Supporting Information

of "Clustering on hierarchical heterogeneous data with prior pairwise relationships" by Wei Han, Sanguo Zhang, Hailong Gao, and Deliang Bu

A Proof of Theorem 1

A.1 Lemmas

Lemma 1. *If $\boldsymbol{U} \in \mathbb{R}^d$ follows sub-Gaussian distribution with variance proxy σ_0^2 and $\mathrm{E}(\boldsymbol{U}) = \mathbf{0}$, then with probability at least $1-\delta$ for $\delta \in (0,1)$, $\|\boldsymbol{U}\|_2 \leqslant 4\sigma_0 d^{\frac{1}{2}} + 2\sigma_0 \left(2\log\left(\delta^{-1}\right)\right)^{\frac{1}{2}}$.*

We define the oracle estimators as follows (that is, the minimizer under the true hierarchical clustering structure):

$$\left(\widehat{\boldsymbol{\xi}}^0, \widehat{\boldsymbol{\alpha}}^0\right) = \arg\min_{\boldsymbol{\xi},\boldsymbol{\alpha}} \left\{ \frac{1}{2}\sum_{k_1=1}^{K_1}\sum_{i\in\mathcal{G}_{k_1}^*} \left\|\boldsymbol{\xi}_{k_1} - \boldsymbol{X}_i\right\|_2^2 + \frac{1}{2}\sum_{k_2=1}^{K_2}\sum_{i\in\mathcal{T}_{k_2}^*} \left\|\boldsymbol{\alpha}_{k_2} - \boldsymbol{Z}_i\right\|_2^2 \right\},$$

where $\widehat{\boldsymbol{\xi}}^0 = \left(\widehat{\boldsymbol{\xi}}_1^0, \cdots, \widehat{\boldsymbol{\xi}}_{K_1}^0\right) \in \mathbb{R}^q \times \mathbb{R}^{K_1}$ and $\widehat{\boldsymbol{\alpha}}^0 = \left(\widehat{\boldsymbol{\alpha}}_1^0, \cdots, \widehat{\boldsymbol{\alpha}}_{K_2}^0\right) \in \mathbb{R}^p \times \mathbb{R}^{K_2}$. Let $\widehat{\boldsymbol{\beta}}_i^0 = \widehat{\boldsymbol{\xi}}_{k_1}^0$ and $\widehat{\boldsymbol{\gamma}}_i^0 = \widehat{\boldsymbol{\alpha}}_{k_2}^0$ if i belongs to $\mathcal{G}_{k_1}^*$ and $\mathcal{T}_{k_2}^*$. Here, we introduce the following Lemmas.

Lemma 2. *Define $\phi_n = (q+p)^{\frac{1}{2}} T_{\min}^{-\frac{1}{2}} (\log n)^{\frac{1}{2}}$. Under Condition 1, there exists a constant C, such that with probability at least $1-2n^{-1}$, the oracle estimators $\left(\widehat{\boldsymbol{\xi}}^0, \widehat{\boldsymbol{\alpha}}^0\right)$ satisfy that*

$$\sup_{1\leqslant k_1\leqslant K_1} \left\|\widehat{\boldsymbol{\xi}}_{k_1}^0 - \boldsymbol{\xi}_{k_1}^*\right\|_2 \leqslant C\phi_n, \quad \sup_{1\leqslant k_2\leqslant K_2} \left\|\widehat{\boldsymbol{\alpha}}_{k_2}^0 - \boldsymbol{\alpha}_{k_2}^*\right\|_2 \leqslant C\phi_n,$$

Lemma 3. *Suppose that the conditions in Theorem 1 hold. There exists a local minimizer $\left(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}}\right)$ of $Q(\boldsymbol{\beta},\boldsymbol{\gamma})$ subject to $\boldsymbol{\beta} \in \mathcal{M}_1$ and $\boldsymbol{\gamma} \in \mathcal{M}_2$ satisfying that as $n \to \infty$,*

$$\Pr\left(\left(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}}\right) = \left(\widehat{\boldsymbol{\beta}}^0, \widehat{\boldsymbol{\gamma}}^0\right)\right) \to 1.$$

