Supplemental Materials for "A concave pairwise fusion approach to subgroup analysis"

Shujie Ma*
Department of Statistics, University of California at Riverside
and
Jian Huang†
Department of Statistics and Actuarial Science, University of Iowa

In this supplement, we give the technical proofs for Proposition 1 and Theorems 1-3. We also provide a detailed estimation procedure for model (2) based on the ADMM algorithm in a way similar to that for model (1).

A.1 Proof of Proposition 1

In this section we show the results in Proposition 1. By the definition of $\boldsymbol{\eta}^{(m+1)}$, we have

$$L(\boldsymbol{\mu}^{(m+1)}, \boldsymbol{\beta}^{(m+1)}, \boldsymbol{\eta}^{(m+1)}, \boldsymbol{\upsilon}^{(m)}) \leq L(\boldsymbol{\mu}^{(m+1)}, \boldsymbol{\beta}^{(m+1)}, \boldsymbol{\eta}, \boldsymbol{\upsilon}^{(m)})$$

for any $\boldsymbol{\eta}$. Define

$$\begin{aligned} f^{(m+1)} &= \inf_{\boldsymbol{\Delta}\boldsymbol{\mu}^{(m+1)}-\boldsymbol{\eta}=\mathbf{0}} \{\frac{1}{2}\left\|\mathbf{y}-\boldsymbol{\mu}^{(m+1)}-\mathbf{X}\boldsymbol{\beta}^{(m+1)}\right\|^2 + \sum\nolimits_{i<j} p_\gamma(|\eta_{ij}|, \lambda)\} \\ &= \inf_{\boldsymbol{\Delta}\boldsymbol{\mu}^{(m+1)}-\boldsymbol{\eta}=\mathbf{0}} L(\boldsymbol{\mu}^{(m+1)}, \boldsymbol{\beta}^{(m+1)}, \boldsymbol{\eta}, \boldsymbol{\upsilon}^{(m)}). \end{aligned}$$

Then

$$L(\boldsymbol{\mu}^{(m+1)}, \boldsymbol{\beta}^{(m+1)}, \boldsymbol{\eta}^{(m+1)}, \boldsymbol{\upsilon}^{(m)}) \leq f^{(m+1)}.$$

*The research of Ma is supported in part by the U.S. NSF grant DMS-13-06972.
†Corresponding author. The research of Huang is supported in part by the U.S. NSF grant DMS-12-08225.

Let t be an integer. Since $\boldsymbol{\upsilon}^{(m+t-1)} = \boldsymbol{\upsilon}^{(m)} + \vartheta \sum_{i=1}^{t-1}(\boldsymbol{\Delta\mu}^{(m+i)} - \boldsymbol{\eta}^{(m+i)})$, we have

$$\begin{aligned}
& L(\boldsymbol{\mu}^{(m+t)}, \boldsymbol{\beta}^{(m+t)}, \boldsymbol{\eta}^{(m+t)}, \boldsymbol{\upsilon}^{(m+t-1)}) \\
= \; & \frac{1}{2}\left\|\mathbf{y} - \boldsymbol{\mu}^{(m+t)} - \mathbf{X}\boldsymbol{\beta}^{(m+t)}\right\|^2 + \boldsymbol{\upsilon}^{(m+t-1)\mathrm{T}}(\boldsymbol{\Delta\mu}^{(m+t)} - \boldsymbol{\eta}^{(m+t)}) \\
& + \frac{\vartheta}{2}||\boldsymbol{\Delta\mu}^{(m+t)} - \boldsymbol{\eta}^{(m+t)}||^2 + \sum_{i<j} p_\gamma(|\eta_{ij}^{(m+t)}|, \lambda) \\
= \; & \frac{1}{2}\left\|\mathbf{y} - \boldsymbol{\mu}^{(m+t)} - \mathbf{X}\boldsymbol{\beta}^{(m+t)}\right\|^2 + \boldsymbol{\upsilon}^{(m)\mathrm{T}}(\boldsymbol{\Delta\mu}^{(m+t)} - \boldsymbol{\eta}^{(m+t)}) \\
& + \vartheta \sum_{i=1}^{t-1}(\boldsymbol{\Delta\mu}^{(m+i)} - \boldsymbol{\eta}^{(m+i)})^{\mathrm{T}}(\boldsymbol{\Delta\mu}^{(m+t)} - \boldsymbol{\eta}^{(m+t)}) \\
& + \frac{\vartheta}{2}||\boldsymbol{\Delta\mu}^{(m+t)} - \boldsymbol{\eta}^{(m+t)}||^2 + \sum_{i<j} p_\gamma(|\eta_{ij}^{(m+t)}|, \lambda) \\
\leq \; & f^{(m+t)}.
\end{aligned}$$

Since the objective function $L(\boldsymbol{\mu}, \boldsymbol{\beta}, \boldsymbol{\eta}, \boldsymbol{\upsilon})$ is differentiable with respect to $(\boldsymbol{\mu}, \boldsymbol{\beta})$ and is convex with respect to $\boldsymbol{\eta}$, by applying the results in Theorem 4.1 of Tseng (1991), the sequence $(\boldsymbol{\mu}^{(m)}, \boldsymbol{\beta}^{(m)}, \boldsymbol{\eta}^{(m)})$ has a limit point, denoted by $(\boldsymbol{\mu}^*, \boldsymbol{\beta}^*, \boldsymbol{\eta}^*)$. Then we have

$$f^* = \lim_{m\to\infty} f^{(m+1)} = \lim_{m\to\infty} f^{(m+t)} = \inf_{\boldsymbol{\Delta\mu}^* - \boldsymbol{\eta} = \mathbf{0}} \{\frac{1}{2}\left\|\mathbf{y} - \boldsymbol{\mu}^* - \mathbf{X}\boldsymbol{\beta}^*\right\|^2 + \sum_{i<j} p_\gamma(|\eta_{ij}|, \lambda)\},$$

and for all $t \geq 0$

$$\begin{aligned}
& \lim_{m\to\infty} L(\boldsymbol{\mu}^{(m+t)}, \boldsymbol{\beta}^{(m+t)}, \boldsymbol{\eta}^{(m+t)}, \boldsymbol{\upsilon}^{(m+t-1)}) \\
= \; & \frac{1}{2}\left\|\mathbf{y} - \boldsymbol{\mu}^* - \mathbf{X}\boldsymbol{\beta}^*\right\|^2 + \sum_{i<j} p_\gamma(|\eta_{ij}^*|, \lambda) + \lim_{m\to\infty} \boldsymbol{\upsilon}^{(m)\mathrm{T}}(\boldsymbol{\Delta\mu}^* - \boldsymbol{\eta}^*) + (t - \frac{1}{2})\vartheta||\boldsymbol{\Delta\mu}^* - \boldsymbol{\eta}^*||^2 \\
\leq \; & f^*.
\end{aligned}$$

Hence $\lim_{m\to\infty} ||\mathbf{r}^{(m)}||^2 = r^* = ||\boldsymbol{\Delta\mu}^* - \boldsymbol{\eta}^*||^2 = 0$.

Since $(\boldsymbol{\mu}^{(m+1)}, \boldsymbol{\beta}^{(m+1)})$ minimize $L(\boldsymbol{\mu}, \boldsymbol{\beta}, \boldsymbol{\eta}^{(m)}, \boldsymbol{\upsilon}^{(m)})$ by definition, we have that

$$\partial L(\boldsymbol{\mu}^{(m+1)}, \boldsymbol{\beta}^{(m+1)}, \boldsymbol{\eta}^{(m)}, \boldsymbol{\upsilon}^{(m)})/\partial \boldsymbol{\mu} = \mathbf{0},$$

and moreover,

$$\begin{aligned}
& \partial L(\boldsymbol{\mu}^{(m+1)}, \boldsymbol{\beta}^{(m+1)}, \boldsymbol{\eta}^{(m)}, \boldsymbol{\upsilon}^{(m)})/\partial \boldsymbol{\mu} \\
= \; & \boldsymbol{\mu}^{(m+1)} + \mathbf{X}\boldsymbol{\beta}^{(m+1)} - \mathbf{y} + \boldsymbol{\Delta}^{\mathrm{T}}\boldsymbol{\upsilon}^{(m)} + \boldsymbol{\Delta}^{\mathrm{T}}\vartheta(\boldsymbol{\Delta\mu}^{(m+1)} - \boldsymbol{\eta}^{(m)}) \\
= \; & \boldsymbol{\mu}^{(m+1)} + \mathbf{X}\boldsymbol{\beta}^{(m+1)} - \mathbf{y} + \boldsymbol{\Delta}^{\mathrm{T}}(\boldsymbol{\upsilon}^{(m)} + \vartheta(\boldsymbol{\Delta\mu}^{(m+1)} - \boldsymbol{\eta}^{(m)})) \\
= \; & \boldsymbol{\mu}^{(m+1)} + \mathbf{X}\boldsymbol{\beta}^{(m+1)} - \mathbf{y} + \boldsymbol{\Delta}^{\mathrm{T}}\boldsymbol{\upsilon}^{(m+1)} + \vartheta\boldsymbol{\Delta}^{\mathrm{T}}(\boldsymbol{\eta}^{(m+1)} - \boldsymbol{\eta}^{(m)}).
\end{aligned}$$

