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 definition of \( \eta^{(m+1)} \), we have

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

for any \( \eta \). Define

\[
f^{(m+1)} = \inf_{\Delta \mu^{(m+1)} - \eta = 0} \left\{ \frac{1}{2} \| y - \mu^{(m+1)} - X' \beta^{(m+1)} \|^2 + \sum_{i<j} p_{ij}(|\eta_{ij}|, \lambda) \right\} = \inf_{\Delta \mu^{(m+1)} - \eta = 0} L(\mu^{(m+1)}, \beta^{(m+1)}, \eta, \upsilon^{(m)}).
\]

Then

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

*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 $\mathbf{v}^{(m+t-1)} = \mathbf{v}^{(m)} + \vartheta \sum_{i=1}^{t-1}(\Delta \mathbf{\mu}^{(m+i)} - \eta^{(m+i)})$, we have

$$L(\mathbf{\mu}^{(m+t)}, \mathbf{\beta}^{(m+t)}, \mathbf{\eta}^{(m+t)}, \mathbf{v}^{(m+t-1)}) = \frac{1}{2} \left\| \mathbf{y} - \mathbf{\mu}^{(m+t)} - \mathbf{X} \mathbf{\beta}^{(m+t)} \right\|^2 + \mathbf{v}^{(m+t-1)\mathsf{T}} (\Delta \mathbf{\mu}^{(m+t)} - \eta^{(m+t)}) + \frac{\vartheta}{2} \| \Delta \mathbf{\mu}^{(m+t)} - \eta^{(m+t)} \|^2 + \sum_{i<j} p_\gamma(|\eta_{ij}|, \lambda)

= \frac{1}{2} \left\| \mathbf{y} - \mathbf{\mu}^{(m+t)} - \mathbf{X} \mathbf{\beta}^{(m+t)} \right\|^2 + \mathbf{v}^{(m)\mathsf{T}} (\Delta \mathbf{\mu}^{(m+t)} - \eta^{(m+t)}) + \frac{\vartheta}{2} \| \Delta \mathbf{\mu}^{(m+t)} - \eta^{(m+t)} \|^2 + \sum_{i<j} p_\gamma(|\eta_{ij}|, \lambda)

\leq f^{(m+t)}.

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

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

and for all $t \geq 0$

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

\leq f^*.$$

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

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

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

and moreover,

$$\partial L(\mathbf{\mu}^{(m+1)}, \mathbf{\beta}^{(m+1)}, \mathbf{\eta}^{(m)}, \mathbf{v}^{(m)}) / \partial \mathbf{\mu} = \mu^{(m+1)} + \mathbf{X} \mathbf{\beta}^{(m+1)} - \mathbf{y} + \Delta^\mathsf{T} \mathbf{v}^{(m)} + \Delta^\mathsf{T} \partial (\Delta \mathbf{\mu}^{(m+1)} - \mathbf{\eta}^{(m)})$$

$$= \mu^{(m+1)} + \mathbf{X} \mathbf{\beta}^{(m+1)} - \mathbf{y} + \Delta^\mathsf{T} \mathbf{v}^{(m)} + \partial (\Delta \mathbf{\mu}^{(m+1)} - \mathbf{\eta}^{(m)})$$

$$= \mu^{(m+1)} + \mathbf{X} \mathbf{\beta}^{(m+1)} - \mathbf{y} + \Delta^\mathsf{T} \mathbf{v}^{(m+1)} + \partial \Delta^\mathsf{T} (\mathbf{\eta}^{(m+1)} - \mathbf{\eta}^{(m)}).$$

A.2
The last step follows from \( \nu^{(m+1)} = \nu^{(m)} + \partial(\Delta\mu^{(m+1)} - \eta^{(m+1)}) \). Therefore,

\[
\begin{align*}
\nu^{(m+1)} &= \partial^T(\eta^{(m+1)} - \eta^{(m)}) = -(\mu^{(m+1)} + X\beta^{(m+1)} - y + \Delta^T\nu^{(m+1)}).
\end{align*}
\]

