Supplementary material for: Mutual Transfer Learning for Massive Data

Ching-Wei Cheng Xingye Qiao Guang Cheng

The supplementary material is organized as follows:

- In Section S.1, we give the formulas for the penalty functions we have used.
- In Section S.2, we prove Theorem 2.1.
- In Section S.3, we prove Theorem 4.1.
- In Section S.4, we derive the bounds related to random matrices, which are needed in Section S.3.
- In Section S.5, we present the detailed derivation of the ADMM Algorithm 1.
- In Section S.6, we prove Proposition 3.1.
- In Section S.7, we prove Theorem 4.2.
- In Section S.8, we prove (ii) in Corollary 4.1.
- In Section S.9, we present technical lemmas.
- In Section S.10, we define the signal-to-noise ratio (SNR) for the proposed method.
- In Section S.11, we present additional simulation and real data results.

S.1. Penalty Functions

More specifically, for $t > 0$, the L_1, MCP, SCAD and TLP are respectively given by

$$p^{L_1}(t;\lambda) = \lambda t, \tag{S.1}$$

$$p_\gamma^{\mathrm{MCP}}(t;\lambda) = \lambda \int_0^t \left(1 - \frac{x}{\lambda\gamma}\right)_+ \mathrm{d}x,\ \gamma > 1, \tag{S.2}$$

$$p_\gamma^{\mathrm{SCAD}}(t;\lambda) = \lambda \int_0^t \min\left\{1, \frac{(\gamma - \frac{x}{\lambda})_+}{\gamma - 1}\right\} \mathrm{d}x,\ \gamma > 2, \tag{S.3}$$

and

$$p_\gamma^{\mathrm{TLP}}(t;\lambda) = \lambda \min\left(t, \lambda\gamma\right),\ \gamma > 0. \tag{S.4}$$

S.2. Proof of Theorem 2.1

To prove $Q_N^{\mathrm{CD}}(\boldsymbol{\beta}, \boldsymbol{\Theta}) - Q_N(\boldsymbol{\beta}, \boldsymbol{\Theta})$ is a constant, first note that

$$Q_N^{\mathrm{CD}}(\boldsymbol{\beta}, \boldsymbol{\Theta}) - Q_N(\boldsymbol{\beta}, \boldsymbol{\alpha}) = L_N^{\mathrm{CD}}(\boldsymbol{\beta}, \boldsymbol{\alpha}) - L_N(\boldsymbol{\beta}, \boldsymbol{\Theta}),$$

where

$$L_N^{\mathrm{CD}}(\boldsymbol{\beta}, \boldsymbol{\Theta}) = \frac{1}{2N}\sum_{i=1}^{M} \begin{pmatrix} \breve{\boldsymbol{\beta}}_i - \boldsymbol{\beta} \\ \breve{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i \end{pmatrix}^\top (\boldsymbol{x}_i, \boldsymbol{z}_i)^\top \boldsymbol{W}_i (\boldsymbol{x}_i, \boldsymbol{z}_i) \begin{pmatrix} \breve{\boldsymbol{\beta}}_i - \boldsymbol{\beta} \\ \breve{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i \end{pmatrix}, \tag{S.5}$$

$$L_N(\boldsymbol{\beta}, \boldsymbol{\Theta}) = \frac{1}{2N}\sum_{i=1}^{M} (\boldsymbol{y}_i - \boldsymbol{x}_i\boldsymbol{\beta} - \boldsymbol{z}_i\boldsymbol{\theta}_i)^\top \boldsymbol{W}_i (\boldsymbol{y}_i - \boldsymbol{x}_i\boldsymbol{\beta} - \boldsymbol{z}_i\boldsymbol{\theta}_i). \tag{S.6}$$

Define $\boldsymbol{r}_i(\boldsymbol{\beta},\boldsymbol{\theta}_i)=(\boldsymbol{x}_i,\boldsymbol{z}_i)\begin{pmatrix}\boldsymbol{\beta}-\boldsymbol{\beta}_0\\ \boldsymbol{\theta}_i-\boldsymbol{\theta}_{i0}\end{pmatrix}$, where $\boldsymbol{\beta}_0$ and $\boldsymbol{\theta}_{i0}$ are true values of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}_i$, respectively. Then we can write

$$\begin{aligned}(\boldsymbol{x}_i,\boldsymbol{z}_i)\begin{pmatrix}\breve{\boldsymbol{\beta}}_i-\boldsymbol{\beta}\\ \breve{\boldsymbol{\theta}}_i-\boldsymbol{\theta}_i\end{pmatrix}&=(\boldsymbol{x}_i,\boldsymbol{z}_i)\begin{pmatrix}\breve{\boldsymbol{\beta}}_i-\boldsymbol{\beta}_0\\ \breve{\boldsymbol{\theta}}_i-\boldsymbol{\theta}_{i0}\end{pmatrix}-(\boldsymbol{x}_i,\boldsymbol{z}_i)\begin{pmatrix}\boldsymbol{\beta}-\boldsymbol{\beta}_0\\ \boldsymbol{\theta}_i-\boldsymbol{\theta}_{i0}\end{pmatrix}\\&=(\boldsymbol{x}_i,\boldsymbol{z}_i)\left[(\boldsymbol{x}_i,\boldsymbol{z}_i)^\top\boldsymbol{W}_i(\boldsymbol{x}_i,\boldsymbol{z}_i)\right]^{-1}(\boldsymbol{x}_i,\boldsymbol{z}_i)^\top\boldsymbol{W}_i(\boldsymbol{z}_i\boldsymbol{u}_i+\boldsymbol{\varepsilon}_i)-\boldsymbol{r}_i(\boldsymbol{\beta},\boldsymbol{\theta}_i)\\&=\boldsymbol{V}_i^{-1}\boldsymbol{P}_{\boldsymbol{V}_i(\boldsymbol{x}_i,\boldsymbol{z}_i)}\boldsymbol{V}_i(\boldsymbol{z}_i\boldsymbol{u}_i+\boldsymbol{\varepsilon}_i)-\boldsymbol{r}_i(\boldsymbol{\beta},\boldsymbol{\theta}_i)\end{aligned}$$

where $\boldsymbol{V}_i$ denote the squared root matrix of $\boldsymbol{W}_i$ such that $\boldsymbol{W}_i=\boldsymbol{V}_i^2$. In addition, we can also write

$$\boldsymbol{y}_i-\boldsymbol{x}_i\boldsymbol{\beta}-\boldsymbol{z}_i\boldsymbol{\theta}_i=\boldsymbol{y}_i-(\boldsymbol{x}_i,\boldsymbol{z}_i)\begin{pmatrix}\boldsymbol{\beta}_0\\ \boldsymbol{\theta}_{i0}\end{pmatrix}-(\boldsymbol{x}_i,\boldsymbol{z}_i)\begin{pmatrix}\boldsymbol{\beta}-\boldsymbol{\beta}_0\\ \boldsymbol{\theta}_i-\boldsymbol{\theta}_{i0}\end{pmatrix}=\boldsymbol{z}_i\boldsymbol{u}_i+\boldsymbol{\varepsilon}_i-\boldsymbol{r}_i(\boldsymbol{\beta},\boldsymbol{\theta}_i).$$

Consequently, we have

$$\begin{aligned}(\boldsymbol{x}_i,\boldsymbol{z}_i)\begin{pmatrix}\breve{\boldsymbol{\beta}}_i-\boldsymbol{\beta}\\ \breve{\boldsymbol{\theta}}_i-\boldsymbol{\theta}_i\end{pmatrix}-(\boldsymbol{y}_i-\boldsymbol{x}_i\boldsymbol{\beta}-\boldsymbol{z}_i\boldsymbol{\theta}_i)&=\left(\boldsymbol{V}_i^{-1}\boldsymbol{P}_{\boldsymbol{V}_i(\boldsymbol{x}_i,\boldsymbol{z}_i)}\boldsymbol{V}_i-\boldsymbol{I}_{n_i}\right)(\boldsymbol{z}_i\boldsymbol{u}_i+\boldsymbol{\varepsilon}_i)\\&=-\boldsymbol{V}_i^{-1}\boldsymbol{P}^{\perp}_{\boldsymbol{V}_i(\boldsymbol{x}_i,\boldsymbol{z}_i)}\boldsymbol{V}_i(\boldsymbol{z}_i\boldsymbol{u}_i+\boldsymbol{\varepsilon}_i).\end{aligned}$$

Note that $\boldsymbol{P}^{\perp}_{\boldsymbol{V}_i(\boldsymbol{x}_i,\boldsymbol{z}_i)}$ is the projection matrix onto the orthogonal complement of the column space of $\boldsymbol{V}_i(\boldsymbol{x}_i,\boldsymbol{z}_i)$ such that $\boldsymbol{P}^{\perp}_{\boldsymbol{V}_i(\boldsymbol{x}_i,\boldsymbol{z}_i)}\boldsymbol{V}_i\boldsymbol{x}_i=\boldsymbol{0}$ and $\boldsymbol{P}^{\perp}_{\boldsymbol{V}_i(\boldsymbol{x}_i,\boldsymbol{z}_i)}\boldsymbol{V}_i\boldsymbol{z}_i=\boldsymbol{0}$. Hence we have

$$(\boldsymbol{x}_i,\boldsymbol{z}_i)\begin{pmatrix}\breve{\boldsymbol{\beta}}_i-\boldsymbol{\beta}\\ \breve{\boldsymbol{\theta}}_i-\boldsymbol{\theta}_i\end{pmatrix}-(\boldsymbol{y}_i-\boldsymbol{x}_i\boldsymbol{\beta}-\boldsymbol{z}_i\boldsymbol{\theta}_i)=-\boldsymbol{V}_i^{-1}\boldsymbol{P}^{\perp}_{\boldsymbol{V}_i(\boldsymbol{x}_i,\boldsymbol{z}_i)}\boldsymbol{V}_i\boldsymbol{\varepsilon}_i.$$

and therefore we can write

$$\begin{aligned}L_N^{\mathrm{CD}}(\boldsymbol{\beta},\boldsymbol{\Theta})&=\frac{1}{2N}\sum_{i=1}^{M}(\boldsymbol{y}_i-\boldsymbol{x}_i\boldsymbol{\beta}-\boldsymbol{z}_i\boldsymbol{\theta}_i-\boldsymbol{V}_i^{-1}\boldsymbol{P}^{\perp}_{\boldsymbol{V}_i(\boldsymbol{x}_i,\boldsymbol{z}_i)}\boldsymbol{V}_i\boldsymbol{\varepsilon}_i)^\top\boldsymbol{W}_i(\boldsymbol{y}_i-\boldsymbol{x}_i\boldsymbol{\beta}-\boldsymbol{z}_i\boldsymbol{\theta}_i-\boldsymbol{V}_i^{-1}\boldsymbol{P}^{\perp}_{\boldsymbol{V}_i(\boldsymbol{x}_i,\boldsymbol{z}_i)}\boldsymbol{V}_i\boldsymbol{\varepsilon}_i)\\&=L_N(\boldsymbol{\beta},\boldsymbol{\Theta})-\frac{1}{N}\sum_{i=1}^{M}(\boldsymbol{y}_i-\boldsymbol{x}_i\boldsymbol{\beta}-\boldsymbol{z}_i\boldsymbol{\theta}_i)^\top\boldsymbol{V}_i\boldsymbol{P}^{\perp}_{\boldsymbol{V}_i(\boldsymbol{x}_i,\boldsymbol{z}_i)}\boldsymbol{V}_i\boldsymbol{\varepsilon}_i+\frac{1}{2N}\sum_{i=1}^{M}\boldsymbol{\varepsilon}_i^\top\boldsymbol{V}_i\boldsymbol{P}^{\perp}_{\boldsymbol{V}_i(\boldsymbol{x}_i,\boldsymbol{z}_i)}\boldsymbol{V}_i\boldsymbol{\varepsilon}_i\\&=L_N(\boldsymbol{\beta},\boldsymbol{\Theta})-\frac{1}{N}\sum_{i=1}^{M}\boldsymbol{y}_i^\top\boldsymbol{V}_i\boldsymbol{P}^{\perp}_{\boldsymbol{V}_i(\boldsymbol{x}_i,\boldsymbol{z}_i)}\boldsymbol{V}_i\boldsymbol{\varepsilon}_i+\frac{1}{2N}\sum_{i=1}^{M}\boldsymbol{\varepsilon}_i^\top\boldsymbol{V}_i\boldsymbol{P}^{\perp}_{\boldsymbol{V}_i(\boldsymbol{x}_i,\boldsymbol{z}_i)}\boldsymbol{V}_i\boldsymbol{\varepsilon}_i,\end{aligned}$$

where the last identity follows from $\boldsymbol{x}_i^\top\boldsymbol{V}_i\boldsymbol{P}^{\perp}_{\boldsymbol{V}_i(\boldsymbol{x}_i,\boldsymbol{z}_i)}=\boldsymbol{0}$ and $\boldsymbol{z}_i^\top\boldsymbol{V}_i\boldsymbol{P}^{\perp}_{\boldsymbol{V}_i(\boldsymbol{x}_i,\boldsymbol{z}_i)}=\boldsymbol{0}$. Hence we conclude $Q_N^{\mathrm{CD}}(\boldsymbol{\beta},\boldsymbol{\Theta})-Q_N(\boldsymbol{\beta},\boldsymbol{\alpha})=L_N^{\mathrm{CD}}(\boldsymbol{\beta},\boldsymbol{\alpha})-L_N(\boldsymbol{\beta},\boldsymbol{\Theta})$ is a constant.

S.3. Proof of Theorem 4.1

To prove (i), we write

$$\begin{pmatrix}\widehat{\boldsymbol{\beta}}_{\mathrm{OR}}-\boldsymbol{\beta}_0\\ \widehat{\boldsymbol{\alpha}}_{\mathrm{OR}}-\boldsymbol{\alpha}_0\end{pmatrix}=\left(\boldsymbol{F}^\top\boldsymbol{W}\boldsymbol{F}\right)^{-1}\boldsymbol{F}^\top\boldsymbol{W}\left(\boldsymbol{Z}\boldsymbol{U}+\boldsymbol{\mathcal{E}}\right),$$

where $\boldsymbol{F}=(\boldsymbol{X},\boldsymbol{Z}\boldsymbol{A})$, and hence

$$\begin{aligned}\left\|\begin{pmatrix}\widehat{\boldsymbol{\beta}}_{\mathrm{OR}}-\boldsymbol{\beta}_0\\ \widehat{\boldsymbol{\alpha}}_{\mathrm{OR}}-\boldsymbol{\alpha}_0\end{pmatrix}\right\|&\le\left\|\left(\boldsymbol{F}^\top\boldsymbol{W}\boldsymbol{F}\right)^{-1}\right\|\left\|\boldsymbol{F}^\top\boldsymbol{W}(\boldsymbol{Z}\boldsymbol{U}+\boldsymbol{\mathcal{E}})\right\|\\&\le\sqrt{p+Sq}\left\|\left(\boldsymbol{F}^\top\boldsymbol{W}\boldsymbol{F}\right)^{-1}\right\|\left\|\boldsymbol{F}^\top\boldsymbol{W}(\boldsymbol{Z}\boldsymbol{U}+\boldsymbol{\mathcal{E}})\right\|_\infty.\end{aligned}$$

We derive the upper bounds for $\left\|\left(\boldsymbol{F}^{\top}\boldsymbol{W}\boldsymbol{F}\right)^{-1}\right\|$ and $\left\|\boldsymbol{F}^{\top}\boldsymbol{W}(\boldsymbol{Z}\boldsymbol{U}+\boldsymbol{\mathcal{E}})\right\|_{\infty}$ under the events

$$E_{CB}=\Big\{\max_{1\le i\le p}\left\|[\boldsymbol{X}]_{\cdot j}\right\|\le\sqrt{5c_fN}\text{ and }\max_{1\le s\le S}\max_{1\le k\le q}\left\|[\widetilde{\mathbb{Z}}_s]_{\cdot k}\right\|\le\sqrt{5c_fg_{\max}}\Big\}$$

and

$$E_{EB}=\Big\{\underline{C}_f\lesssim\lambda_{\min}\big(\boldsymbol{f}_i^{\top}\boldsymbol{f}_i/n_i\big)\le\lambda_{\max}\big(\boldsymbol{f}_i^{\top}\boldsymbol{f}_i/n_i\big)\lesssim\overline{C}_f,i=1,\ldots,M\Big\},$$

where $\boldsymbol{f}_i=(\boldsymbol{x}_i,\boldsymbol{z}_i)$ and $n_{\min}=\min_{1\le i\le M}n_i$. According to Sections S.4.1 and S.4.2, we have $P\big(E_{CB}\big)\ge1-(p+Sq)e^{-g_{\max}}$ and $P\big(E_{EB}\big)\ge1-2Me^{-n_{\min}}$.

We first look at $\left\|(\boldsymbol{F}^{\top}\boldsymbol{W}\boldsymbol{F})^{-1}\right\|$. By definition, we can write

$$\boldsymbol{W}=\left[\sigma_{\varepsilon}^2\boldsymbol{I}_N+\boldsymbol{Z}(\boldsymbol{I}_M\otimes\boldsymbol{\Psi})\boldsymbol{Z}^{\top}\right]^{-1}=\sigma_{\varepsilon}^{-2}\Big[\boldsymbol{I}_N-\sigma_{\varepsilon}^{-2}\boldsymbol{Z}\big(\boldsymbol{I}_M\otimes\boldsymbol{\Psi}^{-1}+\sigma_{\varepsilon}^{-2}\boldsymbol{Z}^{\top}\boldsymbol{Z}\big)^{-1}\boldsymbol{Z}^{\top}\Big],$$

where the second identity holds by applying the Woodbury identity (S.41). It then follows that

$$\begin{aligned}\left\|\left(\boldsymbol{F}^{\top}\boldsymbol{W}\boldsymbol{F}\right)^{-1}\right\|&=\lambda_{\min}^{-1}\left(\boldsymbol{F}^{\top}\boldsymbol{W}\boldsymbol{F}\right)\\&=\sigma_{\varepsilon}^2\lambda_{\min}^{-1}\left[\boldsymbol{F}^{\top}\boldsymbol{F}-\sigma_{\varepsilon}^{-2}\boldsymbol{F}^{\top}\boldsymbol{Z}\big(\boldsymbol{I}_M\otimes\boldsymbol{\Psi}^{-1}+\sigma_{\varepsilon}^{-2}\boldsymbol{Z}^{\top}\boldsymbol{Z}\big)^{-1}\boldsymbol{Z}^{\top}\boldsymbol{F}\right]\\&\le\sigma_{\varepsilon}^2\left\{\lambda_{\min}\left(\boldsymbol{F}^{\top}\boldsymbol{F}\right)+\sigma_{\varepsilon}^2\lambda_{\min}\left[\boldsymbol{F}^{\top}\boldsymbol{Z}\big(\boldsymbol{I}_M\otimes\boldsymbol{\Psi}^{-1}+\sigma_{\varepsilon}^{-2}\boldsymbol{Z}^{\top}\boldsymbol{Z}\big)^{-1}\boldsymbol{Z}^{\top}\boldsymbol{F}\right]\right\}^{-1}\\&\le\sigma_{\varepsilon}^2\lambda_{\min}^{-1}\left(\boldsymbol{F}^{\top}\boldsymbol{F}\right)\\&\lesssim\underline{C}_f^{-1}\sigma_{\varepsilon}^2g_{\min}^{-1},\end{aligned}\tag{S.7}$$

where the last inequality holds from (S.14).

