymbol{\gamma}_{0k}^*\mid\boldsymbol{\Omega}_0^l,\boldsymbol{\gamma}_0\right),\boldsymbol{\Omega}_{0k}^{l\prime}-\boldsymbol{\Omega}_{0k}^{l*}\right\rangle\\
&-\left\langle\nabla_{\boldsymbol{\gamma}_{0k}^*}\mathcal{H}_n\left(\boldsymbol{\Omega}_{0k}^{l*},\boldsymbol{\gamma}_{0k}^*\mid\boldsymbol{\Omega}_0^l,\boldsymbol{\gamma}_0\right),\boldsymbol{\gamma}_{0k}'-\boldsymbol{\gamma}_{0k}^*\right\rangle\\
&\leqslant-\frac{\varrho}{2}\|\boldsymbol{\Omega}_{0k}^{l\prime}-\boldsymbol{\Omega}_{0k}^{l*}\|_2^2-\frac{\varrho}{2}\|\boldsymbol{\gamma}_{0k}'-\boldsymbol{\gamma}_{0k}^*\|_2^2,
\end{aligned}
\tag{S1.23}
$$

with probability at least $1-\frac{1}{K_l p}$, where $\varrho=c_1\min(C_0\kappa_1,\frac{(\kappa_2+\alpha)^{-2}}{2})$ and c_1 is a positive constant. Then, we can simply take the summation, which implies that:

$$\begin{aligned}
&\mathcal{H}_n^l(\boldsymbol{\Omega}_0^{l\prime}, \boldsymbol{\gamma}_0' \mid \boldsymbol{\Omega}_0^l, \boldsymbol{\gamma}_0) - \mathcal{H}_n^l(\boldsymbol{\Omega}_0^{l*}, \boldsymbol{\gamma}_0^* \mid \boldsymbol{\Omega}_0^l, \boldsymbol{\gamma}_0) \\
&\leqslant \left\langle \nabla_{\boldsymbol{\Omega}_0^{l*}} \mathcal{H}_n\left(\boldsymbol{\Omega}_0^{l*}, \boldsymbol{\gamma}_0^* \mid \boldsymbol{\Omega}_0^l, \boldsymbol{\gamma}_0\right), \boldsymbol{\Omega}_0^{l\prime} - \boldsymbol{\Omega}_0^{l*} \right\rangle - \frac{\varrho}{2}\|\boldsymbol{\Omega}_0^{l\prime} - \boldsymbol{\Omega}_0^{l*}\|_2^2 \qquad \text{(S1.24)} \\
&+ \left\langle \nabla_{\boldsymbol{\gamma}_0^*} \mathcal{H}_n\left(\boldsymbol{\Omega}_0^{l*}, \boldsymbol{\gamma}_0^* \mid \boldsymbol{\Omega}_0^l, \boldsymbol{\gamma}_0\right), \boldsymbol{\gamma}_0' - \boldsymbol{\gamma}_0^* \right\rangle - \frac{\varrho}{2}\|\boldsymbol{\gamma}_0' - \boldsymbol{\gamma}_0^*\|_2^2,
\end{aligned}$$

with probability at least $1 - \frac{1}{p}$. The upper-level concavity can be similarly obtained, which means that:

$$\begin{aligned}
\boldsymbol{q}(\boldsymbol{v}) \leqslant &\left\langle \nabla_{\boldsymbol{\Omega}_0^{u*}} \mathcal{H}_n\left(\boldsymbol{\Omega}_0^{u*} \mid \boldsymbol{\Omega}_0^{u(t-1)}\right), \boldsymbol{v}_1 \right\rangle - \frac{\varrho}{2}\|\boldsymbol{v}_1\|_2^2 \\
&+ \left\langle \nabla_{\boldsymbol{\Omega}_0^{l*}} \mathcal{H}_n\left(\boldsymbol{\Omega}_0^{l*}, \boldsymbol{\gamma}_0^* \mid \boldsymbol{\Omega}_0^{l(t-1)}, \boldsymbol{\gamma}_0^{(t-1)}\right), \boldsymbol{v}_2 \right\rangle - \frac{\varrho}{2}\|\boldsymbol{v}_2\|_2^2 \\
&+ \left\langle \nabla_{\boldsymbol{\gamma}_0^*} \mathcal{H}_n\left(\boldsymbol{\Omega}_0^{l*}, \boldsymbol{\gamma}_0^* \mid \boldsymbol{\Omega}_0^{l(t-1)}, \boldsymbol{\gamma}_0^{(t-1)}\right), \boldsymbol{v}_3 \right\rangle - \frac{\varrho}{2}\|\boldsymbol{v}_3\|_2^2, \\
&- \left[\mathcal{P}_1\left(\boldsymbol{\Omega}_0^* + \boldsymbol{v}\right) - \mathcal{P}_1\left(\boldsymbol{\Omega}_0^*\right)\right].
\end{aligned}$$

We focus on the first term in the second line:

$$\begin{aligned}
&\left\langle \nabla_{\boldsymbol{\Omega}_0^{l*}} \mathcal{H}_n\left(\boldsymbol{\Omega}_0^{l*}, \boldsymbol{\gamma}_0^* \mid \boldsymbol{\Omega}_0^{l(t-1)}, \boldsymbol{\gamma}_0^{(t-1)}\right), \boldsymbol{v}_2 \right\rangle \\
=&\left\langle \nabla_{\boldsymbol{\Omega}_0^{l*}} \mathcal{H}_n\left(\boldsymbol{\Omega}_0^{l*}, \boldsymbol{\gamma}_0^* \mid \boldsymbol{\Omega}_0^{l(t-1)}, \boldsymbol{\gamma}_0^{(t-1)}\right) - \nabla_{\boldsymbol{\Omega}_0^{l*}} \mathcal{H}\left(\boldsymbol{\Omega}_0^{l*}, \boldsymbol{\gamma}_0^* \mid \boldsymbol{\Omega}_0^{l(t-1)}, \boldsymbol{\gamma}_0^{(t-1)}\right) \right. \\
&\left. + \nabla_{\boldsymbol{\Omega}_0^{l*}} \mathcal{H}\left(\boldsymbol{\Omega}_0^{l*}, \boldsymbol{\gamma}_0^* \mid \boldsymbol{\Omega}_0^{l(t-1)}, \boldsymbol{\gamma}_0^{(t-1)}\right) - \nabla_{\boldsymbol{\Omega}_0^{l*}} \mathcal{H}\left(\boldsymbol{\Omega}_0^{l*}, \boldsymbol{\gamma}_0^* \mid \boldsymbol{\Omega}_0^{l*}, \boldsymbol{\gamma}_0^*\right), \boldsymbol{v}_2 \right\rangle, \qquad \text{(S1.25)}
\end{aligned}$$

where the equation follows from the self-consistency property of the population version of the objective function (McLachlan and Krishnan, 2007), that is, $\boldsymbol{\Omega}^* = \operatorname{argmax}_{\boldsymbol{\Omega}'} \mathcal{H}\left(\boldsymbol{\Omega}' \mid \boldsymbol{\Omega}^*\right)$.

Following Lemma 1, we have:

$$
\begin{aligned}
&\left\langle \nabla_{\boldsymbol{\Omega}_0^{l*}} \mathcal{H}_n \left(\boldsymbol{\Omega}_0^{l*}, \boldsymbol{\gamma}_0^* \mid \boldsymbol{\Omega}_0^{l(t-1)}, \boldsymbol{\gamma}_0^{(t-1)} \right) - \nabla_{\boldsymbol{\Omega}_0^{l*}} \mathcal{H} \left(\boldsymbol{\Omega}_0^{l*}, \boldsymbol{\gamma}_0^* \mid \boldsymbol{\Omega}_0^{l(t-1)}, \boldsymbol{\gamma}_0^{(t-1)} \right), \boldsymbol{v}_2 \right\rangle \\
&\leqslant \left\| \nabla_{\boldsymbol{\Omega}_0^{l*}} \mathcal{H}_n \left(\boldsymbol{\Omega}_0^{l*}, \boldsymbol{\gamma}_0^* \mid \boldsymbol{\Omega}_0^{l(t-1)}, \boldsymbol{\gamma}_0^{(t-1)} \right) - \nabla_{\boldsymbol{\Omega}_0^{l*}} \mathcal{H} \left(\boldsymbol{\Omega}_0^{l*}, \boldsymbol{\gamma}_0^* \mid \boldsymbol{\Omega}_0^{l(t-1)}, \boldsymbol{\gamma}_0^{(t-1)} \right) \right\|_\infty \|\boldsymbol{v}_2\|_1 \\
&\leqslant \epsilon_2 \sqrt{K_l(s+1)} \, \|\boldsymbol{v}_2\|_2 ,
\end{aligned}
\tag{S1.26}
$$

with probability at least $1 - \frac{8K_l^2+4K_l}{p}$, where $\epsilon_2 = C_2 K_l \sqrt{\frac{\log p}{n}}$ and C_2 is a positive constant.

By Condition 3 and Lemma 2, we have:

$$
\begin{aligned}
&\left\langle \nabla_{\boldsymbol{\Omega}_0^{l*}} \mathcal{H} \left(\boldsymbol{\Omega}_0^{l*}, \boldsymbol{\gamma}_0^* \mid \boldsymbol{\Omega}_0^{l(t-1)}, \boldsymbol{\gamma}_0^{(t-1)} \right) - \nabla_{\boldsymbol{\Omega}_0^{l*}} \mathcal{H} \left(\boldsymbol{\Omega}_0^{l*}, \boldsymbol{\gamma}_0^* \mid \boldsymbol{\Omega}_0^{l*}, \boldsymbol{\gamma}_0^* \right), \boldsymbol{v}_2 \right\rangle \\
&\leqslant \left\| \nabla_{\boldsymbol{\Omega}_0^{l*}} \mathcal{H} \left(\boldsymbol{\Omega}_0^{l*}, \boldsymbol{\gamma}_0^* \mid \boldsymbol{\Omega}_0^{l(t-1)}, \boldsymbol{\gamma}_0^{(t-1)} \right) - \nabla_{\boldsymbol{\Omega}_0^{l*}} \mathcal{H} \left(\boldsymbol{\Omega}_0^{l*}, \boldsymbol{\gamma}_0^* \mid \boldsymbol{\Omega}_0^{l*}, \boldsymbol{\gamma}_0^* \right) \right\|_2 \|\boldsymbol{v}_2\|_2 \\
&\leqslant \tau \left\| \boldsymbol{\Omega}_0^* - \boldsymbol{\Omega}_0^{(t-1)} \right\|_2 \|\boldsymbol{v}_2\|_2 .
\end{aligned}
\tag{S1.27}
$$

For the penalty term, we have:

$$
\begin{aligned}
&\left[\mathcal{P}_1 \left(\boldsymbol{\Omega}_0^* + \boldsymbol{v} \right) - \mathcal{P}_1 \left(\boldsymbol{\Omega}_0^* \right) \right] \\
&\leqslant C_3 \sqrt{\frac{K_l^3 s \log p}{n}} \|\boldsymbol{v}_1\|_2 + C_4 \sqrt{\frac{K_l^3 s \log p}{n}} \|\boldsymbol{v}_2\|_2 + C_5 K_l \sqrt{\frac{K_l^3 s_2 \log p}{n}} \|\boldsymbol{v}_3\|_2 .
\end{aligned}
\tag{S1.28}
$$

With Lasso, we need to set the order of λ_1 as $O(\sqrt{\frac{K_l^3 \log p}{n}})$. Further, note that this term can be negative with MCP and SCAD, combining with the minimal signal condition of the true parameters.

Then, with a properly chosen positive constant C and $\epsilon_u = C\sqrt{\frac{K_l^3 s \log p}{n}}$,

$$C_3\sqrt{\frac{K_l^3 s \log p}{n}} + \epsilon_2\sqrt{K_l(s+1)} \leqslant \epsilon_u. \tag{S1.29}$$

By treating $\left\langle \nabla_{\boldsymbol{\Omega}_0^{u*}} \mathcal{H}_n\left(\boldsymbol{\Omega}_0^{u*} \mid \boldsymbol{\Omega}_0^{u(t-1)}\right), \boldsymbol{v}_1 \right\rangle$ and $\left\langle \nabla_{\boldsymbol{\gamma}_0^*} \mathcal{H}_n\left(\boldsymbol{\Omega}_0^{l*}, \boldsymbol{\gamma}_0^* \mid \boldsymbol{\Omega}_0^{l(t-1)}, \boldsymbol{\gamma}_0^{(t-1)}\right), \boldsymbol{v}_3 \right\rangle$ similarly, there can also be $\epsilon_l = C\sqrt{\frac{K_l^3 s_2 \log p}{n}}$ such that:

$$\begin{aligned}
\boldsymbol{q}(\boldsymbol{v}) \leqslant & -\frac{\varrho}{2}\|\boldsymbol{v}_1\|_2^2 + \left(\epsilon_u + \tau\left\|\boldsymbol{\Omega}_0^* - \boldsymbol{\Omega}_0^{(t-1)}\right\|_2\right)\|\boldsymbol{v}_1\|_2 \\
& -\frac{\varrho}{2}\|\boldsymbol{v}_2\|_2^2 + \left(\epsilon_u + \tau\left\|\boldsymbol{\Omega}_0^* - \boldsymbol{\Omega}_0^{(t-1)}\right\|_2\right)\|\boldsymbol{v}_2\|_2 \\
& -\frac{\varrho}{2}\|\boldsymbol{v}_3\|_2^2 + \left(\epsilon_l + \tau\sqrt{\frac{s_2}{s_1}}\left\|\boldsymbol{\Omega}_0^* - \boldsymbol{\Omega}_0^{(t-1)}\right\|_2\right)\|\boldsymbol{v}_3\|_2,
\end{aligned}$$

with probability at least $1 - \frac{18K_l^2 + 10K_l + 1}{p}$. The term $\sqrt{s_2/s_1}$ is based on Condition 3 and the ratio of $\|\boldsymbol{x}_{\mathcal{S}_2}\|/\|\boldsymbol{x}_{\mathcal{S}_1}\|$ by considering (S1.64).

We can see that $\boldsymbol{q}(\boldsymbol{v}) < 0$ when $\|\boldsymbol{v}_1\|_2 > \frac{2\epsilon_u}{\varrho} + \frac{2\tau}{\varrho}\left\|\boldsymbol{\Omega}_0^{(t-1)} - \boldsymbol{\Omega}_0^*\right\|_2$, $\|\boldsymbol{v}_2\|_2 > \frac{2\epsilon_u}{\varrho} + \frac{2\tau}{\varrho}\left\|\boldsymbol{\Omega}_0^{(t-1)} - \boldsymbol{\Omega}_0^*\right\|_2$, and $\|\boldsymbol{v}_3\|_2 > \frac{2\epsilon_l}{\varrho} + \frac{2\tau}{\varrho}\sqrt{\frac{s_2}{s_1}}\left\|\boldsymbol{\Omega}_0^{(t-1)} - \boldsymbol{\Omega}_0^*\right\|_2$. Note that $\|\boldsymbol{v}_1\| = \chi_1$, $\|\boldsymbol{v}_2\| = \chi_2$, $\|\boldsymbol{v}_\gamma\| = \chi_\gamma$ and $\chi_1 = \frac{4\epsilon_1}{\varrho} + \iota\left\|\boldsymbol{\Omega}_0^{(t-1)} - \boldsymbol{\Omega}_0^*\right\|_2$, $\chi_2 = \frac{4\epsilon_2}{\varrho} + \iota\sqrt{\frac{s_2}{s_1}}\left\|\boldsymbol{\Omega}_0^{(t-1)} - \boldsymbol{\Omega}_0^*\right\|_2$, $\chi_\gamma = \frac{4\epsilon_\gamma}{\varrho} + \iota\left\|\boldsymbol{\Omega}_0^{(t-1)} - \boldsymbol{\Omega}_0^*\right\|_2$, $\frac{1}{6} \leq \iota < 1$, and $\tau \leqslant \varrho/12$. Therefore, there is a local maximizer $\boldsymbol{\Omega}_0^{(t)}$ following $\boldsymbol{\Omega}_0^{(t-1)}$ that satisfies: if $\boldsymbol{\Omega}_0^{(t-1)} \in \mathcal{B}_\alpha(\boldsymbol{\Omega}_0^*)$, then $\left\|\boldsymbol{\Omega}_0^{u(t)} - \boldsymbol{\Omega}_0^{u*}\right\|_2 \leqslant \chi_1$, $\left\|\boldsymbol{\Omega}_0^{l(t)} - \boldsymbol{\Omega}_0^{l*}\right\|_2 \leqslant \chi_2$, and $\left\|\boldsymbol{\gamma}_0^{(t)} - \boldsymbol{\gamma}_0^*\right\|_2 \leqslant \chi_\gamma$ with probability at least $1 - \frac{18K_l^2 + 10K_l + 1}{p}$.

Then, we can show that, if $\boldsymbol{\Omega}_0^{(0)} \in \mathcal{B}_\alpha(\boldsymbol{\Omega}_0^*)$, for any $t \geqslant 1$,

$$\left\|\boldsymbol{\Omega}_0^{(t)} - \boldsymbol{\Omega}_0^*\right\|_2 \leqslant \frac{1-\iota^t}{1-\iota}\frac{4\epsilon_u}{\varrho} + \iota^t\left\|\boldsymbol{\Omega}_0^{(0)} - \boldsymbol{\Omega}_0^*\right\|_2 \leqslant \frac{8\epsilon_u}{(1-\iota)\varrho} + \iota^t\left\|\boldsymbol{\Omega}_0^{(0)} - \boldsymbol{\Omega}_0^*\right\|_2, \tag{S1.30}$$

with probability at least $1 - t\left(18K_l^2 + 10K_l + 1\right)/p$. Note that, when $t \geqslant \hat{t} := \log_{1/\iota}\left(\frac{(1-\iota)\varrho\left\|\boldsymbol{\Omega}_0^{(0)}-\boldsymbol{\Omega}_0^*\right\|_2}{\epsilon_l}\right)$, $\iota^t\left\|\boldsymbol{\Omega}_0^{(0)} - \boldsymbol{\Omega}_0^*\right\|_2$ is dominated by $\frac{8\epsilon_u}{(1-\iota)\varrho}$ and $\frac{8\epsilon_l}{(1-\iota)\varrho}$. So, the final error can be bounded as:

$$\left\|\widehat{\boldsymbol{\Omega}}_0 - \boldsymbol{\Omega}_0^*\right\|_2 = O\left(\sqrt{\frac{K_l^3 s \log p}{n}}\right), \quad \|\widehat{\boldsymbol{\gamma}}_0 - \boldsymbol{\gamma}_0^*\|_2 = O\left(\sqrt{\frac{K_l^3 s_2 \log p}{n}}\right), \tag{S1.31}$$

with probability at least $1-t\left(18K_l^2 + 10K_l + 1\right)/p$. And note that $\hat{t}\left(18K_l^2 + 10K_l + 1\right)/p$ goes to 0 as p and n diverge by Condition 4.

Step 2:

We show that $\widehat{\boldsymbol{\Omega}} = \left(\widehat{\boldsymbol{\Omega}}_0, \widehat{\boldsymbol{\Omega}}_{\mathcal{S}^c}\right) = \left(\widehat{\boldsymbol{\Omega}}_0, 0\right)$ is a local maximizer of the objective function. This indicates that

$$\left\|\widehat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}^*\right\|_2 = O\left(\sqrt{\frac{K_l^3 s \log p}{n}}\right), \quad \|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_2 = O\left(\sqrt{\frac{K_l^3 s_2 \log p}{n}}\right), \tag{S1.32}$$

in Step 1 hold due to the conclusion drawn in this step. Following Theorem 1 in Fan and Lv (2011), it suffices to show that $\left\|\mathrm{vec}(\boldsymbol{z}_\beta, \boldsymbol{z}_\alpha)\right\|_2 \leqslant \lambda_1$ and

$\|\boldsymbol{z}_\gamma\|_2 \leqslant \lambda_2$, where

$$\begin{pmatrix} \nabla_{\boldsymbol{\beta}_{\mathcal{S}_1}} \mathcal{Q}(\widehat{\boldsymbol{\Omega}}) \\ \nabla_{\boldsymbol{\alpha}_{\mathcal{S}_1}} \mathcal{Q}(\widehat{\boldsymbol{\Omega}}) \\ \nabla_{\boldsymbol{\gamma}_{\mathcal{S}_2}} \mathcal{Q}(\widehat{\boldsymbol{\Omega}}) \\ \nabla_{\boldsymbol{\beta}_j} \mathcal{Q}(\widehat{\boldsymbol{\Omega}}) \\ \nabla_{\boldsymbol{\alpha}_j} \mathcal{Q}(\widehat{\boldsymbol{\Omega}}) \\ \nabla_{\boldsymbol{\gamma}_l} \mathcal{Q}(\widehat{\boldsymbol{\Omega}}) \end{pmatrix} = \begin{pmatrix} \nabla_{\boldsymbol{\beta}_{\mathcal{S}_1}} \mathcal{Q}(\boldsymbol{\Omega}^*) \\ \nabla_{\boldsymbol{\alpha}_{\mathcal{S}_1}} \mathcal{Q}(\boldsymbol{\Omega}^*) \\ \nabla_{\boldsymbol{\gamma}_{\mathcal{S}_2}} \mathcal{Q}(\boldsymbol{\Omega}^*) \\ \nabla_{\boldsymbol{\beta}_j} \mathcal{Q}(\boldsymbol{\Omega}^*) \\ \nabla_{\boldsymbol{\alpha}_j} \mathcal{Q}(\boldsymbol{\Omega}^*) \\ \nabla_{\boldsymbol{\gamma}_l} \mathcal{Q}(\boldsymbol{\Omega}^*) \end{pmatrix} + \mathcal{A} \cdot \begin{pmatrix} \widehat{\boldsymbol{\beta}}_{\mathcal{S}_1} - \boldsymbol{\beta}^*_{\mathcal{S}_1} \\ \widehat{\boldsymbol{\alpha}}_{\mathcal{S}_1} - \boldsymbol{\alpha}^*_{\mathcal{S}_1} \\ \widehat{\boldsymbol{\gamma}}_{\mathcal{S}_2} - \boldsymbol{\gamma}^*_{\mathcal{S}_2} \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix} + \begin{pmatrix} \boldsymbol{r}(\boldsymbol{\Delta})_{\boldsymbol{\beta}_{\mathcal{S}_1}} \\ \boldsymbol{r}(\boldsymbol{\Delta})_{\boldsymbol{\alpha}_{\mathcal{S}_1}} \\ \boldsymbol{r}(\boldsymbol{\Delta})_{\boldsymbol{\gamma}_{\mathcal{S}_2}} \\ \boldsymbol{r}(\boldsymbol{\Delta})_{\boldsymbol{\beta}_j} \\ \boldsymbol{r}(\boldsymbol{\Delta})_{\boldsymbol{\alpha}_j} \\ \boldsymbol{r}(\boldsymbol{\Delta})_{\boldsymbol{\gamma}_l} \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \\ \boldsymbol{z}_\beta \\ \boldsymbol{z}_\alpha \\ \boldsymbol{z}_\gamma \end{pmatrix}, \tag{S1.33}$$

and $\boldsymbol{r}(\boldsymbol{\Delta})$ is the residual of the first-order Taylor expansion of the gradient.