The last step follows from $\boldsymbol{\upsilon}^{(m+1)} = \boldsymbol{\upsilon}^{(m)} + \vartheta(\boldsymbol{\Delta\mu}^{(m+1)} - \boldsymbol{\eta}^{(m+1)})$. Therefore,

$$\mathbf{s}^{(m+1)} = \vartheta\boldsymbol{\Delta}^{\mathrm{T}}(\boldsymbol{\eta}^{(m+1)} - \boldsymbol{\eta}^{(m)}) = -(\boldsymbol{\mu}^{(m+1)} + \mathbf{X}\boldsymbol{\beta}^{(m+1)} - \mathbf{y} + \boldsymbol{\Delta}^{\mathrm{T}}\boldsymbol{\upsilon}^{(m+1)}).$$

Since $||\boldsymbol{\Delta\mu}^* - \boldsymbol{\eta}^*||^2 = 0$,

$$\begin{aligned}
& \lim_{m\to\infty} \partial L(\boldsymbol{\mu}^{(m+1)}, \boldsymbol{\beta}^{(m+1)}, \boldsymbol{\eta}^{(m)}, \boldsymbol{\upsilon}^{(m)})/\partial\boldsymbol{\mu} \\
= & \lim_{m\to\infty} \boldsymbol{\mu}^{(m+1)} + \mathbf{X}\boldsymbol{\beta}^{(m+1)} - \mathbf{y} + \boldsymbol{\Delta}^{\mathrm{T}}\boldsymbol{\upsilon}^{(m+1)} \\
= & \boldsymbol{\mu}^* + \mathbf{X}\boldsymbol{\beta}^* - \mathbf{y} + \boldsymbol{\Delta}^{\mathrm{T}}\boldsymbol{\upsilon}^* = \mathbf{0}.
\end{aligned}$$

Therefore, $\lim_{m\to\infty} \mathbf{s}^{(m+1)} = \mathbf{0}$.

A.2 Proof of Theorem 1

In this section we show the results in Theorem 1. Since for every $\boldsymbol{\mu} \in \mathcal{M}_{\mathcal{G}}$, it can be written as $\boldsymbol{\mu} = \mathbf{Z}\boldsymbol{\alpha}$, and hence $\boldsymbol{\alpha} = \mathbf{D}^{-1}\mathbf{Z}^{\mathrm{T}}\boldsymbol{\mu}$. Then $((\widehat{\boldsymbol{\mu}}^{or})^{\mathrm{T}}, (\widehat{\boldsymbol{\beta}}^{or})^{\mathrm{T}})^{\mathrm{T}} = ((\mathbf{Z}\widehat{\boldsymbol{\alpha}}^{or})^{\mathrm{T}}, (\widehat{\boldsymbol{\beta}}^{or})^{\mathrm{T}})^{\mathrm{T}}$, where

$$\begin{pmatrix} \widehat{\boldsymbol{\alpha}}^{or} \\ \widehat{\boldsymbol{\beta}}^{or} \end{pmatrix} = \arg \min_{\boldsymbol{\alpha}\in R^K, \boldsymbol{\beta}\in R^p} \frac{1}{2}||\mathbf{y} - \mathbf{Z}\boldsymbol{\alpha} - \mathbf{X}\boldsymbol{\beta}||^2 = [(\mathbf{Z},\mathbf{X})^{\mathrm{T}}(\mathbf{Z},\mathbf{X})]^{-1}(\mathbf{Z},\mathbf{X})^{\mathrm{T}}\mathbf{y}.$$

Then

$$\begin{pmatrix} \widehat{\boldsymbol{\alpha}}^{or} - \boldsymbol{\alpha}^0 \\ \widehat{\boldsymbol{\beta}}^{or} - \boldsymbol{\beta}^0 \end{pmatrix} = [(\mathbf{Z},\mathbf{X})^{\mathrm{T}}(\mathbf{Z},\mathbf{X})]^{-1}(\mathbf{Z},\mathbf{X})^{\mathrm{T}}\boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^{\mathrm{T}}$ and $\boldsymbol{\alpha}^0 = (\alpha_1^0, \ldots, \alpha_K^0)^{\mathrm{T}}$. Hence

$$\left\| \begin{pmatrix} \widehat{\boldsymbol{\alpha}}^{or} - \boldsymbol{\alpha}^0 \\ \widehat{\boldsymbol{\beta}}^{or} - \boldsymbol{\beta}^0 \end{pmatrix} \right\|_\infty \leq \left\| [(\mathbf{Z},\mathbf{X})^{\mathrm{T}}(\mathbf{Z},\mathbf{X})]^{-1} \right\|_\infty \left\| (\mathbf{Z},\mathbf{X})^{\mathrm{T}}\boldsymbol{\epsilon} \right\|_\infty. \tag{A.1}$$

By Condition (C1), we have $\left\| [(\mathbf{Z},\mathbf{X})^{\mathrm{T}}(\mathbf{Z},\mathbf{X})]^{-1} \right\| \leq C_1^{-1} |\mathcal{G}_{\min}|^{-1}$ and thus

$$\left\| [(\mathbf{Z},\mathbf{X})^{\mathrm{T}}(\mathbf{Z},\mathbf{X})]^{-1} \right\|_\infty \leq \sqrt{K+p}\, C_1^{-1} |\mathcal{G}_{\min}|^{-1}. \tag{A.2}$$

Moreover

$$P(\left\|(\mathbf{Z},\mathbf{X})^{\mathrm{T}}\boldsymbol{\epsilon}\right\|_\infty > C\sqrt{n\log n}) \leq P(\left\|\mathbf{Z}^{\mathrm{T}}\boldsymbol{\epsilon}\right\|_\infty > C\sqrt{n\log n}) + P(\left\|\mathbf{X}^{\mathrm{T}}\boldsymbol{\epsilon}\right\|_\infty > C\sqrt{n\log n}),$$

for some constant $0 < C < \infty$. By Condition (C3) and union bound,

$$\begin{aligned}
& P\left(\left\|\mathbf{Z}^{\mathrm{T}}\boldsymbol{\epsilon}\right\|_\infty > C\sqrt{n\log n}\right) \\
\leq & \sum_{k=1}^K P(|\sum_{i\in\mathcal{G}_k} \epsilon_i| > C\sqrt{n\log n}) \leq \sum_{k=1}^K P(|\sum_{i\in\mathcal{G}_k} \epsilon_i| > \sqrt{|\mathcal{G}_k|} C\sqrt{\log n}) \\
\leq & 2K\exp(-c_1C^2\log n) = 2Kn^{-c_1C^2},
\end{aligned}$$

and by Conditions (C1) and (C3) and union bound,

$$\begin{aligned} & P\left(\left\|\mathbf{X}^{\mathrm{T}}\boldsymbol{\epsilon}\right\|_{\infty} > C\sqrt{n\log n}\right) \\ \leq & \sum_{j=1}^{p} P\left(|\mathbf{X}_j^{\mathrm{T}}\boldsymbol{\epsilon}| > \sqrt{n}C\sqrt{\log n}\right) \\ \leq & \ 2p\exp(-c_1C^2\log n) = 2pn^{-c_1C^2}. \end{aligned}$$

By the above results, we have

$$P(\left\|(\mathbf{Z},\mathbf{X})^{\mathrm{T}}\boldsymbol{\epsilon}\right\|_{\infty} > C\sqrt{n\log n}) \leq 2(K+p)n^{-c_1C^2}. \tag{A.3}$$

Therefore, by (A.1), (A.2) and (A.3), we have with probability at least $1-2(K+p)n^{-c_1C^2}$,

$$\left\|\begin{pmatrix} \widehat{\boldsymbol{\alpha}}^{or} - \boldsymbol{\alpha}^0 \\ \widehat{\boldsymbol{\beta}}^{or} - \boldsymbol{\beta}^0 \end{pmatrix}\right\|_{\infty} \leq CC_1^{-1}\sqrt{K+p}\,|\mathcal{G}_{\min}|^{-1}\sqrt{n\log n},$$

and hence $\|\widehat{\boldsymbol{\mu}}^{or} - \boldsymbol{\mu}^0\|_{\infty} = \|\widehat{\boldsymbol{\alpha}}^{or} - \boldsymbol{\alpha}^0\|_{\infty} \leq CC_1^{-1}\sqrt{K+p}\,|\mathcal{G}_{\min}|^{-1}\sqrt{n\log n}$. The result (8) in Theorem 1 is proved by letting $C = c_1^{-1/2}$, and result (10) follows from Central Limit Theorem.