It remains to show the upper bound for $\left\|\boldsymbol{F}^{\top}\boldsymbol{W}\left(\boldsymbol{Z}\boldsymbol{U}+\boldsymbol{\mathcal{E}}\right)\right\|$. Recall that $\boldsymbol{F}=(\boldsymbol{X},\boldsymbol{Z}\boldsymbol{A})$. Let $\boldsymbol{V}_i$ denote the square root matrix of $\boldsymbol{W}_i$ such that $\boldsymbol{W}_i=\boldsymbol{V}_i^2$, and then define $\boldsymbol{V}=\operatorname{diag}\left[(\boldsymbol{V}_i)_{i=1,\ldots,M}\right]$ and $\mathbb{V}_s=\operatorname{diag}\left[(\boldsymbol{V}_i)_{i:L_i=s}\right]$. It is easy to see from (S.15) that $\|\boldsymbol{V}\|=\|\boldsymbol{W}\|^{1/2}=\max_{1\le i\le M}\|\boldsymbol{W}_i\|^{1/2}\le\sigma_{\varepsilon}^{-1}$ and $\|\mathbb{V}_s\|=\max_{i:L_i=s}\|\boldsymbol{W}_i\|^{1/2}\le\sigma_{\varepsilon}^{-1}$. Let $\widetilde{\mathbb{Z}}_s=(\boldsymbol{z}_i^{\top})_{i:L_i=s}^{\top}$ denote the stacked design matrix for the s-th subgroup. Under the events E_{CB} and E_{EB}, we have

$$\|\boldsymbol{V}[\boldsymbol{X}]_{\cdot j}\|\le\|\boldsymbol{V}\|\|[\boldsymbol{X}]_{\cdot j}\|\le\sigma_{\varepsilon}^{-1}\sqrt{5c_fN}\quad\text{and}\quad\|\mathbb{V}_s[\widetilde{\mathbb{Z}}_s]_{\cdot k}\|\le\|\mathbb{V}_s\|\|[\widetilde{\mathbb{Z}}_s]_{\cdot k}\|\le\sigma_{\varepsilon}^{-1}\sqrt{5c_fg_{\max}}$$

for all $j=1,\ldots,p,s=1,\ldots,S,k=1,\ldots,q$. By union bound and the tail bounds in (S.18) and (S.20), it holds

$$\begin{aligned}P\Big(\left\|\boldsymbol{X}^{\top}\boldsymbol{W}(\boldsymbol{Z}\boldsymbol{U}+\boldsymbol{\mathcal{E}})\right\|_{\infty}>t\Big)&\le\sum_{j=1}^{p}P\Big(\left|[\boldsymbol{X}]_{\cdot j}^{\top}\boldsymbol{W}(\boldsymbol{Z}\boldsymbol{U}+\boldsymbol{\mathcal{E}})\right|>t\Big)\\&\le4p\exp\left(-\frac{c_e\left(\tau\wedge\sigma_{\varepsilon}^2\right)\sigma_{\varepsilon}^2t^2}{5c_fN}\right),\end{aligned}\tag{S.8}$$

$$\begin{aligned}P\Big(\left\|(\boldsymbol{Z}\boldsymbol{A})^{\top}\boldsymbol{W}(\boldsymbol{Z}\boldsymbol{U}+\boldsymbol{\mathcal{E}})\right\|_{\infty}>t\Big)&\le\sum_{s=1}^{S}\sum_{k=1}^{q}P\Big(\left|[\widetilde{\mathbb{Z}}_s]_{\cdot k}^{\top}\mathbb{W}_s(\mathbb{Z}_s\mathbb{U}_s+\boldsymbol{\mathcal{E}}_s)\right|>t\Big)\\&\le4Sq\exp\left(-\frac{c_e\left(\tau\wedge\sigma_{\varepsilon}^2\right)\sigma_{\varepsilon}^2t^2}{5c_fg_{\max}}\right),\end{aligned}\tag{S.9}$$

where $\mathbb{Z}_s=\operatorname{diag}\left[(\boldsymbol{z}_i)_{i:L_i=s}\right]$, $\mathbb{U}_s=(\boldsymbol{u}_i^{\top})_{i:L_i=s}^{\top}$, $\mathbb{W}_s=\operatorname{diag}\left[(\boldsymbol{W}_i)_{i:L_i=s}\right]$, $\boldsymbol{\mathcal{E}}_s=(\boldsymbol{\varepsilon}_i^{\top})_{i:L_i=s}^{\top}$, and $\tau=\lambda_{\min}(\boldsymbol{\Psi})$. Since $\boldsymbol{F}=(\boldsymbol{X},\boldsymbol{Z}\boldsymbol{A})$, it follows that

$$\begin{aligned}P\Big(\left\|\boldsymbol{F}^{\top}\boldsymbol{W}(\boldsymbol{Z}\boldsymbol{U}+\boldsymbol{\mathcal{E}})\right\|_{\infty}>t\Big)&\le P\Big(\left\|\boldsymbol{X}^{\top}\boldsymbol{W}(\boldsymbol{Z}\boldsymbol{U}+\boldsymbol{\mathcal{E}})\right\|_{\infty}>t\Big)+P\Big(\left\|(\boldsymbol{Z}\boldsymbol{A})^{\top}\boldsymbol{W}(\boldsymbol{Z}\boldsymbol{U}+\boldsymbol{\mathcal{E}})\right\|_{\infty}>t\Big)\\&\le2p\exp\left(-\frac{c_e\left(\tau\wedge\sigma_{\varepsilon}^2\right)\sigma_{\varepsilon}^2t^2}{5c_fN}\right)+2Sq\exp\left(-\frac{c_e\left(\tau\wedge\sigma_{\varepsilon}^2\right)\sigma_{\varepsilon}^2t^2}{5c_fg_{\max}}\right)\\&\le4(p+Sq)\exp\left(-\frac{c_e\left(\tau\wedge\sigma_{\varepsilon}^2\right)\sigma_{\varepsilon}^2t^2}{5c_fN}\right).\end{aligned}$$

Taking $t = \sqrt{5}c_e^{-1/2}c_f^{1/2}\left(\tau \wedge \sigma_\varepsilon^2\right)^{-1/2}\sigma_\varepsilon^{-1}\sqrt{N\log N}$ hence yields

$$\left\|\boldsymbol{F}^\top \boldsymbol{W}(\boldsymbol{Z}\boldsymbol{U}+\boldsymbol{\mathcal{E}})\right\|_\infty \le \sqrt{5}c_e^{-1/2}c_f^{1/2}\left(\tau \wedge \sigma_\varepsilon^2\right)^{-1}\sigma_\varepsilon^{-1/2}\sqrt{N\log N} \tag{S.10}$$

with probability at least $1-(p+Sq)(4N^{-1}+e^{-g_{\max}})-2Me^{-n_{\min}}$.

In summary, combining the results in (S.7) and (S.10) leads to

$$\begin{aligned}\left\|\begin{pmatrix}\widehat{\boldsymbol{\beta}}_{\mathrm{OR}}-\boldsymbol{\beta}_0\\ \widehat{\boldsymbol{\alpha}}_{\mathrm{OR}}-\boldsymbol{\alpha}_0\end{pmatrix}\right\| &\le \sqrt{p+Sq}\left\|\left(\boldsymbol{F}^\top\boldsymbol{W}\boldsymbol{F}\right)^{-1}\right\|\left\|\boldsymbol{F}^\top\boldsymbol{W}(\boldsymbol{Z}\boldsymbol{U}+\boldsymbol{\mathcal{E}})\right\|_\infty\\ &\le \sqrt{5}c_e^{-1/2}c_f^{1/2}\underline{C}_f^{-1}\left(\tau\wedge\sigma_\varepsilon^2\right)^{-1/2}\sigma_\varepsilon g_{\min}^{-1}\sqrt{(p+Sq)N\log N},\end{aligned}$$

with probability at least $1-(p+Sq)(4N^{-1}+e^{-g_{\max}})-2Me^{-n_{\min}}$.

To prove (ii), we show the oracle estimator satisfies the Lindeberg-Feller condition as follows. Note that for any $\boldsymbol{a}_N \in \mathbb{R}^{p+Sq}$ with $\|\boldsymbol{a}_N\|=1$,

$$\boldsymbol{a}_N^\top\begin{pmatrix}\widehat{\boldsymbol{\beta}}_{\mathrm{OR}}-\boldsymbol{\beta}_0\\ \widehat{\boldsymbol{\alpha}}_{\mathrm{OR}}-\boldsymbol{\alpha}_0\end{pmatrix} = \boldsymbol{a}_N^\top(\boldsymbol{F}^\top\boldsymbol{W}\boldsymbol{F})^{-1}\boldsymbol{F}^\top\boldsymbol{W}(\boldsymbol{Z}\boldsymbol{U}+\boldsymbol{\mathcal{E}}).$$

It is then clear that

$$\mathsf{E}\left[\boldsymbol{a}_N^\top\begin{pmatrix}\widehat{\boldsymbol{\beta}}_{\mathrm{OR}}-\boldsymbol{\beta}_0\\ \widehat{\boldsymbol{\alpha}}_{\mathrm{OR}}-\boldsymbol{\alpha}_0\end{pmatrix}\right] = \boldsymbol{0}$$

and

$$\begin{aligned}\sigma_N^2(\boldsymbol{a}_N) &= \mathsf{Var}\left[\boldsymbol{a}_N^\top\begin{pmatrix}\widehat{\boldsymbol{\beta}}_{\mathrm{OR}}-\boldsymbol{\beta}_0\\ \widehat{\boldsymbol{\alpha}}_{\mathrm{OR}}-\boldsymbol{\alpha}_0\end{pmatrix}\right] = \boldsymbol{a}_N^\top(\boldsymbol{F}^\top\boldsymbol{W}\boldsymbol{F})^{-1}\boldsymbol{a}_N\\ &\ge \lambda_{\min}\left[(\boldsymbol{F}^\top\boldsymbol{W}\boldsymbol{F})^{-1}\right] = \left\|\boldsymbol{F}^\top\boldsymbol{W}\boldsymbol{F}\right\|^{-1} \ge \|\boldsymbol{F}\|^{-2}\|\boldsymbol{W}\|^{-1} \gtrsim \overline{C}_f\sigma_\varepsilon^2 N^{-1},\end{aligned}$$

where the last inequality is due to (S.13) and (S.16). Let $\xi_{il} = [\boldsymbol{V}_i(\boldsymbol{z}_i\boldsymbol{u}_i+\boldsymbol{\varepsilon}_i)]_l$ denote the l-th coordinate of $\boldsymbol{V}_i(\boldsymbol{z}_i\boldsymbol{u}_i+\boldsymbol{\varepsilon}_i)$, and $\boldsymbol{F}_i$ the $n_i\times(p+Sq)$ submatrix of $\boldsymbol{F}$ that collects the corresponding rows of the i-th unit. Then for any $\epsilon>0$, we can write

$$\begin{aligned}&\sigma_N^{-2}(\boldsymbol{a}_N)\sum_{i=1}^M\sum_{l=1}^{n_i}\mathsf{E}\left\{\left[\boldsymbol{a}_N^\top(\boldsymbol{F}^\top\boldsymbol{W}\boldsymbol{F})^{-1}[\boldsymbol{V}_i\boldsymbol{F}_i]_{l\cdot}^\top\xi_{il}\right]^2\mathbb{1}_{\left|\boldsymbol{a}_N^\top(\boldsymbol{F}^\top\boldsymbol{W}\boldsymbol{F})^{-1}[\boldsymbol{V}_i\boldsymbol{F}_i]_{l\cdot}^\top\xi_{il}\right|>\epsilon\sigma_N(\boldsymbol{a}_N)}\right\}\\ &\le \overline{C}_f^{-1}\sigma_\varepsilon^{-2}N\sum_{i=1}^M\sum_{l=1}^{n_i}\left\{\mathsf{E}\left[\boldsymbol{a}_N^\top(\boldsymbol{F}^\top\boldsymbol{W}\boldsymbol{F})^{-1}[\boldsymbol{V}_i\boldsymbol{F}_i]_{l\cdot}^\top\xi_{il}\right]^4\right\}^{1/2}\\ &\qquad\times\left\{P\Big(\left|\boldsymbol{a}_N^\top(\boldsymbol{F}^\top\boldsymbol{W}\boldsymbol{F})^{-1}[\boldsymbol{V}_i\boldsymbol{F}_i]_{l\cdot}^\top\xi_{il}\right|>\epsilon\sigma_N(\boldsymbol{a}_N)\Big)\right\}^{1/2}.\end{aligned}$$

We are going to show the above quantity is $o(1)$.

From (S.19) we can see that ξ_{il} has sub-Gaussian tails, and thus there exists an a constant $c_3>0$ such that $\mathsf{E}\left[\|\xi_{il}\|^4\right]\le c_3$ and $\mathsf{E}\left[\|\xi_{il}\|^2\right]\le c_3$. By the definition of $\boldsymbol{F}_i$ and Lemma S.9.2, we have $\left\|[\boldsymbol{F}_i]_{l\cdot}\right\|^2 \le 5c_4(p+Sq)$ with probability at least $1-e^{-(p+Sq)}$ for some constant $c_4>0$. It then follows that

$$\begin{aligned}&\left\{\mathsf{E}\left[\boldsymbol{a}_N^\top(\boldsymbol{F}^\top\boldsymbol{W}\boldsymbol{F})^{-1}[\boldsymbol{V}_i\boldsymbol{F}_i]_{l\cdot}^\top\xi_{il}\right]^4\right\}^{1/2}\\ &\quad\le \|\boldsymbol{a}_N\|^2\left\|(\boldsymbol{F}^\top\boldsymbol{W}\boldsymbol{F})^{-1}\right\|^2\|\boldsymbol{V}_i\|^2\left\|[\boldsymbol{F}_i]_{l\cdot}\right\|^2\left\{\mathsf{E}\left[\|\xi_{il}\|^4\right]\right\}^{1/2}\\ &\quad\le 5c_3^{1/2}c_4\underline{C}_f^{-2}\overline{C}_f\sigma_\varepsilon^2(p+Sq)g_{\min}^{-2}\\ &\quad= O\left[(p+Sq)g_{\min}^{-2}\right],\end{aligned}$$

where the second inequality holds based on (S.7) and (S.15) under the event E_{EB}. In addition, similar argument gives

$$\begin{aligned}\mathsf{E}\left[\boldsymbol{a}_N^\top(\boldsymbol{F}^\top\boldsymbol{W}\boldsymbol{F})^{-1}[\boldsymbol{V}_i\boldsymbol{F}_i]_{l\cdot}^\top\xi_{il}\right]^2 &\le \|\boldsymbol{a}_N\|^2\left\|(\boldsymbol{F}^\top\boldsymbol{W}\boldsymbol{F})^{-1}\right\|^2\|\boldsymbol{V}_i\|^2\|[\boldsymbol{F}_i]_{l\cdot}\|^2\,\mathsf{E}\left[\|\xi_{il}\|^2\right]\\ &= O\left[(p+Sq)g_{\min}^{-2}\right].\end{aligned}$$

Then by Markov inequality, it holds

$$\begin{aligned}P\Big(\left|\boldsymbol{a}_N^\top(\boldsymbol{F}^\top\boldsymbol{W}\boldsymbol{F})^{-1}[\boldsymbol{V}_i\boldsymbol{F}_i]_{l\cdot}^\top\xi_{il}\right| > \epsilon\sigma_N(\boldsymbol{a}_N)\Big) &\le \frac{\mathsf{E}\left[\boldsymbol{a}_N^\top(\boldsymbol{F}^\top\boldsymbol{W}\boldsymbol{F})^{-1}[\boldsymbol{V}_i\boldsymbol{F}_i]_{l\cdot}^\top\xi_{il}\right]^2}{\varepsilon^2\sigma_N^2(\boldsymbol{a}_N)}\\ &= O\left[(p+Sq)Ng_{\min}^{-2}\right].\end{aligned}$$

It then can be concluded from the above results that

$$\begin{aligned}&\sigma_N^{-2}(\boldsymbol{a}_N)\sum_{i=1}^{M}\sum_{l=1}^{n_i}\mathsf{E}\left\{\left[\boldsymbol{a}_N^\top(\boldsymbol{F}^\top\boldsymbol{W}\boldsymbol{F})^{-1}[\boldsymbol{V}_i\boldsymbol{F}_i]_{l\cdot}^\top\xi_{il}\right]^2\mathbb{1}_{\left|\boldsymbol{a}_N^\top(\boldsymbol{F}^\top\boldsymbol{W}\boldsymbol{F})^{-1}[\boldsymbol{V}_i\boldsymbol{F}_i]_{l\cdot}^\top\xi_{il}\right|>\epsilon\sigma_N(\boldsymbol{a}_N)}\right\}\\ &\quad= O\left[(p+Sq)^2N^3g_{\min}^{-4}\right],\end{aligned}$$

which is $o(1)$ provided that $g_{\min}\gg(p+Sq)^{1/2}N^{3/4}$. Accordingly, the asymptotic normality assertion (12) follows from the Lindeberg-Feller Central Limit Theorem.

S.4. Properties of Random Matrices

In this section, we derive the bounds for the L_2 norms and eigenvalues of the random design matrices under Assumption 4.1. In addition, we derive the non-asymptotic upper bound for $\boldsymbol{V}(\boldsymbol{Z}\boldsymbol{U}+\boldsymbol{\mathcal{E}})$, where $\boldsymbol{V}$ denotes the square root matrix of $\boldsymbol{W}$ such that $\boldsymbol{W}=\boldsymbol{V}^2$, under Assumption 4.2.

S.4.1. L_2-norm Bounds for Columns of Design Matrices

By union bound and (S.42), there exists a constant $c_1>0$ such that

$$P\Big(\max_{1\le j\le p}\left\|[\boldsymbol{X}]_{\cdot j}\right\|>\sqrt{5c_1N}\Big)\le\sum_{j=1}^{p}P\Big(\left\|[\boldsymbol{X}]_{\cdot j}\right\|>\sqrt{5c_1N}\Big)\le pe^{-N}$$

and similarly

$$P\Big(\max_{1\le s\le S}\max_{1\le k\le q}\left\|[\widetilde{\mathbb{Z}}_s]_{\cdot k}\right\|>\sqrt{5c_1g_{\max}}\Big)\le Sqe^{-g_{\max}},$$

where $\mathbb{Z}_s=(\boldsymbol{z}_i^\top)_{i:L_i=s}^\top$.

We define an event

$$E_{CB}=\left\{\max_{1\le j\le p}\left\|[\boldsymbol{X}]_{\cdot j}\right\|\le\sqrt{5c_1N}\quad\text{and}\quad\max_{1\le s\le S}\max_{1\le k\le q}\left\|[\widetilde{\mathbb{Z}}_s]_{\cdot k}\right\|\le\sqrt{5c_1g_{\max}}\right\}\tag{S.11}$$

with $P(E_{CB})\ge1-(p+Sq)e^{-g_{\max}}$ for later use. (The subscript "CB" stands for "column bounds.")

S.4.2. Eigenvalue Bounds for Design Matrices

In this section, we first derive the eigenvalue bounds for $\boldsymbol{x}_i^\top\boldsymbol{x}_i$ and $\boldsymbol{z}_i^\top\boldsymbol{z}_i$ for $i=1,\ldots,M$, and then those for $(\boldsymbol{X},\boldsymbol{A}\boldsymbol{Z})^\top(\boldsymbol{X},\boldsymbol{A}\boldsymbol{Z})$.