Since \( ||\Delta\mu^* - \eta^*||^2 = 0 \),

\[
\lim_{m \to \infty} \partial L(\mu^{(m+1)}, \beta^{(m+1)}, \eta^{(m)}, \nu^{(m)}) / \partial \mu = \lim_{m \to \infty} \mu^{(m+1)} + X\beta^{(m+1)} - y + \Delta^T\nu^{(m+1)} = \mu^* + X\beta^* - y + \Delta^T\nu^* = 0.
\]

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

### A.2 Proof of Theorem 1

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

\[
\begin{align*}
\left( \begin{array}{c}
\hat{\alpha}^{or} \\
\hat{\beta}^{or}
\end{array} \right) &= \arg \min_{\alpha \in R^K, \beta \in R^P} \frac{1}{2} \|y - Z\alpha - X\beta\|^2 = [(Z, X)^T(Z, X)]^{-1}(Z, X)^Ty.
\end{align*}
\]

Then

\[
\begin{align*}
\left( \begin{array}{c}
\hat{\alpha}^{or} - \alpha^0 \\
\hat{\beta}^{or} - \beta^0
\end{array} \right) &= [(Z, X)^T(Z, X)]^{-1}(Z, X)^T\epsilon,
\end{align*}
\]

where \( \epsilon = (\epsilon_1, \ldots, \epsilon_n)^T \) and \( \alpha^0 = (\alpha_1^0, \ldots, \alpha_K^0)^T \). Hence

\[
\left\| \left( \begin{array}{c}
\hat{\alpha}^{or} - \alpha^0 \\
\hat{\beta}^{or} - \beta^0
\end{array} \right) \right\|_\infty \leq \left\| [(Z, X)^T(Z, X)]^{-1} \right\|_\infty \left\| (Z, X)^T\epsilon \right\|_\infty. \tag{A.1}
\]

By Condition (C1), we have \( \left\| [(Z, X)^T(Z, X)]^{-1} \right\|_\infty \leq C_1^{-1} |G_{\text{min}}|^{-1} \) and thus

\[
\left\| [(Z, X)^T(Z, X)]^{-1} \right\|_\infty \leq \sqrt{K + pC_1^{-1} |G_{\text{min}}|^{-1}}. \tag{A.2}
\]

Moreover

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

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

\[
P \left( \left\| Z^T\epsilon \right\|_\infty > C\sqrt{n \log n} \right) \leq \sum_{k=1}^K P\left( \sum_{i \in G_k} \epsilon_i > C\sqrt{n \log n} \right) \leq \sum_{k=1}^K P\left( \sum_{i \in G_k} \epsilon_i > \sqrt{|G_k|C\sqrt{\log n}} \right) \leq 2K \exp(-c_1C^2 \log n) = 2Kn^{-c_1C^2},
\]

A.3
and by Conditions (C1) and (C3) and union bound,
\[
P \left( \|X^T \epsilon\|_\infty > C \sqrt{n \log n} \right) \\
\leq \sum_{j=1}^{p} P \left( |X_j^T \epsilon| > \sqrt{nC \log n} \right) \\
\leq 2p \exp(-c_1 C^2 \log n) = 2pn^{-c_1 C^2}.
\]

By the above results, we have
\[
P \left( \|Z \cdot X \|_\infty > C \sqrt{n \log n} \right) \leq 2(K + p) n^{-c_1 C^2}.
\]