Here, $\mathcal{A} =$

$$\begin{pmatrix} \nabla^2_{\boldsymbol{\beta}_{\mathcal{S}_1}\boldsymbol{\beta}_{\mathcal{S}_1}} \mathcal{Q}(\boldsymbol{\Omega}^*) & \mathbf{0} & \mathbf{0} & \nabla^2_{\boldsymbol{\beta}_{\mathcal{S}_1}\boldsymbol{\beta}_j} \mathcal{Q}(\boldsymbol{\Omega}^*) & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \nabla^2_{\boldsymbol{\alpha}_{\mathcal{S}_1}\boldsymbol{\alpha}_{\mathcal{S}_1}} \mathcal{Q}(\boldsymbol{\Omega}^*) & \nabla^2_{\boldsymbol{\alpha}_{\mathcal{S}_1}\boldsymbol{\gamma}_{\mathcal{S}_2}} \mathcal{Q}(\boldsymbol{\Omega}^*) & \mathbf{0} & \nabla^2_{\boldsymbol{\alpha}_{\mathcal{S}_1}\boldsymbol{\alpha}_j} \mathcal{Q}(\boldsymbol{\Omega}^*) & \nabla^2_{\boldsymbol{\alpha}_{\mathcal{S}_1}\boldsymbol{\gamma}_l} \mathcal{Q}(\boldsymbol{\Omega}^*) \\ \mathbf{0} & \nabla^2_{\boldsymbol{\gamma}_{\mathcal{S}_2}\boldsymbol{\alpha}_{\mathcal{S}_1}} \mathcal{Q}(\boldsymbol{\Omega}^*) & \nabla^2_{\boldsymbol{\gamma}_{\mathcal{S}_2}\boldsymbol{\gamma}_{\mathcal{S}_2}} \mathcal{Q}(\boldsymbol{\Omega}^*) & \mathbf{0} & \nabla^2_{\boldsymbol{\gamma}_{\mathcal{S}_2}\boldsymbol{\alpha}_j} \mathcal{Q}(\boldsymbol{\Omega}^*) & \nabla^2_{\boldsymbol{\gamma}_{\mathcal{S}_2}\boldsymbol{\gamma}_l} \mathcal{Q}(\boldsymbol{\Omega}^*) \\ \nabla^2_{\boldsymbol{\beta}_j\boldsymbol{\beta}_{\mathcal{S}_1}} \mathcal{Q}(\boldsymbol{\Omega}^*) & \mathbf{0} & \mathbf{0} & \nabla^2_{\boldsymbol{\beta}_j\boldsymbol{\beta}_j} \mathcal{Q}(\boldsymbol{\Omega}^*) & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \nabla^2_{\boldsymbol{\alpha}_j\boldsymbol{\alpha}_{\mathcal{S}_1}} \mathcal{Q}(\boldsymbol{\Omega}^*) & \nabla^2_{\boldsymbol{\alpha}_j\boldsymbol{\gamma}_{\mathcal{S}_2}} \mathcal{Q}(\boldsymbol{\Omega}^*) & \mathbf{0} & \nabla^2_{\boldsymbol{\alpha}_j\boldsymbol{\alpha}_j} \mathcal{Q}(\boldsymbol{\Omega}^*) & \nabla^2_{\boldsymbol{\alpha}_j\boldsymbol{\gamma}_l} \mathcal{Q}(\boldsymbol{\Omega}^*) \\ \mathbf{0} & \nabla^2_{\boldsymbol{\gamma}_l\boldsymbol{\alpha}_{\mathcal{S}_1}} \mathcal{Q}(\boldsymbol{\Omega}^*) & \nabla^2_{\boldsymbol{\gamma}_l\boldsymbol{\gamma}_{\mathcal{S}_2}} \mathcal{Q}(\boldsymbol{\Omega}^*) & \mathbf{0} & \nabla^2_{\boldsymbol{\gamma}_l\boldsymbol{\alpha}_j} \mathcal{Q}(\boldsymbol{\Omega}^*) & \nabla^2_{\boldsymbol{\gamma}_l\boldsymbol{\gamma}_l} \mathcal{Q}(\boldsymbol{\Omega}^*) \end{pmatrix}. \tag{S1.34}$$

Note that here $j \in \mathcal{S}_1^c$ and $l \in \mathcal{S}_2^c \cap \mathcal{S}_1^c = \mathcal{S}^c$, $\boldsymbol{\alpha}_j = (\alpha_{1j}, \alpha_{2j}, \cdots, \alpha_{K_lj},)^\top$, and α_{kj} is the j-th element of $\boldsymbol{\alpha}_k$ for the k-th subgroup. $\boldsymbol{\beta}_j$ and $\boldsymbol{\gamma}_l$ are similarly defined.

First consider $\boldsymbol{z}_\beta$. For the k-th coordinate, we have:

$$
\begin{aligned}
(\boldsymbol{z}_\beta)_k &= (\nabla_{\boldsymbol{\beta}_j}\mathcal{Q}(\boldsymbol{\Omega}^*))_k + (\nabla^2_{\boldsymbol{\beta}_j\boldsymbol{\beta}_{\mathcal{S}_1}}\mathcal{Q}(\boldsymbol{\Omega}^*)(\widehat{\boldsymbol{\beta}}_{\mathcal{S}_1} - \boldsymbol{\beta}^*_{\mathcal{S}_1}))_k + (\boldsymbol{r}(\boldsymbol{\Delta})_{\boldsymbol{\beta}_j})_k \\
&= \frac{\sigma_k^{u*}}{n}\sum_{i=1}^n L_k^u\left(y_i, \boldsymbol{x}_i, \boldsymbol{\Omega}^{u*}\right)\left(y_i - \boldsymbol{x}_{i\mathcal{S}_1}^\top \boldsymbol{\beta}^*_{k\mathcal{S}_1}\right)\boldsymbol{x}_{ij} \\
&+ \frac{\sigma_k^{u*}}{n}\sum_{i=1}^n L_k^u\left(y_i, \boldsymbol{x}_i, \boldsymbol{\Omega}^{u*}\right)\boldsymbol{x}_{i\mathcal{S}_1}^\top \boldsymbol{x}_{ij}(\widehat{\boldsymbol{\beta}}_{k\mathcal{S}_1} - \boldsymbol{\beta}^*_{k\mathcal{S}_1}) + (\boldsymbol{r}(\boldsymbol{\Delta})_{\boldsymbol{\beta}_j})_k \\
&= (\xi_{\beta 1})_k + (\xi_{\beta 2})_k + (\boldsymbol{r}(\boldsymbol{\Delta})_{\boldsymbol{\beta}_j})_k, \quad , k = 1, 2, \cdots, K_u.
\end{aligned}
\tag{S1.35}
$$

According to Condition 1 and with bounded x_{ij}, we have:

$$
\begin{aligned}
|(\xi_{\beta 1})_k| \leqslant & \kappa_2 \left|\frac{1}{n}\sum_{i=1}^n \mathbb{E}(L_k^u\left(y_i, \boldsymbol{x}_i, \boldsymbol{\Omega}^{u*}\right))\left(y_i - \boldsymbol{x}_{i\mathcal{S}_1}^\top \boldsymbol{\beta}^*_{k\mathcal{S}_1}\right)\boldsymbol{x}_{ij}\right| \\
& + \kappa_2 \left|\frac{1}{n}\sum_{i=1}^n (L_k^u\left(y_i, \boldsymbol{x}_i, \boldsymbol{\Omega}^{u*}\right) - \mathbb{E}(L_k^u\left(y_i, \boldsymbol{x}_i, \boldsymbol{\Omega}^{u*}\right))\left(y_i - \boldsymbol{x}_{i\mathcal{S}_1}^\top \boldsymbol{\beta}^*_{k\mathcal{S}_1}\right)\boldsymbol{x}_{ij}\right| \\
\leqslant & \kappa_2 \left|\frac{1}{n}\sum_{i=1}^n \mathbb{E}(L_k^u\left(y_i, \boldsymbol{x}_i, \boldsymbol{\Omega}^{u*}\right))\left(y_i - \boldsymbol{x}_{i\mathcal{S}_1}^\top \boldsymbol{\beta}^*_{k\mathcal{S}_1}\right)\right| \max|\boldsymbol{x}_{ij}| \\
& + \kappa_2 \left|\frac{1}{n}\sum_{i=1}^n (L_k^u\left(y_i, \boldsymbol{x}_i, \boldsymbol{\Omega}^{u*}\right) - \mathbb{E}(L_k^u\left(y_i, \boldsymbol{x}_i, \boldsymbol{\Omega}^{u*}\right))\right| \max\left|\left(y_i - \boldsymbol{x}_{i\mathcal{S}_1}^\top \boldsymbol{\beta}^*_{k\mathcal{S}_1}\right)\boldsymbol{x}_{ij}\right|.
\end{aligned}
\tag{S1.36}
$$

Here, the first term can be bounded by Hoeffding inequality due to the gaussian distribution of $\mathbb{E}(L_k^l\left(y_i, \boldsymbol{x}_i, \boldsymbol{\Omega}^{u*}\right))\left(y_i - \boldsymbol{x}_{i\mathcal{S}_1}^\top \boldsymbol{\beta}^*_{k\mathcal{S}_1}\right)$. The second term can be bounded by Hoeffding inequality similar to (S1.21). So, we can get that there exists a positive constant $c_{\beta 1}$, such that $|(\xi_{\beta 1})_k| \leqslant c_{\beta 1}\sqrt{\frac{\log p}{n}}$ with probability at least $1 - \frac{1}{p}$, which tends to 1. Then by Condition 4, we have $\|\xi_{\beta 1}\|_2 \leqslant c_{\beta 1}\sqrt{\frac{K_u \log p}{n}}$ with probability tending to 1.

According to (S1.33), we have:

$$\widehat{\boldsymbol{\beta}}_{\mathcal{S}_1} - \boldsymbol{\beta}^*_{\mathcal{S}} = -\left[\nabla^2_{\boldsymbol{\beta}_{\mathcal{S}_1}\boldsymbol{\beta}_{\mathcal{S}_1}} \mathcal{Q}(\boldsymbol{\Omega}^*)\right]^{-1}\left[\nabla_{\boldsymbol{\beta}_{\mathcal{S}_1}} \mathcal{Q}(\boldsymbol{\Omega}^*) + \boldsymbol{r}(\boldsymbol{\Delta})_{\boldsymbol{\beta}_{\mathcal{S}_1}}\right]. \tag{S1.37}$$

So, we can get that:

$$\begin{aligned}(\xi_{\beta 2})_k = & -\frac{\sigma_k^{u*}}{n}\sum_{i=1}^{n} L_k^u\left(y_i, \boldsymbol{x}_i, \boldsymbol{\Omega}^{u*}\right)\boldsymbol{x}_{i\mathcal{S}_1}^{\top}\boldsymbol{x}_{ij}\left[\nabla^2_{\boldsymbol{\beta}_{k\mathcal{S}_1}\boldsymbol{\beta}_{k\mathcal{S}_1}} \mathcal{Q}(\boldsymbol{\Omega}^*)\right]^{-1}\left[\nabla_{\boldsymbol{\beta}_{k\mathcal{S}_1}} \mathcal{Q}(\boldsymbol{\Omega}^*) + \boldsymbol{r}(\boldsymbol{\Delta})_{\boldsymbol{\beta}_{k\mathcal{S}_1}}\right] \\ = & -\left[\frac{1}{n}\sum_{i=1}^{n} L_k^u\left(y_i, \boldsymbol{x}_i, \boldsymbol{\Omega}^{u*}\right)\boldsymbol{x}_{i\mathcal{S}_1}^{\top}\boldsymbol{x}_{ij}\right]\left[\frac{1}{n}\sum_{i=1}^{n} L_k^u\left(y_i, \boldsymbol{x}_i, \boldsymbol{\Omega}^{u*}\right)\boldsymbol{x}_{i\mathcal{S}_1}^{\top}\boldsymbol{x}_{i\mathcal{S}_1}\right)\right]^{-1} \\ & \left[\frac{\sigma_k^{u*}}{n}\sum_{i=1}^{n} L_k^u\left(y_i, \boldsymbol{x}_i, \boldsymbol{\Omega}^{u*}\right)\left(y_i - \boldsymbol{x}_{i\mathcal{S}_1}^{\top}\boldsymbol{\beta}^*_{k\mathcal{S}_1}\right)\boldsymbol{x}_{i\mathcal{S}_1} + \boldsymbol{r}(\boldsymbol{\Delta})_{\boldsymbol{\beta}_{k\mathcal{S}_1}}\right].\end{aligned} \tag{S1.38}$$

We bound it by:

$$\begin{aligned}|(\xi_{\beta 2})_k| \leqslant & \left\|\left[\frac{1}{n}\sum_{i=1}^{n} L_k^u\left(y_i, \boldsymbol{x}_i, \boldsymbol{\Omega}^{u*}\right)\boldsymbol{x}_{i\mathcal{S}_1}^{\top}\boldsymbol{x}_{ij}\right]\left[\frac{1}{n}\sum_{i=1}^{n} L_k^u\left(y_i, \boldsymbol{x}_i, \boldsymbol{\Omega}^{u*}\right)\boldsymbol{x}_{i\mathcal{S}_1}^{\top}\boldsymbol{x}_{i\mathcal{S}_1}\right)\right]^{-1}\right\|_1 \\ & \left\|\frac{\sigma_k^{u*}}{n}\sum_{i=1}^{n} L_k^u\left(y_i, \boldsymbol{x}_i, \boldsymbol{\Omega}^{u*}\right)\left(y_i - \boldsymbol{x}_{i\mathcal{S}_1}^{\top}\boldsymbol{\beta}^*_{k\mathcal{S}_1}\right)\boldsymbol{x}_{i\mathcal{S}_1} + \boldsymbol{r}(\boldsymbol{\Delta})_{\boldsymbol{\beta}_{k\mathcal{S}_1}}\right\|_\infty \\ \leqslant & \left\|\left[\frac{1}{n}\boldsymbol{X}_{\mathcal{S}_1^c}^{\top}\boldsymbol{G}_k^u\boldsymbol{X}_{\mathcal{S}_1}\right]\left[\frac{1}{n}\boldsymbol{X}_{\mathcal{S}_1}^{\top}\boldsymbol{G}_k^u\boldsymbol{X}_{\mathcal{S}_1}\right]^{-1}\right\|_\infty \\ & \left(\left\|\frac{\sigma_k^{u*}}{n}\sum_{i=1}^{n} L_k^u\left(y_i, \boldsymbol{x}_i, \boldsymbol{\Omega}^{u*}\right)\left(y_i - \boldsymbol{x}_{i\mathcal{S}_1}^{\top}\boldsymbol{\beta}^*_{k\mathcal{S}_1}\right)\boldsymbol{x}_{i\mathcal{S}_1}\right\|_\infty + \left\|\boldsymbol{r}(\boldsymbol{\Delta})_{\boldsymbol{\beta}_{k\mathcal{S}_1}}\right\|_\infty\right).\end{aligned} \tag{S1.39}$$

Here, the norm $\left\|\frac{\sigma_k^{u*}}{n}\sum_{i=1}^{n} L_k^l\left(y_i, \boldsymbol{x}_i, \boldsymbol{\Omega}^{u*}\right)\left(y_i - \boldsymbol{x}_{i\mathcal{S}_1}^{\top}\boldsymbol{\beta}^*_{k\mathcal{S}_1}\right)\boldsymbol{x}_{i\mathcal{S}_1}\right\|_\infty$ can be bounded similarly as $|(\xi_{\beta 1})_k|$. Overall, there exists a positive constant $c_{\beta 2}$, $\|\xi_{\beta 2}\|_2 \leqslant h_u n^{\eta_1}(\sqrt{\frac{K_u \log p}{n}} + \sqrt{K_u}\left\|\boldsymbol{r}(\boldsymbol{\Delta})_{\boldsymbol{\beta}_{\mathcal{S}_1}}\right\|_\infty)$ with probability tending to 1 by applying Condition 6.

Similarly, we get:

$$
\begin{aligned}
(\boldsymbol{z}_{\alpha})_k &= (\nabla_{\boldsymbol{\alpha}_j} \mathcal{Q}(\boldsymbol{\Omega}^*))_k + (\nabla^2_{\boldsymbol{\alpha}_j \boldsymbol{\alpha}_{\mathcal{S}_1}} \mathcal{Q}(\boldsymbol{\Omega}^*)(\widehat{\boldsymbol{\alpha}}_{\mathcal{S}_1} - \boldsymbol{\alpha}^*_{\mathcal{S}_1}))_k \\
&\quad + (\nabla^2_{\boldsymbol{\alpha}_j \boldsymbol{\gamma}_{\mathcal{S}_2}} \mathcal{Q}(\boldsymbol{\Omega}^*)(\widehat{\boldsymbol{\gamma}}_{\mathcal{S}_2} - \boldsymbol{\gamma}^*_{\mathcal{S}_2}))_k + (\boldsymbol{r}(\boldsymbol{\Delta})_{\boldsymbol{\alpha}_j})_k \\
&= \frac{\sigma_k^{l*}}{n} \sum_{i=1}^n L_k^l \left(y_i, \boldsymbol{x}_i, \boldsymbol{\Omega}^{l*}, \boldsymbol{\gamma}^*\right) \left(y_i - \boldsymbol{x}_{i\mathcal{S}_1}^\top \boldsymbol{\alpha}^*_{k\mathcal{S}_1} - \boldsymbol{x}_{i\mathcal{S}_2}^\top \boldsymbol{\gamma}^*_{k\mathcal{S}_2}\right) \boldsymbol{x}_{ij} \\
&\quad + \frac{\sigma_k^{l*}}{n} \sum_{i=1}^n L_k^l \left(y_i, \boldsymbol{x}_i, \boldsymbol{\Omega}^{l*}, \boldsymbol{\gamma}^*\right) \left[\boldsymbol{x}_{i\mathcal{S}_1}^\top \boldsymbol{x}_{ij} (\widehat{\boldsymbol{\alpha}}_{k\mathcal{S}_1} - \boldsymbol{\alpha}^*_{k\mathcal{S}_1}) + \boldsymbol{x}_{i\mathcal{S}_2}^\top \boldsymbol{x}_{ij} (\widehat{\boldsymbol{\gamma}}_{k\mathcal{S}_2} - \boldsymbol{\gamma}^*_{k\mathcal{S}_2})\right] \\
&\quad + (\boldsymbol{r}(\boldsymbol{\Delta})_{\boldsymbol{\alpha}_j})_k = (\xi_{\alpha 1})_k + (\xi_{\alpha 2})_k + (\boldsymbol{r}(\boldsymbol{\Delta})_{\boldsymbol{\alpha}_j})_k, \quad k = 1, 2, \cdots, K_l,
\end{aligned} \tag{S1.40}
$$

where we can bound $\xi_{\alpha 1}$ as above.