Let $\boldsymbol{f}_i=(\boldsymbol{x}_i,\boldsymbol{z}_i)$. Since the rows of $\boldsymbol{f}_i$ are independently drawn from $\mathbf{F}$, then the rows of $\mathsf{E}\left[\mathbf{F}\mathbf{F}^\top\right]^{-1/2}\boldsymbol{f}_i$ are isotropic, i.e., $\mathsf{Cov}\big(\mathsf{E}\left[\mathbf{F}\mathbf{F}^\top\right]^{-1/2}[\boldsymbol{f}_i]_{k\cdot}\big)=\boldsymbol{I}$. Following Theorem 5.39 of Vershynin (2012), it then can be shown that, on an event with probability at least $1-2\exp(-c_2t^2)$, we have

$$\begin{aligned}\sqrt{n_i}-C\sqrt{p+q}-t&\le\lambda_{\min}^{1/2}\left(\mathsf{E}\left[\mathbf{F}\mathbf{F}^\top\right]^{-1/2}\boldsymbol{f}_i^\top\boldsymbol{f}_i\mathsf{E}\left[\mathbf{F}\mathbf{F}^\top\right]^{-1/2}\right)\\ &\le\lambda_{\max}^{1/2}\left(\mathsf{E}\left[\mathbf{F}\mathbf{F}^\top\right]^{-1/2}\boldsymbol{f}_i^\top\boldsymbol{f}_i\mathsf{E}\left[\mathbf{F}\mathbf{F}^\top\right]^{-1/2}\right)\le\sqrt{n_i}+C\sqrt{p+q}+t,\end{aligned}$$

where C, c_2 are some positive constants. By taking $t = c_2^{-1/2}\sqrt{n_i}$, with probability at least $1 - 2e^{-n_i}$ we have

$$\begin{aligned}\lambda_{\min}^{1/2}\left(\boldsymbol{f}_i^\top \boldsymbol{f}_i/n_i\right) &= n_i^{-1/2}\lambda_{\min}^{1/2}\left(\mathsf{E}\left[\mathbf{F}\mathbf{F}^\top\right]^{1/2}\mathsf{E}\left[\mathbf{F}\mathbf{F}^\top\right]^{-1/2}\boldsymbol{f}_i^\top \boldsymbol{f}_i\mathsf{E}\left[\mathbf{F}\mathbf{F}^\top\right]^{-1/2}\mathsf{E}\left[\mathbf{F}\mathbf{F}^\top\right]^{1/2}\right)\\ &\geq n_i\lambda_{\min}\left(\mathsf{E}\left[\mathbf{F}\mathbf{F}^\top\right]^{1/2}\right)\lambda_{\min}^{1/2}\left(\mathsf{E}\left[\mathbf{F}\mathbf{F}^\top\right]^{-1/2}\boldsymbol{f}_i^\top \boldsymbol{f}_i\mathsf{E}\left[\mathbf{F}\mathbf{F}^\top\right]^{-1/2}\right)\\ &\geq C_f^{1/2}\left(1 - C\sqrt{\frac{p+q}{n_i}} - c_2^{-1/2}\right) \gtrsim (1 - c_2^{-1/2})C_f^{1/2}\end{aligned}$$

and similarly

$$\begin{aligned}\lambda_{\max}^{1/2}\left(\boldsymbol{f}_i^\top \boldsymbol{f}_i/n_i\right) &\leq n_i^{-1/2}\lambda_{\max}\left(\mathsf{E}\left[\mathbf{F}\mathbf{F}^\top\right]^{1/2}\right)\lambda_{\max}^{1/2}\left(\mathsf{E}\left[\mathbf{F}\mathbf{F}^\top\right]^{-1/2}\boldsymbol{f}_i^\top \boldsymbol{f}_i\mathsf{E}\left[\mathbf{F}\mathbf{F}^\top\right]^{-1/2}\right)\\ &\leq C_f^{-1/2}\left(1 + C\sqrt{\frac{p+q}{n_i}} + c_2^{-1/2}\right) \lesssim (1 + c_2^{-1/2})C_f^{-1/2}.\end{aligned}$$

In summary, it holds

$$(1 - c_2^{-1/2})C_f \lesssim \lambda_{\min}\big(\boldsymbol{f}_i^\top \boldsymbol{f}_i/n_i\big) \leq \lambda_{\max}\big(\boldsymbol{f}_i^\top \boldsymbol{f}_i/n_i\big) \lesssim (1 + c_2^{-1/2})C_f^{-1},$$

with probability at least $1 - 2e^{-n_i}$ for $i = 1, \ldots, M$.

By the fact that the largest (smallest) eigenvalue of an submatrix with fewer columns is bounded above (below) from the largest (smallest) eigenvalue of the full matrix, we have

$$\begin{aligned}(1 - c_2^{-1/2})C_f \lesssim \lambda_{\min}\big(\boldsymbol{x}_i^\top \boldsymbol{x}_i/n_i\big) \leq \lambda_{\max}\big(\boldsymbol{x}_i^\top \boldsymbol{x}_i/n_i\big) \lesssim (1 + c_2^{-1/2})C_f^{-1},\\ (1 - c_2^{-1/2})C_f \lesssim \lambda_{\min}\big(\boldsymbol{z}_i^\top \boldsymbol{z}_i/n_i\big) \leq \lambda_{\max}\big(\boldsymbol{z}_i^\top \boldsymbol{z}_i/n_i\big) \lesssim (1 + c_2^{-1/2})C_f^{-1},\end{aligned}$$

with probability at least $1 - 2e^{-n_i}$ for each $i = 1, \ldots, M$.

Define the event

$$E_{EB} = \left\{\underline{C}_f \lesssim \lambda_{\min}\big(\boldsymbol{f}_i^\top \boldsymbol{f}_i/n_i\big) \leq \lambda_{\max}\big(\boldsymbol{f}_i^\top \boldsymbol{f}_i/n_i\big) \lesssim \overline{C}_f, i = 1, \ldots, M\right\} \tag{S.12}$$

with $P(E_{EB}) \geq 1 - 2Me^{-n_{\min}}$, where $\underline{C}_f = (1 - c_2^{-1/2})C_f, \overline{C}_f = (1 + c_2^{-1/2})C_f^{-1}$ and $n_{\min} = \min_{1\leq i\leq M} n_i$. Then, on the event E_{EB}, it follows that

$$\begin{aligned}\lambda_{\max}\big(\boldsymbol{F}^\top \boldsymbol{F}\big) &\leq \max\left\{\lambda_{\max}\big(\boldsymbol{X}^\top \boldsymbol{X}\big), \lambda_{\max}\big[(\boldsymbol{Z}\boldsymbol{A})^\top (\boldsymbol{Z}\boldsymbol{A})\big]\right\}\\ &\leq \max\left\{\sum_{i=1}^{M}\lambda_{\max}\big(\boldsymbol{x}_i^\top \boldsymbol{x}_i\big), \max_{1\leq s\leq S}\sum_{i:L_i=s}\lambda_{\max}\big(\boldsymbol{z}_i^\top \boldsymbol{z}_i\big)\right\}\\ &\lesssim \max\left\{\overline{C}_f N, \overline{C}_f g_{\max}\right\} = \overline{C}_f N\end{aligned} \tag{S.13}$$

and

$$\begin{aligned}\lambda_{\min}\big(\boldsymbol{F}^\top \boldsymbol{F}\big) &\geq \min\left\{\lambda_{\min}\big(\boldsymbol{X}^\top \boldsymbol{X}\big), \lambda_{\min}\big[(\boldsymbol{Z}\boldsymbol{A})^\top (\boldsymbol{Z}\boldsymbol{A})\big]\right\}\\ &\geq \min\left\{\sum_{i=1}^{M}\lambda_{\min}\big(\boldsymbol{x}_i^\top \boldsymbol{x}_i\big), \min_{1\leq s\leq S}\sum_{i:L_i=s}\lambda_{\min}\big(\boldsymbol{z}_i^\top \boldsymbol{z}_i\big)\right\}\\ &\gtrsim \min\left\{\underline{C}_f N, \underline{C}_f g_{\min}\right\} = \underline{C}_f g_{\min}.\end{aligned} \tag{S.14}$$

S.4.3. Weight Matrix

It is easy to see that

$$\lambda_{\max}(\boldsymbol{W}_i) = \lambda_{\min}^{-1}\left(\sigma_\varepsilon^2 \boldsymbol{I}_{n_i} + \boldsymbol{z}_i \boldsymbol{\Psi} \boldsymbol{z}_i^\top\right) \leq \frac{1}{\sigma_\varepsilon^2 + \lambda_{\min}(\boldsymbol{z}_i \boldsymbol{\Psi} \boldsymbol{z}_i^\top)} = \sigma_\varepsilon^{-2}, \tag{S.15}$$

where $\lambda_{\min}(\boldsymbol{z}_i\boldsymbol{\Psi}\boldsymbol{z}_i^\top) = 0$ since $q < n_i$. Note that this is an exact result without any probabilistic nor asymptotic statement. Moreover, since $\boldsymbol{W} = \operatorname{diag}\left[(\boldsymbol{W}_i)_{i=1,\ldots,M}\right]$, we can also conclude

$$\lambda_{\max}(\boldsymbol{W}) = \max_{1\le i\le M} \lambda_{\max}(\boldsymbol{W}_i) \le \sigma_\varepsilon^{-2}. \tag{S.16}$$

S.4.4. Tail bounds of Weighted Random Effects and Noises

On the event E_{EB}, we show the non-asymptotic tail bounds for $\boldsymbol{V}(\boldsymbol{Z}\boldsymbol{U} + \boldsymbol{\mathcal{E}})$. For $\boldsymbol{V}\boldsymbol{Z}\boldsymbol{U}$, we first check $\left\|\boldsymbol{Z}^\top\boldsymbol{V}\right\|$ as follows. By definition, we can write

$$\boldsymbol{W}_i = \left(\sigma_\varepsilon^2\boldsymbol{I}_{n_i} + \boldsymbol{z}_i\boldsymbol{\Psi}\boldsymbol{z}_i^\top\right)^{-1} = \sigma_\varepsilon^{-2}\left[\boldsymbol{I}_{n_i} - \sigma_\varepsilon^{-2}\boldsymbol{z}_i(\boldsymbol{\Psi}^{-1} + \sigma_\varepsilon^{-2}\boldsymbol{z}_i^\top\boldsymbol{z}_i)^{-1}\boldsymbol{z}_i^\top\right],$$

where the second identity holds by applying the Woodbury identity (S.41). It then follows that

$$\begin{aligned}(\boldsymbol{z}_i^\top\boldsymbol{W}_i\boldsymbol{z}_i)^{-1} &= \sigma_\varepsilon^2\left[\boldsymbol{z}_i^\top\boldsymbol{z}_i - \sigma_\varepsilon^{-2}\boldsymbol{z}_i^\top\boldsymbol{z}_i(\boldsymbol{\Psi}^{-1} + \sigma_\varepsilon^{-2}\boldsymbol{z}_i^\top\boldsymbol{z}_i)^{-1}\boldsymbol{z}_i^\top\boldsymbol{z}_i\right]^{-1}\\ &= \sigma_\varepsilon^2(\boldsymbol{z}_i^\top\boldsymbol{z}_i)^{-1} + \boldsymbol{\Psi}^{-1},\end{aligned}$$

where the second and identity holds by applying Woodbury identity (S.41) on the inverse term. It then holds

$$\begin{aligned}\lambda_{\max}\left(\boldsymbol{z}_i^\top\boldsymbol{W}_i\boldsymbol{z}_i\right) &= \lambda_{\min}^{-1}\left[(\boldsymbol{z}_i^\top\boldsymbol{W}_i\boldsymbol{z}_i)^{-1}\right]\\ &= \sigma_\varepsilon^{-2}\lambda_{\min}^{-1}\left[(\boldsymbol{z}_i^\top\boldsymbol{z}_i)^{-1} + \sigma_\varepsilon^{-2}\boldsymbol{\Psi}^{-1}\right]\\ &\le \sigma_\varepsilon^{-2}\left\{\lambda_{\min}\left[(\boldsymbol{z}_i^\top\boldsymbol{z}_i)^{-1}\right] + \sigma_\varepsilon^{-2}\lambda_{\min}(\boldsymbol{\Psi}^{-1})\right\}^{-1}\\ &\le \tau^{-1}.\end{aligned}$$

Accordingly, we have

$$\left\|\boldsymbol{Z}^\top\boldsymbol{V}\right\| \le \|\boldsymbol{V}\boldsymbol{Z}\| = \lambda_{\max}^{1/2}\left(\boldsymbol{Z}^\top\boldsymbol{W}\boldsymbol{Z}\right) = \max_{1\le i\le M}\lambda_{\max}^{1/2}\left(\boldsymbol{z}_i^\top\boldsymbol{W}_i\boldsymbol{z}_i\right) \le \tau^{-1/2}. \tag{S.17}$$

Under events E_{CB} and E_{EB}, it follows for any $\boldsymbol{a} \in \mathbb{R}^N$ that

$$\begin{aligned}P\left(\left|\boldsymbol{a}^\top\boldsymbol{V}\boldsymbol{Z}\boldsymbol{U}\right| > \|\boldsymbol{a}\|t\middle|\boldsymbol{Z}\right) &\le P\left(\left|\boldsymbol{a}^\top\boldsymbol{V}\boldsymbol{Z}\boldsymbol{U}\right| > \|\boldsymbol{Z}^\top\boldsymbol{V}\boldsymbol{a}\|\frac{t}{\|\boldsymbol{V}\boldsymbol{Z}\|}\middle|\boldsymbol{Z}\right)\\ &\le 2\exp\left(-\frac{c_e t^2}{\|\boldsymbol{V}\boldsymbol{Z}\|^2}\right) \le 2\exp\left(-c_e\tau t^2\right).\end{aligned}$$

By law of iterated expectations, we have

$$\begin{aligned}P\left(\left|\boldsymbol{a}^\top\boldsymbol{V}\boldsymbol{Z}\boldsymbol{U}\right| > \|\boldsymbol{a}\|t\right) &= \mathsf{E}_{\boldsymbol{Z}}\left[P\left(\left|\boldsymbol{a}^\top\boldsymbol{V}\boldsymbol{Z}\boldsymbol{U}\right| > \|\boldsymbol{a}\|t\middle|\boldsymbol{Z}\right)\right]\\ &\le \mathsf{E}_{\boldsymbol{Z}}\left\{2\exp\left[-c_e\tau t^2\right]\right\} = 2\exp\left[-c_e\tau t^2\right].\end{aligned}$$

For $\boldsymbol{V}\boldsymbol{\mathcal{E}}$, since $\|\boldsymbol{V}\| \le \sigma_\varepsilon^{-1}$, similar argument yields

$$P\left(\left|\boldsymbol{a}^\top\boldsymbol{V}\boldsymbol{\mathcal{E}}\right| > \|\boldsymbol{a}\|t\right) \le 2\exp\left[-c_e\sigma_\varepsilon^2 t^2\right].$$

In summary, it holds

$$\begin{aligned}P\left(\left|\boldsymbol{a}^\top\boldsymbol{V}(\boldsymbol{Z}\boldsymbol{U} + \boldsymbol{\mathcal{E}})\right| > \|\boldsymbol{a}\|t\right) &\le P\left(\left|\boldsymbol{a}^\top\boldsymbol{V}\boldsymbol{Z}\boldsymbol{U}\right| > \|a\|t\right) + P\left(\left|\boldsymbol{a}^\top\boldsymbol{V}\boldsymbol{\mathcal{E}}\right| > \|\boldsymbol{a}\|t\right)\\ &\le 4\exp\left(-c_e\left(\tau\wedge\sigma_\varepsilon^2\right)t^2\right),\end{aligned} \tag{S.18}$$

where $\tau = \lambda_{\min}(\boldsymbol{\Psi})$.

In addition, similar argument results in

$$P\left(\left|\boldsymbol{a}_i^\top\boldsymbol{V}_i(\boldsymbol{z}_i\boldsymbol{u}_i + \boldsymbol{\varepsilon}_i)\right| > \|\boldsymbol{a}_i\|t\right) \le 4\exp\left\{-c_e\left(\tau\wedge\sigma_\varepsilon^2\right)t^2\right\} \tag{S.19}$$

for any $\boldsymbol{a}_i \in \mathbb{R}^{n_i}$, where $\boldsymbol{V}_i$ is the square root matrix of $\boldsymbol{W}_i$ such that $\boldsymbol{W}_i = \boldsymbol{V}_i^2$. Moreover, let $\mathbb{V}_s = \operatorname{diag}\left[(\boldsymbol{V}_i)_{i:L_i=s}\right]$, $\mathbb{U}_s = (\boldsymbol{u}_i)_{i:L_i=s}^\top$ and $\boldsymbol{\mathcal{E}}_s = (\boldsymbol{\varepsilon}_i)_{i:L_i=s}^\top$. Then we have

$$P\left(\left|\boldsymbol{b}_s^\top\mathbb{V}_s(\mathbb{Z}_s\mathbb{U}_s + \boldsymbol{\mathcal{E}}_s)\right| > \|\boldsymbol{b}_s\|t\right) \le 4\exp\left\{-c_e\left(\tau\wedge\sigma_\varepsilon^2\right)t^2\right\} \tag{S.20}$$

for any $\boldsymbol{b}_s \in \mathbb{R}^{g_s}$.

S.5. Derivation of Algorithm 1

The augmented Lagrangian of (9) can then be formulated by

$$\begin{aligned} L(\boldsymbol{\beta}, \boldsymbol{\Theta}, \boldsymbol{\delta}, \boldsymbol{\nu}) =& L_0(\boldsymbol{\beta}, \boldsymbol{\Theta}, \boldsymbol{\delta}) \\ &+ \sum_{1 \le i < j \le M} \boldsymbol{\nu}_{ij}^\top (\boldsymbol{\theta}_i - \boldsymbol{\theta}_j - \boldsymbol{\delta}_{ij}) \\ &+ \frac{\kappa}{2} \sum_{1 \le i < j \le M} \left\| \boldsymbol{\theta}_i - \boldsymbol{\theta}_j - \boldsymbol{\delta}_{ij} \right\|^2, \end{aligned} \tag{S.21}$$

where the dual variables $\boldsymbol{\nu} = (\boldsymbol{\nu}_{ij}^\top)_{i<j}^\top$ are Lagrange multipliers and $\kappa > 0$ is the penalty parameter.