Therefore, by (A.1), (A.2) and (A.3), we have with probability at least 1 \(- 2(K + p) n^{-c_1 C^2}\),
\[
\left\| \left( \tilde{\alpha}^{or} - \alpha^0 \right) \right\|_\infty \leq CC_{1}^{-1} \sqrt{K + p} \left| \mathcal{G}_{\infty} \right|^{-1} \sqrt{n \log n},
\]
and hence \( \|\tilde{\mu}^{or} - \mu^0\|_\infty = \|\tilde{\alpha}^{or} - \alpha^0\|_\infty \leq CC_{1}^{-1} \sqrt{K + p} \left| \mathcal{G}_{\infty} \right|^{-1} \sqrt{n \log n} \). 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
\[
L_n(\mu, \beta) = \frac{1}{2} \|y - \mu - X \beta\|^2, P_n(\mu) = \lambda \sum_{i<j} \rho(|\mu_i - \mu_j|),
\]
\[
L_n^\mathcal{G}(\alpha, \beta) = \frac{1}{2} \|y - Z \alpha - X \beta\|^2, P_n^\mathcal{G}(\alpha) = \lambda \sum_{k<k'} |G_k||G_{k'}| \rho(|\alpha_k - \alpha_{k'}|),
\]
and let
\[
Q_n(\mu, \beta) = L_n(\mu, \beta) + P_n(\mu), Q_n^\mathcal{G}(\alpha, \beta) = L_n^\mathcal{G}(\alpha, \beta) + P_n^\mathcal{G}(\alpha).
\]

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

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

Consider the neighborhood of \((\mu^0, \beta^0)\):
\[
\Theta = \{ \mu \in R^n, \beta \in R^p : \left\| (\mu - \mu^0)^T, (\beta - \beta^0)^T \right\|_\infty \leq \phi_n \}.
\]

A.4
By the result in Theorem 1, there is an event $E_1$ such that on the event $E_1$,
\[ \left\| (\hat{\mu}^{or} - \mu^0, \hat{\beta}^{or} - \beta^0) \right\|_\infty \leq \phi_n, \]
and $P(E_1^C) \leq 2(K + p)n^{-1}$. Hence $(\hat{\mu}^{or}, \hat{\beta}^{or})^T \in \Theta$ on the event $E_1$. For any $\mu \in \mathbb{R}^n$, let $\mu^* = T^{-1}(T^*(\mu))$. We show that $(\hat{\mu}^{or}, \hat{\beta}^{or})$ is a strictly local minimizer of the objective function (3) with probability approaching 1 through the following two steps.

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

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

Therefore, by the results in (i) and (ii), we have $Q_n(\mu, \beta) > Q_n(\hat{\mu}^{or}, \hat{\beta}^{or})$ for any $(\mu^T, \beta^T)^T \in \Theta_n \cap \Theta$ and $(\mu^T, \beta^T)^T \neq (\hat{\mu}^{or}, \hat{\beta}^{or})^T$, so that $(\hat{\mu}^{or}, \hat{\beta}^{or})^T$ is a strict local minimizer of $Q_n(\mu, \beta)$ 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 $P_n^G(T^*(\mu)) = C_n$ for any $\mu \in \Theta$, where $C_n$ is a constant which does not depend on $\mu$. Let $T^*(\mu) = (\alpha_1, \ldots, \alpha_k)^T$. 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 $P_n^G(T^*(\mu))$ is a constant. Since
\[ |\alpha_k - \alpha_{k'}| \geq |\alpha_k^0 - \alpha_{k'}^0| - 2||\alpha - \alpha^0||_\infty, \]
and
\[ ||\alpha - \alpha^0||_\infty = \sup_k \left| \sum_{i\in G_k} \mu_i / |G_k| - \alpha_k^0 \right| = \sup_k \left| \sum_{i\in G_k} (\mu_i - \mu_i^0) / |G_k| \right| \leq \sup_k \sup_{i\in G_k} |\mu_i - \mu_i^0| = ||\mu - \mu^0||_\infty, \]
then for all $k$ and $k'$
\[ |\alpha_k - \alpha_{k'}| \geq |\alpha_k^0 - \alpha_{k'}^0| - 2||\mu - \mu^0||_\infty \geq b_n - 2\phi_n > a\lambda, \]
where the last inequality follows from the assumption that $b_n > a\lambda \gg \phi_n$. Therefore, we have $P_n^G(T^*(\mu)) = C_n$, and hence $Q_n^G(T^*(\mu), \beta) = L_n^G(T^*(\mu), \beta) + C_n$ for all $(\mu^T, \beta^T)^T \in \Theta$. Since $(\hat{\alpha}^{or}, \hat{\beta}^{or})^T$ is the unique global minimizer of $L_n^G(\alpha, \beta)$, then $L_n^G(T^*(\mu), \beta) > L_n^G(\hat{\alpha}^{or}, \hat{\beta}^{or})$ for all $(T^*(\mu)^T, \beta^T)^T \neq (\hat{\alpha}^{or}, \hat{\beta}^{or})^T$ and thus $Q_n^G(T^*(\mu), \beta) > Q_n^G(\hat{\alpha}^{or}, \hat{\beta}^{or})$ for all $T^*(\mu) \neq \hat{\alpha}^{or}$. By (A.4), we have $Q_n^G(\hat{\alpha}^{or}, \hat{\beta}^{or}) = Q_n(\hat{\mu}^{or}, \hat{\beta}^{or})$ and $Q_n^G(T^*(\mu), \beta) = Q_n(T^{-1}(T^*(\mu)), \beta) = Q_n(\mu^*, \beta)$. Therefore, $Q_n(\mu, \beta) > Q_n(\hat{\mu}^{or}, \hat{\beta}^{or})$ for all $\mu^* \neq \hat{\mu}^{or}$, and the result in (i) is proved.