A.3 Proof of Theorem 2

In this section we show the results in Theorem 2. Define

$$\begin{aligned} L_n(\boldsymbol{\mu},\boldsymbol{\beta}) &= \frac{1}{2}||\mathbf{y}-\boldsymbol{\mu}-\mathbf{X}\boldsymbol{\beta}||^2, P_n(\boldsymbol{\mu}) = \lambda\sum_{i<j}\rho(|\mu_i-\mu_j|), \\ L_n^{\mathcal{G}}(\boldsymbol{\alpha},\boldsymbol{\beta}) &= \frac{1}{2}||\mathbf{y}-\mathbf{Z}\boldsymbol{\alpha}-\mathbf{X}\boldsymbol{\beta}||^2, P_n^{\mathcal{G}}(\boldsymbol{\alpha}) = \lambda\sum_{k<k'}|\mathcal{G}_k||\mathcal{G}_{k'}|\rho(|\alpha_k-\alpha_{k'}|), \end{aligned}$$

and let

$$Q_n(\boldsymbol{\mu},\boldsymbol{\beta}) = L_n(\boldsymbol{\mu},\boldsymbol{\beta}) + P_n(\boldsymbol{\mu}), Q_n^{\mathcal{G}}(\boldsymbol{\alpha},\boldsymbol{\beta}) = L_n^{\mathcal{G}}(\boldsymbol{\alpha},\boldsymbol{\beta}) + P_n^{\mathcal{G}}(\boldsymbol{\alpha}).$$

Let $T:\mathcal{M}_{\mathcal{G}} \to R^K$ be the mapping such that $T(\boldsymbol{\mu})$ is the $K\times 1$ vector whose k^{th} coordinate equals to the common value of μ_i for $i\in\mathcal{G}_k$. Let $T^*: R^n \to R^K$ be the mapping such that $T^*(\boldsymbol{\mu}) = \{|\mathcal{G}_k|^{-1}\sum_{i\in\mathcal{G}_k}\mu_i\}_{k=1}^K$. Clearly, when $\boldsymbol{\mu}\in\mathcal{M}_{\mathcal{G}}$, $T(\boldsymbol{\mu}) = T^*(\boldsymbol{\mu})$.

By calculation, for every $\boldsymbol{\mu}\in\mathcal{M}_{\mathcal{G}}$, we have $P_n(\boldsymbol{\mu}) = P_n^{\mathcal{G}}(T(\boldsymbol{\mu}))$ and for every $\boldsymbol{\alpha}\in R^K$, we have $P_n(T^{-1}(\boldsymbol{\alpha})) = P_n^{\mathcal{G}}(\boldsymbol{\alpha})$. Hence

$$Q_n(\boldsymbol{\mu},\boldsymbol{\beta}) = Q_n^{\mathcal{G}}(T(\boldsymbol{\mu}),\boldsymbol{\beta}), Q_n^{\mathcal{G}}(\boldsymbol{\alpha},\boldsymbol{\beta}) = Q_n(T^{-1}(\boldsymbol{\alpha}),\boldsymbol{\beta}). \tag{A.4}$$

Consider the neighborhood of $(\boldsymbol{\mu}^0,\boldsymbol{\beta}^0)$:

$$\Theta = \{\boldsymbol{\mu}\in R^n, \boldsymbol{\beta}\in R^p: \left\|((\boldsymbol{\mu}-\boldsymbol{\mu}^0)^{\mathrm{T}}, (\boldsymbol{\beta}-\boldsymbol{\beta}^0)^{\mathrm{T}})^{\mathrm{T}}\right\|_{\infty} \leq \phi_n\}.$$

By the result in Theorem 1, there is an event E_1 such that on the event E_1,

$$\left\| ((\widehat{\boldsymbol{\mu}}^{or} - \boldsymbol{\mu}^0)^{\mathrm{T}}, (\widehat{\boldsymbol{\beta}}^{or} - \boldsymbol{\beta}^0)^{\mathrm{T}})^{\mathrm{T}} \right\|_{\infty} \leq \phi_n,$$

and $P(E_1^C) \leq 2(K+p)n^{-1}$. Hence $((\widehat{\boldsymbol{\mu}}^{or})^{\mathrm{T}}, (\widehat{\boldsymbol{\beta}}^{or})^{\mathrm{T}})^{\mathrm{T}} \in \Theta$ on the event E_1. For any $\boldsymbol{\mu} \in R^n$, let $\boldsymbol{\mu}^* = T^{-1}(T^*(\boldsymbol{\mu}))$. We show that $(\widehat{\boldsymbol{\mu}}^{or}, \widehat{\boldsymbol{\beta}}^{or})$ is a strictly local minimizer of the objective function (3) with probability approaching 1 through the following two steps.

(i). On the event E_1, $Q_n(\boldsymbol{\mu}^*, \boldsymbol{\beta}) > Q_n(\widehat{\boldsymbol{\mu}}^{or}, \widehat{\boldsymbol{\beta}}^{or})$ for any $((\boldsymbol{\mu})^{\mathrm{T}}, (\boldsymbol{\beta})^{\mathrm{T}})^{\mathrm{T}} \in \Theta$ and $((\boldsymbol{\mu}^*)^{\mathrm{T}}, (\boldsymbol{\beta})^{\mathrm{T}})^{\mathrm{T}} \neq ((\widehat{\boldsymbol{\mu}}^{or})^{\mathrm{T}}, (\widehat{\boldsymbol{\beta}}^{or})^{\mathrm{T}})^{\mathrm{T}}$.

(ii). There is an event E_2 such that $P(E_2^C) \leq 2n^{-1}$. On $E_1 \cap E_2$, there is a neighborhood of $((\widehat{\boldsymbol{\mu}}^{or})^{\mathrm{T}}, (\widehat{\boldsymbol{\beta}}^{or})^{\mathrm{T}})^{\mathrm{T}}$, denoted by Θ_n, such that $Q_n(\boldsymbol{\mu}, \boldsymbol{\beta}) \geq Q_n(\boldsymbol{\mu}^*, \boldsymbol{\beta})$ for any $((\boldsymbol{\mu})^{\mathrm{T}}, (\boldsymbol{\beta})^{\mathrm{T}})^{\mathrm{T}} \in \Theta_n \cap \Theta$ for sufficiently large n.

Therefore, by the results in (i) and (ii), we have $Q_n(\boldsymbol{\mu}, \boldsymbol{\beta}) > Q_n(\widehat{\boldsymbol{\mu}}^{or}, \widehat{\boldsymbol{\beta}}^{or})$ for any $((\boldsymbol{\mu})^{\mathrm{T}}, (\boldsymbol{\beta})^{\mathrm{T}})^{\mathrm{T}} \in \Theta_n \cap \Theta$ and $((\boldsymbol{\mu})^{\mathrm{T}}, (\boldsymbol{\beta})^{\mathrm{T}})^{\mathrm{T}} \neq ((\widehat{\boldsymbol{\mu}}^{or})^{\mathrm{T}}, (\widehat{\boldsymbol{\beta}}^{or})^{\mathrm{T}})^{\mathrm{T}}$, so that $((\widehat{\boldsymbol{\mu}}^{or})^{\mathrm{T}}, (\widehat{\boldsymbol{\beta}}^{or})^{\mathrm{T}})^{\mathrm{T}}$ is a strict local minimizer of $Q_n(\boldsymbol{\mu}, \boldsymbol{\beta})$ given in (3) on the event $E_1 \cap E_2$ with $P(E_1 \cap E_2) \geq 1 - 2(K+p+1)n^{-1}$ for sufficiently large n.