The minimizer of the augmented Lagrangian (S.21) can be solved by developing an ADMM algorithm. Given the values of $\boldsymbol{\Theta}^{[k]}$, $\boldsymbol{\delta}^{[k]}$ and $\boldsymbol{\nu}^{[k]}$ at the k-th iteration, we update $\boldsymbol{\beta}$ using

$$\begin{aligned} \boldsymbol{\beta}^{[k+1]} = & \left(\sum_{i=1}^M \boldsymbol{x}_i^\top \boldsymbol{W}_i \boldsymbol{x}_i \right)^{-1} \times \\ & \left(\sum_{i=1}^M \boldsymbol{x}_i^\top \boldsymbol{W}_i \left[\boldsymbol{x}_i \breve{\boldsymbol{\beta}}_i + \boldsymbol{z}_i \left(\breve{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i^{[k]} \right) \right] \right). \end{aligned} \tag{S.22}$$

Next, given $\boldsymbol{\beta}^{[k+1]}$, $\boldsymbol{\delta}^{[k]}$ and $\boldsymbol{\nu}^{[k]}$, the updating formula for $\boldsymbol{\Theta}$ is

$$\begin{aligned} \boldsymbol{\Theta}^{[k+1]} = & \left(\boldsymbol{Z}^\top \boldsymbol{W} \boldsymbol{Z} + \kappa N \boldsymbol{B} \boldsymbol{B}^\top \right)^{-1} \times \\ & \left[\boldsymbol{Z}^\top \boldsymbol{W} \boldsymbol{Z} \breve{\boldsymbol{\Theta}} + \boldsymbol{D}_{zx} \left(\breve{\boldsymbol{\beta}} - \mathbf{1}_M \otimes \boldsymbol{\beta}^{[k+1]} \right) \right. \\ & \left. + N \boldsymbol{B} \left(\kappa \boldsymbol{\delta}^{[k]} - \boldsymbol{\nu}^{[k]} \right) \right], \end{aligned} \tag{S.23}$$

where $\widehat{\boldsymbol{\Theta}} = (\breve{\boldsymbol{\theta}}_i^\top)_{i=1,\ldots,M}^\top$, $\breve{\boldsymbol{\beta}} = (\breve{\boldsymbol{\beta}}_i^\top)_{i=1,\ldots,M}^\top$, $\boldsymbol{D}_{zx} = \operatorname{diag}\left[(\boldsymbol{z}_i^\top \boldsymbol{W}_i \boldsymbol{x}_i)_{i=1,\ldots,M} \right]$ and $\boldsymbol{B}$ is a matrix such that $\boldsymbol{B}^\top \boldsymbol{\Theta} = (\boldsymbol{\theta}_i^\top - \boldsymbol{\theta}_j^\top)_{i<j}^\top$. Note that the form of $\boldsymbol{B}$ depends on how we arrange $\boldsymbol{\delta}_{ij}$'s in $\boldsymbol{\delta}$; see Remark S.5.1.

To update $\boldsymbol{\delta}$, we use (S.29), (S.30), (S.31) and (S.32) derived below for L_1, MCP, SCAD and TLP, respectively. Finally, updating the Lagrange multipliers $\boldsymbol{\nu}$ takes a standard form in general ADMM algorithms,

$$\boldsymbol{\nu}^{[k+1]} = \boldsymbol{\nu}^{[k]} + \kappa \left(\boldsymbol{B}^\top \boldsymbol{\Theta}^{[k+1]} - \boldsymbol{\delta}^{[k+1]} \right). \tag{S.24}$$

The above discussions are summarized in Algorithm 1.

As the initial values of $(\boldsymbol{\Theta}^{[0]}, \boldsymbol{\delta}^{[0]}, \boldsymbol{\nu}^{[0]})$, we use $\boldsymbol{\Theta}^{[0]} = \widehat{\boldsymbol{\Theta}}$, $\boldsymbol{\delta}^{[0]} = \boldsymbol{B}^\top \widehat{\boldsymbol{\Theta}}$ and $\boldsymbol{\nu}^{[0]} = \mathbf{0}$ in our implementation. We also fix $\kappa = 1$ because the ADMM method can be shown to converge for all values of the penalty parameter $\kappa > 0$ (e.g., Eckstein (2012); Ghadimi et al. (2015)).

Detailed derivation for (S.29), (S.30), (S.31) and (S.32).

Given the values of $\boldsymbol{\delta}^{[k]}$ and $\boldsymbol{\nu}^{[k]}$ in the k-th interation, the ADMM that solves the augmented Lagrangian (S.21) goes as follows:

$$\left(\boldsymbol{\beta}^{[k+1]}, \boldsymbol{\Theta}^{[k+1]} \right) = \arg\min_{\boldsymbol{\Theta}} L(\boldsymbol{\beta}, \boldsymbol{\Theta}, \boldsymbol{\delta}^{[k]}, \boldsymbol{\nu}^{[k]}), \tag{S.25}$$

$$\boldsymbol{\delta}^{[k+1]} = \arg\min_{\boldsymbol{\delta}} L(\boldsymbol{\beta}^{[k+1]}, \boldsymbol{\Theta}^{[k+1]}, \boldsymbol{\delta}, \boldsymbol{\nu}^{[k]}), \tag{S.26}$$

and $\boldsymbol{\nu}^{[k+1]}$ is updated using (S.24).

Firstly, we rewrite the augmented Lagrangian (S.21) by expanding the first term as

$$
\begin{aligned}
L(\boldsymbol{\beta}, \boldsymbol{\Theta}, \boldsymbol{\delta}, \boldsymbol{\nu}) = & \frac{1}{2N} \sum_{i=1}^{M} (\breve{\boldsymbol{\beta}}_i - \boldsymbol{\beta})^\top \boldsymbol{x}_i^\top \boldsymbol{W}_i \boldsymbol{x}_i (\breve{\boldsymbol{\beta}}_i - \boldsymbol{\beta}) + \frac{1}{2N} \sum_{i=1}^{M} (\breve{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i)^\top \boldsymbol{z}_i^\top \boldsymbol{W}_i \boldsymbol{z}_i (\breve{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i) \\
& + \frac{1}{N} \sum_{i=1}^{M} (\breve{\boldsymbol{\beta}}_i - \boldsymbol{\beta})^\top \boldsymbol{x}_i^\top \boldsymbol{W}_i \boldsymbol{z}_i (\breve{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i) + \sum_{1 \le i < j \le M} p_\gamma \left(\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_j\| ; \lambda \right) \\
& + \sum_{1 \le i < j \le M} \boldsymbol{\nu}_{ij}^\top (\boldsymbol{\theta}_i - \boldsymbol{\theta}_j - \boldsymbol{\delta}_{ij}) + \frac{\kappa}{2} \sum_{1 \le i < j \le M} \|\boldsymbol{\theta}_i - \boldsymbol{\theta}_j - \boldsymbol{\delta}_{ij}\|^2 \qquad \text{(S.27)} \\
= & \frac{1}{2N} (\breve{\boldsymbol{\beta}} - \mathbf{1}_M \otimes \boldsymbol{\beta})^\top \boldsymbol{D}_x (\breve{\boldsymbol{\beta}} - \mathbf{1}_M \otimes \boldsymbol{\beta}) + \frac{1}{2N} (\breve{\boldsymbol{\Theta}} - \boldsymbol{\Theta})^\top \boldsymbol{Z}^\top \boldsymbol{W} \boldsymbol{Z} (\breve{\boldsymbol{\Theta}} - \boldsymbol{\Theta}) \\
& + \frac{1}{N} (\breve{\boldsymbol{\Theta}} - \boldsymbol{\Theta})^\top \boldsymbol{D}_{zx} (\breve{\boldsymbol{\beta}} - \mathbf{1}_M \otimes \boldsymbol{\beta}) + \sum_{i<j} p_\gamma \left(\|\boldsymbol{\delta}_{ij}\| ; \lambda \right) \\
& + (\boldsymbol{B}^\top \boldsymbol{\Theta} - \boldsymbol{\delta})^\top \boldsymbol{\nu} + \frac{\kappa}{2} (\boldsymbol{B}^\top \boldsymbol{\Theta} - \boldsymbol{\delta})^\top (\boldsymbol{B}^\top \boldsymbol{\Theta} - \boldsymbol{\delta}), \qquad \text{(S.28)}
\end{aligned}
$$

where $\boldsymbol{D}_x = \operatorname{diag}\left[(\boldsymbol{x}_i^\top \boldsymbol{W}_i \boldsymbol{x}_i)_{i=1,\ldots,M} \right]$, $\boldsymbol{D}_{zx} = \operatorname{diag}\left[(\boldsymbol{z}_i^\top \boldsymbol{W}_i \boldsymbol{x}_i)_{i=1,\ldots,M} \right]$ and $\boldsymbol{B}$ is a matrix such that $\boldsymbol{B}^\top \boldsymbol{\Theta} = (\boldsymbol{\theta}_i^\top - \boldsymbol{\theta}_j^\top)_{i<j}^\top$. Given $\boldsymbol{\Theta}^{[k]}$, $\boldsymbol{\delta}^{[k]}$ and $\boldsymbol{\nu}^{[k]}$, solving the equation using (S.27)

$$
\frac{\partial L(\boldsymbol{\beta}, \boldsymbol{\Theta}^{[k]}, \boldsymbol{\delta}^{[k]}, \boldsymbol{\nu}^{[k]})}{\partial \boldsymbol{\beta}} = -\frac{1}{N} \sum_{i=1}^{M} \boldsymbol{x}_i^\top \boldsymbol{W}_i \left[\boldsymbol{x}_i (\breve{\boldsymbol{\beta}}_i - \boldsymbol{\beta}) + \boldsymbol{z}_i (\breve{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i^{[k]}) \right] = \mathbf{0}
$$

yields the updating formula (S.27) for $\boldsymbol{\beta}$. Similarly, given $\boldsymbol{\beta}^{[k+1]}$, $\boldsymbol{\delta}^{[k]}$ and $\boldsymbol{\nu}^{[k]}$, setting the partial derivative w.r.t. $\boldsymbol{\Theta}$ using (S.28),

$$
\frac{\partial L(\boldsymbol{\beta}^{[k+1]}, \boldsymbol{\Theta}, \boldsymbol{\delta}^{[k]}, \boldsymbol{\nu}^{[k]})}{\partial \boldsymbol{\Theta}} = -\frac{1}{N} \left[\boldsymbol{Z}^\top \boldsymbol{W} \boldsymbol{Z} (\breve{\boldsymbol{\Theta}} + \boldsymbol{\Theta}) - \boldsymbol{D}_{zx} (\breve{\boldsymbol{\beta}} - \mathbf{1}_M \otimes \boldsymbol{\beta}^{[k+1]}) \right] + \boldsymbol{B} \boldsymbol{\nu}^{[k]} + \kappa \boldsymbol{B} (\boldsymbol{B}^\top \boldsymbol{\Theta} - \boldsymbol{\delta}^{[k]})
$$

equal to zero leads to the updating formula (S.23) for $\boldsymbol{\Theta}$.

Now, given $\boldsymbol{\beta}^{[k+1]}$, $\boldsymbol{\Theta}^{[k+1]}$ and $\boldsymbol{\nu}^{[k]}$, we derive the updating formula for $\boldsymbol{\delta}$. By omitting irrelevant terms w.r.t. $\boldsymbol{\delta}$, (S.27) results in

$$
\sum_{i<j} \left[p_\gamma \left(\|\boldsymbol{\delta}_{ij}\| ; \lambda \right) + \frac{\kappa}{2} \left\| \boldsymbol{\zeta}_{ij}^{[k+1]} - \boldsymbol{\delta}_{ij} \right\|^2 \right],
$$

where $\boldsymbol{\zeta}_{ij}^{[k+1]} = \boldsymbol{\theta}_i^{[k+1]} - \boldsymbol{\theta}_j^{[k+1]} + \kappa^{-1} \boldsymbol{\nu}_{ij}^{[k]}$. Its minimizer is taken as the update of $\boldsymbol{\delta}$. In particular, the solution for the L_1 penalty is

$$
\boldsymbol{\delta}_{ij}^{[k+1]} = \operatorname{ST}\left(\boldsymbol{\zeta}_{ij}^{[k+1]}; \lambda/\kappa \right), \qquad \text{(S.29)}
$$

where $\operatorname{ST}\left(\boldsymbol{v}; t \right) = (1 - t/\|\boldsymbol{v}\|)_+ \boldsymbol{v}$ is the soft thresholding operator. For the MCP given in (S.2) with $\gamma > \kappa^{-1}$, the solution is

$$
\boldsymbol{\delta}_{ij}^{[k+1]} = \begin{cases} \dfrac{\operatorname{ST}\left(\boldsymbol{\zeta}_{ij}^{[k+1]}; \lambda/\kappa \right)}{1 - 1/(\gamma\kappa)} & \text{if } \left\| \boldsymbol{\zeta}_{ij}^{[k+1]} \right\| \le \gamma\lambda, \\ \boldsymbol{\zeta}_{ij}^{[k+1]} & \text{if } \left\| \boldsymbol{\zeta}_{ij}^{[k+1]} \right\| > \gamma\lambda. \end{cases} \qquad \text{(S.30)}
$$

For the SCAD penalty given in (S.3) with $\gamma > \kappa^{-1} + 1$, the solution is

$$
\boldsymbol{\delta}_{ij}^{[k+1]} = \begin{cases} \operatorname{ST}\left(\boldsymbol{\zeta}_{ij}^{[k+1]}; \lambda/\kappa \right) & \text{if } \left\| \boldsymbol{\zeta}_{ij}^{[k+1]} \right\| \le \lambda + \lambda/\kappa, \\ \dfrac{\operatorname{ST}\left(\boldsymbol{\zeta}_{ij}^{[k+1]}, \gamma\lambda/[(\gamma-1)\kappa] \right)}{1 - 1/[(\gamma-1)\kappa]} & \text{if } \lambda + \lambda/\kappa < \left\| \boldsymbol{\zeta}_{ij}^{[k+1]} \right\| \le \gamma\lambda, \\ \boldsymbol{\zeta}_{ij}^{[k+1]} & \text{if } \left\| \boldsymbol{\zeta}_{ij}^{[k+1]} \right\| > \gamma\lambda. \end{cases} \qquad \text{(S.31)}
$$

For the TLP given in (S.4) with $\gamma > \kappa^{-1}$, the solution is

$$\boldsymbol{\delta}_{ij}^{[k+1]} = \begin{cases} \mathrm{ST}\left(\boldsymbol{\zeta}_{ij}^{[k+1]}; \lambda/\kappa\right) & \text{if } \left\|\boldsymbol{\zeta}_{ij}^{[k+1]}\right\| \leq \gamma\lambda, \\ \boldsymbol{\zeta}_{ij}^{[k+1]} & \text{if } \left\|\boldsymbol{\zeta}_{ij}^{[k+1]}\right\| > \gamma\lambda. \end{cases} \tag{S.32}$$

Algorithm 1 ADMM for CD Fusion Approach

Input: Initial values of $\boldsymbol{\Theta}^{[0]}$, $\boldsymbol{\delta}^{[0]}$ and $\boldsymbol{\nu}^{[0]}$
for $k = 1, 2, \ldots$ **do**
  Compute $\boldsymbol{\beta}^{[k+1]}$ using (S.22)
  Compute $\boldsymbol{\Theta}^{[k+1]}$ using (S.23)
  Compute $\boldsymbol{\delta}^{[k+1]}$ using (S.29), (S.30), (S.31) or (S.32)
  Compute $\boldsymbol{\nu}^{[k+1]}$ using (S.24)
  if Convergence criterion is met **then**
    Break
  else
    $k \leftarrow k + 1$
  end if
end for
Input: $\widehat{\boldsymbol{\beta}}(\lambda) = \boldsymbol{\beta}^{[k+1]}$ and $\widehat{\boldsymbol{\Theta}}(\lambda) = \boldsymbol{\Theta}^{[k+1]}$

Remark S.5.1. Here we stress that, though $\boldsymbol{\delta} = (\boldsymbol{\delta}_{ij}^\top)_{i<j}^\top, \boldsymbol{\nu} = (\boldsymbol{\nu}_{ij})_{i<j}^\top$ and the matrix $\boldsymbol{B}$ such that $\boldsymbol{B}^\top\boldsymbol{\Theta} = (\boldsymbol{\theta}_i^\top - \boldsymbol{\theta}_j^\top)_{i<j}^\top$ are just auxiliary parameters in the ADMM algorithm, one has to be cautious with matching their forms. More specifically, $\boldsymbol{\nu}_{ij}$'s are the corresponding Lagrangian multipliers for $\boldsymbol{\delta}_{ij}$, which are made to match $\boldsymbol{\theta}_i - \boldsymbol{\theta}_j$. For example, suppose $M = 4$, and one may have

$$\boldsymbol{\delta} = \begin{pmatrix} \boldsymbol{\delta}_{12} \\ \boldsymbol{\delta}_{13} \\ \boldsymbol{\delta}_{14} \\ \boldsymbol{\delta}_{23} \\ \boldsymbol{\delta}_{24} \\ \boldsymbol{\delta}_{34} \end{pmatrix}, \quad \boldsymbol{\nu} = \begin{pmatrix} \boldsymbol{\nu}_{12} \\ \boldsymbol{\nu}_{13} \\ \boldsymbol{\nu}_{14} \\ \boldsymbol{\nu}_{23} \\ \boldsymbol{\nu}_{24} \\ \boldsymbol{\nu}_{34} \end{pmatrix} \quad \text{and} \quad \boldsymbol{B}^\top = \begin{bmatrix} \boldsymbol{I}_q & -\boldsymbol{I}_q & & \\ \boldsymbol{I}_q & & -\boldsymbol{I}_q & \\ \boldsymbol{I}_q & & & -\boldsymbol{I}_q \\ & \boldsymbol{I}_q & -\boldsymbol{I}_q & \\ & \boldsymbol{I}_q & & -\boldsymbol{I}_q \\ & & \boldsymbol{I}_q & -\boldsymbol{I}_q \end{bmatrix} \text{ s.t. } \boldsymbol{B}^\top\boldsymbol{\Theta} = \begin{pmatrix} \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2 \\ \boldsymbol{\theta}_1 - \boldsymbol{\theta}_3 \\ \boldsymbol{\theta}_1 - \boldsymbol{\theta}_4 \\ \boldsymbol{\theta}_2 - \boldsymbol{\theta}_3 \\ \boldsymbol{\theta}_2 - \boldsymbol{\theta}_4 \\ \boldsymbol{\theta}_3 - \boldsymbol{\theta}_4 \end{pmatrix}.$$

However, it is also possible to set

$$\boldsymbol{\delta} = \begin{pmatrix} \boldsymbol{\delta}_{12} \\ \boldsymbol{\delta}_{23} \\ \boldsymbol{\delta}_{34} \\ \boldsymbol{\delta}_{13} \\ \boldsymbol{\delta}_{24} \\ \boldsymbol{\delta}_{14} \end{pmatrix}, \quad \boldsymbol{\nu} = \begin{pmatrix} \boldsymbol{\nu}_{12} \\ \boldsymbol{\nu}_{23} \\ \boldsymbol{\nu}_{34} \\ \boldsymbol{\nu}_{13} \\ \boldsymbol{\nu}_{24} \\ \boldsymbol{\nu}_{14} \end{pmatrix} \quad \text{and} \quad \boldsymbol{B}^\top = \begin{bmatrix} \boldsymbol{I}_q & -\boldsymbol{I}_q & & \\ & \boldsymbol{I}_q & -\boldsymbol{I}_q & \\ & & \boldsymbol{I}_q & -\boldsymbol{I}_q \\ \boldsymbol{I}_q & & -\boldsymbol{I}_q & \\ & \boldsymbol{I}_q & & -\boldsymbol{I}_q \\ \boldsymbol{I}_q & & & -\boldsymbol{I}_q \end{bmatrix} \text{ s.t. } \boldsymbol{B}^\top\boldsymbol{\Theta} = \begin{pmatrix} \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2 \\ \boldsymbol{\theta}_2 - \boldsymbol{\theta}_3 \\ \boldsymbol{\theta}_3 - \boldsymbol{\theta}_4 \\ \boldsymbol{\theta}_1 - \boldsymbol{\theta}_3 \\ \boldsymbol{\theta}_2 - \boldsymbol{\theta}_4 \\ \boldsymbol{\theta}_1 - \boldsymbol{\theta}_4 \end{pmatrix}.$$

As can be seen that the order of $\boldsymbol{\nu}_{ij}$'s and the form of $\boldsymbol{B}$ both depend on the order of $\boldsymbol{\delta}_{ij}$'s. Although it is not necessary to have guidelines to make their forms unique, while matching them correspondingly is crucial in implementation.