Next we prove the result in (ii). For a positive sequence $t_n$, let $\Theta_n = \{ \mu : ||\mu - \hat{\mu}^{or}|| \leq t_n \}$. For $(\mu^T, \beta^T)^T \in \Theta_n \cap \Theta$, by Taylor’s expansion, we have
\[ Q_n(\mu, \beta) - Q_n(\mu^*, \beta) = \Gamma_1 + \Gamma_2, \]
A.5
where

\[ \Gamma_1 = -(y - (I_n, X)((\mu^m)^T, \beta^T)^T)(\mu - \mu^*), \]
\[ \Gamma_2 = \sum_{i=1}^{n} \frac{\partial P_i(\mu^m)}{\partial \mu_i} (\mu_i - \mu_i^*), \]

in which \( \mu^m = \zeta \mu + (1 - \zeta) \mu^* \) for some \( \zeta \in (0, 1) \). Moreover,

\[ \Gamma_2 = \lambda \sum_{\{j > i\}} \rho(\mu^m_i - \mu^m_j)(\mu_i - \mu_i^*) + \lambda \sum_{\{j < i\}} \rho(\mu^m_i - \mu^m_j)(\mu_i - \mu_i^*) \]
\[ = \lambda \sum_{\{j > i\}} \rho(\mu^m_i - \mu^m_j)(\mu_i - \mu_i^*) + \lambda \sum_{\{i < j\}} \rho(\mu^m_j - \mu^m_i)(\mu_j - \mu_j^*) \]
\[ = \lambda \sum_{\{j > i\}} \rho(\mu^m_i - \mu^m_j)(\mu_i - \mu_i^*) - \lambda \sum_{\{i < j\}} \rho(\mu^m_i - \mu^m_j)(\mu_j - \mu_j^*) \]
\[ = \lambda \sum_{\{j > i\}} \rho(\mu^m_i - \mu^m_j)\{(\mu_i - \mu_i^*) - (\mu_j - \mu_j^*)\}. \quad (A.6) \]

When \( i, j \in G_k, \mu_i^* = \mu_j^* \), and \( \mu^m_i - \mu^m_j \) has the same sign as \( \mu_i - \mu_j \). Hence

\[ \Gamma_2 = \lambda \sum_{k=1}^{K} \sum_{\{i,j \in G_k, i < j\}} \rho'(\|\mu^m_i - \mu^m_j\|)\|\mu_i - \mu_j\|
+ \lambda \sum_{k < k'} \sum_{\{i \in G_k, j \in G_{k'}\}} \rho(\mu^m_i - \mu^m_j)\{(\mu_i - \mu_i^*) - (\mu_j - \mu_j^*)\}. \]

As shown in \( (A.5) \),

\[ ||\mu^* - \mu^0||_{\infty} = ||\alpha - \alpha^0||_{\infty} \leq ||\mu - \mu^0||_{\infty}. \]