In the following we prove the result in (i). We first show $P_n^{\mathcal{G}}(T^*(\boldsymbol{\mu})) = C_n$ for any $\boldsymbol{\mu} \in \Theta$, where C_n is a constant which does not depend on $\boldsymbol{\mu}$. Let $T^*(\boldsymbol{\mu}) = \boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)^{\mathrm{T}}$. It suffices to show that $|\alpha_k - \alpha_{k'}| > a\lambda$ for all k and k'. Then by Condition (C2), $\rho(|\alpha_k - \alpha_{k'}|)$ is a constant, and as a result $P_n^{\mathcal{G}}(T^*(\boldsymbol{\mu}))$ is a constant. Since

$$|\alpha_k - \alpha_{k'}| \geq |\alpha_k^0 - \alpha_{k'}^0| - 2||\boldsymbol{\alpha} - \boldsymbol{\alpha}^0||_{\infty},$$

and

$$\begin{aligned} ||\boldsymbol{\alpha} - \boldsymbol{\alpha}^0||_{\infty} &= \sup_k |\sum\nolimits_{i \in \mathcal{G}_k} \mu_i / |\mathcal{G}_k| - \alpha_k^0| = \sup_k |\sum\nolimits_{i \in \mathcal{G}_k} (\mu_i - \mu_i^0)/|\mathcal{G}_k|| \\ &\leq \sup_k \sup_{i \in \mathcal{G}_k} |\mu_i - \mu_i^0| = ||\boldsymbol{\mu} - \boldsymbol{\mu}^0||_{\infty}, \end{aligned} \quad \text{(A.5)}$$

then for all k and k'

$$|\alpha_k - \alpha_{k'}| \geq |\alpha_k^0 - \alpha_{k'}^0| - 2||\boldsymbol{\mu} - \boldsymbol{\mu}^0||_{\infty} \geq b_n - 2\phi_n > a\lambda,$$

where the last inequality follows from the assumption that $b_n > a\lambda \gg \phi_n$. Therefore, we have $P_n^{\mathcal{G}}(T^*(\boldsymbol{\mu})) = C_n$, and hence $Q_n^{\mathcal{G}}(T^*(\boldsymbol{\mu}), \boldsymbol{\beta}) = L_n^{\mathcal{G}}(T^*(\boldsymbol{\mu}), \boldsymbol{\beta}) + C_n$ for all $(\boldsymbol{\mu}^{\mathrm{T}}, \boldsymbol{\beta}^{\mathrm{T}})^{\mathrm{T}} \in \Theta$. Since $((\widehat{\boldsymbol{\alpha}}^{or})^{\mathrm{T}}, (\widehat{\boldsymbol{\beta}}^{or})^{\mathrm{T}})^{\mathrm{T}}$ is the unique global minimizer of $L_n^{\mathcal{G}}(\boldsymbol{\alpha}, \boldsymbol{\beta})$, then $L_n^{\mathcal{G}}(T^*(\boldsymbol{\mu}), \boldsymbol{\beta}) > L_n^{\mathcal{G}}(\widehat{\boldsymbol{\alpha}}^{or}, \widehat{\boldsymbol{\beta}}^{or})$ for all $(T^*(\boldsymbol{\mu})^{\mathrm{T}}, \boldsymbol{\beta}^{\mathrm{T}})^{\mathrm{T}} \neq ((\widehat{\boldsymbol{\alpha}}^{or})^{\mathrm{T}}, (\widehat{\boldsymbol{\beta}}^{or})^{\mathrm{T}})^{\mathrm{T}}$ and thus $Q_n^{\mathcal{G}}(T^*(\boldsymbol{\mu}), \boldsymbol{\beta}) > Q_n^{\mathcal{G}}(\widehat{\boldsymbol{\alpha}}^{or}, \widehat{\boldsymbol{\beta}}^{or})$ for all $T^*(\boldsymbol{\mu}) \neq \widehat{\boldsymbol{\alpha}}^{or}$. By (A.4), we have $Q_n^{\mathcal{G}}(\widehat{\boldsymbol{\alpha}}^{or}, \widehat{\boldsymbol{\beta}}^{or}) = Q_n(\widehat{\boldsymbol{\mu}}^{or}, \widehat{\boldsymbol{\beta}}^{or})$ and $Q_n^{\mathcal{G}}(T^*(\boldsymbol{\mu}), \boldsymbol{\beta}) = Q_n(T^{-1}(T^*(\boldsymbol{\mu})), \boldsymbol{\beta}) = Q_n(\boldsymbol{\mu}^*, \boldsymbol{\beta})$. Therefore, $Q_n(\boldsymbol{\mu}^*, \boldsymbol{\beta}) > Q_n(\widehat{\boldsymbol{\mu}}^{or}, \widehat{\boldsymbol{\beta}}^{or})$ for all $\boldsymbol{\mu}^* \neq \widehat{\boldsymbol{\mu}}^{or}$, and the result in (i) is proved.

Next we prove the result in (ii). For a positive sequence t_n, let $\Theta_n = \{\boldsymbol{\mu} : ||\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}^{or}|| \leq t_n\}$. For $(\boldsymbol{\mu}^{\mathrm{T}}, \boldsymbol{\beta}^{\mathrm{T}})^{\mathrm{T}} \in \Theta_n \cap \Theta$, by Taylor's expansion, we have

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

where

$$\begin{aligned}\Gamma_1 &= -(\mathbf{y}-(\mathbf{I}_n,\mathbf{X})((\boldsymbol{\mu}^m)^{\mathrm{T}},\boldsymbol{\beta}^{\mathrm{T}})^{\mathrm{T}})^{\mathrm{T}}(\boldsymbol{\mu}-\boldsymbol{\mu}^*),\\ \Gamma_2 &= \sum_{i=1}^{n}\frac{\partial P_n(\boldsymbol{\mu}^m)}{\partial \mu_i}(\mu_i-\mu_i^*),\end{aligned}$$

in which $\boldsymbol{\mu}^m = \varsigma\boldsymbol{\mu} + (1-\varsigma)\boldsymbol{\mu}^*$ for some $\varsigma \in (0,1)$. Moreover,

$$\begin{aligned}\Gamma_2 &= \lambda\sum\nolimits_{\{j>i\}}\overline{\rho}(\mu_i^m-\mu_j^m)(\mu_i-\mu_i^*)+\lambda\sum\nolimits_{\{j<i\}}\overline{\rho}(\mu_i^m-\mu_j^m)(\mu_i-\mu_i^*)\\ &= \lambda\sum\nolimits_{\{j>i\}}\overline{\rho}(\mu_i^m-\mu_j^m)(\mu_i-\mu_i^*)+\lambda\sum\nolimits_{\{i<j\}}\overline{\rho}(\mu_j^m-\mu_i^m)(\mu_j-\mu_j^*)\\ &= \lambda\sum\nolimits_{\{j>i\}}\overline{\rho}(\mu_i^m-\mu_j^m)(\mu_i-\mu_i^*)-\lambda\sum\nolimits_{\{i<j\}}\overline{\rho}(\mu_i^m-\mu_j^m)(\mu_j-\mu_j^*)\\ &= \lambda\sum\nolimits_{\{j>i\}}\overline{\rho}(\mu_i^m-\mu_j^m)\{(\mu_i-\mu_i^*)-(\mu_j-\mu_j^*)\}. \qquad (A.6)\end{aligned}$$

When $i,j\in\mathcal{G}_k$, $\mu_i^*=\mu_j^*$, and $\mu_i^m-\mu_j^m$ has the same sign as $\mu_i-\mu_j$. Hence

$$\begin{aligned}\Gamma_2 &= \lambda\sum\nolimits_{k=1}^{K}\sum\nolimits_{\{i,j\in\mathcal{G}_k,i<j\}}\rho'(|\mu_i^m-\mu_j^m|)|\mu_i-\mu_j|\\ &\quad+\lambda\sum\nolimits_{k<k'}\sum\nolimits_{\{i\in\mathcal{G}_k,j'\in\mathcal{G}_{k'}\}}\overline{\rho}(\mu_i^m-\mu_j^m)\{(\mu_i-\mu_i^*)-(\mu_j-\mu_j^*).\end{aligned}$$

As shown in (A.5),

$$||\boldsymbol{\mu}^*-\boldsymbol{\mu}^0||_\infty = ||\boldsymbol{\alpha}-\boldsymbol{\alpha}^0||_\infty \le ||\boldsymbol{\mu}-\boldsymbol{\mu}^0||_\infty.$$

Since $\boldsymbol{\mu}^m = \varsigma\boldsymbol{\mu}+(1-\varsigma)\boldsymbol{\mu}^*$,

$$||\boldsymbol{\mu}^m-\boldsymbol{\mu}^0||_\infty \le ||\boldsymbol{\mu}-\boldsymbol{\mu}^0||_\infty \le \phi_n, \qquad (A.7)$$

and then for $k\neq k'$, $i\in\mathcal{G}_k$, $j\in\mathcal{G}_{k'}$,

$$\begin{aligned}|\mu_i^m-\mu_j^m| &\ge \min_{i\in\mathcal{G}_k,j\in\mathcal{G}_{k'}}|\mu_i^0-\mu_j^0|-2||\boldsymbol{\mu}^m-\boldsymbol{\mu}^0||_\infty\\ &\ge b_n-2||\boldsymbol{\mu}-\boldsymbol{\mu}^0||_\infty \ge b_n-2\phi_n > a\lambda,\end{aligned}$$

and thus $\overline{\rho}(\mu_i^m-\mu_j^m)=0$. Therefore,

$$\Gamma_2 = \lambda\sum\nolimits_{k=1}^{K}\sum\nolimits_{\{i,j\in\mathcal{G}_k,i<j\}}\rho'(|\mu_i^m-\mu_j^m|)|\mu_i-\mu_j|.$$