S.6. Proof of Proposition 3.1

Here we show $\lim_{k\to\infty} \left\|\boldsymbol{r}^{[k]}\right\|^2 = 0$. Define

$$\begin{aligned} f^{[k+1]} &= \inf_{\boldsymbol{B}^\top\boldsymbol{\Theta}^{[k+1]} - \boldsymbol{\delta} = \boldsymbol{0}} \left\{ L_N^{\mathrm{CD}}(\boldsymbol{\beta}^{[k+1]}, \boldsymbol{\Theta}^{[k+1]}) + \sum_{i<j} p_\gamma(\|\boldsymbol{\delta}\|; \lambda) \right\} \\ &= \inf_{\boldsymbol{B}^\top\boldsymbol{\Theta}^{[k+1]} - \boldsymbol{\delta} = \boldsymbol{0}} L(\boldsymbol{\beta}^{[k+1]}, \boldsymbol{\Theta}^{[k+1]}, \boldsymbol{\delta}, \boldsymbol{\nu}^{[k]}), \end{aligned}$$

where $L_N^{\mathrm{CD}}(\boldsymbol{\beta}, \boldsymbol{\Theta})$ is defined in (S.5). By the definition of $\boldsymbol{\delta}^{[k+1]}$ as in (S.26), we have

$$L(\boldsymbol{\beta}^{[k+1]}, \boldsymbol{\Theta}^{[k+1]}, \boldsymbol{\delta}^{[k+1]}, \boldsymbol{\nu}^{[k]}) \leq f^{[k+1]}.$$

Let t be an integer. Since $\boldsymbol{\nu}^{[k+t-1]} = \boldsymbol{\nu}^{[k]} + \kappa \sum_{i=1}^{t-1}(\boldsymbol{B}^\top \boldsymbol{\Theta}^{[k+i]} - \boldsymbol{\delta}^{[k+i]})$, it follows that

$$\begin{aligned}
L(\boldsymbol{\beta}^{[k+t]}, \boldsymbol{\Theta}^{[k+t]}, \boldsymbol{\delta}^{[k+t]}, \boldsymbol{\nu}^{[k+t-1]}) &= L_N^{\mathrm{CD}}(\boldsymbol{\beta}^{[k+t]}, \boldsymbol{\Theta}^{[k+t]}) + \sum_{i<j} p_\gamma(\|\boldsymbol{\delta}^{[k+t]}\|; \lambda) \\
&\quad + (\boldsymbol{\nu}^{[k+t-1]})^\top (\boldsymbol{B}^\top \boldsymbol{\Theta}^{[k+t]} - \boldsymbol{\delta}^{[k+t]}) + \frac{\kappa}{2} \left\| \boldsymbol{B}^\top \boldsymbol{\Theta}^{[k+t]} - \boldsymbol{\delta}^{[k+t]} \right\|^2 \\
&= L_N^{\mathrm{CD}}(\boldsymbol{\beta}^{[k+t]}, \boldsymbol{\Theta}^{[k+t]}) + \sum_{i<j} p_\gamma(\|\boldsymbol{\delta}^{[k+t]}\|; \lambda) \\
&\quad + (\boldsymbol{\nu}^{[k]})^\top (\boldsymbol{B}^\top \boldsymbol{\Theta}^{[k+t]} - \boldsymbol{\delta}^{[k+t]}) \\
&\quad + \kappa \sum_{i=1}^{t-1} (\boldsymbol{B}^\top \boldsymbol{\Theta}^{[k+i]} - \boldsymbol{\delta}^{[k+i]})^\top (\boldsymbol{B}^\top \boldsymbol{\Theta}^{[k+t]} - \boldsymbol{\delta}^{[k+t]}) \\
&\quad + \frac{\kappa}{2} \left\| \boldsymbol{B}^\top \boldsymbol{\Theta}^{[k+t]} - \boldsymbol{\delta}^{[k+t]} \right\|^2 \\
&\leq f^{[k+t]}.
\end{aligned}$$

We are going to show that $\lim_{k\to\infty} \left\| \boldsymbol{r}^{[k]} \right\|^2 = 0$ is a necessary condition for the inequality

$$\lim_{k\to\infty} L(\boldsymbol{\beta}^{[k+t]}, \boldsymbol{\Theta}^{[k+t]}, \boldsymbol{\delta}^{[k+t]}, \boldsymbol{\nu}^{[k+t-1]}) \leq \lim_{k\to\infty} f^{[k+t]}$$

for all $t \geq 0$.

By applying the results in Theorem 4.1 of (Tseng, 2001), the sequence $(\boldsymbol{\beta}^{[k]}, \boldsymbol{\Theta}^{[k]}, \boldsymbol{\delta}^{[k]})$ has a limit point, denoted by $(\boldsymbol{\beta}^{[*]}, \boldsymbol{\Theta}^{[*]}, \boldsymbol{\delta}^{[*]})$, since the objective function $L(\boldsymbol{\beta}, \boldsymbol{\Theta}, \boldsymbol{\delta}, \boldsymbol{\nu})$ is differentiable w.r.t. $(\boldsymbol{\beta}, \boldsymbol{\Theta})$ and is convex w.r.t. $\boldsymbol{\delta}$. Define

$$f^{[*]} = \lim_{k\to\infty} f^{[k+1]} = \lim_{k\to\infty} f^{[k+t]} = \inf_{\boldsymbol{B}^\top \boldsymbol{\Theta}^{[*]} - \boldsymbol{\delta}^{[*]} = \boldsymbol{0}} \left\{ L_N^{\mathrm{CD}}(\boldsymbol{\beta}^{[*]}, \boldsymbol{\Theta}^{[*]}) + \sum_{i<j} p_\gamma(\|\boldsymbol{\delta}^{[*]}\|; \lambda) \right\}.$$

For all $t \geq 0$, we have

$$\begin{aligned}
\lim_{k\to\infty} L(\boldsymbol{\beta}^{[k+t]}, \boldsymbol{\Theta}^{[k+t]}, \boldsymbol{\delta}^{[k+t]}, \boldsymbol{\nu}^{[k+t-1]}) &= L_N^{\mathrm{CD}}(\boldsymbol{\beta}^{[*]}, \boldsymbol{\Theta}^{[*]}) + \sum_{i<j} p_\gamma(\|\boldsymbol{\delta}_{ij}^{[*]}\|; \lambda) \\
&\quad + \lim_{k\to\infty} (\boldsymbol{\nu}^{[k]})^\top (\boldsymbol{B}^\top \boldsymbol{\Theta}^{[*]} - \boldsymbol{\delta}^{[*]}) + (t - \frac{1}{2})\kappa \left\| \boldsymbol{B}^\top \boldsymbol{\Theta}^{[*]} - \boldsymbol{\delta}^{[*]} \right\|^2 \\
&\leq f^{[*]}.
\end{aligned}$$

Note that $f^{[*]}$ is a constant by definition, and thus we must have $\lim_{k\to\infty} \left\| \boldsymbol{r}^{[k]} \right\|^2 = \left\| \boldsymbol{B}^\top \boldsymbol{\Theta}^{[*]} - \boldsymbol{\delta}^{[*]} \right\|^2 = 0$ to make the last inequality hold for all $t \geq 0$.

S.7. Proof of Theorem 4.2

In this section, we prove the results in Theorem 4.2. We start with introducing the folllowing notations. The objective functions for the proposed and oracle estimators can be written respectively by

$$Q_N(\boldsymbol{\beta}, \boldsymbol{\Theta}) = L_N(\boldsymbol{\beta}, \boldsymbol{\Theta}) + P_N(\boldsymbol{\Theta}) \quad \text{and} \quad Q_N^{\mathcal{G}}(\boldsymbol{\beta}, \boldsymbol{\alpha}) = L_N^{\mathcal{G}}(\boldsymbol{\beta}, \boldsymbol{\alpha}) + P_N^{\mathcal{G}}(\boldsymbol{\alpha}),$$

where

$$L_N(\boldsymbol{\beta}, \boldsymbol{\Theta}) = \frac{1}{2N}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{\Theta})^\top \boldsymbol{W}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{\Theta}),\ P_N(\boldsymbol{\Theta}) = \lambda \sum_{1\leq i<j\leq N} \rho(\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_j\|),$$

$$L_N^{\mathcal{G}}(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \frac{1}{2N}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{A}\boldsymbol{\alpha})^\top \boldsymbol{W}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{A}\boldsymbol{\alpha}),\ P_N^{\mathcal{G}}(\boldsymbol{\Theta}) = \lambda \sum_{1\leq s<s'\leq S} M_s M_{s'} \rho(\|\boldsymbol{\alpha}_s - \boldsymbol{\alpha}_{s'}\|),$$

$\rho(t) = \lambda^{-1} p_\gamma(t;\lambda)$ and M_s is the number of units in the s-th subgroup. Recall that the set

$$\mathcal{M}_{\mathcal{G}} = \left\{ \boldsymbol{\Theta} \in \mathbb{R}^{Mq} : \boldsymbol{\theta}_i = \boldsymbol{\theta}_j \text{ for any } L_i = L_j = s, s = 1, \ldots, S \right\}$$

is the collection of stacked heterogeneous effect vector $\boldsymbol{\Theta} = (\boldsymbol{\theta}_i^\top)_{i=1,\ldots,M}^\top$ with only S distinct values of $\boldsymbol{\theta}_i$'s. For each $\boldsymbol{\theta} \in \mathcal{M}_{\mathcal{G}}$, there exists a $Mq \times Sq$ label matrix $\mathbb{A}$ and $\boldsymbol{\alpha} \in \mathbb{R}^{Sq}$ such that $\boldsymbol{\Theta} = \mathbb{A}\boldsymbol{\alpha}$ and $\boldsymbol{\alpha} = (\mathbb{A}^\top \mathbb{A})^{-1} \mathbb{A}^\top \boldsymbol{\Theta}$. Hence, let $T : \mathcal{M}_{\mathcal{G}} \to \mathbb{R}^{Sq}$ be the mapping such that $T(\boldsymbol{\Theta})$ is the (Sq)-vector stacking the S different $\boldsymbol{\theta}_i$'s in $\boldsymbol{\Theta}$, and its inverse mapping $T^{-1} : \mathbb{R}^{Sq} \to \mathcal{M}_{\mathcal{G}}$ is well-defined. Let $T^* : \mathbb{R}^{Mq} \to \mathbb{R}^{Sq}$ be the mapping such that $T^*(\boldsymbol{\Theta}) = \left\{ M_s^{-1} \sum_{i:L_i=s} \boldsymbol{\theta}_i \right\}_{s=1}^S$. For any $\boldsymbol{\Theta} \in \mathbb{R}^{Mq}$, define $\boldsymbol{\Theta}^* = T^{-1}(T^*(\boldsymbol{\Theta}))$.

By the above definitions, it clearly follows that $T(\boldsymbol{\Theta}) = T^*(\boldsymbol{\Theta})$ and $P_N(\boldsymbol{\Theta}) = P_N^{\mathcal{G}}(T(\boldsymbol{\Theta}))$ for every $\boldsymbol{\Theta} \in \mathcal{M}_{\mathcal{G}}$, and that $P_N(T^{-1}(\boldsymbol{\alpha})) = P_N^{\mathcal{G}}(\boldsymbol{\alpha})$ for every $\boldsymbol{\alpha} \in \mathbb{R}^{Sq}$. Moreover, we can write

$$\begin{aligned} Q_N(\boldsymbol{\beta}, \boldsymbol{\Theta}) &= Q_N^{\mathcal{G}}(\boldsymbol{\beta}, T(\boldsymbol{\Theta})) \quad \text{for every } \boldsymbol{\Theta} \in \mathcal{M}_{\mathcal{G}}, \\ Q_N^{\mathcal{G}}(\boldsymbol{\beta}, \boldsymbol{\alpha}) &= Q_N(\boldsymbol{\beta}, T^{-1}(\boldsymbol{\alpha})) \quad \text{for every } \boldsymbol{\alpha} \in \mathbb{R}^{Sq}. \end{aligned} \tag{S.33}$$

Consider a neighborhood of $(\boldsymbol{\beta}_0, \boldsymbol{\Theta}_0)$:

$$\Upsilon = \left\{ \boldsymbol{\beta} \in \mathbb{R}^p, \boldsymbol{\Theta} \in \mathbb{R}^{Mq} : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| \le \phi_N, \max_{1 \le i \le M} \|\boldsymbol{\theta}_i - \boldsymbol{\theta}_{i0}\| \le \phi_N \right\}.$$

By Theorem 4.1, there is an event E_1 with $P(E_1) \ge 1 - (p+Sq)(4N^{-1} + e^{-g_{\max}}) - 2Me^{-n_{\min}}$ such that $(\widehat{\boldsymbol{\beta}}_{\mathrm{OR}}, \widehat{\boldsymbol{\Theta}}_{\mathrm{OR}}) \in \Upsilon$ on event E_1. We show that $(\widehat{\boldsymbol{\beta}}_{\mathrm{OR}}, \widehat{\boldsymbol{\Theta}}_{\mathrm{OR}})$ is a strict local minimizer of the objective function (5) with probablity tending to 1 through the following two steps.

(i) On the event E_1, $Q_N(\boldsymbol{\beta}, \boldsymbol{\Theta}^*) > Q_N(\widehat{\boldsymbol{\beta}}_{\mathrm{OR}}, \widehat{\boldsymbol{\Theta}}_{\mathrm{OR}})$ for any $(\boldsymbol{\beta}, \boldsymbol{\Theta}) \in \Upsilon$ and $(\boldsymbol{\beta}, \boldsymbol{\Theta}^*) \ne (\widehat{\boldsymbol{\beta}}_{\mathrm{OR}}, \widehat{\boldsymbol{\Theta}}_{\mathrm{OR}})$.

(ii) There is an event E_2 with $P(E_2) \ge 1 - Me^{-g_{\min}}$. On $E_1 \cap E_2$, there is a neighborhood of $(\widehat{\boldsymbol{\beta}}_{\mathrm{OR}}, \widehat{\boldsymbol{\Theta}}_{\mathrm{OR}})$, denoted by Υ_N, such that $Q_N(\boldsymbol{\beta}, \boldsymbol{\Theta}) \ge Q_N(\boldsymbol{\beta}, \boldsymbol{\Theta}^*)$ for any $(\boldsymbol{\beta}, \boldsymbol{\Theta}) \in \Upsilon_N \cap \Upsilon$ for sufficiently large N.

By resutls (i) and (ii), we have $Q_N(\boldsymbol{\beta}, \boldsymbol{\Theta}) > Q_N(\widehat{\boldsymbol{\beta}}_{\mathrm{OR}}, \widehat{\boldsymbol{\Theta}}_{\mathrm{OR}})$ for any $(\boldsymbol{\beta}, \boldsymbol{\Theta}) \in \Upsilon_N \cap \Upsilon$ and $(\boldsymbol{\beta}, \boldsymbol{\Theta}^*) \ne (\widehat{\boldsymbol{\beta}}_{\mathrm{OR}}, \widehat{\boldsymbol{\Theta}}_{\mathrm{OR}})$ on the event $E_1 \cap E_2$, so that $(\widehat{\boldsymbol{\beta}}_{\mathrm{OR}}, \widehat{\boldsymbol{\Theta}}_{\mathrm{OR}})$ is a strict local minimizer of $Q_N(\boldsymbol{\beta}, \boldsymbol{\Theta})$ given in (5) over the event $E_1 \cap E_2$ with $P(E_1 \cap E_2) \ge 1 - (p+Sq)(4N^{-1} + e^{-g_{\max}}) - M(2e^{-n_{\min}} + e^{-g_{\min}})$ for sufficiently large N.

We first show $P_N^{\mathcal{G}}(T^*(\boldsymbol{\Theta})) = K_N$ for any $\boldsymbol{\Theta} \in \Upsilon$, where K_N is a constant independent of $\boldsymbol{\Theta}$. Let $T^*(\boldsymbol{\Theta}) = \boldsymbol{\alpha} = (\boldsymbol{\alpha}_s^\top)_{s=1,\ldots,S}^\top$. Since $\rho(t) = \lambda^{-1} p_\gamma(t;\lambda)$ is constant for $t \ge a\lambda$ according to Assumption 4.3, it suffices to show that $\|\boldsymbol{\alpha}_s - \boldsymbol{\alpha}_{s'}\| > a\lambda$ for all $s, s' \in \{1, \ldots, S\}$. Since

$$\begin{aligned} \max_{1 \le s \le S} \|\boldsymbol{\alpha}_s - \boldsymbol{\alpha}_{s0}\| &= \max_{1 \le s \le S} \left\| \frac{1}{M_s} \sum_{i:L_i=s} (\boldsymbol{\theta}_i - \boldsymbol{\theta}_{i0}) \right\| \\ &\le \max_{1 \le s \le S} \frac{1}{M_s} \sum_{i:L_i=s} \|\boldsymbol{\theta}_i - \boldsymbol{\theta}_{i0}\| \\ &\le \max_{1 \le s \le S} \|\boldsymbol{\theta}_i - \boldsymbol{\theta}_{i0}\| \\ &\le \phi_N, \end{aligned} \tag{S.34}$$

then for all $s, s' \in \{1, \ldots, S\}$ we have

$$\|\boldsymbol{\alpha}_s - \boldsymbol{\alpha}_{s'}\| \ge \|\boldsymbol{\alpha}_{s0} - \boldsymbol{\alpha}_{s'0}\| - 2 \max_{1 \le s \le S} \|\boldsymbol{\alpha}_s - \boldsymbol{\alpha}_{s0}\| \ge \Delta_N - 2\phi_N > a\lambda, \tag{S.35}$$

where the last inequality follows from the assumption that $\Delta_N \gg a\lambda \gg \phi_N$. Accordingly, we have $P_N^{\mathcal{G}}(T^*(\boldsymbol{\Theta})) = K_N$ for some constant $K_N > 0$ and hence $Q_N^{\mathcal{G}}(\boldsymbol{\beta}, T^*(\boldsymbol{\Theta})) = L_N^{\mathcal{G}}(\boldsymbol{\beta}, T^*(\boldsymbol{\Theta})) + K_N$ for all $(\boldsymbol{\beta}, \boldsymbol{\Theta}) \in \Upsilon$. Since $(\widehat{\boldsymbol{\beta}}_{\mathrm{OR}}, \widehat{\boldsymbol{\alpha}}_{\mathrm{OR}})$ is the unique global minimizer of $L_N^{\mathcal{G}}(\boldsymbol{\beta}, \boldsymbol{\alpha})$, we have $L_N^{\mathcal{G}}(\boldsymbol{\beta}, T^*(\boldsymbol{\Theta})) > L_N^{\mathcal{G}}(\widehat{\boldsymbol{\beta}}_{\mathrm{OR}}, \widehat{\boldsymbol{\alpha}}_{\mathrm{OR}})$ for all $(\boldsymbol{\beta}, T^*(\boldsymbol{\Theta})) \ne (\widehat{\boldsymbol{\beta}}_{\mathrm{OR}}, \widehat{\boldsymbol{\alpha}}_{\mathrm{OR}})$.