Here, we proof Theorem 1 by Lemmas 2 and 3. For any $\delta \in (0,1)$, there exists sufficiently large n, such that there exists a local minimizer $\left(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}}\right)$ of $Q(\boldsymbol{\beta}, \boldsymbol{\gamma})$, with probability at least $1-\frac{\delta}{2}$, $\left(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}}\right) = \left(\widehat{\boldsymbol{\beta}}^0, \widehat{\boldsymbol{\gamma}}^0\right)$ holds, and with probability at least $1-2n^{-1} > 1-\frac{\delta}{2}$,

$$\sup_{1 \leqslant k_1 \leqslant K_1} \left\| \widehat{\boldsymbol{\xi}}_{k_1}^0 - \boldsymbol{\xi}_{k_1}^* \right\|_2 \leqslant C\phi_n, \quad \sup_{1 \leqslant k_2 \leqslant K_2} \left\| \widehat{\boldsymbol{\alpha}}_{k_2}^0 - \boldsymbol{\alpha}_{k_2}^* \right\|_2 \leqslant C\phi_n.$$

Hence, with probability at least $1-\delta$, the above hold simultaneously. Observe that oracle estimators can recover true clustering structure, hence $\left(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}}\right) = \left(\widehat{\boldsymbol{\beta}}^0, \widehat{\boldsymbol{\gamma}}^0\right)$ implies that $\widehat{K}_1 = K_1$, $\widehat{K}_2 = K_2$, $\widehat{\mathcal{G}}_{k_1} = \mathcal{G}_{k_1}^*$, and $\widehat{\mathcal{T}}_{k_2} = \mathcal{T}_{k_2}^*$. Therefore, Theorem 1 is proved.

A.2 Proofs of Lemmas

Lemma 1 can be referred to Theorem 1.19 in [1] (https://math.mit.edu/~rigollet/PDFs/RigNotes17.pdf), which is a classical textbook in high dimensional statistics. We firstly give a rigorous proof of Lemma 2. For $k_1 = 1, \cdots, K_1$ and $k_2 = 1, \cdots, K_2$, the oracle estimators have analytic solutions,

$$\widehat{\boldsymbol{\xi}}_{k_1}^0 = \left|\mathcal{G}_{k_1}^*\right|^{-1} \sum_{i \in \mathcal{G}_{k_1}^*} \boldsymbol{X}_i, \quad \widehat{\boldsymbol{\alpha}}_{k_2}^0 = \left|\mathcal{T}_{k_2}^*\right|^{-1} \sum_{i \in \mathcal{T}_{k_2}^*} \boldsymbol{Z}_i.$$

Directly, for $k_1 = 1, \cdots, K_1$, we have that

$$\widehat{\boldsymbol{\xi}}_{k_1}^0 - \boldsymbol{\xi}_{k_1}^* = \left|\mathcal{G}_{k_1}^*\right|^{-1} \sum_{i \in \mathcal{G}_{k_1}^*} \boldsymbol{X}_i - \boldsymbol{\xi}_{k_1}^* = \left|\mathcal{G}_{k_1}^*\right|^{-1} \sum_{i \in \mathcal{G}_{k_1}^*} \left(\boldsymbol{\xi}_{k_1}^* + \boldsymbol{\epsilon}_{1i}\right) - \boldsymbol{\xi}_{k_1}^* = \left|\mathcal{G}_{k_1}^*\right|^{-1} \sum_{i \in \mathcal{G}_{k_1}^*} \boldsymbol{\epsilon}_{1i}.$$

Note that $\left|\mathcal{G}_{k_1}^*\right|^{-1} \sum_{i \in \mathcal{G}_{k_1}^*} \boldsymbol{\epsilon}_{1i}$ has mean $\mathbf{0}$ and follows sub-Gaussian distribution with variance proxy $\left|\mathcal{G}_{k_1}^*\right|^{-1} \sigma_0^2$. By Lemma 1, for fixed k_1, with probability at least $1-n^{-2}$,

$$\left\| \widehat{\boldsymbol{\xi}}_{k_1}^0 - \boldsymbol{\xi}_{k_1}^* \right\|_2 \leqslant 4 \left|\mathcal{G}_{k_1}^*\right|^{-\frac{1}{2}} \sigma_0 q^{\frac{1}{2}} + 2 \left|\mathcal{G}_{k_1}^*\right|^{-\frac{1}{2}} \sigma_0 \left(4 \log n\right)^{\frac{1}{2}} \leqslant 8 \sigma_0 q^{\frac{1}{2}} T_{\min}^{-\frac{1}{2}} \left(\log n\right)^{\frac{1}{2}}.$$