Furthermore, by the same reasoning as (A.5), we have

$$||\boldsymbol{\mu}^*-\widehat{\boldsymbol{\mu}}^{or}||_\infty \le ||\boldsymbol{\mu}-\widehat{\boldsymbol{\mu}}^{or}||_\infty.$$

Then

$$\begin{aligned}|\mu_i^m-\mu_j^m| &\le 2||\boldsymbol{\mu}^m-\boldsymbol{\mu}^*||_\infty \le 2||\boldsymbol{\mu}-\boldsymbol{\mu}^*||_\infty\\ &\le 2(||\boldsymbol{\mu}-\widehat{\boldsymbol{\mu}}^{or}||_\infty+||\boldsymbol{\mu}^*-\widehat{\boldsymbol{\mu}}^{or}||_\infty)\\ &\le 4||\boldsymbol{\mu}-\widehat{\boldsymbol{\mu}}^{or}||_\infty \le 4t_n.\end{aligned}$$

Hence $\rho'(|\mu_i^m - \mu_j^m|) \geq \rho'(4t_n)$ by concavity of $\rho(\cdot)$. As a result,

$$\Gamma_2 \geq \lambda \sum\nolimits_{k=1}^{K} \sum\nolimits_{\{i,j\in\mathcal{G}_k, i<j\}} \rho'(4t_n)|\mu_i - \mu_j|. \tag{A.8}$$

Let

$$\mathbf{w} = (w_1, \ldots, w_n)^{\mathrm{T}} = \mathbf{y} - (\mathbf{I}_n, \mathbf{X})((\boldsymbol{\mu}^m)^{\mathrm{T}}, \boldsymbol{\beta}^{\mathrm{T}})^{\mathrm{T}}.$$

Then

$$\begin{aligned}
\Gamma_1 &= -\mathbf{w}^{\mathrm{T}}(\boldsymbol{\mu} - \boldsymbol{\mu}^*) = -\sum\nolimits_{k=1}^{K} \sum\nolimits_{\{i,j\in\mathcal{G}_k\}} \frac{w_i(\mu_i - \mu_j)}{|\mathcal{G}_k|} \\
&= -\sum\nolimits_{k=1}^{K} \sum\nolimits_{\{i,j\in\mathcal{G}_k\}} \frac{w_i(\mu_i - \mu_j)}{2|\mathcal{G}_k|} - \sum\nolimits_{k=1}^{K} \sum\nolimits_{\{i,j\in\mathcal{G}_k\}} \frac{w_i(\mu_i - \mu_j)}{2|\mathcal{G}_k|} \\
&= -\sum\nolimits_{k=1}^{K} \sum\nolimits_{\{i,j\in\mathcal{G}_k\}} \frac{(w_j - w_i)(\mu_j - \mu_i)}{2|\mathcal{G}_k|} \\
&= -\sum\nolimits_{k=1}^{K} \sum\nolimits_{\{i,j\in\mathcal{G}_k, i<j\}} \frac{(w_j - w_i)(\mu_j - \mu_i)}{|\mathcal{G}_k|}. \qquad (A.9)
\end{aligned}$$

Since

$$\mathbf{w} = \boldsymbol{\epsilon} + \mathbf{X}(\boldsymbol{\beta}^0 - \boldsymbol{\beta}) + \boldsymbol{\mu}^0 - \boldsymbol{\mu}^m,$$

then

$$\max_{i,j} |w_j - w_i| \leq 2||\mathbf{w}||_\infty \leq 2||\boldsymbol{\epsilon}||_\infty + 2||\mathbf{X}||_\infty ||\boldsymbol{\beta}^0 - \boldsymbol{\beta}||_\infty + 2||\boldsymbol{\mu}^0 - \boldsymbol{\mu}^m||_\infty.$$

Hence by (A.7) and Condition (C1),

$$\max_{i,j} |w_j - w_i| \leq 2||\boldsymbol{\epsilon}||_\infty + 2C_2 p\phi_n + 2\phi_n.$$

By Condition (C3),

$$P(||\boldsymbol{\epsilon}||_\infty > \sqrt{2c_1^{-1}}\sqrt{\log n}) \leq \sum\nolimits_{i=1}^{n} P(|\varepsilon_i| > \sqrt{2c_1^{-1}}\sqrt{\log n}) \leq 2n^{-1}.$$

Thus there is an event E_2 such that $P(E_2^C) \leq 2n^{-1}$, and on the event E_2,

$$\max_{i,j} |w_j - w_i| \leq 2\sqrt{2c_1^{-1}}\sqrt{\log n} + 2(C_2 p + 1)\phi_n. \tag{A.10}$$

Hence

$$|\mathcal{G}_{\min}|^{-1} \max_{i,j} |w_j - w_i| \leq |\mathcal{G}_{\min}|^{-1}\{2\sqrt{2c_1^{-1}}\sqrt{\log n} + 2(C_2 p + 1)\phi_n\}.$$

Since $|\mathcal{G}_{\min}| \gg \sqrt{(K+p)n\log n}$ and $p = o(n)$, then $|\mathcal{G}_{\min}|^{-1}p = o(1)$. Thus $\lambda \gg \phi_n \gg |\mathcal{G}_{\min}|^{-1}2(C_2 p + 1)\phi_n$. Moreover, $\lambda \gg \phi_n \gg |\mathcal{G}_{\min}|^{-1}\sqrt{\log n}$. Hence

$$\lambda \gg |\mathcal{G}_{\min}|^{-1} \max_{i,j} |w_j - w_i|. \tag{A.11}$$

Let $t_n = o(1)$, then $\rho'(4t_n) \to 1$. Therefore, by (A.8), (A.9), and (A.11),

$$\begin{aligned}
&Q_n(\boldsymbol{\mu}, \boldsymbol{\beta}) - Q_n(\boldsymbol{\mu}^*, \boldsymbol{\beta}) = \Gamma_1 + \Gamma_2 \\
&\geq \sum\nolimits_{k=1}^{K} \sum\nolimits_{\{i,j\in\mathcal{G}_k, i<j\}} [\lambda\rho'(4t_n) - |\mathcal{G}_{\min}|^{-1} \max_{i,j} |w_j - w_i|]|\mu_i - \mu_j| \geq 0,
\end{aligned}$$

for sufficiently large n, so that the result in (ii) is proved.

A.4 Proof of Theorem 3

In this section we show the results in Theorem 3. The proofs of (12) and (13) follow the same arguments as the proof of Theorem 1 by letting $\mathbf{Z} = \mathbf{1}_n$ and $|\mathcal{G}_{\min}| = n$, and thus they are omitted. Next, we will show (14). It follows similar procedures as the proof of Theorem 2 with the details given below. Let $\mathcal{M}$ be the subspace of R^n, defined as

$$\mathcal{M} = \{\boldsymbol{\mu} \in R^n : \mu_1 = \cdots = \mu_n\}.$$

For each $\boldsymbol{\mu} \in \mathcal{M}$, it can be written as $\boldsymbol{\mu} = \mathbf{1}_n \alpha$, where α is the common value of $\boldsymbol{\mu}$. Let $T : \mathcal{M} \to R$ be the mapping such that $T(\boldsymbol{\mu})$ is the scalar that equals to the common value of μ_i's. Let $T^* : R^n \to R$ be the mapping such that $T^*(\boldsymbol{\mu}) = n^{-1} \sum_{i=1}^n \mu_i$. Clearly, when $\boldsymbol{\mu} \in \mathcal{M}$, $T(\boldsymbol{\mu}) = T^*(\boldsymbol{\mu})$. Consider the neighborhood of $(\boldsymbol{\mu}^0, \boldsymbol{\beta}^0)$:

$$\Theta = \{\boldsymbol{\mu} \in R^n, \boldsymbol{\beta} \in R^p: \left\|((\boldsymbol{\mu} - \boldsymbol{\mu}^0)^{\mathrm{T}}, (\boldsymbol{\beta} - \boldsymbol{\beta}^0)^{\mathrm{T}})^{\mathrm{T}}\right\|_\infty \leq \phi_n\},$$

where $\phi_n = c_1^{-1/2} C_1^{-1} \sqrt{1+p} \sqrt{n^{-1} \log n}$. By the result in (12), there is an event E_1 such that on the even E_1,

$$\left\|((\widehat{\boldsymbol{\mu}}^{or} - \boldsymbol{\mu}^0)^{\mathrm{T}}, (\widehat{\boldsymbol{\beta}}^{or} - \boldsymbol{\beta}^0)^{\mathrm{T}})^{\mathrm{T}}\right\|_\infty \leq \phi_n,$$

and $P(E_1^C) \leq 2(1+p)n^{-1}$. Hence $((\widehat{\boldsymbol{\mu}}^{or})^{\mathrm{T}}, (\widehat{\boldsymbol{\beta}}^{or})^{\mathrm{T}})^{\mathrm{T}} \in \Theta$ on the event E_1. For any $\boldsymbol{\mu} \in R^n$, let $\boldsymbol{\mu}^* = T^{-1}(T^*(\boldsymbol{\mu}))$. We show that $(\widehat{\boldsymbol{\mu}}^{or}, \widehat{\boldsymbol{\beta}}^{or})$ is a strictly local minimizer of the objective function (3) with probability approaching 1 through the following two steps.