It then follows that $Q_N^{\mathcal{G}}(\boldsymbol{\beta}, T^*(\boldsymbol{\Theta})) > Q_N^{\mathcal{G}}(\widehat{\boldsymbol{\beta}}_{\mathrm{OR}}, \widehat{\boldsymbol{\alpha}}_{\mathrm{OR}})$ for all $T^*(\boldsymbol{\Theta}) \neq \widehat{\boldsymbol{\alpha}}_{\mathrm{OR}}$. Let $\widehat{\boldsymbol{\Theta}}_{\mathrm{OR}} = \boldsymbol{A}\widehat{\boldsymbol{\alpha}}_{\mathrm{OR}}$, and by (S.33) we have $Q_N^{\mathcal{G}}(\widehat{\boldsymbol{\beta}}_{\mathrm{OR}}, \widehat{\boldsymbol{\alpha}}_{\mathrm{OR}}) = Q_N(\widehat{\boldsymbol{\beta}}_{\mathrm{OR}}, \widehat{\boldsymbol{\Theta}}_{\mathrm{OR}})$ and $Q_N^{\mathcal{G}}(\boldsymbol{\beta}, T^*(\boldsymbol{\Theta})) = Q_N(\boldsymbol{\beta}, T^{-1}(T^*(\boldsymbol{\Theta}))) = Q_N(\boldsymbol{\beta}, \boldsymbol{\Theta}^*)$. Consequently, we have $Q_N(\boldsymbol{\beta}, \boldsymbol{\Theta}^*) > Q_N(\widehat{\boldsymbol{\beta}}_{\mathrm{OR}}, \widehat{\boldsymbol{\Theta}}_{\mathrm{OR}})$ for all $(\boldsymbol{\beta}, \boldsymbol{\Theta}) \in \Upsilon$ and $(\boldsymbol{\beta}, \boldsymbol{\Theta}^*) \neq (\widehat{\boldsymbol{\beta}}_{\mathrm{OR}}, \widehat{\boldsymbol{\Theta}}_{\mathrm{OR}})$, and the result (i) follows.

Next we prove the result in (ii). For positive sequence $\{t_N\}$, let

$$\Upsilon_N = \{\boldsymbol{\Theta} : \max_{1\leq i\leq M} \|\boldsymbol{\theta}_i - \widehat{\boldsymbol{\Theta}}_{\mathrm{OR},i}\| \leq t_N\}.$$

For $(\boldsymbol{\beta}, \boldsymbol{\Theta}) \in \Upsilon_N \cap \Upsilon$, Taylor's expansion leads to

$$Q_N(\boldsymbol{\beta}, \boldsymbol{\Theta}) - Q_N(\boldsymbol{\beta}, \boldsymbol{\Theta}^*) = \underbrace{-N^{-1}\sum_{i=1}^{M}(\boldsymbol{y}_i - \boldsymbol{x}_i\boldsymbol{\beta} - \boldsymbol{z}_i\widetilde{\boldsymbol{\theta}}_i)^\top \boldsymbol{W}_i\boldsymbol{z}_i(\boldsymbol{\theta}_i - \boldsymbol{\theta}_i^*)}_{\Gamma_1} + \underbrace{\sum_{i=1}^{M}\frac{\partial P_N(\widetilde{\boldsymbol{\Theta}})}{\partial \boldsymbol{\theta}_i}(\boldsymbol{\theta}_i - \boldsymbol{\theta}_i^*)}_{\Gamma_2},$$

where $\widetilde{\boldsymbol{\Theta}} = \varsigma\boldsymbol{\Theta} + (1-\varsigma)\boldsymbol{\Theta}^*$ for some $\varsigma \in (0,1)$.

We firstly deal with Γ_2. Observe that

$$\begin{aligned}
\Gamma_2 &= \lambda\sum_{i\neq j}\frac{\rho'(\|\widetilde{\boldsymbol{\theta}}_i - \widetilde{\boldsymbol{\theta}}_j\|)}{\|\widetilde{\boldsymbol{\theta}}_i - \widetilde{\boldsymbol{\theta}}_j\|}(\widetilde{\boldsymbol{\theta}}_i - \widetilde{\boldsymbol{\theta}}_j)^\top(\boldsymbol{\theta}_i - \boldsymbol{\theta}_i^*) \\
&= \lambda\sum_{i<j}\frac{\rho'(\|\widetilde{\boldsymbol{\theta}}_i - \widetilde{\boldsymbol{\theta}}_j\|)}{\|\widetilde{\boldsymbol{\theta}}_i - \widetilde{\boldsymbol{\theta}}_j\|}(\widetilde{\boldsymbol{\theta}}_i - \widetilde{\boldsymbol{\theta}}_j)^\top(\boldsymbol{\theta}_i - \boldsymbol{\theta}_i^*) + \lambda\sum_{i<j}\frac{\rho'(\|\widetilde{\boldsymbol{\theta}}_j - \widetilde{\boldsymbol{\theta}}_i\|)}{\|\widetilde{\boldsymbol{\theta}}_j - \widetilde{\boldsymbol{\theta}}_i\|}(\widetilde{\boldsymbol{\theta}}_j - \widetilde{\boldsymbol{\theta}}_i)^\top(\boldsymbol{\theta}_j - \boldsymbol{\theta}_j^*) \\
&= \lambda\sum_{i<j}\frac{\rho'(\|\widetilde{\boldsymbol{\theta}}_i - \widetilde{\boldsymbol{\theta}}_j\|)}{\|\widetilde{\boldsymbol{\theta}}_i - \widetilde{\boldsymbol{\theta}}_j\|}(\widetilde{\boldsymbol{\theta}}_i - \widetilde{\boldsymbol{\theta}}_j)^\top(\boldsymbol{\theta}_i - \boldsymbol{\theta}_i^*) - \lambda\sum_{i<j}\frac{\rho'(\|\widetilde{\boldsymbol{\theta}}_i - \widetilde{\boldsymbol{\theta}}_j\|)}{\|\widetilde{\boldsymbol{\theta}}_i - \widetilde{\boldsymbol{\theta}}_j\|}(\widetilde{\boldsymbol{\theta}}_i - \widetilde{\boldsymbol{\theta}}_j)^\top(\boldsymbol{\theta}_j - \boldsymbol{\theta}_j^*) \\
&= \lambda\sum_{i<j}\frac{\rho'(\|\widetilde{\boldsymbol{\theta}}_i - \widetilde{\boldsymbol{\theta}}_j\|)}{\|\widetilde{\boldsymbol{\theta}}_i - \widetilde{\boldsymbol{\theta}}_j\|}(\widetilde{\boldsymbol{\theta}}_i - \widetilde{\boldsymbol{\theta}}_j)^\top\left[(\boldsymbol{\theta}_i - \boldsymbol{\theta}_i^*) - (\boldsymbol{\theta}_j - \boldsymbol{\theta}_j^*)\right].
\end{aligned} \tag{S.36}$$

When $L_i = L_j = s$, we have $\boldsymbol{\theta}_i^* = \boldsymbol{\theta}_j^*$ and $\widetilde{\boldsymbol{\theta}}_i - \widetilde{\boldsymbol{\theta}}_j = \varsigma(\boldsymbol{\theta}_i - \boldsymbol{\theta}_j)$, and therefore

$$\begin{aligned}
\Gamma_2 = \lambda\sum_{s=1}^{S}\sum_{\substack{L_i=L_j=s\\ i<j}}\rho'(\|\widetilde{\boldsymbol{\theta}}_i - \widetilde{\boldsymbol{\theta}}_j\|)\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_j\| \\
+ \lambda\sum_{s<s'}\sum_{\substack{L_i=s\\ L_j=s'}}\frac{\rho'(\|\widetilde{\boldsymbol{\theta}}_i - \widetilde{\boldsymbol{\theta}}_j\|)}{\|\widetilde{\boldsymbol{\theta}}_i - \widetilde{\boldsymbol{\theta}}_j\|}(\widetilde{\boldsymbol{\theta}}_i - \widetilde{\boldsymbol{\theta}}_j)^\top\left[(\boldsymbol{\theta}_i - \boldsymbol{\theta}_i^*) - (\boldsymbol{\theta}_j - \boldsymbol{\theta}_j^*)\right].
\end{aligned}$$

We look at the second term in Γ_2. Since $\boldsymbol{\Theta} \in \Upsilon$ and $\boldsymbol{\Theta}^* = T^{-1}(\boldsymbol{\alpha}) \in \mathcal{M}_{\mathcal{G}}$, it holds

$$\max_{1\leq i\leq M}\left\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_{i0}\right\| \leq \phi_N \quad \text{and} \quad \max_{1\leq i\leq M}\left\|\boldsymbol{\theta}_i^* - \boldsymbol{\theta}_{i0}\right\| = \max_{1\leq s\leq S}\left\|\boldsymbol{\alpha}_s - \boldsymbol{\alpha}_{s0}\right\| \leq \phi_N,$$

where the inequality follows from (S.34). Since $\widetilde{\boldsymbol{\Theta}} = \varsigma\boldsymbol{\Theta} + (1-\varsigma)\boldsymbol{\Theta}^*$, we have

$$\max_{1\leq i\leq M}\left\|\widetilde{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_{i0}\right\| \leq \varsigma\max_{1\leq i\leq M}\left\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_{i0}\right\| + (1-\varsigma)\max_{1\leq i\leq M}\left\|\boldsymbol{\theta}_i^* - \boldsymbol{\theta}_{i0}\right\| \leq \varsigma\phi_N + (1-\varsigma)\phi_N = \phi_N.$$

Then for $L_i = s$ and $L_j = s'$ with $s \neq s'$, it follows that

$$\left\|\widetilde{\boldsymbol{\theta}}_i - \widetilde{\boldsymbol{\theta}}_j\right\| \geq \min_{L_i=s, L_j=s'}\left\|\boldsymbol{\theta}_{i0} - \boldsymbol{\theta}_{j0}\right\| - 2\max_{1\leq i\leq M}\left\|\widetilde{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_{i0}\right\| \geq \Delta_N - 2\phi_N \gg a\lambda,$$

and accordingly $\rho'\big(\|\widetilde{\boldsymbol{\theta}}_i - \widetilde{\boldsymbol{\theta}}_j\|\big) = 0$. As a result, the second term in Γ_2 is zero and we have

$$\Gamma_2 = \lambda \sum_{s=1}^{S} \sum_{\substack{L_i = L_j = s \\ i<j}} \rho'\big(\|\widetilde{\boldsymbol{\theta}}_i - \widetilde{\boldsymbol{\theta}}_j\|\big) \|\boldsymbol{\theta}_i - \boldsymbol{\theta}_j\|.$$

For the first term in Γ_2, since $\boldsymbol{\Theta} \in \Upsilon_N$, the same reasoning as (S.34) yields

$$\max_{1\le i\le M} \big\|\boldsymbol{\theta}_i^* - \widehat{\boldsymbol{\theta}}_{\mathrm{OR},i}\big\| = \max_{1\le i\le M} \big\|\boldsymbol{\alpha}_s - \widehat{\boldsymbol{\alpha}}_{\mathrm{OR},s}\big\| \le \max_{1\le i\le M} \big\|\boldsymbol{\theta}_i - \widehat{\boldsymbol{\theta}}_{\mathrm{OR},i}\big\| \le t_N.$$

Since $\widetilde{\boldsymbol{\theta}}_i = \varsigma\boldsymbol{\theta}_i + (1-\varsigma)\boldsymbol{\theta}_i^*$ for some $\varsigma \in (0,1)$, it holds $\widetilde{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i^* = \varsigma(\boldsymbol{\theta}_i - \boldsymbol{\theta}_i^*)$. For $L_i = L_j$, due to $\boldsymbol{\theta}_i^* = \boldsymbol{\theta}_j^*$, we have

$$\begin{aligned}
\max_{1\le i\le M} \big\|\widetilde{\boldsymbol{\theta}}_i - \widetilde{\boldsymbol{\theta}}_j\big\| &\le \max_{1\le i\le M} \Big(\big\|\widetilde{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i^*\big\| + \big\|\widetilde{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j^*\big\|\Big) \\
&\le 2 \max_{1\le i\le M} \big\|\widetilde{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i^*\big\| \\
&\le 2 \max_{1\le i\le M} \big\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_i^*\big\| \\
&\le 2 \Big(\max_{1\le i\le M} \big\|\boldsymbol{\theta}_i - \widehat{\boldsymbol{\theta}}_{\mathrm{OR},i}\big\| + \max_{1\le i\le M} \big\|\boldsymbol{\theta}_i^* - \widehat{\boldsymbol{\theta}}_{\mathrm{OR},i}\big\|\Big) \\
&\le 4t_N.
\end{aligned}$$

The concavity of $\rho(\cdot)$ then implies that $\rho'(\|\widetilde{\boldsymbol{\theta}}_i - \widetilde{\boldsymbol{\theta}}_j\|) \ge \rho'(4t_N)$. In summary, we have

$$\Gamma_2 \ge \lambda \sum_{s=1}^{S} \sum_{\substack{L_i = L_j = s \\ i<j}} \rho'(4t_N) \|\boldsymbol{\theta}_i - \boldsymbol{\theta}_j\|. \tag{S.37}$$

Now we analyze Γ_1. For $i = 1, \ldots, M$, let

$$\boldsymbol{h}_i = (\boldsymbol{y}_i - \boldsymbol{x}_i\boldsymbol{\beta} - \boldsymbol{z}_i\widetilde{\boldsymbol{\theta}}_i)^\top \boldsymbol{W}_i\boldsymbol{z}_i = \boldsymbol{z}_i^\top \boldsymbol{W}_i \Big[\boldsymbol{z}_i\boldsymbol{u}_i + \boldsymbol{\varepsilon}_i + \boldsymbol{x}_i(\boldsymbol{\beta}_0 - \boldsymbol{\beta}) + \boldsymbol{z}_i(\boldsymbol{\theta}_{i0} - \widetilde{\boldsymbol{\theta}}_i)\Big].$$

Recall that M_s is the number of units in the s-th subgroup. By the definition of $\boldsymbol{\theta}_i^*$, it holds

$$\boldsymbol{\theta}_i - \boldsymbol{\theta}_i^* = \boldsymbol{\theta}_i - \sum_{j:L_j = L_i} \frac{\boldsymbol{\theta}_i}{M_s} = \sum_{j:L_j = L_i} \frac{(\boldsymbol{\theta}_i - \boldsymbol{\theta}_j)}{M_s}.$$

Then we have

$$\begin{aligned}
\Gamma_1 &= -N^{-1}\sum_{i=1}^{M}(\boldsymbol{y}_i - \boldsymbol{x}_i\boldsymbol{\beta} - \boldsymbol{z}_i\widetilde{\boldsymbol{\theta}}_i)^\top \boldsymbol{W}_i\boldsymbol{z}_i(\boldsymbol{\theta}_i - \boldsymbol{\theta}_i^*) \\
&= -N^{-1}\sum_{i=1}^{M}\boldsymbol{h}_i^\top(\boldsymbol{\theta}_i - \boldsymbol{\theta}_i^*) \\
&= -N^{-1}\sum_{s=1}^{S}\sum_{L_i=L_j=s}\frac{\boldsymbol{h}_i^\top(\boldsymbol{\theta}_i - \boldsymbol{\theta}_j)}{M_s} \\
&= -N^{-1}\sum_{s=1}^{S}\sum_{L_i=L_j=s}\frac{\boldsymbol{h}_i^\top(\boldsymbol{\theta}_i - \boldsymbol{\theta}_j)}{2M_s} - \sum_{s=1}^{S}\sum_{L_i=L_j=s}\frac{\boldsymbol{h}_j^\top(\boldsymbol{\theta}_i - \boldsymbol{\theta}_j)}{2M_s} \\
&= -N^{-1}\sum_{s=1}^{S}\sum_{L_i=L_j=s}\frac{(\boldsymbol{h}_i - \boldsymbol{h}_j)^\top(\boldsymbol{\theta}_i - \boldsymbol{\theta}_j)}{2M_s} \\
&= -N^{-1}\sum_{s=1}^{S}\sum_{\substack{L_i=L_j=s\\ i<j}}\frac{(\boldsymbol{h}_i - \boldsymbol{h}_j)^\top(\boldsymbol{\theta}_i - \boldsymbol{\theta}_j)}{M_s} \\
&\geq -N^{-1}\sum_{s=1}^{S}\sum_{\substack{L_i=L_j=s\\ i<j}}\max_{i,j}\|\boldsymbol{h}_i - \boldsymbol{h}_j\|\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_j\|. \qquad \text{(S.38)}
\end{aligned}$$

Combining (S.37) and (S.38) yields

$$\begin{aligned}
Q_N(\boldsymbol{\beta},\boldsymbol{\Theta}) - Q_N(\boldsymbol{\beta},\boldsymbol{\Theta}^*) = \Gamma_1 + \Gamma_2 &\geq \sum_{s=1}^{S}\sum_{\substack{L_i=L_j=s\\ i<j}}\left[\lambda\rho'(4t_N) - N^{-1}\max_{i,j}\|\boldsymbol{h}_i - \boldsymbol{h}_j\|\right]\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_j\| \\
&\gtrsim \sum_{s=1}^{S}\sum_{\substack{L_i=L_j=s\\ i<j}}\left[\lambda - N^{-1}\max_{i,j}\|\boldsymbol{h}_i - \boldsymbol{h}_j\|\right]\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_j\|,
\end{aligned}$$

where the last inequality holds by taking $t_N = o(1)$ so that $\rho'(4t_N) \to 1$. It is then left to show $\lambda > N^{-1}\max_{i,j}\|\boldsymbol{h}_i - \boldsymbol{h}_j\|$ to conclude the results in (ii).

Recall that $\boldsymbol{W}_i = \boldsymbol{V}_i^2$. On the event E_1, we can write

$$\begin{aligned}
\max_{i,j}\|\boldsymbol{h}_i - \boldsymbol{h}_j\| &\leq 2\max_{1\leq i\leq M}\|\boldsymbol{h}_i\| \\
&\leq 2\max_{1\leq i\leq M}\Big\{\|\boldsymbol{V}_i\boldsymbol{z}_i\|\|\boldsymbol{V}_i(\boldsymbol{z}_i\boldsymbol{u}_i + \boldsymbol{\varepsilon}_i)\| + \big(\|\boldsymbol{z}_i^\top\boldsymbol{W}_i\boldsymbol{x}_i\| + \|\boldsymbol{z}_i^\top\boldsymbol{W}_i\boldsymbol{z}_i\|\big)\phi_N\Big\} \\
&\leq 2\max_{1\leq i\leq M}\Big\{\|\boldsymbol{V}_i\boldsymbol{z}_i\|\|\boldsymbol{V}_i(\boldsymbol{z}_i\boldsymbol{u}_i + \boldsymbol{\varepsilon}_i)\| + \big(\|\boldsymbol{V}_i\boldsymbol{z}_i\|\|\boldsymbol{V}_i\|\|\boldsymbol{x}_i\| + \|\boldsymbol{z}_i^\top\boldsymbol{W}_i\boldsymbol{z}_i\|\big)\phi_N\Big\} \\
&\lesssim 2\max_{1\leq i\leq M}\Big\{\tau^{-1}\big\|\boldsymbol{V}_i(\boldsymbol{z}_i\boldsymbol{u}_i + \boldsymbol{\varepsilon}_i)\big\| + \big(\overline{C}_f^{1/2}\sigma_\varepsilon^{-1}\tau^{-1}n_i^{1/2} + \tau^{-2}\big)\phi_N\Big\},
\end{aligned}$$

where the last inequality follows from (S.17) and (S.15) under events E_{CB} and E_{EB}. Moreover, since $\boldsymbol{V}_i(\boldsymbol{z}_i\boldsymbol{u}_i + \boldsymbol{\varepsilon}_i)$ has sub-Gaussian tails according to (S.19), applying union bound and (S.42) yields

$$P\Big(\max_{1\leq i\leq M}\big\|\boldsymbol{V}_i(\boldsymbol{z}_i\boldsymbol{u}_i + \boldsymbol{\varepsilon}_i)\big\| > \sqrt{5c_5 g_{\min}}\Big) \leq Me^{-g_{\min}},$$

where $c_5 > 0$ is a constant. Then we have

$$\begin{aligned}\max_{i,j}\|\boldsymbol{h}_i-\boldsymbol{h}_j\| &\lesssim 2\left[\sqrt{5}c_5^{1/2}\tau^{-1}g_{\min}^{1/2}+\left(\overline{C}_f^{1/2}\sigma_\varepsilon^{-1}\tau^{-1}\max_{1\le i\le M}n_i^{1/2}+\tau^{-2}\right)\phi_N\right]\\ &\le 2g_{\min}^{1/2}\left[\sqrt{5}c_5^{1/2}\tau^{-1}+\left(\overline{C}_f^{1/2}\sigma_\varepsilon^{-1}\tau^{-1}+\tau^{-2}\right)\phi_N\right]\end{aligned}$$

with probability at least $1-Me^{-g_{\min}}$. Hence, there is an event E_2 with $P(E_2)\ge 1-Me^{-g_{\min}}$ and, on the event E_2, we have

$$\frac{N^{-1}\max_{i,j}\|\boldsymbol{h}_i-\boldsymbol{h}_j\|}{\phi_N}\lesssim\frac{2N^{-1}g_{\min}^{1/2}\left[\sqrt{5}c_5^{1/2}\tau^{-1}+\left(\overline{C}_f^{1/2}\sigma_\varepsilon^{-1}\tau^{-1}+\tau^{-2}\right)\phi_N\right]}{\phi_N}=O\big(N^{-1}g_{\min}^{1/2}\big)=o(1),$$

which indicates that

$$\lambda\gg\phi_N\gg N^{-1}\max_{i,j}\|\boldsymbol{h}_i-\boldsymbol{h}_j\|. \tag{S.39}$$

In summary, we conclude the results in (ii) and the proof is complete.