Then, according to the second line of (S1.33), we have:

$$
\begin{pmatrix} \widehat{\boldsymbol{\alpha}}_{\mathcal{S}_1} - \boldsymbol{\alpha}^*_{\mathcal{S}_1} \\ \\ \widehat{\boldsymbol{\gamma}}_{\mathcal{S}_2} - \boldsymbol{\gamma}^*_{\mathcal{S}_2} \end{pmatrix} = \begin{pmatrix} \nabla^2_{\boldsymbol{\alpha}_{\mathcal{S}_1} \boldsymbol{\alpha}_{\mathcal{S}_1}} \mathcal{Q}(\boldsymbol{\Omega}^*) & \nabla^2_{\boldsymbol{\alpha}_{\mathcal{S}_1} \boldsymbol{\gamma}_{\mathcal{S}_2}} \mathcal{Q}(\boldsymbol{\Omega}^*) \\ \\ \nabla^2_{\boldsymbol{\gamma}_{\mathcal{S}_2} \boldsymbol{\alpha}_{\mathcal{S}_1}} \mathcal{Q}(\boldsymbol{\Omega}^*) & \nabla^2_{\boldsymbol{\gamma}_{\mathcal{S}_2} \boldsymbol{\gamma}_{\mathcal{S}_2}} \mathcal{Q}(\boldsymbol{\Omega}^*) \end{pmatrix}^{-1} \begin{pmatrix} \nabla_{\boldsymbol{\alpha}_{\mathcal{S}_1}} \mathcal{Q}(\boldsymbol{\Omega}^*) + \boldsymbol{r}(\boldsymbol{\Delta})_{\boldsymbol{\alpha}_{\mathcal{S}_1}} \\ \\ \nabla_{\boldsymbol{\gamma}_{\mathcal{S}_2}} \mathcal{Q}(\boldsymbol{\Omega}^*) + \boldsymbol{r}(\boldsymbol{\Delta})_{\boldsymbol{\gamma}_{\mathcal{S}_2}} \end{pmatrix}. \tag{S1.41}
$$

Then, we can similarly obtain that:

$$
\begin{aligned}
|(\xi_{\alpha 2})_k| \leqslant &\left\| \left[\frac{1}{n} \boldsymbol{X}_{\mathcal{S}_1^c}^\top \boldsymbol{G}_k^l \boldsymbol{X}_{\mathcal{S}}\right] \left[\frac{1}{n} \boldsymbol{X}_{\mathcal{S}}^\top \boldsymbol{G}_k^l \boldsymbol{X}_{\mathcal{S}}\right]^{-1} \right\|_\infty \\
&\left(\left\| \frac{\sigma_k^{l*}}{n} \sum_{i=1}^n L_k^l \left(y_i, \boldsymbol{x}_i, \boldsymbol{\Omega}^{l*}, \boldsymbol{\gamma}^*\right) \left(y_i - \boldsymbol{x}_{i\mathcal{S}_1}^\top \boldsymbol{\alpha}^*_{k\mathcal{S}_1} - \boldsymbol{x}_{i\mathcal{S}_2}^\top \boldsymbol{\gamma}^*_{k\mathcal{S}_2}\right) \boldsymbol{x}_{i\mathcal{S}} \right\|_\infty \right. \\
&\quad \left. + \left\| \mathrm{vec}(\boldsymbol{r}(\boldsymbol{\Delta})_{\boldsymbol{\alpha}_{k\mathcal{S}_1}}, \boldsymbol{r}(\boldsymbol{\Delta})_{\boldsymbol{\gamma}_{k\mathcal{S}_1}}) \right\|_\infty \right),
\end{aligned} \tag{S1.42}
$$

$$
\begin{aligned}
|(\xi_{\gamma 2})_k| \leqslant & \left\| \left[\frac{1}{n} \boldsymbol{X}_{\mathcal{S}^c}^\top \boldsymbol{G}_k^l \boldsymbol{X}_{\mathcal{S}} \right] \left[\frac{1}{n} \boldsymbol{X}_{\mathcal{S}}^\top \boldsymbol{G}_k^l \boldsymbol{X}_{\mathcal{S}} \right]^{-1} \right\|_\infty \\
& \left(\left\| \frac{\sigma_k^{l*}}{n} \sum_{i=1}^n L_k^l \left(y_i, \boldsymbol{x}_i, \boldsymbol{\Omega}^{l*}, \boldsymbol{\gamma}^* \right) \left(y_i - \boldsymbol{x}_{i\mathcal{S}_1}^\top \boldsymbol{\alpha}_{k\mathcal{S}_1}^* - \boldsymbol{x}_{i\mathcal{S}_2}^\top \boldsymbol{\gamma}_{k\mathcal{S}_2}^* \right) \boldsymbol{x}_{i\mathcal{S}} \right\|_\infty \right. \\
& \left. + \left\| \mathrm{vec}(\boldsymbol{r}(\boldsymbol{\Delta})_{\boldsymbol{\alpha}_{k\mathcal{S}_1}}, \boldsymbol{r}(\boldsymbol{\Delta})_{\boldsymbol{\gamma}_{k\mathcal{S}_1}}) \right\|_\infty \right).
\end{aligned} \tag{S1.43}
$$

Then, by $\|\mathrm{vec}(\boldsymbol{z}_\beta, \boldsymbol{z}_\alpha)\|_2 \leqslant \|\boldsymbol{z}_\beta\|_2 + \|\boldsymbol{z}_\alpha\|_2$, $\|\boldsymbol{z}_\beta\|_2 \leqslant \|\xi_{\beta 1}\|_2 + \|\xi_{\beta 2}\|_2 + \left\|\boldsymbol{r}(\boldsymbol{\Delta})_{\boldsymbol{\beta}_j}\right\|_2$, and combining with Lemma 3 in Wytock and Kolter (2013), there exists a positive constant c_c such that:

$$
\begin{aligned}
\|\mathrm{vec}(\boldsymbol{z}_\beta, \boldsymbol{z}_\alpha)\|_2 \leqslant & c_c \left(\sqrt{\frac{K_u \log p}{n}} + \left\| \left[\frac{1}{n} \boldsymbol{X}_{\mathcal{S}_1^c}^\top \boldsymbol{G}_k^u \boldsymbol{X}_{\mathcal{S}_1} \right] \left[\frac{1}{n} \boldsymbol{X}_{\mathcal{S}_1}^\top \boldsymbol{G}_k^u \boldsymbol{X}_{\mathcal{S}_1} \right]^{-1} \right\|_\infty \right. \\
& \left(\sqrt{\frac{K_u \log p}{n}} + \left\| \widehat{\boldsymbol{\Omega}}_0 - \boldsymbol{\Omega}_0^* \right\|_\infty^2 \right) + \left\| \widehat{\boldsymbol{\Omega}}_0 - \boldsymbol{\Omega}_0^* \right\|_\infty^2 \\
& + \sqrt{\frac{K_l \log p}{n}} + \left\| \left[\frac{1}{n} \boldsymbol{X}_{\mathcal{S}_1^c}^\top \boldsymbol{G}_k^l \boldsymbol{X}_{\mathcal{S}} \right] \left[\frac{1}{n} \boldsymbol{X}_{\mathcal{S}}^\top \boldsymbol{G}_k^l \boldsymbol{X}_{\mathcal{S}} \right]^{-1} \right\|_\infty \\
& \left. \left(\sqrt{\frac{K_l \log p}{n}} + \left\| \widehat{\boldsymbol{\Omega}}_0 - \boldsymbol{\Omega}_0^* \right\|_\infty^2 \right) + \left\| \widehat{\boldsymbol{\Omega}}_0 - \boldsymbol{\Omega}_0^* \right\|_\infty^2 \right);
\end{aligned} \tag{S1.44}
$$

$$
\begin{aligned}
\|\boldsymbol{z}_\gamma\|_2 \leqslant & c_c \left(\sqrt{\frac{K_l \log p}{n}} + \left\| \left[\frac{1}{n} \boldsymbol{X}_{\mathcal{S}^c}^\top \boldsymbol{G}_k^l \boldsymbol{X}_{\mathcal{S}} \right] \left[\frac{1}{n} \boldsymbol{X}_{\mathcal{S}}^\top \boldsymbol{G}_k^l \boldsymbol{X}_{\mathcal{S}} \right]^{-1} \right\|_\infty \right. \\
& \left. \left(\sqrt{\frac{K_l \log p}{n}} + \left\| \widehat{\boldsymbol{\Omega}}_0 - \boldsymbol{\Omega}_0^* \right\|_\infty^2 \right) + \left\| \widehat{\boldsymbol{\Omega}}_0 - \boldsymbol{\Omega}_0^* \right\|_\infty^2 \right),
\end{aligned} \tag{S1.45}
$$

with probability tending to 1.

Combining Condition 6 and the results in Step 1, we can show that $\|\mathrm{vec}(\boldsymbol{z}_\beta, \boldsymbol{z}_\alpha)\|_2 \leqslant \lambda_1$ and $\|\boldsymbol{z}_\gamma\|_2 \leqslant \lambda_2$ with probability tending to 1. Then, by the minimal signal of the true parameters, Theorem 1 can be established.

Remarks With Lasso, λ_1 and λ_2 need to be much large than $\sqrt{\frac{K_l^3 \log p}{n}}$, which is a violation of (S1.28), leading to $\left\|\widehat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}^*\right\|_2 \gg \left(\sqrt{\frac{K_l^3 s \log p}{n}}\right)$. This re-establishes the advantages of concave penalties like MCP.

Related Lemmas

Lemma 1. *Suppose that conditions (1) and (4) hold. With probability tending to 1,*

$$\begin{aligned}&\left\|\nabla_{\boldsymbol{\Omega}_0^{l*}} \mathcal{H}_n\left(\boldsymbol{\Omega}_0^{l*}, \boldsymbol{\gamma}_0^* \mid \boldsymbol{\Omega}_0^{l(t-1)}, \boldsymbol{\gamma}_0^{(t-1)}\right)\right.\\ &\qquad\left.-\nabla_{\boldsymbol{\Omega}_0^{l*}} \mathcal{H}\left(\boldsymbol{\Omega}_0^{l*}, \boldsymbol{\gamma}_0^* \mid \boldsymbol{\Omega}_0^{l(t-1)}, \boldsymbol{\gamma}_0^{(t-1)}\right)\right\|_\infty = O_p(K_l\sqrt{\frac{\log p}{n}}).\end{aligned} \tag{S1.46}$$

Proof. Recall that:

$$\begin{aligned}&\left\|\nabla_{\boldsymbol{\Omega}_0^{l*}} \mathcal{H}_n\left(\boldsymbol{\Omega}_0^{l*}, \boldsymbol{\gamma}_0^* \mid \boldsymbol{\Omega}_0^{l(t-1)}, \boldsymbol{\gamma}_0^{(t-1)}\right) - \nabla_{\boldsymbol{\Omega}_0^{l*}} \mathcal{H}\left(\boldsymbol{\Omega}_0^{l*}, \boldsymbol{\gamma}_0^* \mid \boldsymbol{\Omega}_0^{l(t-1)}, \boldsymbol{\gamma}_0^{(t-1)}\right)\right\|_\infty \\ &= \max_{1 \leqslant k \leqslant K_l} \left\|\nabla_{\boldsymbol{\alpha}_{k\mathcal{S}_1}^*} \mathcal{H}_n\left(\boldsymbol{\Omega}_0^{l*}, \boldsymbol{\gamma}_0^* \mid \boldsymbol{\Omega}_0^{l(t-1)}, \boldsymbol{\gamma}_0^{(t-1)}\right) - \nabla_{\boldsymbol{\alpha}_{k\mathcal{S}_1}^*} \mathcal{H}\left(\boldsymbol{\Omega}_0^{l*}, \boldsymbol{\gamma}_0^* \mid \boldsymbol{\Omega}_0^{l(t-1)}, \boldsymbol{\gamma}_0^{(t-1)}\right)\right\|_\infty \\ &+ \max_{1 \leqslant k \leqslant K_l} \left(\nabla_{\boldsymbol{\sigma}_k^{l*}} \mathcal{H}_n\left(\boldsymbol{\Omega}_0^{l*}, \boldsymbol{\gamma}_0^* \mid \boldsymbol{\Omega}_0^{l(t-1)}, \boldsymbol{\gamma}_0^{(t-1)}\right) - \nabla_{\boldsymbol{\sigma}_k^{l*}} \mathcal{H}\left(\boldsymbol{\Omega}_0^{l*}, \boldsymbol{\gamma}_0^* \mid \boldsymbol{\Omega}_0^{l(t-1)}, \boldsymbol{\gamma}_0^{(t-1)}\right)\right).\end{aligned} \tag{S1.47}$$

And we have:

$$\begin{aligned}&\left\|\nabla_{\boldsymbol{\alpha}_{k\mathcal{S}_1}^*} \mathcal{H}_n\left(\boldsymbol{\Omega}_0^{l*}, \boldsymbol{\gamma}_0^* \mid \boldsymbol{\Omega}_0^{l(t-1)}, \boldsymbol{\gamma}_0^{(t-1)}\right) - \nabla_{\boldsymbol{\alpha}_{k\mathcal{S}_1}^*} \mathcal{H}\left(\boldsymbol{\Omega}_0^{l*}, \boldsymbol{\gamma}_0^* \mid \boldsymbol{\Omega}_0^{l(t-1)}, \boldsymbol{\gamma}_0^{(t-1)}\right)\right\|_\infty \\ &\leqslant \kappa_2 \left\|\frac{1}{n}\sum_{i=1}^n L_k^l\left(y_i, \boldsymbol{x}_i, \boldsymbol{\Omega}_0^l, \boldsymbol{\gamma}_0\right) y_i \boldsymbol{x}_{i\mathcal{S}_1} - \mathbb{E}\left(L_k^l\left(y, \boldsymbol{x}, \boldsymbol{\Omega}_0^l, \boldsymbol{\gamma}_0\right) y \boldsymbol{x}_{\mathcal{S}_1}\right)\right\|_\infty \\ &+ \kappa_2 \mu_k \left|\frac{1}{n}\sum_{i=1}^n L_k^l\left(y_i, \boldsymbol{x}_i, \boldsymbol{\Omega}_0^l, \boldsymbol{\gamma}_0\right) - \mathbb{E}\left(L_k^l\left(y, \boldsymbol{x}, \boldsymbol{\Omega}_0^l, \boldsymbol{\gamma}_0\right)\right)\right| = \kappa_2 \left\|I_1\right\|_\infty + \kappa_2 \mu_k \left|I_2\right|,\end{aligned} \tag{S1.48}$$

where $\mu_k = \left\|(\boldsymbol{x}_{\mathcal{S}_1}^\top \boldsymbol{\alpha}_{k\mathcal{S}_1}^* + \boldsymbol{x}_{\mathcal{S}_2}^\top \boldsymbol{\gamma}_{k\mathcal{S}_2}^*)\boldsymbol{x}_{\mathcal{S}}\right\|_\infty$ which can be treated similarly to Hao et al. (2018). And I_2 can be bounded similarly to (S1.21), which implies that $I_2 \leq \sqrt{\frac{1}{2n}\log(2p)}$ with probability at least $1-\frac{1}{p}$. Then considering I_1 by following Lemma S1 in Li et al. (2023), we have:

$$\|I_1\|_\infty \leqslant D_1 K_l \sqrt{\frac{\log p}{n}}, \tag{S1.49}$$

with probability at least $1-\frac{2K_l}{p}$, where D_1 is a positive constant.

By applying union bound, we can obtain that:

$$\begin{aligned}&\max_{1\leqslant k\leqslant K_l}\left\|\nabla_{\boldsymbol{\alpha}_{k\mathcal{S}_1}^*}\mathcal{H}_n\left(\boldsymbol{\Omega}_0^{l*},\boldsymbol{\gamma}_0^* \mid \boldsymbol{\Omega}_0^{l(t-1)},\boldsymbol{\gamma}_0^{(t-1)}\right)-\nabla_{\boldsymbol{\alpha}_{k\mathcal{S}_1}^*}\mathcal{H}\left(\boldsymbol{\Omega}_0^{l*},\boldsymbol{\gamma}_0^* \mid \boldsymbol{\Omega}_0^{l(t-1)},\boldsymbol{\gamma}_0^{(t-1)}\right)\right\|_\infty\\&\leqslant \sqrt{\frac{1}{2n}\log(2p)}+D_1K_l\sqrt{\frac{\log p}{n}}\leqslant D_2K_l\sqrt{\frac{\log p}{n}},\end{aligned} \tag{S1.50}$$

with probability at least $1-\frac{2K_l^2+K_l}{p}$, where D_2 is a positive constant.

Consider:

$$\begin{aligned}&\left(\nabla_{\boldsymbol{\sigma}_k^{l*}}\mathcal{H}_n\left(\boldsymbol{\Omega}_0^{l*},\boldsymbol{\gamma}_0^* \mid \boldsymbol{\Omega}_0^{l(t-1)},\boldsymbol{\gamma}_0^{(t-1)}\right)-\nabla_{\boldsymbol{\sigma}_k^{l*}}\mathcal{H}\left(\boldsymbol{\Omega}_0^{l*},\boldsymbol{\gamma}_0^* \mid \boldsymbol{\Omega}_0^{l(t-1)},\boldsymbol{\gamma}_0^{(t-1)}\right)\right)\\&\leqslant \frac{1}{2\kappa_1}\left|\frac{1}{n}\sum_{i=1}^n L_k^l\left(y_i,\boldsymbol{x}_i,\boldsymbol{\Omega}_0^l,\boldsymbol{\gamma}_0\right)-\mathbb{E}\left(L_k^l\left(y,\boldsymbol{x},\boldsymbol{\Omega}_0^l,\boldsymbol{\gamma}_0\right)\right)\right|\\&+\frac{1}{2}\left|\frac{1}{n}\sum_{i=1}^n\left[L_k^l\left(y_i,\boldsymbol{x}_i,\boldsymbol{\Omega}_0^l,\boldsymbol{\gamma}_0\right)\left(y_i-\boldsymbol{x}_{i\mathcal{S}_1}^\top\boldsymbol{\alpha}_{k\mathcal{S}_1}^*-\boldsymbol{x}_{i\mathcal{S}_2}^\top\boldsymbol{\gamma}_{k\mathcal{S}_2}^*\right)^2\right]\right.\\&\left.-\mathbb{E}\left[L_k^l\left(y,\boldsymbol{x},\boldsymbol{\Omega}_0^l,\boldsymbol{\gamma}_0\right)\left(y-\boldsymbol{x}_{\mathcal{S}_1}^\top\boldsymbol{\alpha}_{k\mathcal{S}_1}^*-\boldsymbol{x}_{\mathcal{S}_2}^\top\boldsymbol{\gamma}_{k\mathcal{S}_2}^*\right)^2\right]\right|=II_1+II_2.\end{aligned} \tag{S1.51}$$

II_1 can be similarly bounded as above with probability at least $1-\frac{1}{p}$. Then, we only need to bound II_2. Take $\boldsymbol{x}_{\mathcal{S}_1}^\top\boldsymbol{\alpha}_{k\mathcal{S}_1}^*-\boldsymbol{x}_{\mathcal{S}_2}^\top\boldsymbol{\gamma}_{k\mathcal{S}_2}^*=\widetilde{\boldsymbol{x}}_{\mathcal{S}}^\top\widetilde{\boldsymbol{\alpha}}_{k\mathcal{S}}^*$, where $\widetilde{\boldsymbol{x}}_{i\mathcal{S}}$

is a rearrangement of $\boldsymbol{x}_{i\mathcal{S}}$, and $\widetilde{\boldsymbol{\alpha}}_{k\mathcal{S}}^{*}$ contains the corresponding coefficients.

Then, we have:

$$\begin{aligned}
II_2 \leqslant & \frac{1}{2}\left|\frac{1}{n}\sum_{i=1}^{n} L_k^l\left(y_i, \boldsymbol{x}_i, \boldsymbol{\Omega}_0^l, \boldsymbol{\gamma}_0\right) y_i^2 - \mathbb{E}\left[L_k^l\left(y, \boldsymbol{x}, \boldsymbol{\Omega}_0^l, \boldsymbol{\gamma}_0\right) y^2\right]\right| \\
& + \|\widetilde{\boldsymbol{\alpha}}_{k\mathcal{S}}^{*}\| \left\|\frac{1}{n}\sum_{i=1}^{n} L_k^l\left(y_i, \boldsymbol{x}_i, \boldsymbol{\Omega}_0^l, \boldsymbol{\gamma}_0\right) y_i \widetilde{\boldsymbol{x}}_{i\mathcal{S}} - \mathbb{E}\left[L_k^l\left(y, \boldsymbol{x}, \boldsymbol{\Omega}_0^l, \boldsymbol{\gamma}_0\right) y_i \widetilde{\boldsymbol{x}}_{\mathcal{S}}\right]\right\|_{\infty} \\
& + \frac{1}{2}\|\widetilde{\boldsymbol{\alpha}}_{k\mathcal{S}}^{*}\|^2 \left\|\frac{1}{n}\sum_{i=1}^{n} L_k^l\left(y_i, \boldsymbol{x}_i, \boldsymbol{\Omega}_0^l, \boldsymbol{\gamma}_0\right) \widetilde{\boldsymbol{x}}_{i\mathcal{S}} \widetilde{\boldsymbol{x}}_{i\mathcal{S}}^{\top} - \mathbb{E}\left[L_k^l\left(y, \boldsymbol{x}, \boldsymbol{\Omega}_0^l, \boldsymbol{\gamma}_0\right) \widetilde{\boldsymbol{x}}_{\mathcal{S}} \widetilde{\boldsymbol{x}}_{\mathcal{S}}^{\top}\right]\right\|_{\infty} \\
& = II_{21} + II_{22} + II_{23},
\end{aligned} \tag{S1.52}$$

where II_{22} can be bounded similarly to I_1. Following Li et al. (2023), with probability at least $1 - \frac{4K_l}{p}$, $II_{21} \leqslant D_3\sqrt{\frac{1}{n}}$, and with probability at least $1 - \frac{1}{p}$, $II_{24} \leqslant D_4\sqrt{\frac{1}{n}}$, where D_3 and D_4 are positive constants.