(i). On the event E_1, $Q_n(\boldsymbol{\mu}^*, \boldsymbol{\beta}) > Q_n(\widehat{\boldsymbol{\mu}}^{or}, \widehat{\boldsymbol{\beta}}^{or})$ for any $(\boldsymbol{\mu}^{\mathrm{T}}, \boldsymbol{\beta}^{\mathrm{T}})^{\mathrm{T}} \in \Theta$ and $((\boldsymbol{\mu}^*)^{\mathrm{T}}, \boldsymbol{\beta}^{\mathrm{T}})^{\mathrm{T}} \neq ((\widehat{\boldsymbol{\mu}}^{or})^{\mathrm{T}}, (\widehat{\boldsymbol{\beta}}^{or})^{\mathrm{T}})^{\mathrm{T}}$.

(ii). There is an event E_2 such that $P(E_2^C) \leq 2n^{-1}$. On $E_1 \cap E_2$, there is a neighborhood of $((\widehat{\boldsymbol{\mu}}^{or})^{\mathrm{T}}, (\widehat{\boldsymbol{\beta}}^{or})^{\mathrm{T}})^{\mathrm{T}}$, denoted by Θ_n, such that $Q_n(\boldsymbol{\mu}, \boldsymbol{\beta}) \geq Q_n(\boldsymbol{\mu}^*, \boldsymbol{\beta})$ for any $(\boldsymbol{\mu}^{\mathrm{T}}, \boldsymbol{\beta}^{\mathrm{T}})^{\mathrm{T}} \in \Theta_n \cap \Theta$ for sufficiently large n.

Therefore, by the results in (i) and (ii), we have $Q_n(\boldsymbol{\mu}, \boldsymbol{\beta}) > Q_n(\widehat{\boldsymbol{\mu}}^{or}, \widehat{\boldsymbol{\beta}}^{or})$ for any $(\boldsymbol{\mu}^{\mathrm{T}}, \boldsymbol{\beta}^{\mathrm{T}})^{\mathrm{T}} \in \Theta_n \cap \Theta$ and $(\boldsymbol{\mu}^{\mathrm{T}}, \boldsymbol{\beta}^{\mathrm{T}})^{\mathrm{T}} \neq ((\widehat{\boldsymbol{\mu}}^{or})^{\mathrm{T}}, (\widehat{\boldsymbol{\beta}}^{or})^{\mathrm{T}})^{\mathrm{T}}$, so that $((\widehat{\boldsymbol{\mu}}^{or})^{\mathrm{T}}, (\widehat{\boldsymbol{\beta}}^{or})^{\mathrm{T}})^{\mathrm{T}}$ is a strict local minimizer of $Q_n(\boldsymbol{\mu}, \boldsymbol{\beta})$ on the event $E_1 \cap E_2$ with $P(E_1 \cap E_2) \geq 1 - 2(p+2)n^{-1}$ for sufficiently large n.

By the definition of $((\widehat{\boldsymbol{\mu}}^{or})^{\mathrm{T}}, (\widehat{\boldsymbol{\beta}}^{or})^{\mathrm{T}})^{\mathrm{T}}$, we have $\frac{1}{2}\sum_{i=1}^n (y_i - \boldsymbol{\mu}^* - \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta})^2 > \frac{1}{2}\sum_{i=1}^n (y_i - \widehat{\boldsymbol{\mu}}^{or} - \mathbf{x}_i^{\mathrm{T}} \widehat{\boldsymbol{\beta}}^{or})^2$ for any $(\boldsymbol{\mu}^{\mathrm{T}}, \boldsymbol{\beta}^{\mathrm{T}})^{\mathrm{T}} \in \Theta$ and $((\boldsymbol{\mu}^*)^{\mathrm{T}}, \boldsymbol{\beta}^{\mathrm{T}})^{\mathrm{T}} \neq ((\widehat{\boldsymbol{\mu}}^{or})^{\mathrm{T}}, (\widehat{\boldsymbol{\beta}}^{or})^{\mathrm{T}})^{\mathrm{T}}$. Moreover, since

$p_\gamma(|\widehat{\mu}_i^{or} - \widehat{\mu}_j^{or}|, \lambda) = p_\gamma(|\mu_i^* - \mu_j^*|, \lambda) = 0$ for $1 \le i, j \le n$, we have $Q_n(\boldsymbol{\mu}^*, \boldsymbol{\beta}) = \frac{1}{2}\sum_{i=1}^n (y_i - \boldsymbol{\mu}^* - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})^2$ and $Q_n(\widehat{\boldsymbol{\mu}}^{or}, \widehat{\boldsymbol{\beta}}^{or}) = \frac{1}{2}\sum_{i=1}^n (y_i - \widehat{\boldsymbol{\mu}}^{or} - \mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}^{or})^2$. Therefore, $Q_n(\boldsymbol{\mu}^*, \boldsymbol{\beta}) > Q_n(\widehat{\boldsymbol{\mu}}^{or}, \widehat{\boldsymbol{\beta}}^{or})$.

Next we prove the result in (ii). For a positive sequence t_n, let $\Theta_n = \{\boldsymbol{\mu} : ||\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}^{or}|| \le t_n\}$. For $(\boldsymbol{\mu}^{\mathrm{T}}, \boldsymbol{\beta}^{\mathrm{T}})^{\mathrm{T}} \in \Theta_n \cap \Theta$, by Taylor's expansion, we have

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

where

$$\begin{aligned} \Gamma_1 &= -(\mathbf{y} - (\mathbf{I}_n, \mathbf{X})((\boldsymbol{\mu}^m)^{\mathrm{T}}, \boldsymbol{\beta}^{\mathrm{T}})^{\mathrm{T}})^{\mathrm{T}}(\boldsymbol{\mu} - \boldsymbol{\mu}^*), \\ \Gamma_2 &= \sum_{i=1}^n \frac{\partial P_n(\boldsymbol{\mu}^m)}{\partial \mu_i}(\mu_i - \mu_i^*). \end{aligned}$$

in which $\boldsymbol{\mu}^m = \varsigma\boldsymbol{\mu} + (1-\varsigma)\boldsymbol{\mu}^*$ for some $\varsigma \in (0,1)$. Moreover, by (A.6), we have

$$\begin{aligned} \Gamma_2 &= \lambda \sum\nolimits_{i<j} \overline{\rho}(\mu_i^m - \mu_j^m)\{(\mu_i - \mu_i^*) - (\mu_j - \mu_j^*)\} \\ &= \lambda \sum\nolimits_{i<j} \rho'(|\mu_i^m - \mu_j^m|)|\mu_i - \mu_j|, \end{aligned}$$

where the second equality holds due to the fact that $\mu_i^* = \mu_j^*$ and $\mu_i^m - \mu_j^m$ has the same sign as $\mu_i - \mu_j$. Let $T^*(\boldsymbol{\mu}) = \alpha$. Following the same reasoning as the proof for (A.5), we have

$$||\boldsymbol{\mu}^* - \widehat{\boldsymbol{\mu}}^{or}||_\infty = |\alpha - \widehat{\alpha}^{or}| \le ||\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}^{or}||_\infty.$$

Then

$$\begin{aligned} |\mu_i^m - \mu_j^m| &\le 2||\boldsymbol{\mu}^m - \boldsymbol{\mu}^*||_\infty \le 2||\boldsymbol{\mu} - \boldsymbol{\mu}^*||_\infty \\ &\le 2(||\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}^{or}||_\infty + ||\boldsymbol{\mu}^* - \widehat{\boldsymbol{\mu}}^{or}||_\infty) \\ &\le 4||\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}^{or}||_\infty \le 4t_n. \end{aligned}$$

Hence $\rho'(|\mu_i^m - \mu_j^m|) \ge \rho'(4t_n)$ by concavity of $\rho(\cdot)$. As a result,

$$\Gamma_2 \ge \lambda \sum\nolimits_{i<j} \rho'(4t_n)|\mu_i - \mu_j|. \tag{A.12}$$