S.8. Proof of (ii) in Corollary 4.1

Since the asymptotic equivalence between the proposed estimator and the oracle estimator has been shown in Theorem 4.2, it suffices to show, for any p-vector $\boldsymbol{v}$ and q-vector $\boldsymbol{v}_q$,

$$\boldsymbol{v}_p^\top\mathsf{Cov}\big(\widehat{\boldsymbol{\beta}}_{\mathrm{OR}}\big)\boldsymbol{v}_p\le\boldsymbol{v}_p^\top\mathsf{Cov}\big(\breve{\boldsymbol{\beta}}_i\big)\boldsymbol{v}_p\quad\text{and}\quad\boldsymbol{v}_q^\top\mathsf{Cov}\big(\widehat{\boldsymbol{\theta}}_{\mathrm{OR},i}\big)\boldsymbol{v}_q\le\boldsymbol{v}_q^\top\mathsf{Cov}\big(\breve{\boldsymbol{\theta}}_i\big)\boldsymbol{v}_q, \tag{S.40}$$

for all $i=1,\ldots,M$. Recall that

$$\mathsf{Cov}\begin{pmatrix}\widehat{\boldsymbol{\beta}}_{\mathrm{OR}}\\ \widehat{\boldsymbol{\alpha}}_{\mathrm{OR}}\end{pmatrix}=\Big[(\boldsymbol{X},\boldsymbol{ZA})^\top\boldsymbol{W}(\boldsymbol{X},\boldsymbol{ZA})\Big]^{-1}\quad\text{and}\quad\mathsf{Cov}\begin{pmatrix}\breve{\boldsymbol{\beta}}_i\\ \breve{\boldsymbol{\theta}}_i\end{pmatrix}=\big[(\boldsymbol{x}_i,\boldsymbol{z}_i)^\top\boldsymbol{W}_i(\boldsymbol{x}_i,\boldsymbol{z}_i)\big]^{-1}.$$

With straightforward matrix algebra, we have

$$(\boldsymbol{X},\boldsymbol{ZA})^\top\boldsymbol{W}(\boldsymbol{X},\boldsymbol{ZA})=\sum_{i=1}^M\boldsymbol{G}_{L_i}^\top(\boldsymbol{x}_i,\boldsymbol{z}_i)^\top\boldsymbol{W}_i(\boldsymbol{x}_i,\boldsymbol{z}_i)\boldsymbol{G}_{L_i},$$

where $\boldsymbol{G}_s$ are the $(p+q)\times(p+Sq)$ matrices such that $(\boldsymbol{\beta}_0^\top,\boldsymbol{\theta}_{i0}^\top)^\top=\boldsymbol{G}_{L_i}(\boldsymbol{\beta}_0^\top,\boldsymbol{\alpha}_{s0}^\top)^\top$, for any parameter dimensions.

Without loss of generality, assume $M=2$. When $S=1$, we have $\boldsymbol{G}_{L_1}=\boldsymbol{G}_{L_2}=\boldsymbol{I}_{p+q}$ and $\widehat{\boldsymbol{\theta}}_{1,0}=\widehat{\boldsymbol{\theta}}_{2,0}=\widehat{\boldsymbol{\alpha}}_0$. Hence, we have for any $(p+q)$-vector $\boldsymbol{a}$,

$$\begin{aligned}\boldsymbol{a}^\top\mathsf{Cov}\begin{pmatrix}\widehat{\boldsymbol{\beta}}_{\mathrm{OR}}\\ \widehat{\boldsymbol{\alpha}}_{\mathrm{OR}}\end{pmatrix}\boldsymbol{a}&=\boldsymbol{a}^\top\Big[(\boldsymbol{X},\boldsymbol{ZA})^\top\boldsymbol{W}(\boldsymbol{X},\boldsymbol{ZA})\Big]^{-1}\boldsymbol{a}\\ &=\boldsymbol{a}^\top\left[(\boldsymbol{x}_1,\boldsymbol{z}_1)^\top\boldsymbol{W}_1(\boldsymbol{x}_1,\boldsymbol{z}_1)+(\boldsymbol{x}_2,\boldsymbol{z}_2)^\top\boldsymbol{W}_2(\boldsymbol{x}_2,\boldsymbol{z}_2)\right]^{-1}\boldsymbol{a}\\ &\le\boldsymbol{a}^\top\left[(\boldsymbol{x}_i,\boldsymbol{z}_i)^\top\boldsymbol{W}_i(\boldsymbol{x}_i,\boldsymbol{z}_i)\right]^{-1}\boldsymbol{a}\\ &=\boldsymbol{a}^\top\mathsf{Cov}\begin{pmatrix}\breve{\boldsymbol{\beta}}_i\\ \breve{\boldsymbol{\theta}}_i\end{pmatrix}\boldsymbol{a},\quad\text{for } i=1,2.\end{aligned}$$

where the inequality follows from Lemma S.9.3.

When $S=2$, we calculate the expressions of $\mathsf{Cov}\big(\breve{\boldsymbol{\beta}}_i\big),\mathsf{Cov}\big(\breve{\boldsymbol{\theta}}_i\big),\mathsf{Cov}\big(\widehat{\boldsymbol{\beta}}_{\mathrm{OR}}\big)$ and $\mathsf{Cov}\big(\widehat{\boldsymbol{\theta}}_{\mathrm{OR},i}\big)$ in the following so as to establish the inequalities in (S.40). Firstly, for the unit estimators, let $\boldsymbol{D}_i=\mathsf{Cov}\big(\breve{\boldsymbol{\beta}}_i\big)$ and $\boldsymbol{H}_i=\mathsf{Cov}\big(\breve{\boldsymbol{\theta}}_i\big)$, and then we can write

$$\mathsf{Cov}\begin{pmatrix}\breve{\boldsymbol{\beta}}_i\\ \breve{\boldsymbol{\theta}}_i\end{pmatrix}=\big[(\boldsymbol{x}_i,\boldsymbol{z}_i)^\top\boldsymbol{W}_i(\boldsymbol{x}_i,\boldsymbol{z}_i)\big]^{-1}=\begin{bmatrix}\boldsymbol{D}_i & \boldsymbol{B}_i\\ \boldsymbol{B}_i^\top & \boldsymbol{H}_i\end{bmatrix},\quad i=1,2,$$

where $\boldsymbol{B}_i$ are some $p \times q$ matrices.

For the oracle estimators, suppose $L_1 = 1$ and $L_2 = 2$. Then we have $(\boldsymbol{\beta}_0^\top, \boldsymbol{\alpha}_0^\top)^\top = (\boldsymbol{\beta}_0^\top, \boldsymbol{\theta}_{1,0}, \boldsymbol{\theta}_{2,0})^\top$,

$$\boldsymbol{G}_1 = \begin{bmatrix} \boldsymbol{I}_p & \boldsymbol{O}_{p\times q} & \boldsymbol{O}_{p\times q} \\ \boldsymbol{O}_{p\times q}^\top & \boldsymbol{I}_q & \boldsymbol{O}_{q\times q} \end{bmatrix} \quad \text{and} \quad \boldsymbol{G}_2 = \begin{bmatrix} \boldsymbol{I}_p & \boldsymbol{O}_{p\times q} & \boldsymbol{O}_{p\times q} \\ \boldsymbol{O}_{p\times q}^\top & \boldsymbol{O}_{q\times q} & \boldsymbol{I}_q \end{bmatrix},$$

where $\boldsymbol{O}_{d_1\times d_2}$ stands for the $d_1 \times d_2$ zero matrix. By blockwise matrix inversion, we can write

$$(\boldsymbol{x}_i, \boldsymbol{z}_i)^\top \boldsymbol{W}_i (\boldsymbol{x}_i, \boldsymbol{z}_i) = \begin{bmatrix} \boldsymbol{D}_i^{-1} + \boldsymbol{D}_i^{-1}\boldsymbol{B}_i\boldsymbol{K}_i\boldsymbol{B}_i^\top\boldsymbol{D}_i^{-1} & -\boldsymbol{D}_i^{-1}\boldsymbol{B}_i\boldsymbol{K}_i \\ -\boldsymbol{K}_i\boldsymbol{B}_i^\top\boldsymbol{D}_i^{-1} & \boldsymbol{K}_i \end{bmatrix}, \quad i = 1, 2,$$

where $\boldsymbol{K}_i = (\boldsymbol{H}_i - \boldsymbol{B}_i^\top \boldsymbol{D}_i^{-1}\boldsymbol{B}_i)^{-1}$. Accordingly, it follows that

$$\big(\boldsymbol{X}, \boldsymbol{Z}\boldsymbol{A}\big)^\top \boldsymbol{W}\big(\boldsymbol{X}, \boldsymbol{Z}\boldsymbol{A}\big) = \begin{bmatrix} \sum_{i=1}^2 \{\boldsymbol{D}_i^{-1} + \boldsymbol{D}_i^{-1}\boldsymbol{B}_i\boldsymbol{K}_i\boldsymbol{B}_i^\top\boldsymbol{D}_i^{-1}\} & -\boldsymbol{D}_1^{-1}\boldsymbol{B}_1\boldsymbol{K}_1 & -\boldsymbol{D}_2^{-1}\boldsymbol{B}_2\boldsymbol{K}_2 \\ -\boldsymbol{K}_1\boldsymbol{B}_1^\top\boldsymbol{D}_1^{-1} & \boldsymbol{K}_1 & \boldsymbol{O}_{q\times q} \\ -\boldsymbol{K}_2\boldsymbol{B}_2^\top\boldsymbol{D}_2^{-1} & \boldsymbol{O}_{q\times q} & \boldsymbol{K}_2 \end{bmatrix}.$$

By the blockwise matrix inversion, we have

$$\begin{aligned} \mathsf{Cov}\begin{pmatrix} \widehat{\boldsymbol{\beta}}_{\mathrm{OR}} \\ \widehat{\boldsymbol{\alpha}}_{\mathrm{OR}} \end{pmatrix} &= \Big[\big(\boldsymbol{X}, \boldsymbol{Z}\boldsymbol{A}\big)^\top \boldsymbol{W}\big(\boldsymbol{X}, \boldsymbol{Z}\boldsymbol{A}\big)\Big]^{-1} \\ &= \begin{bmatrix} \boldsymbol{Q} & \boldsymbol{Q}\boldsymbol{D}_1^{-1}\boldsymbol{B}_1 & -\boldsymbol{Q}\boldsymbol{D}_2^{-1}\boldsymbol{B}_2 \\ \boldsymbol{B}_1^\top\boldsymbol{D}_1^{-1}\boldsymbol{Q} & \boldsymbol{H}_1 - \boldsymbol{B}_1\boldsymbol{D}_1^{-1}\boldsymbol{B}_1 + \boldsymbol{B}_1^\top\boldsymbol{D}_1^{-1}\boldsymbol{Q}\boldsymbol{D}_1^{-1}\boldsymbol{B}_1 & \boldsymbol{B}_1^\top\boldsymbol{D}_1^{-1}\boldsymbol{Q}\boldsymbol{D}_2^{-1}\boldsymbol{B}_2 \\ -\boldsymbol{B}_2^\top\boldsymbol{D}_2^{-1}\boldsymbol{Q} & \boldsymbol{B}_2^\top\boldsymbol{D}_2^{-1}\boldsymbol{Q}\boldsymbol{D}_1^{-1}\boldsymbol{B}_1 & \boldsymbol{H}_2 - \boldsymbol{B}_2\boldsymbol{D}_2^{-1}\boldsymbol{B}_2 + \boldsymbol{B}_2^\top\boldsymbol{D}_2^{-1}\boldsymbol{Q}\boldsymbol{D}_2^{-1}\boldsymbol{B}_2 \end{bmatrix}, \end{aligned}$$

where $\boldsymbol{Q} = (\boldsymbol{D}_1^{-1} + \boldsymbol{D}_2^{-1})^{-1}$. As a result, we have $\mathsf{Cov}\big(\widehat{\boldsymbol{\beta}}_{\mathrm{OR}}\big) = \boldsymbol{Q} = (\boldsymbol{D}_1^{-1} + \boldsymbol{D}_2^{-1})^{-1}$ and $\mathsf{Cov}\big(\widehat{\boldsymbol{\theta}}_{\mathrm{OR},i}\big) = \boldsymbol{H}_i - \boldsymbol{B}_i\boldsymbol{D}_i^{-1}\boldsymbol{B}_i + \boldsymbol{B}_i^\top\boldsymbol{D}_i^{-1}\boldsymbol{Q}\boldsymbol{D}_i^{-1}\boldsymbol{B}_i, i = 1, 2$.

With the above results, it is then straightforward that, for any p-vector $\boldsymbol{v}_p$,

$$\boldsymbol{v}_p^\top \mathsf{Cov}\big(\widehat{\boldsymbol{\beta}}_{\mathrm{OR}}\big)\boldsymbol{v}_p = \boldsymbol{v}_p^\top \left(\boldsymbol{D}_1^{-1} + \boldsymbol{D}_2^{-1}\right)^{-1} \boldsymbol{v}_p \le \boldsymbol{v}_p^\top \boldsymbol{D}_i \boldsymbol{v}_p = \boldsymbol{v}_p^\top \mathsf{Cov}\big(\breve{\boldsymbol{\beta}}_i\big)\boldsymbol{v}_p, \quad \text{for } i = 1, 2,$$

where the inequality follows from Lemma S.9.3. In addition, for any q-vector $\boldsymbol{v}_q$,

$$\boldsymbol{v}_q^\top \mathsf{Cov}\big(\widehat{\boldsymbol{\theta}}_{\mathrm{OR},i}\big)\boldsymbol{v}_q = \boldsymbol{v}_q^\top \left[\boldsymbol{H}_i - \boldsymbol{B}_i^\top\boldsymbol{D}_i^{-1}\boldsymbol{B}_i + \boldsymbol{B}_i^\top\boldsymbol{D}_i^{-1}(\boldsymbol{D}_1^{-1} + \boldsymbol{D}_2^{-1})^{-1}\boldsymbol{D}_i^{-1}\boldsymbol{B}_i\right] \boldsymbol{v}_q.$$

For the third term of the R.H.S., it holds

$$\boldsymbol{v}_q^\top \boldsymbol{B}_i^\top\boldsymbol{D}_i^{-1}(\boldsymbol{D}_1^{-1} + \boldsymbol{D}_2^{-1})^{-1}\boldsymbol{D}_i^{-1}\boldsymbol{B}_i\boldsymbol{v}_q \le \boldsymbol{v}_q^\top \boldsymbol{B}_i^\top\boldsymbol{D}_i^{-1}\boldsymbol{D}_i\boldsymbol{D}_i^{-1}\boldsymbol{B}_i\boldsymbol{v}_q = \boldsymbol{v}_q^\top \boldsymbol{B}_i^\top\boldsymbol{D}_i^{-1}\boldsymbol{B}_i\boldsymbol{v}_q,$$

where the ineqaulity follows from Lemma S.9.3. Consequently, we have

$$\boldsymbol{v}_q^\top \mathsf{Cov}\big(\widehat{\boldsymbol{\theta}}_{\mathrm{OR},i}\big)\boldsymbol{v}_q \le \boldsymbol{v}_q^\top \boldsymbol{H}_i \boldsymbol{v}_q = \mathsf{Cov}\big(\breve{\boldsymbol{\theta}}_i\big), \quad i = 1, 2.$$

The proof is thus complete.

S.9. Auxiliary Lemmas

In this section we provide technical lemmas, with a bit abuse of notations.

Lemma S.9.1 (Woodbury matrix identity, Theorem 18.2.8 of Harville (2000))**.** *Let $\boldsymbol{A}, \boldsymbol{U}, \boldsymbol{B}$ and $\boldsymbol{V}$ denote $n \times n, n \times m, m \times m$ and $m \times n$ matrices, respectively. Suppose $\boldsymbol{A}$ and $\boldsymbol{B}$ are nonsingular. Then we have*

$$\big(\boldsymbol{A} \pm \boldsymbol{U}\boldsymbol{B}\boldsymbol{V}\big)^{-1} = \boldsymbol{A}^{-1} \mp \boldsymbol{A}^{-1}\boldsymbol{U}\left(\boldsymbol{B}^{-1} \pm \boldsymbol{V}\boldsymbol{A}^{-1}\boldsymbol{U}\right)^{-1}\boldsymbol{V}\boldsymbol{A}^{-1}. \tag{S.41}$$

Lemma S.9.2 (Lemma 8 of Hsu et al. (2014)). *Suppose $\boldsymbol{\xi}$ is a sub-Gaussian random n-vector. For all symmetric positive semidefinite matrices $\boldsymbol{M} \succeq \boldsymbol{O}$ and all $t > 0$, we have*

$$P\left[\boldsymbol{\xi}^\top \boldsymbol{M}\boldsymbol{\xi} > c\left(\mathrm{tr}(\boldsymbol{M}) + 2\sqrt{\mathrm{tr}(\boldsymbol{M}^2)t} + 2\|\boldsymbol{M}\|t\right)\right] \leq e^{-t},$$

where $c > 0$ is an absolute constant. With $\boldsymbol{M} = \boldsymbol{I}$ and $t = n$, it follows that

$$P\big(\|\boldsymbol{\xi}\|^2 > 5cn\big) \leq e^{-n}. \tag{S.42}$$

Lemma S.9.3 (Lemma A.3 of Liu et al. (2015)). *Suppose $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$ are $d \times d$ positive definite matrices. Then, for any d-vector $\boldsymbol{v}$,*

$$\left(\boldsymbol{v}^\top \boldsymbol{M}_1^{-1}\boldsymbol{v}\right)^{-1} + \left(\boldsymbol{v}^{-1}\boldsymbol{M}_1^\top \boldsymbol{v}\right)^{-1} \leq \left[\boldsymbol{v}^\top \left(\boldsymbol{M}_1 + \boldsymbol{M}_2\right)^{-1}\boldsymbol{v}\right]^{-1}. \tag{S.43}$$

This implies that $\boldsymbol{v}^\top \left(\boldsymbol{M}_1 + \boldsymbol{M}_2\right)^{-1}\boldsymbol{v} \leq \boldsymbol{v}^\top \boldsymbol{M}_1^{-1}\boldsymbol{v}$.