Similarly, for fixed k_2, with probability at least $1-n^{-2}$,

$$\left\|\widehat{\boldsymbol{\alpha}}_{k_2}^{0}-\boldsymbol{\alpha}_{k_2}^{*}\right\|_2 \leqslant 4\left|\mathcal{T}_{k_2}^{*}\right|^{-\frac{1}{2}}\sigma_0 p^{\frac{1}{2}}+2\left|\mathcal{T}_{k_2}^{*}\right|^{-\frac{1}{2}}\sigma_0\left(4\log n\right)^{\frac{1}{2}} \leqslant 8\sigma_0 p^{\frac{1}{2}} T_{\min}^{-\frac{1}{2}}\left(\log n\right)^{\frac{1}{2}}.$$

Note that $\phi_n=(q+p)^{\frac{1}{2}}T_{\min}^{-\frac{1}{2}}(\log n)^{\frac{1}{2}}$. There exists a constant C, such that with probability at least $1-\left(K_1+K_2\right)n^{-2}>1-2n^{-1}$, it holds that

$$\sup_{1\leqslant k_1\leqslant K_1}\left\|\widehat{\boldsymbol{\xi}}_{k_1}^{0}-\boldsymbol{\xi}_{k_1}^{*}\right\|_2 \leqslant C\phi_n, \quad \sup_{1\leqslant k_2\leqslant K_2}\left\|\widehat{\boldsymbol{\alpha}}_{k_2}^{0}-\boldsymbol{\alpha}_{k_2}^{*}\right\|_2 \leqslant C\phi_n. \tag{*}$$

Hence, Lemma 2 is proved. Then we give a rigorous proof of Lemma 3. Define the following two sets,

$$\mathcal{M}_{\mathcal{G}}=\left\{\boldsymbol{\beta}\in\mathbb{R}^q\times\mathbb{R}^n:\boldsymbol{\beta}_j=\boldsymbol{\beta}_m, j,m\in\mathcal{G}_{k_1}^{*}, k_1=1,\cdots,K_1\right\},$$
$$\mathcal{M}_{\mathcal{T}}=\left\{\boldsymbol{\gamma}\in\mathbb{R}^p\times\mathbb{R}^n:\boldsymbol{\gamma}_j=\boldsymbol{\gamma}_m, j,m\in\mathcal{T}_{k_2}^{*}, k_2=1,\cdots,K_2\right\}.$$

Recall that

$$\mathcal{M}_1=\left\{\boldsymbol{\beta}\in\mathbb{R}^q\times\mathbb{R}^n:\boldsymbol{\beta}_j=\boldsymbol{\beta}_m,(j,m)\in\mathcal{A}^{\mathrm{p}}\right\},$$
$$\mathcal{M}_2=\left\{\boldsymbol{\gamma}\in\mathbb{R}^p\times\mathbb{R}^n:\boldsymbol{\gamma}_j=\boldsymbol{\gamma}_m,(j,m)\in\mathcal{A}^{\mathrm{p}}\right\}.$$

Since the prior information does not violate the true heterogeneous structure, it is obvious that $\mathcal{M}_{\mathcal{G}}\subset\mathcal{M}_1$ and $\mathcal{M}_{\mathcal{T}}\subset\mathcal{M}_2$.

Let $\Omega_1:\mathcal{M}_{\mathcal{G}}\to\mathbb{R}^q\times\mathbb{R}^{K_1}$ be the one-to-one mapping that $\Omega_1(\boldsymbol{\beta})$ is the $q\times K_1$ matrix consisting of K_1 column vectors with dimension q and its k_1-th column equals to the common value of $\boldsymbol{\beta}_i$ for $i\in\mathcal{G}_{k_1}^{*}$. Let $\breve{\Omega}_1:\mathcal{M}_1\to\mathbb{R}^q\times\mathbb{R}^{K_1}$ be the mapping that

$$\breve{\Omega}_1(\boldsymbol{\beta})=\left\{\left|\mathcal{G}_{k_1}^{*}\right|^{-1}\sum_{i\in\mathcal{G}_{k_1}^{*}}\boldsymbol{\beta}_i, k_1=1,\cdots,K_1\right\}.$$