Since \( \mu^m = \zeta \mu + (1 - \zeta) \mu^* \),

\[ ||\mu^m - \mu^0||_{\infty} \leq ||\mu - \mu^0||_{\infty} \leq \phi_n, \quad \text{(A.7)} \]

and then for \( k \neq k', i \in G_k, j \in G_{k'} \),

\[ ||\mu^m_i - \mu^m_j|| \geq \min_{i \in G_k, j \in G_{k'}} ||\mu^0_i - \mu^0_j|| - 2||\mu^m - \mu^0||_{\infty} \]
\[ \geq b_n - 2||\mu - \mu^0||_{\infty} \geq b_n - 2\phi_n > a\lambda, \]

and thus \( \rho(\mu^m_i - \mu^m_j) = 0 \). Therefore,

\[ \Gamma_2 = \lambda \sum_{k=1}^{K} \sum_{\{i,j \in G_k, i < j\}} \rho'(\|\mu^m_i - \mu^m_j\|)\|\mu_i - \mu_j\|. \]

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

\[ ||\mu^* - \hat{\mu}^o_r||_{\infty} \leq ||\mu - \hat{\mu}^o_r||_{\infty}. \]

Then

\[ ||\mu^m_i - \mu^m_j|| \leq 2||\mu^m - \mu^*||_{\infty} \leq 2||\mu - \mu^*||_{\infty} \]
\[ \leq 2(||\mu - \hat{\mu}^o_r||_{\infty} + ||\mu^* - \hat{\mu}^o_r||_{\infty}) \]
\[ \leq 4||\mu - \hat{\mu}^o_r||_{\infty} \leq 4t_n. \]
Hence $\rho'(\|\mu_i^m - \mu_j^m\|) \geq \rho'(4t_n)$ by concavity of $\rho(\cdot)$. As a result,

$$\Gamma_2 \geq \lambda \sum_{k=1}^{K} \sum_{\{i,j \in G_k, i < j\}} \rho'(4t_n)|\mu_i - \mu_j|.$$  \hspace{1cm} (A.8)

Let

$$w = (w_1, \ldots, w_n)^T = y - (I_n, X)((\mu^m)^T, \beta^T)^T.$$  

Then

$$\Gamma_1 = -w^T(\mu - \mu^*) = -\sum_{k=1}^{K} \sum_{\{i,j \in G_k\}} \frac{w_i(\mu_i - \mu_j)}{|G_k|} = -\sum_{k=1}^{K} \sum_{\{i,j \in G_k\}} \frac{(w_i - w_j)(\mu_j - \mu_i)}{|G_k|}.$$  \hspace{1cm} (A.9)

Since

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

then

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

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

$$\max_{i,j} |w_j - w_i| \leq 2||\epsilon||_\infty + 2C_2p\phi_n + 2\phi_n.$$  

By Condition (C3),

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

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

$$\max_{i,j} |w_j - w_i| \leq 2\sqrt{2c_1^{-1} \sqrt{\log n}} + 2(C_2p + 1)\phi_n.$$  \hspace{1cm} (A.10)

Hence

$$|G_{\min}|^{-1} \max_{i,j} |w_j - w_i| \leq |G_{\min}|^{-1}\left\{ 2\sqrt{2c_1^{-1} \sqrt{\log n}} + 2(C_2p + 1)\phi_n \right\}.$$  

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

$$\lambda \gg |G_{\min}|^{-1}\max_{i,j} |w_j - w_i|.$$  \hspace{1cm} (A.11)

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

$$Q_n(\mu, \beta) - Q_n(\mu^*, \beta) = \Gamma_1 + \Gamma_2 \geq \sum_{k=1}^{K} \sum_{\{i,j \in G_k, i < j\}} [\lambda \rho'(4t_n) - |G_{\min}|^{-1} \max_{i,j} |w_j - w_i||\mu_i - \mu_j| \geq 0,$$

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 $Z = 1_n$ and $|G_{\text{min}}| = n$, 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

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

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

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

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 $E_1$,

$$\|(\hat{\mu}^{or} - \mu^0)^T, (\hat{\beta}^{or} - \beta^0)^T\|_\infty \leq \phi_n,$$

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

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

(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 $((\hat{\mu}^{or})^T, (\hat{\beta}^{or})^T)^T$, denoted by $\Theta_n$, such that $Q_n(\mu, \beta) \geq Q_n(\mu^*, \beta)$ for any $(\mu^T, \beta^T)^T \in \Theta_n \cap \Theta$ for sufficiently large $n$.