Also, by applying union bound, we can obtain that:

$$\begin{aligned}
\max_{1\leqslant k\leqslant K_l} & \left(\nabla_{\boldsymbol{\sigma}_k^{l*}} \mathcal{H}_n\left(\boldsymbol{\Omega}_0^{l*}, \boldsymbol{\gamma}_0^{*} \mid \boldsymbol{\Omega}_0^{l(t-1)}, \boldsymbol{\gamma}_0^{(t-1)}\right) - \nabla_{\boldsymbol{\sigma}_k^{l*}} \mathcal{H}\left(\boldsymbol{\Omega}_0^{l*}, \boldsymbol{\gamma}_0^{*} \mid \boldsymbol{\Omega}_0^{l(t-1)}, \boldsymbol{\gamma}_0^{(t-1)}\right)\right) \\
& \leqslant D_5 K_l \sqrt{\frac{1}{n}},
\end{aligned} \tag{S1.53}$$

with probability at least $1 - \frac{6K_l^2 + 2K_l}{p}$, where D_5 is a positive constant.

Overall, we have:

$$\begin{aligned}
& \left\|\nabla_{\boldsymbol{\Omega}_0^{l*}} \mathcal{H}_n\left(\boldsymbol{\Omega}_0^{l*}, \boldsymbol{\gamma}_0^{*} \mid \boldsymbol{\Omega}_0^{l(t-1)}, \boldsymbol{\gamma}_0^{(t-1)}\right) - \nabla_{\boldsymbol{\Omega}_0^{l*}} \mathcal{H}\left(\boldsymbol{\Omega}_0^{l*}, \boldsymbol{\gamma}_0^{*} \mid \boldsymbol{\Omega}_0^{l(t-1)}, \boldsymbol{\gamma}_0^{(t-1)}\right)\right\|_{\infty} \\
& = O_p(K_l \sqrt{\frac{\log p}{n}}).
\end{aligned} \tag{S1.54}$$

Lemma 2. *Under Condition 3 and Condition 4, and with* $\mathbf{\Omega}_0 \in \mathcal{B}_\alpha\left(\mathbf{\Omega}_0^*\right)$,

$$\left\|\nabla_{\mathbf{\Omega}_0^{l*}} \mathcal{H}\left(\mathbf{\Omega}_0^{l*}, \boldsymbol{\gamma}_0^* \mid \mathbf{\Omega}_0^l, \boldsymbol{\gamma}_0\right) - \nabla_{\mathbf{\Omega}_0^{l*}} \mathcal{H}\left(\mathbf{\Omega}_0^{l*}, \boldsymbol{\gamma}_0^* \mid \mathbf{\Omega}_0^{l*}, \boldsymbol{\gamma}_0^*\right)\right\|_2 \leqslant \tau \left\|\mathbf{\Omega}_0^* - \mathbf{\Omega}_0\right\|_2, \tag{S1.55}$$

where $\tau \leqslant \frac{\varrho}{12}$ according to Condition 3.

Proof. Consider the k-th subgroup. We have:

$$\begin{aligned}
&\left\|\nabla_{\mathbf{\Omega}_{0k}^{l*}} \mathcal{H}\left(\mathbf{\Omega}_0^{l*}, \boldsymbol{\gamma}_0^* \mid \mathbf{\Omega}_0^l, \boldsymbol{\gamma}_0\right) - \nabla_{\mathbf{\Omega}_{0k}^{l*}} \mathcal{H}\left(\mathbf{\Omega}_0^{l*}, \boldsymbol{\gamma}_0^* \mid \mathbf{\Omega}_0^{l*}, \boldsymbol{\gamma}_0^*\right)\right\|_2^2 = I + II \\
&= \left\|\sigma_k^{l*} \mathbb{E}\left[F_k^l\left(\mathbf{\Omega}_0^l, \boldsymbol{\gamma}_0, \mathbf{\Omega}_0^{l*}, \boldsymbol{\gamma}_0^*\right)\left(y - \boldsymbol{x}_{\mathcal{S}_1}^\top \boldsymbol{\alpha}_{k\mathcal{S}_1}^* - \boldsymbol{x}_{\mathcal{S}_2}^\top \boldsymbol{\gamma}_{k\mathcal{S}_2}^*\right) \boldsymbol{x}_{i\mathcal{S}_1}\right]\right\|_2^2 \\
&+ \left|\frac{1}{2\sigma_k^{l*}} \mathbb{E}\left[F_k^l\left(\mathbf{\Omega}_0^l, \boldsymbol{\gamma}_0, \mathbf{\Omega}_0^{l*}, \boldsymbol{\gamma}_0^*\right)\right] - \frac{1}{2} \mathbb{E}\left[F_k^l\left(\mathbf{\Omega}_0^l, \boldsymbol{\gamma}_0, \mathbf{\Omega}_0^{l*}, \boldsymbol{\gamma}_0^*\right)\left(y - \boldsymbol{x}_{\mathcal{S}_1}^\top \boldsymbol{\alpha}_{k\mathcal{S}_1}^* - \boldsymbol{x}_{\mathcal{S}_2}^\top \boldsymbol{\gamma}_{k\mathcal{S}_2}^*\right)^2\right]\right|_2^2,
\end{aligned} \tag{S1.56}$$

where $F_k^l\left(\mathbf{\Omega}_0^l, \mathbf{\Omega}_0^{l*}\right) = L_k^l\left(y, \boldsymbol{x}, \mathbf{\Omega}_0^l, \boldsymbol{\gamma}_0\right) - L_k^l\left(y, \boldsymbol{x}, \mathbf{\Omega}_0^{l*}, \boldsymbol{\gamma}_0^*\right)$. It is sufficient to show that $I \leqslant \tau_1 \left\|\mathbf{\Omega}_0^{l*} - \mathbf{\Omega}_0^l\right\|_2$ and $II \leqslant \tau_2 \left\|\mathbf{\Omega}_0^{l*} - \mathbf{\Omega}_0^l\right\|_2$. Here we take $\pi_k^l = 1/K_l$ as a much simpler balanced setting by following Condition 4 and Lemma 7 of Hao et al. (2018).

Then, the Taylor expansion of $L_k^l\left(y, \boldsymbol{x}, \mathbf{\Omega}_0^l, \boldsymbol{\gamma}_0\right)$ at $(\mathbf{\Omega}_0^{l*}, \boldsymbol{\gamma}_0^*)$ leads to:

$$L_k^l\left(y, \boldsymbol{x}, \mathbf{\Omega}_0^l, \boldsymbol{\gamma}_0\right) = L_k^l\left(y, \boldsymbol{x}, \mathbf{\Omega}_0^{l*}, \boldsymbol{\gamma}_0^*\right) + \nabla L_k^l\left(y, \boldsymbol{x}, \mathbf{\Omega}_0^{lz}, \boldsymbol{\gamma}_0^z\right) \operatorname{vec}(\mathbf{\Omega}_0^l - \mathbf{\Omega}_0^{l*}, \boldsymbol{\gamma}_0 - \boldsymbol{\gamma}_0^*), \tag{S1.57}$$

where $(\mathbf{\Omega}_0^{lz}, \boldsymbol{\gamma}_0^z) = (z\mathbf{\Omega}_0^l + (1-z)\mathbf{\Omega}_0^{l*}, z\boldsymbol{\gamma}_0 + (1-z)\boldsymbol{\gamma}_0^*)$. To simplify notation, we use $\widetilde{\boldsymbol{x}}_{\mathcal{S}}$ and $\widetilde{\boldsymbol{\alpha}}^*_{k\mathcal{S}}$ defined in Lemma 1 and take $\mathrm{vec}(\boldsymbol{\Omega}^l_0, \boldsymbol{\gamma}_0) = \boldsymbol{\Upsilon}$. Then,

$$\nabla L^l_k\left(y, \boldsymbol{x}, \boldsymbol{\Upsilon}\right) = \left([\nabla_{\boldsymbol{\Upsilon}_1} L^l_k\left(y, \boldsymbol{x}, \boldsymbol{\Upsilon}\right)]^\top, \cdots, [\nabla_{\boldsymbol{\Upsilon}_{K_l}} L^l_k\left(y, \boldsymbol{x}, \boldsymbol{\Upsilon}\right)]^\top\right)^\top, \tag{S1.58}$$

$$\nabla_{\boldsymbol{\Upsilon}_{k'}} L^l_k\left(y, \boldsymbol{x}, \boldsymbol{\Upsilon}\right) = \begin{cases} -L^l_k\left(y, \boldsymbol{x}, \boldsymbol{\Upsilon}\right) L^l_{k'}\left(y, \boldsymbol{x}, \boldsymbol{\Upsilon}\right) \boldsymbol{\delta}_{\boldsymbol{\Upsilon}_{k'}}, k' \neq k; \\ L^l_{k'}\left(y, \boldsymbol{x}, \boldsymbol{\Upsilon}\right) [1 - L^l_{k'}\left(y, \boldsymbol{x}, \boldsymbol{\Upsilon}\right)] \boldsymbol{\delta}_{\boldsymbol{\Upsilon}_k}, k' = k. \end{cases} \tag{S1.59}$$

Then, for $k = 1, 2, \cdots, K_l$,

$$\boldsymbol{\delta}_{\boldsymbol{\Upsilon}_k} = \begin{pmatrix} \mathrm{vec}\left(\sigma^l_k (y - \widetilde{\boldsymbol{x}}^\top_{\mathcal{S}} \widetilde{\boldsymbol{\alpha}}_{k\mathcal{S}}) \widetilde{\boldsymbol{x}}_{\mathcal{S}}\right) \\ \dfrac{1}{2}\left(\dfrac{1}{\sigma^l_k} - (y - \widetilde{\boldsymbol{x}}^\top_{\mathcal{S}} \widetilde{\boldsymbol{\alpha}}_{k\mathcal{S}})^2\right) \end{pmatrix}. \tag{S1.60}$$

Next, we apply Taylor expansion to bound I:

$$\begin{aligned} I &= \left\| \mathbb{E}\left[\sigma^{l*}_k \left(y - \widetilde{\boldsymbol{x}}^\top_{\mathcal{S}} \widetilde{\boldsymbol{\alpha}}^*_{k\mathcal{S}}\right) \boldsymbol{x}_{i\mathcal{S}_1} \nabla L^l_k\left(y, \boldsymbol{x}, \boldsymbol{\Upsilon}^z\right) \left(\boldsymbol{\Upsilon} - \boldsymbol{\Upsilon}^*\right)\right] \right\|^2_2 \\ &= \left\| \mathbb{E}\left[\sigma^{l*}_k \left(y - \widetilde{\boldsymbol{x}}^\top_{\mathcal{S}} \widetilde{\boldsymbol{\alpha}}^*_{k\mathcal{S}}\right) \boldsymbol{x}_{i\mathcal{S}_1} \nabla L^l_k\left(y, \boldsymbol{x}, \boldsymbol{\Upsilon}^z\right)\right] \right\|^2_2 \left\|\boldsymbol{\Upsilon} - \boldsymbol{\Upsilon}^*\right\|^2_2 \\ &\leqslant \sup_{z \in [0,1]} \mathbb{E}\left[\left\| \sigma^{l*}_k \left(y - \widetilde{\boldsymbol{x}}^\top_{\mathcal{S}} \widetilde{\boldsymbol{\alpha}}^*_{k\mathcal{S}}\right) \boldsymbol{x}_{i\mathcal{S}_1} \right\|^2_2 \left\|\nabla L^l_k\left(y, \boldsymbol{x}, \boldsymbol{\Upsilon}^z\right)\right\|^2_2\right] \left\|\boldsymbol{\Upsilon} - \boldsymbol{\Upsilon}^*\right\|^2_2. \end{aligned} \tag{S1.61}$$

Then, by the definition of $\nabla L^l_k\left(y, \boldsymbol{x}, \boldsymbol{\Upsilon}^z\right)$, we can obtain that:

$$\begin{aligned} \left\|\nabla L^l_k\left(y, \boldsymbol{x}, \boldsymbol{\Upsilon}^z\right)\right\|^2_2 = &\sum_{k' \neq k} \left[L^l_k\left(y, \boldsymbol{x}, \boldsymbol{\Upsilon}^z\right) L^l_{k'}\left(y, \boldsymbol{x}, \boldsymbol{\Upsilon}^z\right)\right]^2 \cdot \boldsymbol{\delta}^\top_{\boldsymbol{\Upsilon}^z_{k'}} \boldsymbol{\delta}_{\boldsymbol{\Upsilon}^z_{k'}} \\ &+ \left[L^l_k\left(y, \boldsymbol{x}, \boldsymbol{\Upsilon}^z\right) \left(1 - L^l_k\left(y, \boldsymbol{x}, \boldsymbol{\Upsilon}^z\right)\right)\right]^2 \cdot \boldsymbol{\delta}^\top_{\boldsymbol{\Upsilon}^z_k} \boldsymbol{\delta}_{\boldsymbol{\Upsilon}^z_k} = A_1 + A_2. \end{aligned} \tag{S1.62}$$

For $k=1,\cdots,K_l$, define:

$$\begin{aligned}W_{k1}:=&\sup_{z\in[0,1]}\mathbb{E}\left[\boldsymbol{\delta}_{\boldsymbol{\Upsilon}_k^z}^{\top}\boldsymbol{\delta}_{\boldsymbol{\Upsilon}_k^z}\cdot\left\|\sigma_k^{l*}\left(y-\widetilde{\boldsymbol{x}}_{\mathcal{S}}^{\top}\widetilde{\boldsymbol{\alpha}}_{k\mathcal{S}}^{*}\right)\boldsymbol{x}_{i\mathcal{S}_1}\right\|_2^2\right]\\ W_{k2}:=&\sup_{z\in[0,1]}\mathbb{E}\left[\boldsymbol{\delta}_{\boldsymbol{\Upsilon}_k^z}^{\top}\boldsymbol{\delta}_{\boldsymbol{\Upsilon}_k^z}\cdot\left\|\sigma_k^{l*}\left(y-\widetilde{\boldsymbol{x}}_{\mathcal{S}}^{\top}\widetilde{\boldsymbol{\alpha}}_{k\mathcal{S}}^{*}\right)\boldsymbol{x}_{i\mathcal{S}_2}\right\|_2^2\right],\end{aligned}\tag{S1.63}$$

and $W=\max_{1\leqslant k\leqslant K_l}\{W_{k1},W_{k2}\}$. Then, under Condition 3, it is sufficient to bound τ_1:

$$\tau_1\leqslant\sup_{z\in[0,1]}\mathbb{E}\left[\left\|\sigma_k^{l*}\left(y-\widetilde{\boldsymbol{x}}_{\mathcal{S}}^{\top}\widetilde{\boldsymbol{\alpha}}_{k\mathcal{S}}^{*}\right)\boldsymbol{x}_{i\mathcal{S}_1}\right\|_2^2A_1\right]+\sup_{z\in[0,1]}\mathbb{E}\left[\left\|\sigma_k^{l*}\left(y-\widetilde{\boldsymbol{x}}_{\mathcal{S}}^{\top}\widetilde{\boldsymbol{\alpha}}_{k\mathcal{S}}^{*}\right)\boldsymbol{x}_{i\mathcal{S}_1}\right\|_2^2A_2\right].\tag{S1.64}$$

For II and $k=1,2,\cdots,K_l$, define:

$$\begin{aligned}&W_k':=\sup_{z\in[0,1]}\mathbb{E}\left[\boldsymbol{\delta}_{\boldsymbol{\Upsilon}_k^z}^{\top}\boldsymbol{\delta}_{\boldsymbol{\Upsilon}_k^z}\cdot\sigma_k^{l*-2}\right];\quad W_k'':=\sup_{z\in[0,1]}\mathbb{E}\left[\boldsymbol{\delta}_{\boldsymbol{\Upsilon}_k^z}^{\top}\boldsymbol{\delta}_{\boldsymbol{\Upsilon}_k^z}\cdot(y-\widetilde{\boldsymbol{x}}_{\mathcal{S}}^{\top}\widetilde{\boldsymbol{\alpha}}_{k\mathcal{S}})^2\right]\\ &W'=\max_{1\leqslant k\leqslant K_l}W_k',\quad W''=\max_{1\leqslant k\leqslant K_l}W_k''.\end{aligned}\tag{S1.65}$$

Similar to I, we can bound II under Condition 3, that is,

$$\left\|\nabla_{\boldsymbol{\Omega}_{0k}^{l*}}\mathcal{H}\left(\boldsymbol{\Omega}_0^{l*},\boldsymbol{\gamma}_0^*\mid\boldsymbol{\Omega}_0^{l},\boldsymbol{\gamma}_0\right)-\nabla_{\boldsymbol{\Omega}_{0k}^{l*}}\mathcal{H}\left(\boldsymbol{\Omega}_0^{l*},\boldsymbol{\gamma}_0^*\mid\boldsymbol{\Omega}_0^{l*},\boldsymbol{\gamma}_0^*\right)\right\|_2^2\leqslant\frac{\varrho^2}{144K_l}\left\|\boldsymbol{\Omega}_0^*-\boldsymbol{\Omega}_0\right\|_2^2.\tag{S1.66}$$

Now we take the summation, and there exists $\tau\leqslant\frac{\varrho}{12}$ satisfying:

$$\left\|\nabla_{\boldsymbol{\Omega}_0^{l*}}\mathcal{H}\left(\boldsymbol{\Omega}_0^{l*},\boldsymbol{\gamma}_0^*\mid\boldsymbol{\Omega}_0^{l},\boldsymbol{\gamma}_0\right)-\nabla_{\boldsymbol{\Omega}_0^{l*}}\mathcal{H}\left(\boldsymbol{\Omega}_0^{l*},\boldsymbol{\gamma}_0^*\mid\boldsymbol{\Omega}_0^{l*},\boldsymbol{\gamma}_0^*\right)\right\|_2\leqslant\tau\left\|\boldsymbol{\Omega}_0^*-\boldsymbol{\Omega}_0\right\|_2.\tag{S1.67}$$

S2 Additional Numerical Results

S2.1 Additional Simulation Results

Table S1: Simulation results for the balanced subgroup design. $\rho = 1$ and $\xi = 1$. In each cell: mean (sd).