Let $\Omega_2:\mathcal{M}_{\mathcal{T}}\to\mathbb{R}^p\times\mathbb{R}^{K_2}$ be the one-to-one mapping that $\Omega_2(\boldsymbol{\gamma})$ is the $p\times K_2$ matrix consisting of K_2 column vectors with dimension p and its k_2-th column equals to the common value of $\boldsymbol{\gamma}_i$ for $i \in \mathcal{T}_{k_2}^*$. Let $\breve{\Omega}_2 : \mathcal{M}_2 \to \mathbb{R}^p \times \mathbb{R}^{K_2}$ be the mapping that

$$\breve{\Omega}_2(\boldsymbol{\gamma}) = \left\{ \left|\mathcal{T}_{k_2}^*\right|^{-1} \sum_{i \in \mathcal{T}_{k_2}^*} \boldsymbol{\gamma}_i, k_2 = 1, \cdots, K_2 \right\}.$$

Clearly, when $\boldsymbol{\beta} \in \mathcal{M}_{\mathcal{G}}$ and $\boldsymbol{\gamma} \in \mathcal{M}_{\mathcal{T}}$, $\Omega_1(\boldsymbol{\beta}) = \breve{\Omega}_1(\boldsymbol{\beta})$ and $\Omega_2(\boldsymbol{\gamma}) = \breve{\Omega}_2(\boldsymbol{\gamma})$. For any $\boldsymbol{\beta} \in \mathcal{M}_1$ and $\boldsymbol{\gamma} \in \mathcal{M}_2$, define $\breve{\boldsymbol{\beta}} = \Omega_1^{-1}\left(\breve{\Omega}_1(\boldsymbol{\beta})\right)$ and $\breve{\boldsymbol{\gamma}} = \Omega_2^{-1}\left(\breve{\Omega}_2(\boldsymbol{\gamma})\right)$. Then, we consider the neighborhoods of $(\boldsymbol{\beta}^*, \boldsymbol{\gamma}^*)$ and $\left(\widehat{\boldsymbol{\beta}}^0, \widehat{\boldsymbol{\gamma}}^0\right)$,

$$\Theta = \left\{(\boldsymbol{\beta}, \boldsymbol{\gamma}) : \boldsymbol{\beta} \in \mathcal{M}_1, \boldsymbol{\gamma} \in \mathcal{M}_2, \|\boldsymbol{\beta}_i - \boldsymbol{\beta}_i^*\|_2 \leqslant C\phi_n, \|\boldsymbol{\gamma}_i - \boldsymbol{\gamma}_i^*\|_2 \leqslant C\phi_n\right\},$$

$$\Theta_n = \left\{(\boldsymbol{\beta}, \boldsymbol{\gamma}) : \boldsymbol{\beta} \in \mathcal{M}_1, \boldsymbol{\gamma} \in \mathcal{M}_2, \left\|\boldsymbol{\beta}_i - \widehat{\boldsymbol{\beta}}_i^0\right\|_2 \leqslant \psi_n, \|\boldsymbol{\gamma}_i - \widehat{\boldsymbol{\gamma}}_i^0\|_2 \leqslant \psi_n\right\},$$

where sequence $\{\psi_n\}$ satisfies $\psi_n \to 0$. By Lemma 2, with probability at least $1 - 2n^{-1}$, (*) holds. The proof of Lemma 3 contains two steps, (1) When (*) holds, for any $(\boldsymbol{\beta}, \boldsymbol{\gamma}) \in \Theta$, we have $\left(\breve{\boldsymbol{\beta}}, \breve{\boldsymbol{\gamma}}\right) \in \Theta$ and $Q\left(\breve{\boldsymbol{\beta}}, \breve{\boldsymbol{\gamma}}\right) \geqslant Q\left(\widehat{\boldsymbol{\beta}}^0, \widehat{\boldsymbol{\gamma}}^0\right)$; (2) With probability at least $1 - 2n^{-1}$, for any $(\boldsymbol{\beta}, \boldsymbol{\gamma}) \in \Theta \cap \Theta_n$, we have that $Q(\boldsymbol{\beta}, \boldsymbol{\gamma}) \geqslant Q\left(\breve{\boldsymbol{\beta}}, \breve{\boldsymbol{\gamma}}\right)$.