Then, by the same reasoning as the proof for (A.9), we have

$$\Gamma_1 = -\mathbf{w}^{\mathrm{T}}(\boldsymbol{\mu} - \boldsymbol{\mu}^*) = -n^{-1}\sum\nolimits_{i<j}(w_j - w_i)(\mu_j - \mu_i), \tag{A.13}$$

where $\mathbf{w} = (w_1, \ldots, w_n)^{\mathrm{T}} = \mathbf{y} - (\mathbf{I}_n, \mathbf{X})((\boldsymbol{\mu}^m)^{\mathrm{T}}, \boldsymbol{\beta}^{\mathrm{T}})^{\mathrm{T}}$. By the same reasoning as the proof for (A.10), we have that there is an event E_2 such that $P(E_2^C) \leq 2n^{-1}$, and on the event E_2,

$$\max_{i,j} |w_j - w_i| \leq 2\sqrt{2c_1^{-1}}\sqrt{\log n} + 2(C_2 p + 1)\phi_n.$$

Hence

$$n^{-1} \max_{i,j} |w_j - w_i| \leq n^{-1}\{2\sqrt{2c_1^{-1}}\sqrt{\log n} + 2(C_2 p + 1)\phi_n\}.$$

Since $n^{-1}p = o(1)$, then $\lambda \gg \phi_n \gg n^{-1}2(C_2p+1)\phi_n$. Moreover, $\lambda \gg \phi_n \gg n^{-1}\sqrt{\log n}$. Hence

$$\lambda \gg n^{-1} \max_{i,j} |w_j - w_i|. \tag{A.14}$$

Let $t_n = o(1)$, then $\rho'(4t_n) \to 1$. Therefore, by (A.12), (A.13), and (A.14),

$$\begin{aligned} Q_n(\boldsymbol{\mu}, \boldsymbol{\beta}) - Q_n(\boldsymbol{\mu}^*, \boldsymbol{\beta}) &= \Gamma_1 + \Gamma_2 \\ \geq \sum\nolimits_{i<j} [\lambda\rho'(4t_n) &- n^{-1} \max_{i,j} |w_j - w_i|]|\mu_i - \mu_j| \geq 0, \end{aligned}$$

for sufficiently large n, so that the result in (ii) is proved.

A.5 Estimation procedure for model (2)

We let $\widetilde{\mathbf{x}}_i = (1, \mathbf{x}_i^{\mathrm{T}})^{\mathrm{T}}$ and $\boldsymbol{\beta}^* = (\mu, \boldsymbol{\beta}^{\mathrm{T}})^{\mathrm{T}}$. The model (2) can be written as $y_i = \mathbf{z}_i^{\mathrm{T}}\boldsymbol{\theta}_i + \widetilde{\mathbf{x}}_i^{\mathrm{T}}\boldsymbol{\beta}^* + \epsilon_i, i = 1, \ldots, n$. Similar to the assumption for model (1), we assume that observations can be divided into K different subgroups with $K < n$. Let $\mathcal{G} = (\mathcal{G}_1, \ldots, \mathcal{G}_K)$ be a partition of $\{1, \ldots, n\}$, and we assume $\boldsymbol{\theta}_i = \boldsymbol{\alpha}_k$ for all $i \in \mathcal{G}_k$, where $\boldsymbol{\alpha}_k$ is the common value for the $\boldsymbol{\theta}_i$'s from group $\mathcal{G}_k$. Then the estimates of $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^{\mathrm{T}}, \ldots, \boldsymbol{\theta}_n^{\mathrm{T}})^{\mathrm{T}}$ and $\boldsymbol{\beta}^*$ can be obtained by minimizing

$$Q_n(\boldsymbol{\theta}, \boldsymbol{\beta}^*; \lambda) = \frac{1}{2}\sum\nolimits_{i=1}^{n} (y_i - \mathbf{z}_i^{\mathrm{T}}\boldsymbol{\theta}_i - \widetilde{\mathbf{x}}_i^{\mathrm{T}}\boldsymbol{\beta}^*)^2 + \sum\nolimits_{1 \leq i < j \leq n} p(||\boldsymbol{\theta}_i - \boldsymbol{\theta}_j||, \lambda), \tag{A.15}$$

where $p(\cdot, \lambda)$ is a concave penalty function with a tuning parameter λ, such as MCP or SCAD as described in Section 2. Then for a given $\lambda > 0$, define

$$(\widehat{\boldsymbol{\theta}}(\lambda), \widehat{\boldsymbol{\beta}}^*(\lambda)) = \operatorname{argmin} Q_n(\boldsymbol{\theta}, \boldsymbol{\beta}^*; \lambda).$$

The penalty shrinks some of $||\boldsymbol{\theta}_i - \boldsymbol{\theta}_j||$ to zero. Based on this, we can partition the treatment effects into subgroups. Specifically, let $\widehat{\lambda}$ be the value of the tuning parameter selected based on a data-driven procedure such as the BIC. For simplicity, write $(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\beta}}^*) \equiv (\widehat{\boldsymbol{\theta}}(\lambda), \widehat{\boldsymbol{\beta}}^*(\lambda))$. Let $\{\widehat{\boldsymbol{\alpha}}_1, \ldots, \widehat{\boldsymbol{\alpha}}_{\widehat{K}}\}$ be the distinct values of $\widehat{\boldsymbol{\theta}}$. Let $\widehat{\mathcal{G}}_k = \{i : \widehat{\boldsymbol{\theta}}_i = \widehat{\boldsymbol{\alpha}}_k, 1 \le i \le n\}, 1 \le k \le \widehat{K}$. Then $\{\widehat{\mathcal{G}}_1, \ldots, \widehat{\mathcal{G}}_{\widehat{K}}\}$ constitutes a partition of $\{1, \ldots, n\}$. Then we apply our proposed ADMM algorithm to obtain the estimates of $\boldsymbol{\theta}$ and $\boldsymbol{\beta}^*$ described as follows.

We reparametrize by introducing a new set of parameters $\boldsymbol{\delta}_{ij} = \boldsymbol{\theta}_i - \boldsymbol{\theta}_j$, and hence minimization of (A.15) is equivalent to the constraint optimization problem:

$$S(\boldsymbol{\theta}, \boldsymbol{\beta}^*, \boldsymbol{\delta}) = \frac{1}{2} \sum\nolimits_{i=1}^{n} (y_i - \mathbf{z}_i^{\mathrm{T}} \boldsymbol{\theta}_i - \widetilde{\mathbf{x}}_i^{\mathrm{T}} \boldsymbol{\beta}^*)^2 + \sum\nolimits_{i<j} p_\gamma(||\boldsymbol{\delta}_{ij}||, \lambda),$$
$$\text{subject to } \boldsymbol{\theta}_i - \boldsymbol{\theta}_j - \boldsymbol{\delta}_{ij} = \mathbf{0},$$

where $\boldsymbol{\delta} = \{\boldsymbol{\delta}_{ij}^{\mathrm{T}}, i < j\}^{\mathrm{T}}$. By the augmented Lagrangian method (ALM), the estimates of the parameters can be obtained by minimizing

$$L(\boldsymbol{\theta}, \boldsymbol{\beta}^*, \boldsymbol{\delta}, \boldsymbol{\upsilon}) = S(\boldsymbol{\theta}, \boldsymbol{\beta}^*, \boldsymbol{\delta}) + \sum\nolimits_{i<j} \langle \boldsymbol{\upsilon}_{ij}, \boldsymbol{\theta}_i - \boldsymbol{\theta}_j - \boldsymbol{\delta}_{ij} \rangle + \frac{\vartheta}{2} \sum\nolimits_{i<j} ||\boldsymbol{\theta}_i - \boldsymbol{\theta}_j - \boldsymbol{\delta}_{ij}||^2,$$

where the dual variables $\boldsymbol{\upsilon} = \{\boldsymbol{\upsilon}_{ij}^{\mathrm{T}}, i < j\}^{\mathrm{T}}$ are Lagrange multipliers and ϑ is the penalty parameter. We then can obtain the estimators of $(\boldsymbol{\theta}, \boldsymbol{\beta}^*, \boldsymbol{\delta}, \boldsymbol{\upsilon})$ through iterations by the ADMM.