	Method	TPRL	FPRL	MSEL	RIL	ARIL	TPRU	FPRU	MSEU	RIU	ARIU
S1	Proposed	0.934(0.070)	0.044(0.019)	0.386(0.101)	0.847(0.010)	0.592(0.028)	1(0)	0(0)	0.209(0.072)	0.805(0.019)	0.610(0.035)
	S-FMR	0.725(0.153)	0.539(0.118)	14.158(13.187)	0.671(0.026)	0.137(0.064)	–	–	–	–	–
	(separately)	–	–	–	–	–	0.825(0.355)	0.205(0.091)	5.102(4.376)	0.730(0.116)	0.460(0.232)
	G-FMR	0.755(0.062)	0.156(0.092)	2.876(2.513)	0.715(0.048)	0.281(0.106)	–	–	–	–	–
	(separately)	–	–	–	–	–	1(0)	0.015(0.012)	1.130(0.301)	0.793(0.223)	0.586(0.046)
	MG-FMR	0.889(0.067)	0.041(0.022)	0.661(0.095)	0.842(0.014)	0.579(0.036)	–	–	–	–	–
	(separately)	–	–	–	–	–	1(0)	0.004(0.005)	0.252(0.073)	0.796(0.023)	0.593(0.045)
	Order-A	0.832(0.060)	0.046(0.013)	0.507(0.112)	0.855(0.022)	0.612(0.057)	1(0)	0.008(0.009)	0.209(0.066)	0.806(0.017)	0.612(0.034)
	Order-D	0.879(0.085)	0.047(0.016)	0.652(0.079)	0.846(0.015)	0.589(0.040)	1(0)	0.047(0.019)	0.080(0.033)	0.819(0.022)	0.639(0.043)
S2	Proposed	0.885(0.082)	0.048(0.023)	0.432(0.089)	0.856(0.012)	0.616(0.031)	1(0)	0.001(0.002)	0.149(0.056)	0.798(0.019)	0.597(0.037)
	S-FMR	0.714(0.076)	0.554(0.078)	16.225(14.659)	0.675(0.019)	0.147(0.054)	–	–	–	–	–
	(separately)	–	–	–	–	–	0.752(0.409)	0.208(0.102)	5.291(7.416)	0.702(0.127)	0.404(0.255)
	G-FMR	0.722(0.098)	0.066(0.083)	2.870(3.306)	0.802(0.052)	0.481(0.125)	–	–	–	–	–
	(separately)	–	–	–	–	–	1(0)	0.006(0.008)	0.690(0.236)	0.781(0.021)	0.561(0.042)
	MG-FMR	0.826(0.109)	0.033(0.014)	0.476(0.111)	0.840(0.012)	0.601(0.032)	–	–	–	–	–
	(separately)	–	–	–	–	–	1(0)	0.006(0.009)	0.173(0.073)	0.801(0.016)	0.599(0.032)
	Order-A	0.718(0.068)	0.044(0.014)	0.585(0.114)	0.857(0.019)	0.619(0.052)	1(0)	0.005(0.007)	0.158(0.068)	0.797(0.020)	0.593(0.041)
	Order-D	0.829(0.087)	0.035(0.015)	0.466(0.101)	0.851(0.011)	0.607(0.029)	1(0)	0.061(0.017)	0.083(0.029)	0.822(0.017)	0.643(0.034)
S3	Proposed	0.828(0.069)	0.045(0.024)	0.406(0.104)	0.864(0.011)	0.637(0.028)	1(0)	0.004(0.005)	0.129(0.043)	0.781(0.022)	0.563(0.045)
	S-FMR	0.726(0.049)	0.537(0.080)	15.023(14.179)	0.674(0.015)	0.153(0.047)	–	–	–	–	–
	(separately)	–	–	–	–	–	0.859(0.333)	0.204(0.074)	5.306(3.122)	0.729(0.101)	0.459(0.201)
	G-FMR	0.672(0.105)	0.105(0.055)	4.115(3.269)	0.738(0.047)	0.345(0.097)	–	–	–	–	–
	(separately)	–	–	–	–	–	1(0)	0.003(0.005)	0.801(0.308)	0.764(0.019)	0.528(0.039)
	MG-FMR	0.765(0.081)	0.034(0.025)	0.563(0.814)	0.854(0.031)	0.614(0.069)	–	–	–	–	–
	(separately)	–	–	–	–	–	1(0)	0.002(0.004)	0.128(0.047)	0.789(0.018)	0.569(0.035)
	Order-A	0.754(0.066)	0.048(0.014)	0.496(0.120)	0.865(0.020)	0.640(0.052)	1(0)	0.003(0.006)	0.130(0.048)	0.782(0.022)	0.565(0.041)
	Order-D	0.774(0.072)	0.035(0.018)	0.498(0.741)	0.859(0.027)	0.627(0.064)	1(0)	0.071(0.016)	0.103(0.183)	0.811(0.019)	0.622(0.038)

Table S2: Simulation results for the unbalanced subgroup design. $\rho = 1$ and $\xi = 1$. In each cell: mean (sd).

		TPRL	FPRL	MSEL	RIL	ARIL	TPRU	FPRU	MSEU	RIU	ARIU
S1	Proposed	0.914(0.066)	0.053(0.021)	0.561(0.654)	0.844(0.028)	0.602(0.059)	1(0)	0.001(0.002)	0.238(0.071)	0.818(0.023)	0.634(0.047)
	S-FMR	0.805(0.073)	0.551(0.066)	19.509(11.925)	0.682(0.021)	0.184(0.052)	-	-	-	-	-
	(separately)	-	-	-	-	-	0.953(0.160)	0.245(0.097)	2.844(4.890)	0.779(0.084)	0.557(0.168)
	G-FMR	0.800(0.078)	0.167(0.068)	9.197(6.276)	0.713(0.056)	0.316(0.137)	-	-	-	-	-
	(separately)	-	-	-	-	-	1(0)	0.016(0.016)	1.033(0.231)	0.793(0.021)	0.585(0.043)
	MG-FMR	0.847(0.086)	0.031(0.025)	0.545(0.388)	0.842(0.040)	0.598(0.087)	-	-	-	-	-
	(separately)	-	-	-	-	-	1(0)	0.005(0.008)	0.231(0.117)	0.820(0.024)	0.639(0.048)
	Order-A	0.799(0.067)	0.047(0.012)	0.569(0.113)	0.807(0.039)	0.507(0.101)	1(0)	0.007(0.008)	0.223(0.092)	0.823(0.019)	0.645(0.042)
	Order-D	0.860(0.097)	0.057(0.040)	0.812(0.580)	0.828(0.072)	0.573(0.147)	1(0)	0.083(0.065)	0.592(1.174)	0.829(0.036)	0.657(0.073)
S2	Proposed	0.929(0.049)	0.056(0.022)	0.485(0.232)	0.852(0.013)	0.619(0.034)	1(0)	0.001(0.003)	0.163(0.070)	0.813(0.019)	0.624(0.038)
	S-FMR	0.739(0.075)	0.537(0.080)	16.588(18.840)	0.682(0.023)	0.183(0.059)	-	-	-	-	-
	(separately)	-	-	-	-	-	0.925(0.158)	0.247(0.086)	1.701(1.822)	0.790(0.045)	0.579(0.092)
	G-FMR	0.679(0.109)	0.071(0.079)	2.842(3.158)	0.803(0.052)	0.504(0.127)	-	-	-	-	-
	(separately)	-	-	-	-	-	1(0)	0.007(0.009)	0.718(0.271)	0.785(0.020)	0.594(0.041)
	MG-FMR	0.833(0.098)	0.041(0.031)	0.738(1.410)	0.844(0.034)	0.604(0.064)	-	-	-	-	-
	(separately)	-	-	-	-	-	1(0)	0.006(0.009)	0.172(0.078)	0.815(0.019)	0.627(0.038)
	Order-A	0.747(0.071)	0.048(0.016)	0.657(0.180)	0.808(0.041)	0.511(0.105)	1(0)	0.008(0.009)	0.189(0.088)	0.814(0.020)	0.627(0.039)
	Order-D	0.821(0.120)	0.039(0.023)	0.668(0.984)	0.842(0.035)	0.600(0.068)	1(0)	0.067(0.033)	0.141(0.193)	0.824(0.023)	0.647(0.046)
S3	Proposed	0.827(0.082)	0.046(0.029)	0.767(2.439)	0.865(0.019)	0.653(0.048)	1(0)	0.003(0.004)	0.132(0.058)	0.802(0.018)	0.603(0.037)
	S-FMR	0.734(0.069)	0.531(0.063)	16.791(16.194)	0.671(0.008)	0.168(0.029)	-	-	-	-	-
	(separately)	-	-	-	-	-	0.966(0.159)	0.239(0.081)	1.746(2.865)	0.766(0.067)	0.530(0.136)
	G-FMR	0.687(0.089)	0.091(0.053)	5.765(3.354)	0.749(0.065)	0.395(0.122)	-	-	-	-	-
	(separately)	-	-	-	-	-	1(0)	0.004(0.006)	0.809(0.275)	0.777(0.022)	0.553(0.046)
	MG-FMR	0.772(0.066)	0.037(0.026)	0.902(1.929)	0.847(0.056)	0.617(0.115)	-	-	-	-	-
	(separately)	-	-	-	-	-	1(0)	0.002(0.005)	0.141(0.055)	0.803(0.019)	0.604(0.039)
	Order-A	0.724(0.056)	0.047(0.017)	0.597(0.344)	0.810(0.044)	0.513(0.113)	1(0)	0.004(0.007)	0.143(0.058)	0.800(0.017)	0.598(0.035)
	Order-D	0.768(0.098)	0.041(0.028)	0.917(1.912)	0.847(0.054)	0.618(0.107)	1(0)	0.081(0.043)	0.211(0.497)	0.807(0.020)	0.613(0.041)

Table S3: Simulation results for the balanced subgroup design. $\rho = 1.5$ and $\xi = 1.5$. In each cell: mean (sd).

		TPRL	FPRL	MSEL	RIL	ARIL	TPRU	FPRU	MSEU	RIU	ARIU
S1	Proposed	0.953(0.063)	0.011(0.017)	0.960(1.359)	0.839(0.027)	0.575(0.058)	1(0)	0(0)	0.246(0.075)	0.821(0.018)	0.642(0.037)
	S-FMR	0.778(0.094)	0.533(0.075)	17.457(21.237)	0.681(0.027)	0.164(0.076)	-	-	-	-	-
	(separately)	-	-	-	-	-	0.775(0.438)	0.184(0.078)	9.593(10.274)	0.716(0.147)	0.432(0.295)
	G-FMR	0.780(0.100)	0.197(0.030)	10.113(7.645)	0.677(0.036)	0.169(0.109)	-	-	-	-	-
	(separately)	-	-	-	-	-	1(0)	0.006(0.003)	1.514(0.365)	0.808(0.019)	0.615(0.037)
	MG-FMR	0.912(0.076)	0.053(0.051)	2.612(2.378)	0.790(0.037)	0.469(0.085)	-	-	-	-	-
	(separately)	-	-	-	-	-	1(0)	0.001(0.002)	0.253(0.083)	0.823(0.017)	0.646(0.035)
	Order-A	0.874(0.071)	0.023(0.012)	0.857(0.153)	0.837(0.015)	0.564(0.039)	1(0)	0.001(0.003)	0.247(0.078)	0.824(0.018)	0.648(0.036)
	Order-D	0.909(0.091)	0.034(0.035)	2.384(3.335)	0.799(0.050)	0.496(0.098)	1(0)	0.052(0.035)	0.523(0.761)	0.839(0.024)	0.678(0.047)
S2	Proposed	0.967(0.046)	0.015(0.016)	0.540(0.585)	0.841(0.016)	0.577(0.036)	1(0)	0(0)	0.166(0.072)	0.809(0.022)	0.619(0.045)
	S-FMR	0.738(0.137)	0.509(0.096)	13.435(13.548)	0.675(0.039)	0.161(0.058)	-	-	-	-	-
	(separately)	-	-	-	-	-	0.756(0.452)	0.190(0.070)	7.577(9.199)	0.692(0.148)	0.384(0.296)
	G-FMR	0.720(0.067)	0.064(0.062)	7.870(10.848)	0.739(0.057)	0.354(0.130)	-	-	-	-	-
	(separately)	-	-	-	-	-	0.985(0.050)	0.003(0.002)	2.096(3.288)	0.764(0.075)	0.529(0.149)
	MG-FMR	0.904(0.080)	0.023(0.023)	1.643(1.683)	0.808(0.034)	0.504(0.069)	-	-	-	-	-
	(separately)	-	-	-	-	-	1(0)	0.001(0.002)	0.182(0.063)	0.810(0.023)	0.620(0.046)
	Order-A	0.780(0.085)	0.024(0.011)	0.890(0.156)	0.836(0.018)	0.565(0.047)	1(0)	0.001(0.003)	0.178(0.081)	0.806(0.019)	0.612(0.039)
	Order-D	0.911(0.103)	0.025(0.029)	2.103(3.550)	0.816(0.040)	0.521(0.081)	1(0)	0.068(0.020)	0.489(0.866)	0.838(0.020)	0.677(0.039)
S3	Proposed	0.859(0.069)	0.023(0.024)	1.248(1.688)	0.837(0.038)	0.576(0.083)	1(0)	0.001(0.003)	0.222(0.045)	0.803(0.022)	0.606(0.043)
	S-FMR	0.737(0.079)	0.498(0.069)	17.952(11.170)	0.745(0.105)	0.172(0.044)	-	-	-	-	-
	(separately)	-	-	-	-	-	0.867(0.335)	0.184(0.098)	5.072(8.583)	0.745(0.104)	0.491(0.209)
	G-FMR	0.688(0.052)	0.068(0.072)	8.627(8.180)	0.696(0.044)	0.263(0.122)	-	-	-	-	-
	(separately)	-	-	-	-	-	0.972(0.064)	0.001(0.002)	1.077(0.329)	0.784(0.021)	0.568(0.043)
	MG-FMR	0.782(0.086)	0.038(0.032)	3.386(4.059)	0.792(0.045)	0.490(0.083)	-	-	-	-	-
	(separately)	-	-	-	-	-	1(0)	0(0)	0.330(0.073)	0.795(0.025)	0.593(0.052)
	Order-A	0.788(0.048)	0.023(0.010)	0.781(0.126)	0.853(0.023)	0.609(0.054)	1(0)	0(0)	0.238(0.055)	0.796(0.018)	0.593(0.035)
	Order-D	0.779(0.108)	0.040(0.031)	4.509(5.347)	0.801(0.027)	0.506(0.064)	1(0)	0.065(0.028)	1.637(4.104)	0.807(0.051)	0.614(0.102)

Table S4: Simulation results for the unbalanced subgroup design. $\rho = 1.5$ and $\xi = 1.5$. In each cell: mean (sd).

		TPRL	FPRL	MSEL	RIL	ARIL	TPRU	FPRU	MSEU	RIU	ARIU
S1	Proposed	0.969(0.067)	0.017(0.021)	1.323(1.745)	0.825(0.048)	0.564(0.089)	1(0)	0(0)	0.251(0.093)	0.840(0.017)	0.679(0.035)
	S-FMR	0.828(0.086)	0.494(0.081)	13.095(21.563)	0.694(0.021)	0.223(0.047)	–	–	–	–	–
	(separately)	–	–	–	–	–	0.897(0.276)	0.236(0.082)	4.354(6.820)	0.778(0.117)	0.555(0.235)
	G-FMR	0.743(0.070)	0.173(0.077)	11.527(4.860)	0.661(0.043)	0.184(0.108)	–	–	–	–	–
	(separately)	–	–	–	–	–	0.996(0.021)	0.002(0.003)	1.543(0.453)	0.807(0.038)	0.612(0.076)
	MG-FMR	0.892(0.101)	0.040(0.025)	3.028(3.298)	0.798(0.060)	0.507(0.104)	–	–	–	–	–
	(separately)	–	–	–	–	–	1(0)	0.001(0.005)	0.320(0.120)	0.839(0.014)	0.676(0.028)
	Order-A	0.859(0.056)	0.019(0.011)	1.191(0.415)	0.793(0.037)	0.471(0.093)	1(0)	0.001(0.003)	0.277(0.108)	0.836(0.016)	0.670(0.034)
	Order-D	0.901(0.097)	0.039(0.040)	2.054(2.718)	0.811(0.072)	0.543(0.147)	1(0)	0.084(0.031)	0.482(0.705)	0.865(0.022)	0.729(0.045)
S2	Proposed	0.944(0.073)	0.019(0.030)	0.890(1.517)	0.827(0.037)	0.567(0.074)	1(0)	0(0)	0.167(0.073)	0.830(0.023)	0.659(0.046)
	S-FMR	0.768(0.073)	0.503(0.065)	14.360(14.651)	0.690(0.022)	0.211(0.051)	–	–	–	–	–
	(separately)	–	–	–	–	–	0.993(0.040)	0.230(0.089)	2.096(5.130)	0.803(0.051)	0.606(0.104)
	G-FMR	0.705(0.054)	0.084(0.080)	7.231(9.232)	0.718(0.061)	0.323(0.137)	–	–	–	–	–
	(separately)	–	–	–	–	–	1(0)	0.002(0.004)	1.120(0.371)	0.812(0.019)	0.622(0.038)
	MG-FMR	0.897(0.100)	0.037(0.038)	3.045(4.221)	0.800(0.053)	0.496(0.095)	–	–	–	–	–
	(separately)	–	–	–	–	–	1(0)	0(0)	0.197(0.87)	0.826(0.019)	0.652(0.037)
	Order-A	0.811(0.065)	0.021(0.011)	0.980(0.170)	0.784(0.032)	0.449(0.083)	1(0)	0.001(0.001)	0.187(0.079)	0.825(0.018)	0.651(0.037)
	Order-D	0.890(0.113)	0.027(0.032)	4.344(5.627)	0.805(0.055)	0.525(0.099)	1(0)	0.071(0.023)	1.145(2.492)	0.841(0.023)	0.680(0.046)
S3	Proposed	0.837(0.071)	0.029(0.023)	1.167(1.300)	0.796(0.072)	0.515(0.134)	1(0)	0.001(0.002)	0.147(0.062)	0.818(0.015)	0.634(0.031)
	S-FMR	0.753(0.075)	0.515(0.077)	13.747(17.233)	0.679(0.013)	0.180(0.031)	–	–	–	–	–
	(separately)	–	–	–	–	–	0.964(0.157)	0.202(0.079)	2.962(3.842)	0.788(0.076)	0.575(0.156)
	G-FMR	0.685(0.046)	0.073(0.119)	9.315(8.258)	0.677(0.055)	0.286(0.104)	–	–	–	–	–
	(separately)	–	–	–	–	–	1(0)	0.001(0.002)	1.136(0.333)	0.801(0.084)	0.601(0.169)
	MG-FMR	0.755(0.066)	0.039(0.026)	4.756(4.836)	0.760(0.061)	0.446(0.101)	–	–	–	–	–
	(separately)	–	–	–	–	–	1(0)	0(0)	0.158(0.070)	0.817(0.020)	0.631(0.040)
	Order-A	0.773(0.055)	0.030(0.009)	0.889(0.155)	0.803(0.041)	0.501(0.103)	1(0)	0(0)	0.140(0.063)	0.814(0.019)	0.628(0.038)
	Order-D	0.750(0.106)	0.038(0.040)	5.787(6.293)	0.757(0.074)	0.441(0.134)	1(0)	0.068(0.029)	1.630(1.858)	0.813(0.031)	0.625(0.063)

Table S5: Simulation results for the balanced subgroup design. $\rho = 1.5$ and $\xi = 1$. In each cell: mean (sd).

		TPRL	FPRL	MSEL	RIL	ARIL	TPRU	FPRU	MSEU	RIU	ARIU
S1	Proposed	0.935(0.074)	0.046(0.038)	0.710(1.177)	0.837(0.025)	0.569(0.059)	1(0)	0(0)	0.234(0.074)	0.832(0.019)	0.664(0.037)
	S-FMR	0.779(0.092)	0.497(0.114)	15.529(19.249)	0.689(0.021)	0.191(0.057)	–	–	–	–	–
	(separately)	–	–	–	–	–	0.898(0.415)	0.174(0.093)	2.062(1.526)	0.815(0.141)	0.630(0.281)
	G-FMR	0.704(0.075)	0.108(0.056)	7.218(16.012)	0.670(0.041)	0.170(0.116)	–	–	–	–	–
	(separately)	–	–	–	–	–	0.998(0)	0(0)	1.638(2.745)	0.804(0.034)	0.608(0.068)
	MG-FMR	0.750(0.061)	0.049(0.045)	6.593(6.970)	0.786(0.051)	0.467(0.100)	–	–	–	–	–
	(separately)	–	–	–	–	–	1(0)	0(0)	0.279(0.072)	0.822(0.020)	0.644(0.041)
	Order-A	0.802(0.057)	0.044(0.018)	0.992(0.082)	0.841(0.014)	0.577(0.037)	1(0)	0(0)	0.256(0.087)	0.824(0.019)	0.647(0.039)
	Order-D	0.766(0.080)	0.047(0.043)	5.146(7.685)	0.803(0.044)	0.505(0.087)	1(0)	0.042(0.038)	1.323(2.060)	0.847(0.023)	0.694(0.045)
S2	Proposed	0.894(0.075)	0.036(0.019)	0.397(0.079)	0.846(0.010)	0.588(0.026)	1(0)	0(0)	0.156(0.079)	0.811(0.020)	0.623(0.040)
	S-FMR	0.726(0.094)	0.527(0.102)	12.480(11.772)	0.688(0.019)	0.184(0.054)	–	–	–	–	–
	(separately)	–	–	–	–	–	0.908(0.308)	0.236(0.094)	1.913(1.476)	0.798(0.028)	0.596(0.046)
	G-FMR	0.593(0.045)	0.090(0.044)	9.918(11.640)	0.753(0.065)	0.296(0.151)	–	–	–	–	–
	(separately)	–	–	–	–	–	1(0)	0(0)	1.242(2.599)	0.806(0.047)	0.611(0.095)
	MG-FMR	0.694(0.092)	0.040(0.056)	3.409(4.037)	0.795(0.040)	0.485(0.073)	–	–	–	–	–
	(separately)	–	–	–	–	–	1(0)	0(0)	0.927(2.348)	0.804(0.039)	0.608(0.077)
	Order-A	0.712(0.086)	0.042(0.018)	0.599(0.096)	0.841(0.018)	0.576(0.047)	1(0)	0(0)	0.161(0.080)	0.813(0.020)	0.626(0.040)
	Order-D	0.732(0.131)	0.038(0.023)	1.216(2.206)	0.829(0.029)	0.553(0.061)	1(0)	0.040(0.023)	0.247(0.610)	0.830(0.018)	0.661(0.036)
S3	Proposed	0.847(0.086)	0.050(0.026)	0.555(0.350)	0.859(0.035)	0.629(0.080)	1(0)	0.001(0.004)	0.131(0.074)	0.795(0.018)	0.590(0.036)
	S-FMR	0.706(0.179)	0.494(0.152)	10.281(9.512)	0.675(0.039)	0.165(0.067)	–	–	–	–	–
	(separately)	–	–	–	–	–	0.955(0.258)	0.195(0.095)	2.746(3.512)	0.750(0.114)	0.499(0.228)
	G-FMR	0.571(0.054)	0.064(0.030)	12.531(8.475)	0.668(0.028)	0.238(0.083)	–	–	–	–	–
	(separately)	–	–	–	–	–	1(0)	0(0)	2.200(4.176)	0.762(0.062)	0.523(0.124)
	MG-FMR	0.627(0.088)	0.063(0.041)	7.756(6.054)	0.740(0.057)	0.396(0.102)	–	–	–	–	–
	(separately)	–	–	–	–	–	1(0)	0(0)	0.121(0.048)	0.799(0.021)	0.598(0.042)
	Order-A	0.767(0.061)	0.059(0.014)	0.718(0.086)	0.870(0.015)	0.653(0.041)	1(0)	0(0)	0.114(0.049)	0.800(0.020)	0.600(0.039)
	Order-D	0.680(0.093)	0.037(0.036)	3.586(4.819)	0.804(0.071)	0.520(0.142)	1(0)	0.047(0.028)	0.847(1.203)	0.812(0.023)	0.625(0.047)