Now we give the proofs of (1) and (2). With respect to $\boldsymbol{\beta} \in \mathcal{M}_1$ and $\boldsymbol{\gamma} \in \mathcal{M}_2$, we decompose $Q(\boldsymbol{\beta}, \boldsymbol{\gamma})$ into two parts $L(\boldsymbol{\beta}, \boldsymbol{\gamma})$ and $P(\boldsymbol{\beta}, \boldsymbol{\gamma})$, where the loss is

$$L(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \frac{1}{2} \sum_{i=1}^n \left(\|\boldsymbol{X}_i - \boldsymbol{\beta}_i\|_2^2 + \|\boldsymbol{Z}_i - \boldsymbol{\gamma}_i\|_2^2\right),$$

and the penalties are

$$\begin{aligned} P(\boldsymbol{\beta}, \boldsymbol{\gamma}) &= \sum_{(j,m) \in \mathcal{A}} p\left(\left(\left\|\boldsymbol{\beta}_j - \boldsymbol{\beta}_m\right\|_2^2 + \left\|\boldsymbol{\gamma}_j - \boldsymbol{\gamma}_m\right\|_2^2\right)^{\frac{1}{2}}; \lambda_1\right) + \sum_{(j,m) \in \mathcal{A}} p\left(\left\|\boldsymbol{\beta}_j - \boldsymbol{\beta}_m\right\|_2; \lambda_2\right) \\ &= \sum_{(j,m) \in \mathcal{A} \backslash \mathcal{A}^{\mathrm{p}}} p\left(\left(\left\|\boldsymbol{\beta}_j - \boldsymbol{\beta}_m\right\|_2^2 + \left\|\boldsymbol{\gamma}_j - \boldsymbol{\gamma}_m\right\|_2^2\right)^{\frac{1}{2}}; \lambda_1\right) + \sum_{(j,m) \in \mathcal{A} \backslash \mathcal{A}^{\mathrm{p}}} p\left(\left\|\boldsymbol{\beta}_j - \boldsymbol{\beta}_m\right\|_2; \lambda_2\right). \end{aligned}$$

For any $(\boldsymbol{\beta}, \boldsymbol{\gamma}) \in \Theta$, we have $\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_i^*\|_2 \leqslant C\phi_n$ and $\|\boldsymbol{\gamma}_i - \boldsymbol{\gamma}_i^*\|_2 \leqslant C\phi_n$, then

$$\|\breve{\gamma}_i - \boldsymbol{\gamma}_i^*\|_2 = \left\| \left|\mathcal{T}_{k_2}^*\right|^{-1} \sum_{i \in \mathcal{T}_{k_2}^*} (\boldsymbol{\gamma}_i - \boldsymbol{\gamma}_i^*) \right\|_2 \leqslant \left|\mathcal{T}_{k_2}^*\right|^{-1} \sum_{i \in \mathcal{T}_{k_2}^*} \|\boldsymbol{\gamma}_i - \boldsymbol{\gamma}_i^*\|_2 \leqslant C\phi_n.$$

Similarly, $\left\|\breve{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i^*\right\|_2 \leqslant C\phi_n$. Hence, $\left(\breve{\boldsymbol{\beta}}, \breve{\boldsymbol{\gamma}}\right) \in \Theta$. Let $\Omega_1(\boldsymbol{\beta}) = \boldsymbol{\xi} = (\boldsymbol{\xi}_1, \cdots, \boldsymbol{\xi}_{K_1})$, by definition, $\boldsymbol{\xi}_{k_1} = \left|\mathcal{G}_{k_1}^*\right|^{-1} \sum_{i \in \mathcal{G}_{k_1}^*} \boldsymbol{\beta}_i$ for $k_1 = 1, \cdots, K_1$ and $\breve{\boldsymbol{\beta}} = \Omega_1^{-1}\left(\breve{\Omega}_1(\boldsymbol{\beta})\right)$ whose i-th column vector is $\boldsymbol{\xi}_{k_1}$ for $i \in \mathcal{G}_{k_1}^*$. Since $\left(\breve{\boldsymbol{\beta}}, \breve{\boldsymbol{\gamma}}\right) \in \Theta$, we have that $\left\|\boldsymbol{\xi}_{k_1} - \boldsymbol{\xi}_{k_1}^*\right\|_2 \leqslant C\phi_n$ for $k_1 = 1, \cdots, K_1$. Hence, for any $1 \leqslant k_1 \neq k_1' \leqslant K_1$,