For given $(\boldsymbol{\theta}, \boldsymbol{\beta}^*, \boldsymbol{\upsilon})$, the minimizer of $L(\boldsymbol{\theta}, \boldsymbol{\beta}^*, \boldsymbol{\delta}, \boldsymbol{\upsilon})$ with respect to $\boldsymbol{\delta}_{ij}$ is unique and has a closed-form expression for the L_1, MCP and SCAD penalties, respectively. Specifically, for given $(\boldsymbol{\theta}, \boldsymbol{\beta}^*, \boldsymbol{\upsilon})$, the minimization problem is the same as minimizing

$$\frac{\vartheta}{2} \sum\nolimits_{i<j} ||\boldsymbol{\zeta}_{ij} - \boldsymbol{\delta}_{ij}||^2 + \sum\nolimits_{i<j} p_\gamma(||\boldsymbol{\delta}_{ij}||, \lambda)$$

with respect to $\boldsymbol{\delta}_{ij}$, where $\boldsymbol{\zeta}_{ij} = \boldsymbol{\theta}_i - \boldsymbol{\theta}_j + \vartheta^{-1} \boldsymbol{\upsilon}_{ij}$. Hence, the closed-form solution for the L_1 penalty is

$$\widehat{\boldsymbol{\delta}}_{ij} = S(\boldsymbol{\zeta}_{ij}, \lambda/\vartheta), \tag{A.16}$$

where $S(\mathbf{z}, t) = (1 - t/||\mathbf{z}||)_+ \mathbf{z}$ is the groupwise soft thresholding rule, and $(x)_+ = x$ if $x > 0$ and 0, otherwise. For the MCP penalty with $\gamma > 1/\vartheta$, it is

$$\widehat{\boldsymbol{\delta}}_{ij} = \begin{cases} \frac{S(\boldsymbol{\zeta}_{ij}, \lambda/\vartheta)}{1 - 1/(\gamma\vartheta)} & \text{if } ||\boldsymbol{\zeta}_{ij}|| \le \gamma\lambda \\ \boldsymbol{\zeta}_{ij} & \text{if } ||\boldsymbol{\zeta}_{ij}|| > \gamma\lambda. \end{cases} \tag{A.17}$$

For the SCAD penalty with $\gamma > 1/\vartheta + 1$, it is

$$\widehat{\boldsymbol{\delta}}_{ij} = \begin{cases} \mathrm{ST}(\boldsymbol{\zeta}_{ij}, \lambda/\vartheta) & \text{if } ||\boldsymbol{\zeta}_{ij}|| \leq \lambda + \lambda/\vartheta \\ \frac{\mathrm{ST}(\boldsymbol{\zeta}_{ij}, \gamma\lambda/((\gamma-1)\vartheta))}{1-1/((\gamma-1)\vartheta)} & \text{if } \lambda + \lambda/\vartheta < \ ||\boldsymbol{\zeta}_{ij}|| \leq \gamma\lambda \\ \boldsymbol{\zeta}_{ij} & \text{if } ||\boldsymbol{\zeta}_{ij}|| > \gamma\lambda. \end{cases} \tag{A.18}$$

ADMM algorithm for (A.15). We now describe the computational algorithm based on the ADMM for minimizing (A.15). It consists of iteratively updating $\boldsymbol{\theta}, \boldsymbol{\beta}^*, \boldsymbol{\delta}$ and $\boldsymbol{\upsilon}$. The main ingredients of the algorithm are as follows.

First, for a given $(\boldsymbol{\delta}, \boldsymbol{\upsilon})$, to obtain an update of $\boldsymbol{\theta}$ and $\boldsymbol{\beta}^*$, we set the derivatives $\partial L(\boldsymbol{\theta}, \boldsymbol{\beta}^*, \boldsymbol{\delta}, \boldsymbol{\upsilon})/\partial\boldsymbol{\theta}$ and $\partial L(\boldsymbol{\theta}, \boldsymbol{\beta}^*, \boldsymbol{\delta}, \boldsymbol{\upsilon})/\partial\boldsymbol{\beta}^*$ to zero, where

$$\begin{aligned} L(\boldsymbol{\theta}, \boldsymbol{\beta}^*, \boldsymbol{\delta}, \boldsymbol{\upsilon}) &= \frac{1}{2}\sum\nolimits_{i=1}^{n}(y_i - \mathbf{z}_i^{\mathrm{T}}\boldsymbol{\theta}_i - \widetilde{\mathbf{x}}_i^{\mathrm{T}}\boldsymbol{\beta}^*)^2 + \frac{\vartheta}{2}\sum\nolimits_{i<j}||\boldsymbol{\theta}_i - \boldsymbol{\theta}_j - \boldsymbol{\delta}_{ij} + \vartheta^{-1}\boldsymbol{\upsilon}_{ij}||^2 + C \\ &= \frac{1}{2}||\mathbf{Z}\boldsymbol{\theta} + \widetilde{\mathbf{X}}\boldsymbol{\beta}^* - \mathbf{y}||^2 + \frac{\vartheta}{2}||\mathbf{A}\boldsymbol{\beta} - \boldsymbol{\delta} + \vartheta^{-1}\boldsymbol{\upsilon}||^2 + C. \end{aligned}$$

Here C is a constant independent of $\boldsymbol{\theta}$ and $\boldsymbol{\beta}^*$, $\mathbf{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$, $\mathbf{Z} = \mathrm{diag}(\mathbf{z}_1^{\mathrm{T}}, \ldots, \mathbf{z}_n^{\mathrm{T}})$ and $\widetilde{\mathbf{X}} = (\widetilde{\mathbf{x}}_1, \ldots, \widetilde{\mathbf{x}}_n)^{\mathrm{T}}$. Moreover, e_i is the $n \times 1$ vector whose i^{th} element is 1 and the remaining ones are 0, $\Delta = \{(e_i - e_j), i < j\}^{\mathrm{T}}$ and $\mathbf{A} = \Delta \otimes \mathbf{I}_p$, where $\mathbf{I}_d$ denotes the $d \times d$ identity matrix and $\otimes$ denotes the Kronecker product.

Thus for given $\boldsymbol{\delta}^{(m)}$ and $\boldsymbol{\upsilon}^{(m)}$ at the m^{th} step, the updates $\boldsymbol{\theta}^{(m+1)}$ and $\boldsymbol{\beta}^{*(m+1)}$, which are the minimizers of $L(\boldsymbol{\theta}, \boldsymbol{\beta}^*, \boldsymbol{\delta}^{(m)}, \boldsymbol{\upsilon}^{(m)})$, are

$$\begin{aligned} \boldsymbol{\theta}^{(m+1)} &= (\mathbf{Z}^{\mathrm{T}}(\mathbf{I}_n - \mathbf{Q}_{\widetilde{\mathbf{X}}})\mathbf{Z} + \vartheta\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1}[\mathbf{Z}^{\mathrm{T}}(\mathbf{I}_n - \mathbf{Q}_{\widetilde{\mathbf{X}}})\mathbf{y} + \vartheta\mathbf{A}^{\mathrm{T}}(\boldsymbol{\delta}^{(m)} - \vartheta^{-1}\boldsymbol{\upsilon}^{(m)})], \\ \boldsymbol{\beta}^{*(m+1)} &= (\widetilde{\mathbf{X}}^{\mathrm{T}}\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}^{\mathrm{T}}(\mathbf{y} - \mathbf{Z}\boldsymbol{\theta}^{(m+1)}), \end{aligned}$$

where $\mathbf{Q}_{\widetilde{\mathbf{X}}} = \widetilde{\mathbf{X}}(\widetilde{\mathbf{X}}^{\mathrm{T}}\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}^{\mathrm{T}}$.

Second, the update of $\boldsymbol{\delta}_{ij}$ at the $(m+1)^{\text{th}}$ iteration is obtained by the formula given in (A.16), (A.17) and (A.18), respectively, by the Lasso, MCP and SCAD penalties with $\boldsymbol{\zeta}_{ij}$ replaced by $\boldsymbol{\zeta}_{ij}^{(m+1)} = \boldsymbol{\beta}_i^{(m+1)} - \boldsymbol{\beta}_j^{(m+1)} + \vartheta^{-1}\boldsymbol{\upsilon}_{ij}^{(m+1)}$.

Finally, the estimate of $\boldsymbol{\upsilon}_{ij}$ is updated as

$$\boldsymbol{\upsilon}_{ij}^{(m+1)} = \boldsymbol{\upsilon}_{ij}^{(m)} + \vartheta(\boldsymbol{\beta}_i^{(m+1)} - \boldsymbol{\beta}_j^{(m+1)} - \boldsymbol{\delta}_{ij}^{(m+1)}).$$

We iteratively update the estimates of $\boldsymbol{\theta}, \boldsymbol{\beta}^*, \boldsymbol{\delta}$ and $\boldsymbol{\upsilon}$ until the stopping rule is met. We track the progress of the ADMM based on the primal residual $\mathbf{r}^{(m+1)} = \mathbf{A}\boldsymbol{\theta}^{(m+1)} - \boldsymbol{\delta}^{(m+1)}$. We stop the algorithm when $\mathbf{r}^{(m+1)}$ is close to zero such that $\left\|\mathbf{r}^{(m+1)}\right\| < \epsilon$ for some small value ϵ.

References

Tseng, P. (2001). Convergence of a block coordinate descent method for nondifferentiable minimization, *Journal of Optimization Theory and Applications*, 109, 475-494.