Therefore, by the results in (i) and (ii), we have $Q_n(\mu, \beta) > Q_n(\hat{\mu}^{or}, \hat{\beta}^{or})$ for any $(\mu^T, \beta^T)^T \in \Theta_n \cap \Theta$ and $(\mu^T, \beta^T)^T \neq ((\hat{\mu}^{or})^T, (\hat{\beta}^{or})^T)^T$, so that $((\hat{\mu}^{or})^T, (\hat{\beta}^{or})^T)^T$ is a strict local minimizer of $Q_n(\mu, \beta)$ 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 $((\hat{\mu}^{or})^T, (\hat{\beta}^{or})^T)^T$, we have $\frac{1}{2} \sum_{i=1}^n (y_i^* - \mu^* - x_i^T \beta)^2 > \frac{1}{2} \sum_{i=1}^n (y_i - \hat{\mu}^{or} - x_i^T \hat{\beta}^{or})^2$ for any $(\mu^T, \beta^T)^T \in \Theta$ and $((\mu^*)^T, (\beta^*)^T)^T \neq ((\hat{\mu}^{or})^T, (\hat{\beta}^{or})^T)^T$. Moreover, since
Let $\Theta_n = \{ \mu : ||\mu - \mu^\text{or}|| \leq t_n \}$.

For $(\mu^T, \beta^T)^T \in \Theta_n \cap \Theta$, by Taylor’s expansion, we have

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

where

$$\Gamma_1 = -(y - (I_n, X)((\mu^m)^T, (\beta^m)^T)^T(\mu - \mu^*),$$

$$\Gamma_2 = \sum_{i=1}^n \frac{\partial P_n(\mu^m)}{\partial \mu_i}(\mu_i - \mu^*_i).$$

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

$$\Gamma_2 = \lambda \sum_{i<j} \mathcal{P}(\mu^m_i - \mu^m_j)(\mu_i - \mu^*_i)(\mu_j - \mu^*_j)$$

$$= \lambda \sum_{i<j} \rho(\|\mu^m_i - \mu^m_j\|)||\mu_i - \mu_j||,$$

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

$$||\mu^* - \tilde{\mu}^\text{or}||_\infty = |\alpha - \tilde{\alpha}^\text{or}| \leq ||\mu - \tilde{\mu}^\text{or}||_\infty.$$

Then

$$||\mu^m_i - \mu^m_j|| \leq 2||\mu^m - \mu^*||_\infty \leq 2||\mu - \mu^*||_\infty$$

$$\leq 2(||\mu - \tilde{\mu}^\text{or}||_\infty + ||\mu^* - \tilde{\mu}^\text{or}||_\infty)$$

$$\leq 4||\mu - \tilde{\mu}^\text{or}||_\infty \leq 4t_n.$$

Hence $\rho(\|\mu^m_i - \mu^m_j\|) \geq \rho(4t_n)$ by concavity of $\rho(\cdot)$. As a result,

$$\Gamma_2 \geq \lambda \sum_{i<j} \rho(4t_n)||\mu_i - \mu_j||.$$

(A.12)

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

$$\Gamma_1 = -w^T(\mu - \mu^*) = -n^{-1} \sum_{i<j} (w_j - w_i)(\mu_j - \mu_i),$$

(A.13)
where \( w = (w_1, \ldots, w_n)^T = y - (I_n, X)((\mu^m)^T, \beta^T)^T \). 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 \( E_2 \),
\[
\max_{i,j} |w_j - w_i| \leq 2 \sqrt{2c_1^{-1} \sqrt{\log n} + 2(C_2p + 1)\phi_n}.
\]
Hence
\[
n^{-1} \max_{i,j} |w_j - w_i| \leq n^{-1}\{2 \sqrt{2c_1^{-1} \sqrt{\log n} + 2(C_2p + 1)\phi_n}\}.
\]
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
\[
\lambda \gg n^{-1} \max_{i,j} |w_j - w_i|.
\]
(A.14)