S2.2 Additional Data Analysis Results

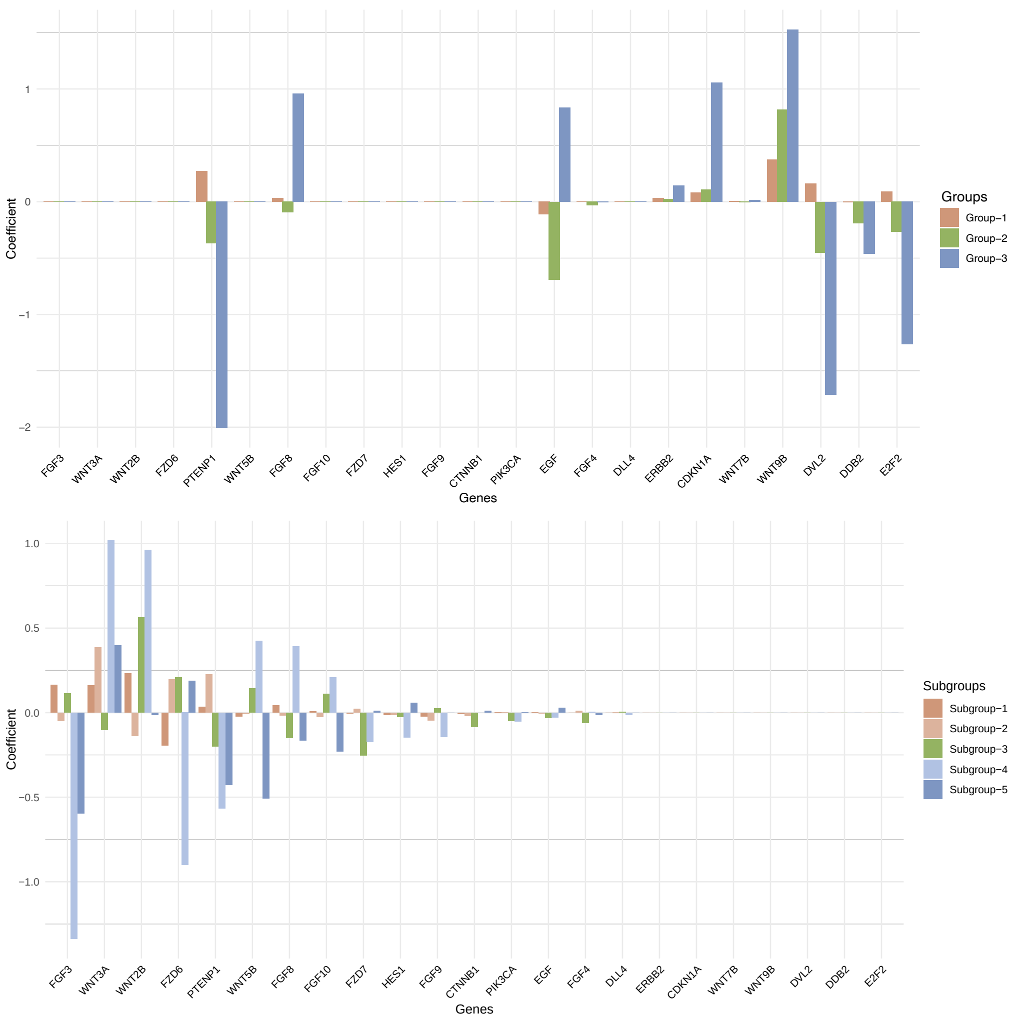

Figure S1: Data analysis using the MG-FMR method: genes sorted in descending order of the magnitudes of the coefficients within the subgroups. Top/bottom: three upper/five lower levels.

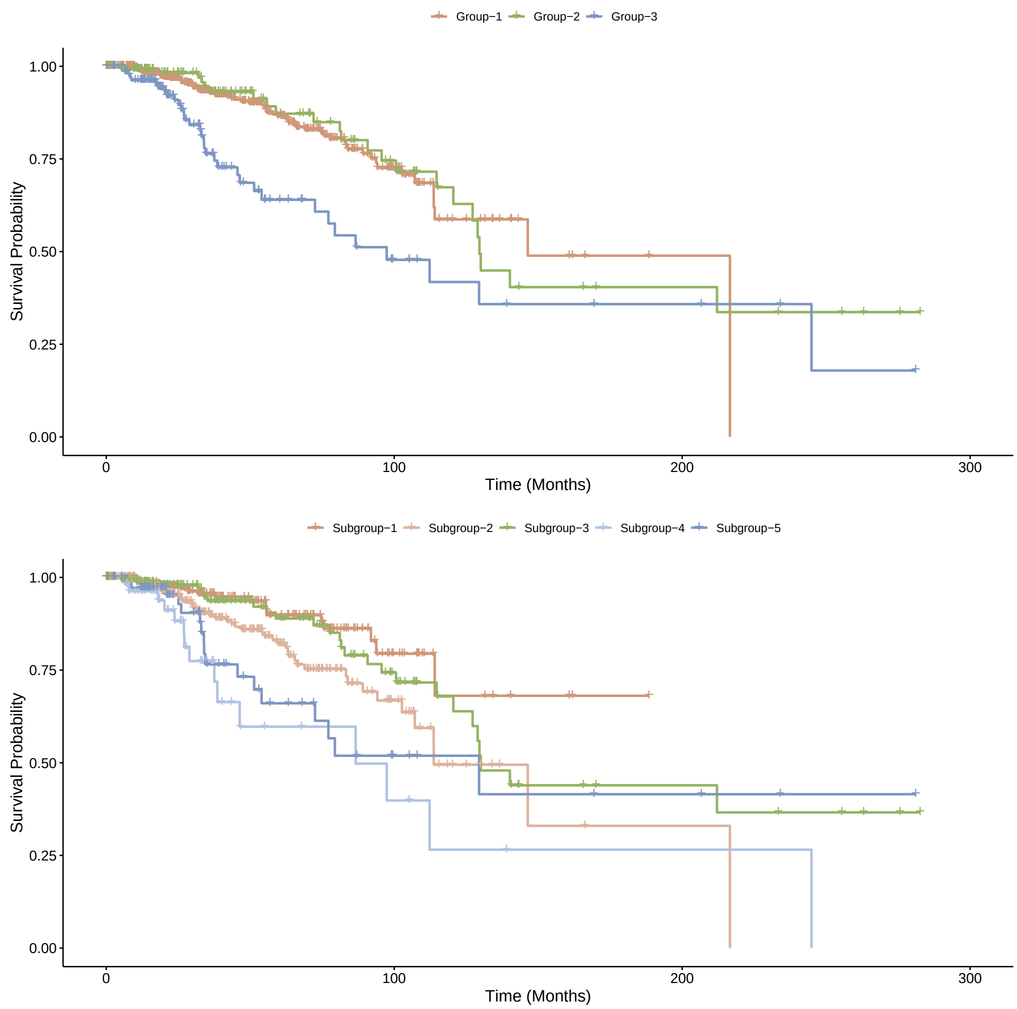

Figure S2: Top: survival curves for the three upper-level groups (p-value of the log-rank test $= 1 \times 10^{-4}$). Bottom: survival curves for the five lower-level subgroups (p-value $= 4 \times 10^{-5}$).

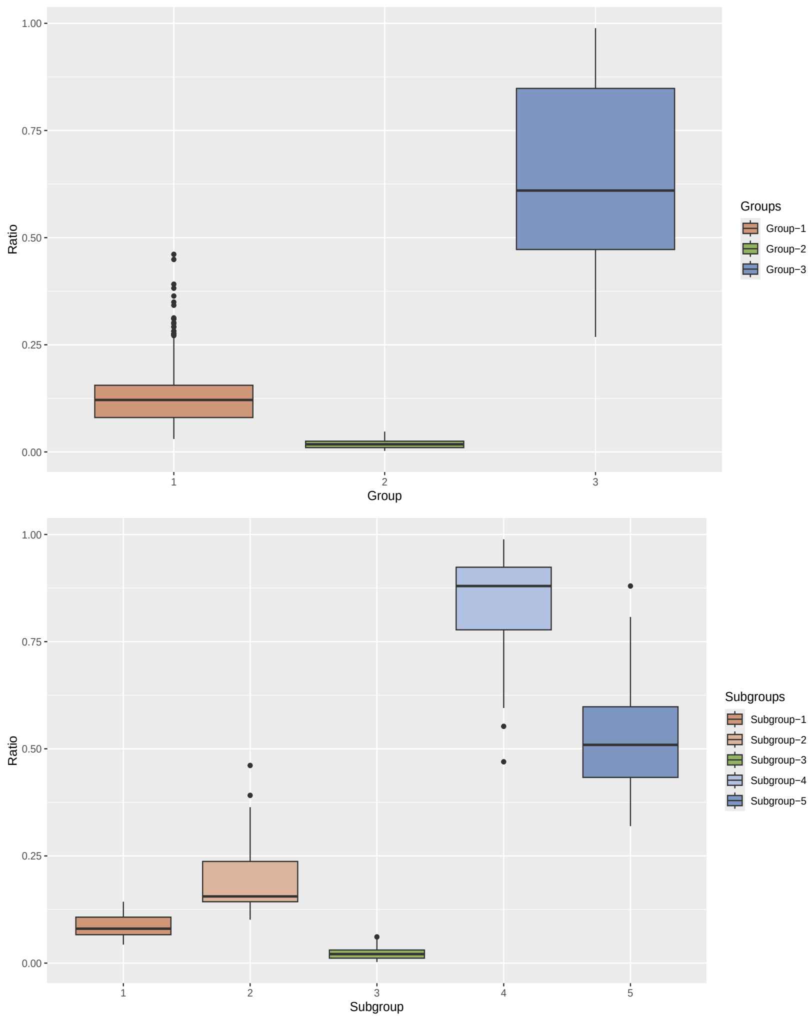

Figure S3: Top/bottom: distribution of lymph node positive ratio for the upper-level groups/lower-level subgroups.

S3 Extension

S3.1 Modeling Homogeneity in Effects Across Levels

With the proposed approach, there is no special attention to the alignment of the estimated effects between the upper and lower levels. When the upper level heterogeneity structure is determined by some strong effects and the lower level has additional contributions from some weak effects, intuitively, it may be sensible to expect that the estimates for the strong effects at the upper and lower levels have certain alignment. Consider the special case of linear regression and orthogonality between the strong and weak effects at the lower level. Then, those estimates are strictly equal. Here, we further explore the aspect of modeling homogeneity in the estimated effects (at multiple levels). This has some connections with the existing literature (Tang et al., 2021; Yang et al., 2019).

Consider the scenario with an index partition $\{\mathcal{G}_m\}_{m=1,2,\cdots,K_u}$ that describes how the lower-level subgroups have effects aligned with those of the groups at the upper level:

$$k \in \mathcal{G}_m, \quad \text{if} \quad |\boldsymbol{\beta}_m^u - \boldsymbol{\alpha}_k^l| \leq r, \quad m = 1, 2, \cdots, K_u, \quad k = 1, 2, \cdots, K_l, \tag{S3.68}$$

where r is small enough to ensure that $\mathcal{G}_m \cap \mathcal{G}_{m'} = \varnothing$ for $m \neq m'$. Note that $r = 0$ leads to the vector-based form of the multidirectional separation penalty (MDSP) (Tang et al., 2021). Here, we consider its addition to the proposed approach:

$$\mathcal{P}_{mdsp}\left(\boldsymbol{\beta}^u, \boldsymbol{\alpha}^l\right) = \sum_{k=1}^{K_l} P\left(\min\left(\sum_{j=1}^{p} P(|\boldsymbol{\beta}^u_{1j} - \boldsymbol{\alpha}^l_{kj}|), \sum_{j=1}^{p} P(|\boldsymbol{\beta}^u_{2j} - \boldsymbol{\alpha}^l_{kj}|), \cdots, \sum_{j=1}^{p} P(|\boldsymbol{\beta}^u_{K_u j} - \boldsymbol{\alpha}^l_{kj}|)\right)\right). \tag{S3.69}$$

Here, the upper-level FMR model automatically estimates the K_u directions. This penalty shares spirits and advantages similar to the original MDSP, which adopts Lasso penalization. It is noted that the zero direction for sparsity is redundant here. Additionally, unlike the commonly adopted pairwise penalization for heterogeneity analysis, we only need to identify a "nested" hierarchy (in terms of the aligned estimated effects). As such, it is not necessary to make pairs of parameters exactly the same for grouping. As shown below, this can lead to relaxed assumptions.

Overall, the objective function with the additional MDSP is:

$$\mathcal{L}_{mdsp}(\boldsymbol{\Omega}) = \mathcal{L}(\boldsymbol{\Omega}^u) + \mathcal{L}(\widetilde{\boldsymbol{\Omega}}^l) - \mathcal{P}_{ma}\left(\boldsymbol{\beta}^u, \boldsymbol{\alpha}^l, \boldsymbol{\gamma}^l\right) - \mathcal{P}_{mdsp}\left(\boldsymbol{\beta}^u, \boldsymbol{\alpha}^l\right). \tag{S3.70}$$

It is noted that the MDSP term, in a similar way as the original penalty, can be extended to multiple levels. The "nesting" hierarchy between different levels can also be adjusted with tuning parameters, and "nesting" here does not pertain to individuals (but rather estimated effects).

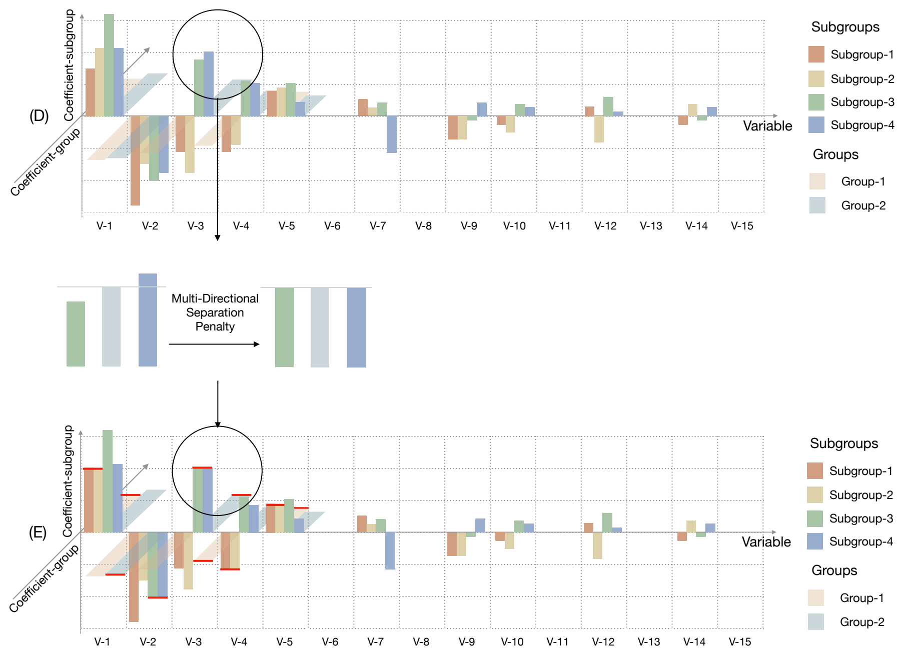

Figure S4: Scheme of the proposed extension. **Top** (**D**): the original proposed two-level analysis; **Bottom** (**E**): Aligning lower-level effects with upper-level ones promotes some subgroups to have equal effects via Multi-Directional Separation Penalty (MDSP).

S3.2 Statistical Properties

Consider oracle estimator $\widehat{\mathbf{\Omega}}^o$, where the oracle MDSP term is:

$$\mathcal{P}_2 = \sum_{l=1}^{K_u} \sum_{k \in \mathcal{G}_l^*} P\left(\sum_{j=1}^{p} P(|\beta_{mj} - \alpha_{kj}|) \right).$$

Theorem 2. *Assume that the conditions in Theorem 1 hold. Additionally, assume Condition 7 (described below) and* $\sqrt{\frac{K_l^3 s \log p}{n}} = O(\lambda_3)$*, b is much larger than* $(a+0.5)\cdot\lambda_3$*. Then, there exists an oracle estimator* $\widehat{\mathbf{\Omega}}^o$ *that shares the same properties as that in Theorem 1, and there exists an MDSP estimator* $\widehat{\mathbf{\Omega}}$ *that satisfies:*

$$Pr(\widehat{\mathbf{\Omega}} = \widehat{\mathbf{\Omega}}^o) \to 1.$$

This indicates:

$$Pr(\bigcap_{l=1,2,\cdots,K_u}\{\widehat{\mathcal{G}}_l = \mathcal{G}_l^*\}) \to 1, \quad Pr(\bigcap_{j\in\mathcal{D}_k^{*c}}\bigcap_{1\leqslant k\leqslant K_l}\widehat{\beta}_{lj} = \widehat{\alpha}_{kj}) \to 1.$$

S3.3 Proof of Theorem 2

Additional notations and condition

Suppose that we have the partition of $\{1,\cdots,K_l\}$, which represents the subordination of the upper-level groups and lower-level subgroups. $k \in \mathcal{G}_m, k=1,\cdots,K_l, m=1,\cdots,K_u$ indicates that the l-th subgroup at the lower level is subordinate to the k-th group at the upper level. This subordination is determined through the multidirectional separation penalty term, that is, $k\in\mathcal{G}_m$ if and only if

$$m = \text{argmin}\left(\sum_{j=1}^{p} P(|\beta_{1j}-\alpha_{kj}|), \sum_{j=1}^{p} P(|\beta_{2j}-\alpha_{kj}|), \cdots, \sum_{j=1}^{p} P(|\beta_{K_uj}-\alpha_{kj}|)\right). \tag{S3.71}$$

More discussions are provided below. Define $\mathcal{D}_k = \{j : |\boldsymbol{\beta}^*_{mj} - \boldsymbol{\alpha}^*_{kj}| \neq 0, k \in \mathcal{G}_m\}, \mathcal{D} = \{(k,j) : |\boldsymbol{\beta}^*_{mj} - \boldsymbol{\alpha}^*_{kj}| \neq 0, k \in \mathcal{G}_m\}$ is the nonzero set for pairs of subordination and $\mathcal{D}_k \in \mathcal{S}_1, d_k = |\mathcal{D}_k|, k = 1, \cdots, K_l$. Then we set $\boldsymbol{v} = (\boldsymbol{v}_1^\top, \boldsymbol{v}_2^\top, \boldsymbol{v}_\gamma^\top)^\top$ corresponding to $\boldsymbol{\Omega}_0^o = (\boldsymbol{\Omega}_0^{ou\top}, \boldsymbol{\Omega}_0^{ol\top}, \boldsymbol{\gamma}_0^{o\top})^\top$, where $\boldsymbol{\Omega}_0^{ou} = \left(\boldsymbol{\Omega}_{01}^{ou\top}, \cdots, \boldsymbol{\Omega}_{0K_u}^{ou\top}\right)^\top$, $\boldsymbol{\Omega}_{0k}^{ou} = \operatorname{vec}\left(\boldsymbol{\beta}^o_{k\mathcal{S}_1}, \sigma_k^{ou}\right)$ with $\boldsymbol{\beta}^o_{k\mathcal{S}_1} = \{\boldsymbol{\beta}^o_{kj} : j \in \mathcal{S}_1\}$ contains the nonzero parameters in the coefficients, $\boldsymbol{\gamma}_0^o$ is defined similarly. $\boldsymbol{\Omega}_0^{ol} = \left(\boldsymbol{\Omega}_{01}^{ol\top}, \cdots, \boldsymbol{\Omega}_{0K_l}^{ol\top}\right)^\top$, $\boldsymbol{\Omega}_{0k}^{ol} = \operatorname{vec}\left(\boldsymbol{\alpha}^o_{k\mathcal{D}_k}, \sigma_k^{ou}\right)$, and the true parameters can be similarly defined by $\boldsymbol{\Omega}_0^{o*}$.

The following condition is additionally assumed.