S.10. Signal-to-Noise Ratio for the proposed MTL method

Here we define the signal-to-noise ratio (SNR) for the proposed method. In order to supply information of how strong the subgroup effects compared to the noise, we first compute the signal for different subgroups i and j as

$$\begin{aligned}\mathrm{Signal}(i,j) &= \mathsf{Var}\big[(\mathbf{X}^\top\boldsymbol{\beta}_0 + \mathbf{Z}^\top\boldsymbol{\alpha}_i) - (\mathbf{X}^\top\boldsymbol{\beta}_0 + \mathbf{Z}^\top\boldsymbol{\alpha}_j)\big] \\ &= \mathsf{Var}\big[\mathbf{Z}^\top(\boldsymbol{\alpha}_i - \boldsymbol{\alpha}_j)\big] \\ &= (\boldsymbol{\alpha}_i - \boldsymbol{\alpha}_j)^\top \boldsymbol{\Sigma}_{\mathbf{Z}}(\boldsymbol{\alpha}_i - \boldsymbol{\alpha}_j),\end{aligned}$$

and then define the desired signal as the minimal signal variance between different subgroups, i.e.,

$$\mathrm{Signal} = \min_{1\leq i<j\leq S} \mathrm{Signal}(i,j).$$

For noise, we calculate

$$\begin{aligned}\mathrm{Noise} &= \mathsf{Var}\big[(\mathbf{Z}^\top\boldsymbol{u}_i + \varepsilon_i) - (\mathbf{Z}^\top\boldsymbol{u}_i + \varepsilon_i)\big] \\ &= \mathsf{Var}\big[\mathbf{Z}^\top(\boldsymbol{u}_j - \boldsymbol{u}_j)\big] + 2\mathsf{Var}(\varepsilon) = 2\big[\,\mathrm{tr}(\mathsf{E}[\mathbf{Z}\mathbf{Z}^\top]\boldsymbol{\Psi}) + \sigma_\varepsilon^2\big].\end{aligned}$$

Accordingly, we have

$$\mathrm{SNR} = \frac{\mathrm{Signal}}{\mathrm{Noise}} = \frac{\min_{1\leq i<j\leq S}\mathrm{Signal}(i,j)}{2\big[\,\mathrm{tr}(\mathsf{E}[\mathbf{Z}\mathbf{Z}^\top]\boldsymbol{\Psi}) + \sigma_\varepsilon^2\big]}.$$

S.11. Additional Simulation and Real Data Results.

S.11.1. Simulation Results for all the cases.

We present the all-case version of Table 2.

Table S.1. Evaluation of subgroup recovery. Complete version of Table 2.

		$\widehat{S}$			Perfect
	Method	Mean (SD)	Median (Min,Max)	NMI	Recovery
Case 1	MCP	2.00 (0.0000)	2 (2, 2)	0.9988	0.99
($S = 2$)	SCAD	2.00 (0.0000)	2 (2, 2)	0.9988	0.99
	TLP	2.00 (0.0000)	2 (2, 2)	0.9988	0.99
	L_1	2.39 (0.6651)	2 (1, 4)	0.9476	0.62
	K-Means	2.05 (0.2179)	2 (2, 3)	0.9896	0.95
Case 2	MCP	3.00 (0.0000)	3 (3, 3)	1.0000	1.00
($S = 3$)	SCAD	3.00 (0.0000)	3 (3, 3)	1.0000	1.00
	TLP	3.00 (0.0000)	3 (3, 3)	1.0000	1.00
	L_1	2.40 (1.0731)	3 (1, 4)	0.6467	0.55
	K-Means	3.00 (0.0000)	3 (3, 3)	1.0000	1.00
Case 3	MCP	2.00 (0.0000)	2 (2, 2)	0.9988	0.99
($S = 2$)	SCAD	2.00 (0.0000)	2 (2, 2)	0.9988	0.99
	TLP	2.00 (0.0000)	2 (2, 2)	0.9988	0.99
	L_1	2.33 (0.5870)	2 (1, 4)	0.9718	0.70
	K-Means	2.01 (0.0995)	2 (2, 3)	0.9979	0.99
Case 4	MCP	3.00 (0.0000)	3 (3, 3)	1.0000	1.00
($S = 3$)	SCAD	3.00 (0.0000)	3 (3, 3)	1.0000	1.00
	TLP	3.00 (0.0000)	3 (3, 3)	1.0000	1.00
	L_1	2.43 (1.2248)	3 (1, 6)	0.6141	0.47
	K-Means	3.00 (0.0000)	3 (3, 3)	1.0000	1.00
Case 5	MCP	2.00 (0.0000)	2 (2, 2)	1.0000	1.00
($S = 2$)	SCAD	2.01 (0.1000)	2 (2, 3)	0.9990	0.99
	TLP	2.01 (0.1000)	2 (2, 3)	0.9990	0.99
	L_1	2.27 (0.5835)	2 (1, 4)	0.9807	0.77
	K-Means	3.36 (1.0538)	4 (2, 5)	0.7877	0.36
Case 6	MCP	3.00 (0.0000)	3 (3, 3)	1.0000	1.00
($S = 3$)	SCAD	3.00 (0.0000)	3 (3, 3)	1.0000	1.00
	TLP	3.00 (0.0000)	3 (3, 3)	1.0000	1.00
	L_1	4.09 (9.7152)	3 (1, 100)	0.9303	0.72
	K-Means	3.76 (0.8261)	4 (3, 6)	0.9303	0.48
Case 7	MCP	5.01 (0.1000)	5 (5, 6)	0.9996	0.98
($S = 5$)	SCAD	5.00 (0.0000)	5 (5, 5)	0.9997	0.99
	TLP	5.01 (0.1000)	5 (5, 6)	0.9996	0.99
	L_1	63.24 (47.1983)	100 (1, 100)	0.5878	0.00
	K-Means	5.00 (0.0000)	5 (5, 5)	0.9768	0.37
Case 8	MCP	5.01 (0.1000)	5 (5, 6)	0.9998	0.99
($S = 5$)	SCAD	5.02 (0.1407)	5 (5, 6)	0.9998	0.98
	TLP	5.02 (0.1407)	5 (5, 6)	0.9998	0.98
	L_1	97.26 (70.6762)	150 (1, 150)	0.6137	0.00
	K-Means	5.43 (0.5875)	5 (5, 7)	0.9550	0.16
Case 9	MCP	7.00 (0.0000)	7 (7, 7)	0.9999	0.99
($S = 7$)	SCAD	7.26 (0.5245)	7 (7, 9)	0.9981	0.77
	TLP	7.22 (0.5041)	7 (7, 9)	0.9982	0.81
	L_1	150.00 (0.0000)	150 (150, 150)	0.6205	0.00
	K-Means	7.00 (0.0000)	7 (7, 7)	0.9786	0.20

S.11.2. NOAA Data Features

In our analysis, each climate division is taken as a data unit. The monthly average temperature is the response of interest, and we include precipitation (PCPN), Palmer Drought Severity Index (PSDI, based on the principles of a balance between moisture supply and demand without considering man-made changes), Palmer Hydrological Drought Index (PHDI, based on principles similar to PDSI, but with consideration of some man-made changes) and Palmer Z Index (ZNDX, a moisture anomaly index) into the covariate pool. Negative and positive values of PDSI, PHDI and ZNDX indicate dry and wet spells, respectively. See the database documentation[4] for more details about these drought indices. To account for seasonal effects, we include dummy variables for the four seasons (March, April and May as spring; June, July and August as summer;

[4] Available on ftp://ftp.ncdc.noaa.gov/pub/data/cirs/climdiv/divisional-readme.txt.

Table S.2. Average $\mathrm{ASD}(\widehat{\alpha}_{1,1})$ and $\mathrm{ESD}(\widehat{\alpha}_{1,1})$, based on replicates with perfect subgroup structure recovery.

	Average $\mathrm{ASD}(\widehat{\alpha}_{1,1})$				$\mathrm{ESD}(\widehat{\alpha}_{1,1})$			
	MCP	SCAD	TLP	Oracle	MCP	SCAD	TLP	Oracle
Case 1	0.1123	0.1123	0.1123	0.1205	0.1166	0.1166	0.1166	0.1166
Case 2	0.1508	0.1508	0.1508	0.1531	0.1569	0.1569	0.1569	0.1569
Case 3	0.1244	0.1244	0.1244	0.1200	0.1117	0.1117	0.1117	0.1177
Case 4	0.1591	0.1591	0.1591	0.1525	0.1728	0.1728	0.1728	0.1728
Case 5	0.0781	0.0781	0.0781	0.0823	0.0871	0.0871	0.0871	0.0871
Case 6	0.0988	0.0988	0.0988	0.1025	0.1131	0.1131	0.1131	0.1131
Case 7	0.1319	0.1318	0.1319	0.1380	0.1285	0.1147	0.1282	0.1285
Case 8	0.1069	0.1066	0.1167	0.1104	0.1023	0.1028	0.1034	0.1030
Case 9	0.1284	0.1280	0.1284	0.1339	0.1414	0.1348	0.1448	0.1409

September, October and November as Fall; December, January and February as winter), and the spring effect is taken as the baseline intercept term. In summary, there are 8 covariates in our analysis.

We next determine heterogeneous effects by observing the kernel densities of the ordinary least-squares (OLS) estimates of the candidate effects, as shown in Figure S.1. Intuitively, the distributions of heterogeneous effects are likely to form a multimodal or wide-spread shapes. As a result, we choose the intercept, PCPN and ZNDX as the heterogeneous effects (i.e., $q = 3$ and hence $p = 5$).

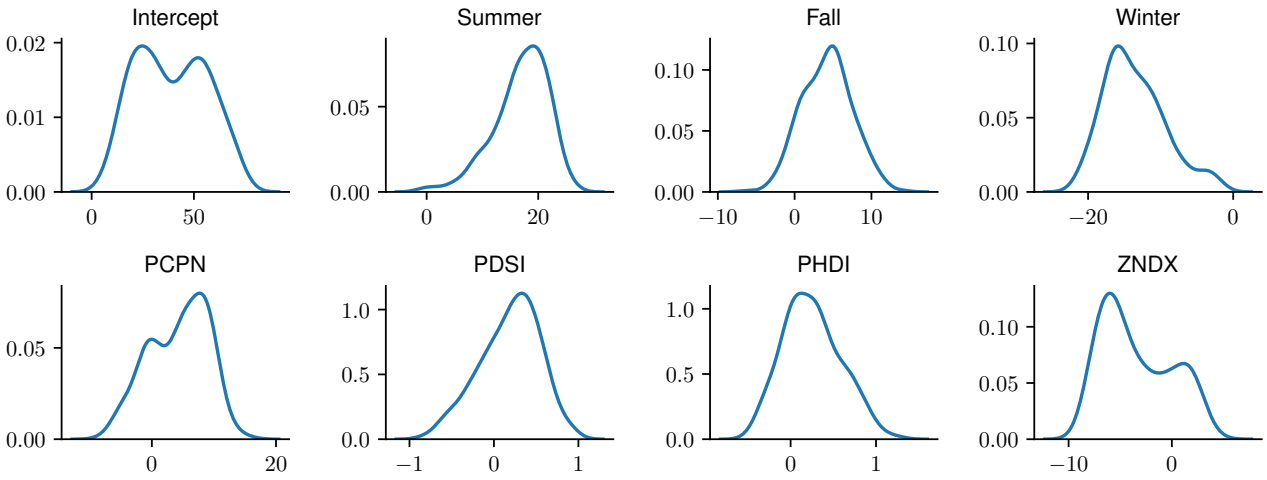

Figure S.1. Kernel densities of the 344 OLS estimates obtained from the nClimDiv database.

S.11.3. Asymptotic covariance approximation

We assess how close the estimated asymptotic covariance is to its oracle and empirical counterparts. To this end, we observe $\|\widehat{\mathbf{\Sigma}} - \widehat{\mathbf{\Sigma}}_{\mathrm{OR}}\|_{\max}$, where $\widehat{\mathbf{\Sigma}}$ and $\widehat{\mathbf{\Sigma}}_{\mathrm{OR}}$ stand for the asymptotic covariance matrices of $(\widehat{\boldsymbol{\beta}}^\top, \widehat{\boldsymbol{\alpha}}^\top)^\top$ and $(\widehat{\boldsymbol{\beta}}_{\mathrm{OR}}^\top, \widehat{\boldsymbol{\alpha}}_{\mathrm{OR}}^\top)^\top$, respectively. Furthermore, we compare the average asymptotic standard deviation (ASD) with the empirical standard deviation (ESD) to evaluate the finite-sample second moment approximation. Since the ASDs and ESDs of all coordinates in $\widehat{\boldsymbol{\alpha}}$ behave similarly, we only show the results for $\widehat{\alpha}_{1,1}$.

Results of $\|\widehat{\mathbf{\Sigma}} - \widehat{\mathbf{\Sigma}}_{\mathrm{OR}}\|_{\max}$, $\mathrm{ASD}(\widehat{\alpha}_{1,1})$ and $\mathrm{ESD}(\widehat{\alpha}_{1,1})$ are shown in Figure S.2 and Table S.2, based on the replicates with perfect subgroup structure recovery. It can be seen that the $\widehat{\mathbf{\Sigma}}$'s are fairly close to $\widehat{\mathbf{\Sigma}}_{\mathrm{OR}}$'s, and the average $\mathrm{ASD}(\widehat{\alpha}_{1,1})$'s are decently comparable to their empirical counterpart $\mathrm{ESD}(\widehat{\alpha}_{1,1})$'s. Accordingly, we confirm that the empirical covariance matrix of the proposed estimator can be properly approximated by the asymptotic covariance matrix given in Corollary 4.1.

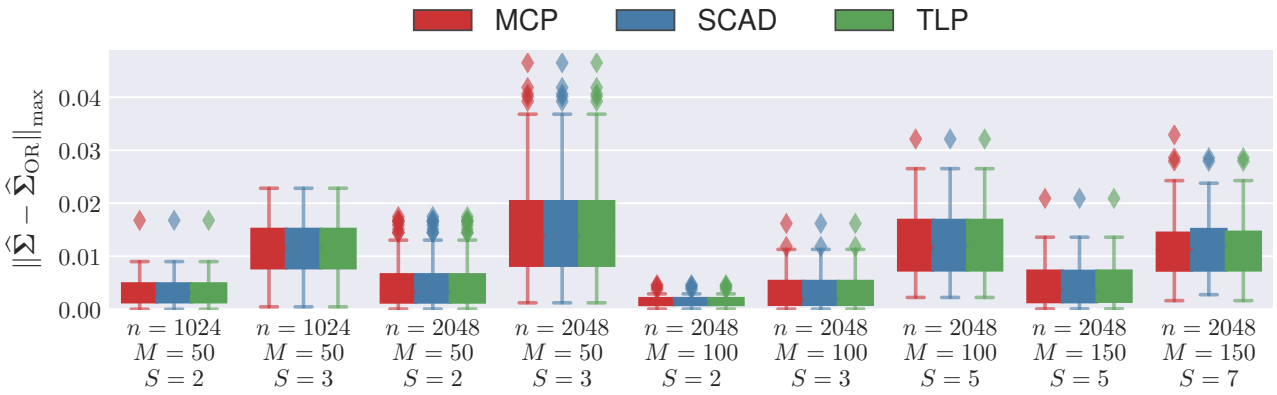

Figure S.2. Boxplots of $\left\|\widehat{\boldsymbol{\Sigma}} - \widehat{\boldsymbol{\Sigma}}_{\mathrm{OR}}\right\|_{\max}$, based on replicates with perfect subgroup structure recovery.

S.11.4. Inferential Accuracy

In this section, we study the statistical inferential accuracy based on the asymptotic normality result in Corollary 4.1. Let $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\alpha}} = (\boldsymbol{Z}\widehat{\boldsymbol{A}})^\top \boldsymbol{W}(\boldsymbol{Z}\widehat{\boldsymbol{A}}) - [(\boldsymbol{Z}\widehat{\boldsymbol{A}})^\top \boldsymbol{W}\boldsymbol{X}](\boldsymbol{X}^\top \boldsymbol{W}\boldsymbol{X})^{-1}[\boldsymbol{X}^\top \boldsymbol{W}(\boldsymbol{Z}\widehat{\boldsymbol{A}})]$ denote the estimated asymptotic covariance matrix for $\widehat{\boldsymbol{\alpha}}$, where $\widehat{\boldsymbol{A}}$ is the estimated label matrix $\boldsymbol{A}$ by substituting the estimated subgroups for the true ones. For the first subgroup effect $\boldsymbol{\alpha}_1$, a 95% confidence region can be constructed by

$$\mathrm{CR}_{\boldsymbol{\alpha}_1} = \big\{\boldsymbol{v} \in \mathbb{R}^q : (\widehat{\boldsymbol{\alpha}}_1 - \boldsymbol{v})^\top \left(\boldsymbol{L}_1 \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\alpha}} \boldsymbol{L}_1^\top\right)^{-1} (\widehat{\boldsymbol{\alpha}}_1 - \boldsymbol{v}) \le \chi_q^2(0.95)\big\},$$

where $\boldsymbol{L}_1 = \left[\boldsymbol{I}_q, \boldsymbol{O}_q, \cdots, \boldsymbol{O}_q\right]_{q \times Sq}$ such that $\boldsymbol{L}_1 \widehat{\boldsymbol{\alpha}} = \widehat{\boldsymbol{\alpha}}_1$ and $\chi_q^2(0.95)$ is the 95% percentile of the χ_q^2 distribution. Table S.3 displays the empirical coverage probabilities obtained from replications with perfect subgroup recovery. It can be seen that the empirical coverages, including that by the oracle estimates, are close to the nominal level 95% except for Case 7, which again indicates that $S = 5$ could be too large for $M = 100$.

Table S.3. Empirical coverage probabilities for $\mathrm{CR}_{\boldsymbol{\alpha}_1}$.

	MCP	SCAD	TLP	Oracle
Case 1	0.9600	0.9600	0.9596	0.9500
Case 2	0.9200	0.9200	0.9200	0.9200
Case 3	0.9394	0.9394	0.9394	0.9100
Case 4	0.9300	0.9300	0.9300	0.9400
Case 5	0.9400	0.9556	0.9596	0.9100
Case 6	0.9200	0.9200	0.9200	0.9300
Case 7	0.8800	0.8800	0.8800	0.8800
Case 8	0.9596	0.9468	0.9600	0.9500
Case 9	0.9286	0.9286	0.9310	0.9200

We further formulate a heterogeneous test $H_0 : \boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_2$ between the first two subgroups. Since $\boldsymbol{\alpha}_1 - \boldsymbol{\alpha}_2 = \boldsymbol{L}_{12}\boldsymbol{\alpha}$, where $\boldsymbol{L}_{12} = \left[\boldsymbol{I}_q, -\boldsymbol{I}_q, \boldsymbol{O}_q, \cdots, \boldsymbol{O}_q\right]_{q \times Sq}$, we define the F test statistic as

$$T = (\widehat{\boldsymbol{\alpha}}_1 - \widehat{\boldsymbol{\alpha}}_2)^\top \left(\boldsymbol{L}_{12} \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\alpha}} \boldsymbol{L}_{12}^\top\right)^{-1} (\widehat{\boldsymbol{\alpha}}_1 - \widehat{\boldsymbol{\alpha}}_2)/q.$$

Under H_0, the test statistic T asymptotically follows the $F_{q, N-p-\widehat{S}q-2}$ distribution. For replications with perfect subgroup recovery, the p-values of the heterogeneity test are all less than 0.0001. This indicates that the estimated subgroups obtained from the proposed method cannot be further combined.