Let \( t_n = o(1) \), then \( \rho'(4t_n) \to 1 \). Therefore, by (A.12), (A.13), and (A.14),
\[
Q_n(\mu, \beta) - Q_n(\mu^*, \beta) = \Gamma_1 + \Gamma_2 \geq \sum_{1 \leq i<j \leq n} [\lambda \rho'(4t_n) - n^{-1} \max_{i,j} |w_j - w_i||\mu_i - \mu_j|] \geq 0,
\]
for sufficiently large \( n \), so that the result in (ii) is proved.

A.5 Estimation procedure for model (2)

We let \( \tilde{x}_i = (1, x_i^T)^T \) and \( \beta^* = (\mu, \beta^T)^T \). The model (2) can be written as \( y_i = z_i^T \theta_i + \tilde{x}_i^T \beta^* + \epsilon_i, i = 1, \ldots, n \). Similar to the assumption for model (1), we assume that observations can be divided into \( K \) different subgroups with \( K < n \). Let \( G = (G_1, \ldots, G_K) \) be a partition of \( \{1, \ldots, n\} \), and we assume \( \theta_i = \alpha_k \) for all \( i \in G_k \), where \( \alpha_k \) is the common value for the \( \theta_i \)'s from group \( G_k \). Then the estimates of \( \theta = (\theta_1^T, \ldots, \theta_n^T)^T \) and \( \beta^* \) can be obtained by minimizing
\[
Q_n(\theta, \beta^*; \lambda) = \frac{1}{2} \sum_{i=1}^n (y_i - z_i^T \theta_i - \tilde{x}_i^T \beta^*)^2 + \sum_{1 \leq i<j \leq n} p(||\theta_i - \theta_j||; \lambda), \tag{A.15}
\]
where \( p(\cdot, \lambda) \) is a concave penalty function with a tuning parameter \( \lambda \), such as MCP or SCAD as described in Section 2. Then for a given \( \lambda > 0 \), define
\[
(\hat{\theta}(\lambda), \hat{\beta}^*(\lambda)) = \arg\min Q_n(\theta, \beta^*; \lambda).
\]
The penalty shrinks some of $||\theta_i - \theta_j||$ to zero. Based on this, we can partition the treatment effects into subgroups. Specifically, let $\hat{\lambda}$ be the value of the tuning parameter selected based on a data-driven procedure such as the BIC. For simplicity, write $(\hat{\theta}, \hat{\beta}^*) \equiv (\hat{\theta}(\lambda), \hat{\beta}^*(\lambda))$. Let $\{\hat{\alpha}_1, \ldots, \hat{\alpha}_{\hat{K}}\}$ be the distinct values of $\hat{\theta}$. Let $\hat{G}_k = \{i : \hat{\theta}_i = \hat{\alpha}_k, 1 \leq i \leq n\}, 1 \leq k \leq \hat{K}$. Then $\{\hat{G}_1, \ldots, \hat{G}_{\hat{K}}\}$ constitutes a partition of $\{1, \ldots, n\}$. Then we apply our proposed ADMM algorithm to obtain the estimates of $\theta$ and $\beta^*$ described as follows.