Condition 7.

$$\left\| \left[\frac{1}{n} \boldsymbol{X}^\top_{\mathcal{S}_1 \setminus \mathcal{D}_k} \boldsymbol{G}_k^l \boldsymbol{X}_{\mathcal{S}} \right] \left[\frac{1}{n} \boldsymbol{X}^\top_{\mathcal{S}} \boldsymbol{G}_k^l \boldsymbol{X}_{\mathcal{S}} + \mathcal{J} \right]^{-1} \right\|_\infty \leqslant O(h_o), \tag{S3.72}$$

where $h_o \leqslant \min(\sqrt{\frac{n}{sK_l^3 \log p}}, \sqrt{K_l^2 s})$ and $\mathcal{J} \in \mathbb{R}^{s \times s}$ is semi-positive-definite and defined as:

$$\mathcal{J} = \begin{bmatrix} \frac{1}{n} \boldsymbol{X}^\top_{\mathcal{S}_1 \setminus \mathcal{D}_k} \boldsymbol{G}_k^u \boldsymbol{X}_{\mathcal{S}_1 \setminus \mathcal{D}_k} & 0 \\ 0 & 0 \end{bmatrix}.$$

This is an adjusted irrepresentable condition and postulates that the variables corresponding to the same coefficients at both the lower and upper levels (subordinate) should be more independent.

Step 1:

We first consider an oracle estimator based on the true subgroup subordinate and all the important variables. This step is similar to Step 1 in the previous proof, while we set the MDSP term $\mathcal{P}_2$ as:

$$\mathcal{P}_2 = \sum_{m=1}^{K_u} \sum_{k \in \mathcal{G}_m} P\left(\sum_{j=1}^{p} P(|\beta_{mj} - \alpha_{kj}|) \right), \tag{S3.73}$$

where $P(\cdot)$ can be a Lasso penalty like Tang et al. (2021) with parameter λ_3 as well as an MCP or SCAD penalty like the composite penalties in Huang et al. (2012) with more parameters.

We continue to use the definition of $\boldsymbol{v}$ and adapt it to the new variables. With the previous strategies, we can show that:

$$\begin{aligned}
\boldsymbol{q}(\boldsymbol{v}) \leqslant & \left\langle \nabla_{\boldsymbol{\Omega}_0^{ou*}} \mathcal{H}_n \left(\boldsymbol{\Omega}_0^{ou*}, \boldsymbol{\Omega}_0^{ol*}, \boldsymbol{\gamma}_0^{o*} \mid \boldsymbol{\Omega}_0^{u}, \boldsymbol{\Omega}_0^{ol(t-1)}, \boldsymbol{\gamma}_0^{o(t-1)} \right), \boldsymbol{v}_1 \right\rangle - \frac{\varrho}{2} \|\boldsymbol{v}_1\|_2^2 \\
& + \left\langle \nabla_{\boldsymbol{\Omega}_0^{ol*}} \mathcal{H}_n \left(\boldsymbol{\Omega}_0^{ou*}, \boldsymbol{\Omega}_0^{ol*}, \boldsymbol{\gamma}_0^{o*} \mid \boldsymbol{\Omega}_0^{ou(t-1)}, \boldsymbol{\Omega}_0^{ol(t-1)}, \boldsymbol{\gamma}_0^{o(t-1)} \right), \boldsymbol{v}_2 \right\rangle - \frac{\varrho}{2} \|\boldsymbol{v}_2\|_2^2 \\
& + \left\langle \nabla_{\boldsymbol{\gamma}_0^{o*}} \mathcal{H}_n \left(\boldsymbol{\Omega}_0^{ou*}, \boldsymbol{\Omega}_0^{ol*}, \boldsymbol{\gamma}_0^{o*} \mid \boldsymbol{\Omega}_0^{ou(t-1)}, \boldsymbol{\Omega}_0^{ol(t-1)}, \boldsymbol{\gamma}_0^{o(t-1)} \right), \boldsymbol{v}_3 \right\rangle - \frac{\varrho}{2} \|\boldsymbol{v}_3\|_2^2, \\
& - \left[\mathcal{P}_1 \left(\boldsymbol{\Omega}_0^{o*} + \boldsymbol{v} \right) - \mathcal{P}_1 \left(\boldsymbol{\Omega}_0^{o*} \right) \right] - \left[\mathcal{P}_2 \left(\boldsymbol{\Omega}_0^{o*} + \boldsymbol{v} \right) - \mathcal{P}_2 \left(\boldsymbol{\Omega}_0^{o*} \right) \right],
\end{aligned} \tag{S3.74}$$

where considering $k = 1, \cdots, K_l$ individually, and $k \in \mathcal{G}_m$:

$$\nabla_{\boldsymbol{\alpha}^{o*}_{k\mathcal{S}_1}} \mathcal{H}_n\left(\boldsymbol{\Omega}_0^{ou*}, \boldsymbol{\Omega}_0^{ol*}, \boldsymbol{\gamma}_0^{o*} \mid \boldsymbol{\Omega}_0^{ou}, \boldsymbol{\Omega}_0^{ol}, \boldsymbol{\gamma}_0^{o}\right)$$
$$= \frac{\sigma_k^{l*}}{n} \sum_{i=1}^{n} \left[L_k^l\left(y_i, \boldsymbol{x}_i, \boldsymbol{\Omega}_0^{ou}, \boldsymbol{\Omega}_0^{ol}, \boldsymbol{\gamma}_0^{o}\right)\left(y_i - \boldsymbol{x}_{i\mathcal{S}_1\backslash\mathcal{D}_k}^{\top}\boldsymbol{\beta}^*_{m\mathcal{S}_1\backslash\mathcal{D}_k} - \boldsymbol{x}_{i\mathcal{D}_k}^{\top}\boldsymbol{\alpha}^*_{k\mathcal{D}_k} - \boldsymbol{x}_{i\mathcal{S}_2}^{\top}\boldsymbol{\gamma}^*_{k\mathcal{S}_2}\right)\boldsymbol{x}_{i\mathcal{S}_1}\right]$$

$$\nabla_{\boldsymbol{\gamma}^{*}_{k\mathcal{S}_2}} \mathcal{H}_n\left(\boldsymbol{\Omega}_0^{ou*}, \boldsymbol{\Omega}_0^{ol*}, \boldsymbol{\gamma}_0^{o*} \mid \boldsymbol{\Omega}_0^{ou}, \boldsymbol{\Omega}_0^{ol}, \boldsymbol{\gamma}_0^{o}\right)$$
$$= \frac{\sigma_k^{l*}}{n} \sum_{i=1}^{n} \left[L_k^l\left(y_i, \boldsymbol{x}_i, \boldsymbol{\Omega}_0^{ou}, \boldsymbol{\Omega}_0^{ol}, \boldsymbol{\gamma}_0^{o}\right)\left(y_i - \boldsymbol{x}_{i\mathcal{S}_1\backslash\mathcal{D}_k}^{\top}\boldsymbol{\beta}^*_{m\mathcal{S}_1\backslash\mathcal{D}_k} - \boldsymbol{x}_{i\mathcal{D}_k}^{\top}\boldsymbol{\alpha}^*_{k\mathcal{D}_k} - \boldsymbol{x}_{i\mathcal{S}_2}^{\top}\boldsymbol{\gamma}^*_{k\mathcal{S}_2}\right)\boldsymbol{x}_{i\mathcal{S}_2}\right]$$

$$\nabla_{\boldsymbol{\sigma}^{l*}_{k}} \mathcal{H}_n\left(\boldsymbol{\Omega}_0^{ou*}, \boldsymbol{\Omega}_0^{ol*}, \boldsymbol{\gamma}_0^{o*} \mid \boldsymbol{\Omega}_0^{ou}, \boldsymbol{\Omega}_0^{ol}, \boldsymbol{\gamma}_0^{o}\right)$$
$$= \frac{1}{2n\sigma_k^{l*}} \sum_{i=1}^{n} \left[L_k^l\left(y_i, \boldsymbol{x}_i, \boldsymbol{\Omega}_0^{ou}, \boldsymbol{\Omega}_0^{ol}, \boldsymbol{\gamma}_0^{o}\right)\right]$$
$$-\frac{1}{2n} \sum_{i=1}^{n} \left[L_k^l\left(y_i, \boldsymbol{x}_i, \boldsymbol{\Omega}_0^{ou}, \boldsymbol{\Omega}_0^{ol}, \boldsymbol{\gamma}_0^{o}\right)\left(y_i - \boldsymbol{x}_{i\mathcal{S}_1\backslash\mathcal{D}_k}^{\top}\boldsymbol{\beta}^*_{m\mathcal{S}_1\backslash\mathcal{D}_k} - \boldsymbol{x}_{i\mathcal{D}_k}^{\top}\boldsymbol{\alpha}^*_{k\mathcal{D}_k} - \boldsymbol{x}_{i\mathcal{S}_2}^{\top}\boldsymbol{\gamma}^*_{k\mathcal{S}_2}\right)^2\right], \tag{S3.75}$$

and for $m = 1, \cdots, K_l$, $j \in \mathcal{S}_1$ we have:

$$\nabla_{\boldsymbol{\beta}^{o*}_{mj}} \mathcal{H}_n\left(\boldsymbol{\Omega}_0^{ou*}, \boldsymbol{\Omega}_0^{ol*}, \boldsymbol{\gamma}_0^{o*} \mid \boldsymbol{\Omega}_0^{ou}, \boldsymbol{\Omega}_0^{ol}, \boldsymbol{\gamma}_0^{o}\right) = \frac{\sigma_k^{u*}}{n} \sum_{i=1}^{n} \left[L_k^l\left(y_i, \boldsymbol{x}_i, \boldsymbol{\Omega}_0^{ou}\right)\left(y_i - \boldsymbol{x}_{i\mathcal{S}_1}^{\top}\boldsymbol{\beta}^*_{m\mathcal{S}_1}\right)x_{ij}\right]$$
$$+\sum_{k=1}^{K_u} I\left(k \in \mathcal{G}_m, j \in \mathcal{S}_1\backslash\mathcal{D}_k\right)\frac{\sigma_k^{l*}}{n}\sum_{i=1}^{n}\left[L_k^l\left(y_i, \boldsymbol{x}_i, \boldsymbol{\Omega}_0^{ou}, \boldsymbol{\Omega}_0^{ol}, \boldsymbol{\gamma}_0^{o}\right)\right.$$
$$\left.\left(y_i - \boldsymbol{x}_{i\mathcal{S}_1\backslash\mathcal{D}_k}^{\top}\boldsymbol{\beta}^*_{m\mathcal{S}_1\backslash\mathcal{D}_k} - \boldsymbol{x}_{i\mathcal{D}_k}^{\top}\boldsymbol{\alpha}^*_{k\mathcal{D}_k} - \boldsymbol{x}_{i\mathcal{S}_2}^{\top}\boldsymbol{\gamma}^*_{k\mathcal{S}_2}\right)x_{ij}\right]$$

$$\nabla_{\boldsymbol{\sigma}^{ou*}_{k}} \mathcal{H}_n\left(\boldsymbol{\Omega}_0^{ou*}, \boldsymbol{\Omega}_0^{ol*}, \boldsymbol{\gamma}_0^{o*} \mid \boldsymbol{\Omega}_0^{ou}, \boldsymbol{\Omega}_0^{ol}, \boldsymbol{\gamma}_0^{o}\right)$$
$$= \frac{1}{2n\sigma_k^{l*}} \sum_{i=1}^{n} \left[L_k^l\left(y_i, \boldsymbol{x}_i, \boldsymbol{\Omega}_0^{ou}\right)\right] - \frac{1}{2n}\sum_{i=1}^{n}\left[L_k^l\left(y_i, \boldsymbol{x}_i, \boldsymbol{\Omega}_0^{ou}\right)\left(y_i - \boldsymbol{x}_{i\mathcal{S}_1}^{\top}\boldsymbol{\beta}^*_{m\mathcal{S}_1}\right)^2\right], \tag{S3.76}$$

where we can get (S3.74) with a similar strategy as in S1.2. Also, $\left[\mathcal{P}_1\left(\boldsymbol{\Omega}_0^{o*} + \boldsymbol{v}\right) - \mathcal{P}_1\left(\boldsymbol{\Omega}_0^{o*}\right)\right]$ has the same conclusion as in (S1.28). The minor differences come from the fact that we replace $\boldsymbol{\alpha}_{k\mathcal{S}_1\setminus\mathcal{D}_k}$ by $\boldsymbol{\beta}_{m\mathcal{S}_1\setminus\mathcal{D}_k}$ if $k \in \mathcal{G}_m$, and so we can get very similar conclusions via this oracle version.

This oracle version of the penalty and triangle inequality implies that:

$$[\mathcal{P}_2\left(\boldsymbol{\Omega}_0^{o*}+\boldsymbol{v}\right)-\mathcal{P}_2\left(\boldsymbol{\Omega}_0^{o*}\right)] \leqslant C_1\sqrt{\frac{K_l^3 s\log p}{n}}\left\|\boldsymbol{v}_1\right\|_2+C_2\sqrt{\frac{K_l^3 s\log p}{n}}\left\|\boldsymbol{v}_2\right\|_2, \tag{S3.77}$$

where C_1 and C_2 are positive constants. Note that, with Lasso, λ_3 needs to have the order of $O(\sqrt{\frac{K_l^3 s\log p}{n}})$. Further, this term can be negative with MCP and SCAD by combining with the minimal signal conditions. Then, with the two lemmas and the same steps as above, we can show that the oracle estimator error can be bounded as:

$$\left\|\widehat{\boldsymbol{\Omega}}_0^o-\boldsymbol{\Omega}_0^*\right\|_2=O\left(\sqrt{\frac{K_l^3 s\log p}{n}}\right), \quad \left\|\widehat{\boldsymbol{\gamma}}_0^o-\boldsymbol{\gamma}_0^*\right\|_2=O\left(\sqrt{\frac{K_l^3 s_2\log p}{n}}\right), \tag{S3.78}$$

with probability tending to 1. Here we replace $\widehat{\boldsymbol{\alpha}}_{k\mathcal{S}_1\setminus\mathcal{D}_k}$ by $\widehat{\boldsymbol{\beta}}_{m\mathcal{S}_1\setminus\mathcal{D}_k}$ if $k \in \mathcal{G}_m$ in $\widehat{\boldsymbol{\Omega}}_0^o$.

Step 2:

We show that $\widehat{\boldsymbol{\Omega}}_o=\left(\widehat{\boldsymbol{\Omega}}_0^o, 0\right)$ is a local maximizer of the objective function. This indicates that:

$$\left\|\widehat{\boldsymbol{\Omega}}^o-\boldsymbol{\Omega}^*\right\|_2=O\left(\sqrt{\frac{K_l^3 s\log p}{n}}\right), \quad \left\|\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*\right\|_2=O\left(\sqrt{\frac{K_l^3 s_2\log p}{n}}\right), \tag{S3.79}$$

in Step 1 holds due to the conclusion drawn in this step. Following Step 2 in Theorem 1, we can select $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ in the same way. Here we define

$$\boldsymbol{\alpha}_{k\mathcal{S}_1\backslash\mathcal{D}_k} + \boldsymbol{\varkappa}_{k\mathcal{S}_1\backslash\mathcal{D}_k} = \boldsymbol{\beta}_{m\mathcal{S}_1\backslash\mathcal{D}_k}, \quad k \in \mathcal{G}_m, k = 1, \cdots, K_l, m = 1, \cdots, K_u, \tag{S3.80}$$

where $\boldsymbol{\varkappa}_{k\mathcal{S}_1\backslash\mathcal{D}_k}$ is the new variable and $\boldsymbol{\varkappa}^*_{k\mathcal{S}_1\backslash\mathcal{D}_k} = 0$. So it suffices to show that $\|\boldsymbol{z}_o\|_\infty \leqslant \lambda_3$, where $\boldsymbol{z}_o = \{\boldsymbol{z}_{ok}\}_{k=1,\cdots,K_l}$ is a vector and $\boldsymbol{z}_{ok} = \nabla_{\boldsymbol{\varkappa}_{k\mathcal{S}_1\backslash\mathcal{D}_k}} \mathcal{Q}(\widehat{\boldsymbol{\Omega}}^o)$. For each $k = 1, \cdots, K_l$ and $k \in \mathcal{G}_m$, it can be obtained that:

$$\begin{pmatrix} \nabla_{\boldsymbol{\beta}_{m\mathcal{D}_k}} \mathcal{Q}(\widehat{\boldsymbol{\Omega}}^o) \\ \nabla_{\boldsymbol{\beta}_{m\mathcal{S}_1\backslash\mathcal{D}_k}} \mathcal{Q}(\widehat{\boldsymbol{\Omega}}^o) \\ \nabla_{\boldsymbol{\alpha}_{k\mathcal{D}_k}} \mathcal{Q}(\widehat{\boldsymbol{\Omega}}^o) \\ \nabla_{\boldsymbol{\gamma}_{k\mathcal{S}_2}} \mathcal{Q}(\widehat{\boldsymbol{\Omega}}^o) \\ \nabla_{\boldsymbol{\varkappa}_{k\mathcal{S}_1\backslash\mathcal{D}_k}} \mathcal{Q}(\widehat{\boldsymbol{\Omega}}^o) \end{pmatrix} = \begin{pmatrix} \nabla_{\boldsymbol{\beta}_{m\mathcal{D}_k}} \mathcal{Q}(\boldsymbol{\Omega}^*) \\ \nabla_{\boldsymbol{\beta}_{m\mathcal{S}_1\backslash\mathcal{D}_k}} \mathcal{Q}(\boldsymbol{\Omega}^*) \\ \nabla_{\boldsymbol{\alpha}_{k\mathcal{D}_k}} \mathcal{Q}(\boldsymbol{\Omega}^*) \\ \nabla_{\boldsymbol{\gamma}_{k\mathcal{S}_2}} \mathcal{Q}(\boldsymbol{\Omega}^*) \\ \nabla_{\boldsymbol{\varkappa}_{\mathcal{S}_1\backslash\mathcal{D}_k}} \mathcal{Q}(\boldsymbol{\Omega}^*) \end{pmatrix} + \mathcal{B} \cdot \begin{pmatrix} \widehat{\boldsymbol{\beta}}^o_{m\mathcal{S}_1\mathcal{D}_k} - \boldsymbol{\beta}^*_{\mathcal{S}_1\mathcal{D}_k} \\ \widehat{\boldsymbol{\beta}}^o_{m\mathcal{S}_1\backslash\mathcal{D}_k} - \boldsymbol{\beta}^*_{m\mathcal{S}_1\backslash\mathcal{D}_k} \\ \widehat{\boldsymbol{\alpha}}^o_{k\mathcal{D}_k} - \boldsymbol{\alpha}^*_{k\mathcal{D}_k} \\ \widehat{\boldsymbol{\gamma}}^o_{k\mathcal{S}_2} - \boldsymbol{\gamma}^*_{k\mathcal{S}_2} \\ \mathbf{0} \end{pmatrix} + \begin{pmatrix} \boldsymbol{r}(\boldsymbol{\Delta})_{\boldsymbol{\beta}_{m\mathcal{D}_k}} \\ \boldsymbol{r}(\boldsymbol{\Delta})_{\boldsymbol{\beta}_{m\mathcal{S}_1\backslash\mathcal{D}_k}} \\ \boldsymbol{r}(\boldsymbol{\Delta})_{\boldsymbol{\alpha}_{k\mathcal{D}_k}} \\ \boldsymbol{r}(\boldsymbol{\Delta})_{\boldsymbol{\gamma}_{k\mathcal{S}_2}} \\ \boldsymbol{r}(\boldsymbol{\Delta})_{\boldsymbol{\varkappa}_{k\mathcal{S}_1\backslash\mathcal{D}_k}} \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \\ \boldsymbol{z}_{ok} \end{pmatrix}, \tag{S3.81}$$

where $\boldsymbol{r}(\boldsymbol{\Delta})$ is the residual of the first-order Taylor expansion of the gradient. And $\mathcal{A} =$