We reparametrize by introducing a new set of parameters $\delta_{ij} = \theta_i - \theta_j$, and hence minimization of (A.15) is equivalent to the constraint optimization problem:

$$S(\theta, \beta^*, \delta) = \frac{1}{2} \sum_{i=1}^{n} (y_i - z_i^T \theta_i - \tilde{x}_i^T \beta^*)^2 + \sum_{i<j} p_\gamma(||\delta_{ij}||, \lambda),$$

subject to $\theta_i - \theta_j - \delta_{ij} = 0$,

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

$$L(\theta, \beta^*, \delta, \nu) = S(\theta, \beta^*, \delta) + \sum_{i<j} \langle v_{ij}, \theta_i - \theta_j - \delta_{ij} \rangle + \frac{\vartheta}{2} \sum_{i<j} ||\theta_i - \theta_j - \delta_{ij}||^2,$$

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

For given $(\theta, \beta^*, \nu)$, the minimizer of $L(\theta, \beta^*, \delta, \nu)$ with respect to $\delta_{ij}$ is unique and has a closed-form expression for the L1, MCP and SCAD penalties, respectively. Specifically, for given $(\theta, \beta^*, \nu)$, the minimization problem is the same as minimizing

$$\frac{\vartheta}{2} \sum_{i<j} ||\zeta_{ij} - \delta_{ij}||^2 + \sum_{i<j} p_\gamma(||\delta_{ij}||, \lambda),$$

with respect to $\delta_{ij}$, where $\zeta_{ij} = \theta_i - \theta_j + \vartheta^{-1} v_{ij}$. Hence, the closed-form solution for the L1 penalty is

$$\hat{\delta}_{ij} = S(\zeta_{ij}, \lambda/\vartheta),$$

(A.16)

where $S(z, t) = (1 - t/||z||)_+ z$ 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

$$\hat{\delta}_{ij} = \begin{cases} \frac{S(\zeta_{ij}; \lambda/\vartheta)}{1 - 1/(\gamma \vartheta)} \zeta_{ij} & \text{if } ||\zeta_{ij}|| \leq \gamma \lambda \\ \zeta_{ij} & \text{if } ||\zeta_{ij}|| > \gamma \lambda. \end{cases}$$

(A.17)

A.11
For the SCAD penalty with $\gamma > 1/\vartheta + 1$, it is

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

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

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

$$
L(\theta, \beta^*, \delta, \upsilon) = \frac{1}{2} \sum_{i=1}^{n} (y_i - z_i^T \theta_i - \tilde{X}_i^T \beta_i^*)^2 + \frac{\vartheta}{2} \sum_{i<j} ||\theta_i - \theta_j - \delta_{ij} + \vartheta^{-1} \upsilon_{ij}||^2 + C
$$

$$
= \frac{1}{2} ||Z \theta + \tilde{X} \beta^* - y||^2 + \frac{\vartheta}{2} ||A \beta - \delta + \vartheta^{-1} \upsilon||^2 + C.
$$

Here $C$ is a constant independent of $\theta$ and $\beta^*$, $y = (y_1, \ldots, y_n)^T$, $Z = \text{diag}(z_1^T, \ldots, z_n^T)$ and $\tilde{X} = (\tilde{x}_1, \ldots, \tilde{x}_n)^T$. Moreover, $e_i$ is the $n \times 1$ vector whose $i^{th}$ element is 1 and the remaining ones are 0, $\Delta = \{(e_i - e_j), i < j\}^T$ and $A = \Delta \otimes I_p$, where $I_d$ denotes the $d \times d$ identity matrix and $\otimes$ denotes the Kronecker product.

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

$$
\theta^{(m+1)} = (Z^T (I_n - Q_{\tilde{X}}) Z + \vartheta A^T A)^{-1} [Z^T (I_n - Q_{\tilde{X}}) y + \vartheta A^T (\delta^{(m)} - \vartheta^{-1} \upsilon^{(m)})],
$$

$$
\beta^{*(m+1)} = (\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T (y - Z \theta^{(m+1)}),
$$

where $Q_{\tilde{X}} = \tilde{X} (\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T$.

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

Finally, the estimate of $\upsilon_{ij}$ is updated as

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

A.12
We iteratively update the estimates of $\theta, \beta^*, \delta$ and $v$ until the stopping rule is met. We track the progress of the ADMM based on the primal residual $r^{(m+1)} = A\theta^{(m+1)} - \delta^{(m+1)}$. We stop the algorithm when $r^{(m+1)}$ is close to zero such that $\|r^{(m+1)}\| < \epsilon$ for some small value $\epsilon$.

References