$$\begin{pmatrix} \nabla^2_{\boldsymbol{\beta}_{m\mathcal{D}_k}\boldsymbol{\beta}_{m\mathcal{D}_k}} \mathcal{Q}(\boldsymbol{\Omega}^*) & \nabla^2_{\boldsymbol{\beta}_{m\mathcal{D}_k}\boldsymbol{\beta}_{m\mathcal{S}_1\backslash\mathcal{D}_k}} \mathcal{Q}(\boldsymbol{\Omega}^*) & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \nabla^2_{\boldsymbol{\beta}_{m\mathcal{S}_1\backslash\mathcal{D}_k}\boldsymbol{\beta}_{m\mathcal{D}_k}} \mathcal{Q}(\boldsymbol{\Omega}^*) & \nabla^2_{\boldsymbol{\beta}_{m\mathcal{S}_1\backslash\mathcal{D}_k}\boldsymbol{\beta}_{m\mathcal{S}_1\backslash\mathcal{D}_k}} \mathcal{Q}(\boldsymbol{\Omega}^*) & \nabla^2_{\boldsymbol{\beta}_{m\mathcal{S}_1\backslash\mathcal{D}_k}\boldsymbol{\alpha}_{k\mathcal{D}_k}} \mathcal{Q}(\boldsymbol{\Omega}^*) & \nabla^2_{\boldsymbol{\beta}_{m\mathcal{S}_1\backslash\mathcal{D}_k}\boldsymbol{\gamma}_{k\mathcal{S}_2}} \mathcal{Q}(\boldsymbol{\Omega}^*) & \nabla^2_{\boldsymbol{\beta}_{m\mathcal{S}_1\backslash\mathcal{D}_k}\boldsymbol{\varkappa}_{k\mathcal{S}_1\backslash\mathcal{D}_k}} \mathcal{Q}(\boldsymbol{\Omega}^*) \\ \mathbf{0} & \nabla^2_{\boldsymbol{\alpha}_{k\mathcal{D}_k}\boldsymbol{\beta}_{m\mathcal{S}_1\backslash\mathcal{D}_k}} \mathcal{Q}(\boldsymbol{\Omega}^*) & \nabla^2_{\boldsymbol{\alpha}_{k\mathcal{D}_k}\boldsymbol{\alpha}_{k\mathcal{D}_k}} \mathcal{Q}(\boldsymbol{\Omega}^*) & \nabla^2_{\boldsymbol{\alpha}_{k\mathcal{D}_k}\boldsymbol{\gamma}_{k\mathcal{S}_2}} \mathcal{Q}(\boldsymbol{\Omega}^*) & \nabla^2_{\boldsymbol{\alpha}_{k\mathcal{D}_k}\boldsymbol{\varkappa}_{k\mathcal{S}_1\backslash\mathcal{D}_k}} \mathcal{Q}(\boldsymbol{\Omega}^*) \\ \mathbf{0} & \nabla^2_{\boldsymbol{\gamma}_{k\mathcal{S}_2}\boldsymbol{\beta}_{m\mathcal{S}_1\backslash\mathcal{D}_k}} \mathcal{Q}(\boldsymbol{\Omega}^*) & \nabla^2_{\boldsymbol{\gamma}_{k\mathcal{S}_2}\boldsymbol{\alpha}_{k\mathcal{D}_k}} \mathcal{Q}(\boldsymbol{\Omega}^*) & \nabla^2_{\boldsymbol{\gamma}_{k\mathcal{S}_2}\boldsymbol{\gamma}_{k\mathcal{S}_2}} \mathcal{Q}(\boldsymbol{\Omega}^*) & \nabla^2_{\boldsymbol{\gamma}_{k\mathcal{S}_2}\boldsymbol{\varkappa}_{k\mathcal{S}_1\backslash\mathcal{D}_k}} \mathcal{Q}(\boldsymbol{\Omega}^*) \\ \mathbf{0} & \nabla^2_{\boldsymbol{\varkappa}_{k\mathcal{S}_1\backslash\mathcal{D}_k}\boldsymbol{\beta}_{m\mathcal{S}_1\backslash\mathcal{D}_k}} \mathcal{Q}(\boldsymbol{\Omega}^*) & \nabla^2_{\boldsymbol{\varkappa}_{k\mathcal{S}_1\backslash\mathcal{D}_k}\boldsymbol{\alpha}_{k\mathcal{D}_k}} \mathcal{Q}(\boldsymbol{\Omega}^*) & \nabla^2_{\boldsymbol{\varkappa}_{k\mathcal{S}_1\backslash\mathcal{D}_k}\boldsymbol{\gamma}_{k\mathcal{S}_2}} \mathcal{Q}(\boldsymbol{\Omega}^*) & \nabla^2_{\boldsymbol{\varkappa}_{k\mathcal{S}_1\backslash\mathcal{D}_k}\boldsymbol{\varkappa}_{k\mathcal{S}_1\backslash\mathcal{D}_k}} \mathcal{Q}(\boldsymbol{\Omega}^*) \end{pmatrix}. \tag{S3.82}$$

Then, we have

$$\begin{aligned}
\boldsymbol{z}_{ok} &= \nabla_{\varkappa_{k\mathcal{S}_1\setminus\mathcal{D}_k}} \mathcal{Q}(\boldsymbol{\Omega}^*) + \boldsymbol{r}(\boldsymbol{\Delta})_{\varkappa_{k\mathcal{S}_1\setminus\mathcal{D}_k}} \\
&+ \nabla^2_{\varkappa_{k\mathcal{S}_1\setminus\mathcal{D}_k}\boldsymbol{\beta}_{m\mathcal{S}_1\setminus\mathcal{D}_k}} \mathcal{Q}(\boldsymbol{\Omega}^*)(\widehat{\boldsymbol{\beta}}^o_{m\mathcal{S}_1\setminus\mathcal{D}_k} - \boldsymbol{\beta}^*_{m\mathcal{S}_1\setminus\mathcal{D}_k}) + \nabla^2_{\varkappa_{k\mathcal{S}_1\setminus\mathcal{D}_k}\boldsymbol{\alpha}_{k\mathcal{S}_1\setminus\mathcal{D}_k}} \mathcal{Q}(\boldsymbol{\Omega}^*)(\widehat{\boldsymbol{\alpha}}^o_{k\mathcal{S}_1\setminus\mathcal{D}_k} - \boldsymbol{\alpha}^*_{k\mathcal{S}_1\setminus\mathcal{D}_k}) \\
&+ \nabla^2_{\varkappa_{k\mathcal{S}_1\setminus\mathcal{D}_k}\boldsymbol{\gamma}_{k\mathcal{S}_2}} \mathcal{Q}(\boldsymbol{\Omega}^*)(\widehat{\boldsymbol{\gamma}}^o_{k\mathcal{S}_2} - \boldsymbol{\gamma}^*_{k\mathcal{S}_2}) \\
&= \frac{\sigma_k^{l*}}{n}\sum_{i=1}^n L_k^{l*}\left(y_i - \boldsymbol{x}^\top_{i\mathcal{S}_1\setminus\mathcal{D}_k}\boldsymbol{\beta}^*_{m\mathcal{S}_1\setminus\mathcal{D}_k} - \boldsymbol{x}^\top_{i\mathcal{D}_k}\boldsymbol{\alpha}^*_{k\mathcal{D}_k} - \boldsymbol{x}^\top_{i\mathcal{S}_2}\boldsymbol{\gamma}^*_{k\mathcal{S}_2}\right)\boldsymbol{x}_{i\mathcal{S}_1\setminus\mathcal{D}_k} + \boldsymbol{r}(\boldsymbol{\Delta})_{\varkappa_{k\mathcal{S}_1\setminus\mathcal{D}_k}} \\
&+ \frac{\sigma_k^{l*}}{n}\sum_{i=1}^n L_k^{l*}\left[\boldsymbol{x}^\top_{i\mathcal{S}_1\setminus\mathcal{D}_k}(\widehat{\boldsymbol{\beta}}^o_{m\mathcal{S}_1\setminus\mathcal{D}_k} - \boldsymbol{\beta}^*_{m\mathcal{S}_1\setminus\mathcal{D}_k}) + \boldsymbol{x}^\top_{i\mathcal{D}_k}(\widehat{\boldsymbol{\alpha}}^o_{k\mathcal{D}_k} - \boldsymbol{\alpha}^*_{k\mathcal{D}_k}) + \boldsymbol{x}^\top_{i\mathcal{S}_2}(\widehat{\boldsymbol{\gamma}}^o_{k\mathcal{S}_2} - \boldsymbol{\gamma}^*_{k\mathcal{S}_2})\right]\boldsymbol{x}_{i\mathcal{S}_1\setminus\mathcal{D}_k} \\
&= \xi_{ok1} + \boldsymbol{r}(\boldsymbol{\Delta})_{\varkappa_{k\mathcal{S}_1\setminus\mathcal{D}_k}} + \xi_{ok2}, \quad , k = 1, 2, \cdots, K_l.
\end{aligned} \tag{S3.83}$$

According to (S1.36), we can use the same inequality to show that $\|\xi_{ok1}\|_\infty = O(\sqrt{\frac{K_l \log p}{n}})$ with probability at least $1 - \frac{1}{p}$, which tends to 1.

For ξ_{ok2}, we have:

$$\begin{pmatrix} \widehat{\boldsymbol{\beta}}^o_{m\mathcal{S}_1\setminus\mathcal{D}_k} - \boldsymbol{\beta}^*_{m\mathcal{S}_1\setminus\mathcal{D}_k} \\ \widehat{\boldsymbol{\alpha}}^o_{k\mathcal{D}_k} - \boldsymbol{\alpha}^*_{k\mathcal{D}_k} \\ \widehat{\boldsymbol{\gamma}}^o_{k\mathcal{S}_2} - \boldsymbol{\gamma}^*_{k\mathcal{S}_2} \end{pmatrix} = -\mathcal{C}\cdot\begin{pmatrix} \nabla_{\boldsymbol{\beta}_{m\mathcal{S}_1\setminus\mathcal{D}_k}}\mathcal{Q}(\boldsymbol{\Omega}^*) & +\boldsymbol{r}(\boldsymbol{\Delta})_{\boldsymbol{\beta}_{m\mathcal{S}_1\setminus\mathcal{D}_k}} \\ \nabla_{\boldsymbol{\alpha}_{k\mathcal{D}_k}}\mathcal{Q}(\boldsymbol{\Omega}^*) & +\boldsymbol{r}(\boldsymbol{\Delta})_{\boldsymbol{\alpha}_{k\mathcal{D}_k}} \\ \nabla_{\boldsymbol{\gamma}_{k\mathcal{S}_2}}\mathcal{Q}(\boldsymbol{\Omega}^*) & +\boldsymbol{r}(\boldsymbol{\Delta})_{\boldsymbol{\gamma}_{k\mathcal{S}_2}} \end{pmatrix}, \tag{S3.84}$$

where $\mathcal{C} =$

$$\begin{pmatrix} \nabla^2_{\boldsymbol{\beta}_{m\mathcal{S}_1\setminus\mathcal{D}_k}\boldsymbol{\beta}_{m\mathcal{S}_1\setminus\mathcal{D}_k}}\mathcal{Q}(\boldsymbol{\Omega}^*) & \nabla^2_{\boldsymbol{\beta}_{m\mathcal{S}_1\setminus\mathcal{D}_k}\boldsymbol{\alpha}_{k\mathcal{D}_k}}\mathcal{Q}(\boldsymbol{\Omega}^*) & \nabla^2_{\boldsymbol{\beta}_{m\mathcal{S}_1\setminus\mathcal{D}_k}\boldsymbol{\gamma}_{k\mathcal{S}_2}}\mathcal{Q}(\boldsymbol{\Omega}^*) \\ \nabla^2_{\boldsymbol{\alpha}_{k\mathcal{D}_k}\boldsymbol{\beta}_{m\mathcal{S}_1\setminus\mathcal{D}_k}}\mathcal{Q}(\boldsymbol{\Omega}^*) & \nabla^2_{\boldsymbol{\alpha}_{k\mathcal{D}_k}\boldsymbol{\alpha}_{k\mathcal{D}_k}}\mathcal{Q}(\boldsymbol{\Omega}^*) & \nabla^2_{\boldsymbol{\alpha}_{k\mathcal{D}_k}\boldsymbol{\gamma}_{k\mathcal{S}_2}}\mathcal{Q}(\boldsymbol{\Omega}^*) \\ \nabla^2_{\boldsymbol{\gamma}_{k\mathcal{S}_2}\boldsymbol{\beta}_{m\mathcal{S}_1\setminus\mathcal{D}_k}}\mathcal{Q}(\boldsymbol{\Omega}^*) & \nabla^2_{\boldsymbol{\gamma}_{k\mathcal{S}_2}\boldsymbol{\alpha}_{k\mathcal{D}_k}}\mathcal{Q}(\boldsymbol{\Omega}^*) & \nabla^2_{\boldsymbol{\gamma}_{k\mathcal{S}_2}\boldsymbol{\gamma}_{k\mathcal{S}_2}}\mathcal{Q}(\boldsymbol{\Omega}^*) \end{pmatrix}^{-1}. \tag{S3.85}$$

Then we bound $\|\xi_{ok2}\|_\infty$ by Condition 7,

$$\|\xi_{ok2}\|_\infty \leqslant \left\| \left[\frac{1}{n} \boldsymbol{X}_{\mathcal{S}_1 \backslash \mathcal{D}_k}^\top \boldsymbol{G}_k^l \boldsymbol{X}_{\mathcal{S}} \right] \left[\frac{1}{n} \boldsymbol{X}_{\mathcal{S}}^\top \boldsymbol{G}_k^l \boldsymbol{X}_{\mathcal{S}} + \mathcal{J} \right]^{-1} \right\|_\infty$$
$$\left(\left\| \frac{\sigma_k^{l*}}{n} \sum_{i=1}^n L_k^{l*} \left(y_i - \boldsymbol{x}_{i\mathcal{S}_1 \backslash \mathcal{D}_k}^\top \boldsymbol{\beta}_{m\mathcal{S}_1 \backslash \mathcal{D}_k}^* - \boldsymbol{x}_{i\mathcal{D}_k}^\top \boldsymbol{\alpha}_{k\mathcal{D}_k}^* - \boldsymbol{x}_{i\mathcal{S}_2}^\top \boldsymbol{\gamma}_{k\mathcal{S}_2}^* \right) \boldsymbol{x}_{i\mathcal{S}} \right\|_\infty \right.$$
$$\left. + \left\| \mathrm{vec}(\boldsymbol{r}(\boldsymbol{\Delta})_{\boldsymbol{\beta}_{m\mathcal{S}_1 \backslash \mathcal{D}_k}}, \boldsymbol{r}(\boldsymbol{\Delta})_{\boldsymbol{\alpha}_{k\mathcal{D}_k}}, \boldsymbol{r}(\boldsymbol{\Delta})_{\boldsymbol{\gamma}_{k\mathcal{S}_2}}) \right\|_\infty \right), \tag{S3.86}$$

where the second line can be bounded as above, and the third line is $O(\frac{K_l^3 s \log p}{n})$. Then we have:

$$\|\xi_{ok2}\|_\infty \leqslant c_o \sqrt{\frac{K_l^3 s \log p}{n}}, \tag{S3.87}$$

with probability tending to 1, where c_o is a positive constant.

Overall, we can get $\|\boldsymbol{z}_o\|_\infty \leqslant \lambda_3$. Thus, by combining with the minimal signal condition of the true parameters, we have:

$$\left\| \widehat{\boldsymbol{\Omega}}^o - \boldsymbol{\Omega}^* \right\|_2 = O\left(\sqrt{\frac{K_l^3 s \log p}{n}} \right), \quad \|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_2 = O\left(\sqrt{\frac{K_l^3 s_2 \log p}{n}} \right), \tag{S3.88}$$

with probability tending to 1.

Step 3:

We show that, for $k = 1, \cdots, K_l, m = 1, \cdots, K_u$ given $\widehat{\boldsymbol{\alpha}}_k \in \mathcal{B}(\boldsymbol{\alpha}_k^*, c_a \tau)$ and $\widehat{\boldsymbol{\beta}}_m \in \mathcal{B}(\boldsymbol{\beta}_m^*, c_b \tau)$, where $\tau = O\left(\sqrt{\frac{K_l^3 s_2 \log p}{n}} \right)$ and c_a, c_b are positive constants, we have:

$$\begin{aligned}\sup_{\widehat{\boldsymbol{\alpha}}_k\in\mathcal{B}(\boldsymbol{\alpha}_k^*),\widehat{\boldsymbol{\beta}}_m\in\mathcal{B}(\boldsymbol{\beta}_m^*)}\left\|\widehat{\boldsymbol{\beta}}_m-\widehat{\boldsymbol{\alpha}}_k\right\| &\leqslant \sup_{\widehat{\boldsymbol{\beta}}_m\in\mathcal{B}(\boldsymbol{\beta}_m^*)}\left\|\boldsymbol{\beta}_m^*-\widehat{\boldsymbol{\beta}}_m\right\| \\ &\quad+\sup_{\widehat{\boldsymbol{\alpha}}_k\in\mathcal{B}(\boldsymbol{\alpha}_k^*)}\left\|\boldsymbol{\alpha}_k^*-\widehat{\boldsymbol{\alpha}}_k\right\|+\left\|\boldsymbol{\beta}_m^*-\boldsymbol{\alpha}_k^*\right\| \qquad \text{(S3.89)} \\ &\leqslant (c_a+c_b)\tau+\left\|\boldsymbol{\beta}_m^*-\boldsymbol{\alpha}_k^*\right\|,\end{aligned}$$

and

$$\begin{aligned}\inf_{\widehat{\boldsymbol{\alpha}}_k\in\mathcal{B}(\boldsymbol{\alpha}_k^*),\widehat{\boldsymbol{\beta}}_{m'}\in\mathcal{B}(\boldsymbol{\beta}_{m'}^*)}\left\|\widehat{\boldsymbol{\beta}}_{m'}-\widehat{\boldsymbol{\alpha}}_k\right\| &\geqslant \left\|\boldsymbol{\beta}_{m'}^*-\boldsymbol{\alpha}_k^*\right\| \\ &\quad-\left(\sup_{\widehat{\boldsymbol{\beta}}_{m'}\in\mathcal{B}(\boldsymbol{\beta}_{m'}^*)}\left\|\boldsymbol{\beta}_{m'}^*-\widehat{\boldsymbol{\beta}}_m\right\|+\sup_{\widehat{\boldsymbol{\alpha}}_k\in\mathcal{B}(\boldsymbol{\alpha}_k^*)}\left\|\boldsymbol{\alpha}_k^*-\widehat{\boldsymbol{\alpha}}_k\right\|\right) \\ &\geqslant \left\|\boldsymbol{\beta}_m^*-\boldsymbol{\alpha}_k^*\right\|-(c_a+c_b)\tau.\end{aligned} \qquad \text{(S3.90)}$$

While $k\in\mathcal{G}_m$ and $m\neq m'$, it follows that:

$$Pr\left(\sup_{\widehat{\boldsymbol{\alpha}}_k\in\mathcal{B}(\boldsymbol{\alpha}_k^*),\widehat{\boldsymbol{\beta}}_m\in\mathcal{B}(\boldsymbol{\beta}_m^*)}\left\|\widehat{\boldsymbol{\beta}}_m-\widehat{\boldsymbol{\alpha}}_k\right\|\leqslant\inf_{\widehat{\boldsymbol{\alpha}}_k\in\mathcal{B}(\boldsymbol{\alpha}_k^*),\widehat{\boldsymbol{\beta}}_{m'}\in\mathcal{B}(\boldsymbol{\beta}_{m'}^*)}\left\|\widehat{\boldsymbol{\beta}}_{m'}-\widehat{\boldsymbol{\alpha}}_k\right\|\right)\to 1. \qquad \text{(S3.91)}$$

Recall

$$\mathcal{P}_2=\sum_{m=1}^{K_u}\sum_{k\in\mathcal{G}_m}P\left(\sum_{j=1}^{p}P(|\beta_{mj}-\alpha_{kj}|)\right). \qquad \text{(S3.92)}$$

Following Tang et al. (2021), when p goes to the L_1 norm function, we have $\widetilde{\mathcal{L}}_{mdsp}=\mathcal{L}_{ma}+\mathcal{P}_2$. We have $\widetilde{\mathcal{L}}_{mdsp}=\mathcal{L}_{mdsp}$ at $\widehat{\boldsymbol{\alpha}}_k\in\mathcal{B}(\boldsymbol{\alpha}_k^*,c_a\tau)$ and $\widehat{\boldsymbol{\beta}}_m\in\mathcal{B}(\boldsymbol{\beta}_m^*,c_b\tau)$, and thus $\operatorname{argmin}\widetilde{\mathcal{L}}_{mdsp}=\operatorname{argmin}\mathcal{L}_{mdsp}$.

There are minor differences for different penalties. For example, following Breheny and Huang (2009), the composite MCP can select important groups of variables and important variables within these variable groups with highly desirable properties. Additionally, following Huang et al. (2012), the internal and external penalties can differ.

Bibliography

Balakrishnan, S., M. J. Wainwright, and B. Yu (2017). Statistical guarantees for the em algorithm: From population to sample-based analysis. *The Annals of Statistics 45*(1), 77–120.

Breheny, P. and J. Huang (2009). Penalized methods for bi-level variable selection. *Statistics and its interface 2*(3), 369.

Fan, J. and J. Lv (2011). Nonconcave penalized likelihood with np-dimensionality. *IEEE Transactions on Information Theory 57*(8), 5467–5484.

Hao, B., W. W. Sun, Y. Liu, and G. Cheng (2018). Simultaneous clustering and estimation of heterogeneous graphical models. *Journal of Machine Learning Research 18*(217), 1–58.

Huang, J., P. Breheny, and S. Ma (2012). A selective review of group selection in high-dimensional models. *Statistical Science 27*(4), 481–499.

Li, R., Q. Zhang, and S. Ma (2023). Regulation-incorporated gene expression network-based heterogeneity analysis. *arXiv preprint arXiv:2308.03946*.

McLachlan, G. J. and T. Krishnan (2007). *The EM algorithm and extensions*. John Wiley & Sons.

Tang, X., F. Xue, and A. Qu (2021). Individualized multidirectional variable selection. *Journal of the American Statistical Association 116*(535), 1280–1296.

Wytock, M. and Z. Kolter (2013). Sparse gaussian conditional random fields: Algorithms, theory, and application to energy forecasting. In *International conference on machine learning*, pp. 1265–1273. PMLR.

Yang, X., X. Yan, and J. Huang (2019). High-dimensional integrative analysis with homogeneity and sparsity recovery. *Journal of Multivariate Analysis 174*, 104529.