$$\begin{aligned}
\left\|\boldsymbol{\xi}_{k_1} - \boldsymbol{\xi}_{k_1'}\right\|_2 &\geqslant \left\|\boldsymbol{\xi}_{k_1}^* - \boldsymbol{\xi}_{k_1'}^*\right\|_2 - \left\|\boldsymbol{\xi}_{k_1}^* - \boldsymbol{\xi}_{k_1}\right\|_2 - \left\|\boldsymbol{\xi}_{k_1'} - \boldsymbol{\xi}_{k_1'}^*\right\|_2 \\
&\geqslant b_n - 2C\phi_n > (a + \kappa)\lambda_2 - 2C\phi_n > a\lambda_2.
\end{aligned}$$

Therefore, $p\left(\left\|\boldsymbol{\xi}_{k_1} - \boldsymbol{\xi}_{k_1'}\right\|_2; \lambda_2\right)$ is a constant. When (*) holds, in the same discussion, $p\left(\left\|\widehat{\boldsymbol{\xi}}_{k_1}^0 - \widehat{\boldsymbol{\xi}}_{k_1'}^0\right\|_2; \lambda_2\right)$ is the same constant. Hence,

$$\sum_{(j,m) \in \mathcal{A}} p\left(\left\|\breve{\boldsymbol{\beta}}_j - \breve{\boldsymbol{\beta}}_m\right\|_2; \lambda_2\right) = \sum_{(j,m) \in \mathcal{A}} p\left(\left\|\widehat{\boldsymbol{\beta}}_j^0 - \widehat{\boldsymbol{\beta}}_m^0\right\|_2; \lambda_2\right).$$

In a similar way, $\left\|\boldsymbol{\alpha}_{k_2} - \boldsymbol{\alpha}_{k_2'}\right\|_2 \geqslant d_n - 2C\phi_n > a\lambda_1$. Therefore, for any $1 \leqslant k_2 \neq k_2' \leqslant K_2$,

$$p\left(\left(\left\|\boldsymbol{\xi}_{h(k_2)} - \boldsymbol{\xi}_{h(k_2\prime)}\right\|_2^2 + \left\|\boldsymbol{\alpha}_{k_2} - \boldsymbol{\alpha}_{k_2'}\right\|_2^2\right)^{\frac{1}{2}}; \lambda_1\right)$$

is a constant, where $h(k_2) = \sum_{k_1=1}^{K_1} k_1 \mathbb{I}\left(k_2 \in \mathcal{H}_{k_1}^*\right)$. Hence,

$$\begin{aligned}
&\sum_{(j,m) \in \mathcal{A}} p\left(\left(\left\|\breve{\boldsymbol{\beta}}_j - \breve{\boldsymbol{\beta}}_m\right\|_2^2 + \left\|\breve{\gamma}_j - \breve{\gamma}_m\right\|_2^2\right)^{\frac{1}{2}}; \lambda_1\right) \\
&= \sum_{(j,m) \in \mathcal{A}} p\left(\left(\left\|\widehat{\boldsymbol{\beta}}_j^0 - \widehat{\boldsymbol{\beta}}_m^0\right\|_2^2 + \left\|\widehat{\gamma}_j^0 - \widehat{\gamma}_m^0\right\|_2^2\right)^{\frac{1}{2}}; \lambda_1\right).
\end{aligned}$$

Hence, $P\left(\breve{\boldsymbol{\beta}}, \breve{\boldsymbol{\gamma}}\right) = P\left(\widehat{\boldsymbol{\beta}}^0, \widehat{\boldsymbol{\gamma}}^0\right)$. Since $\left(\widehat{\boldsymbol{\beta}}^0, \widehat{\boldsymbol{\gamma}}^0\right)$ is the global minimizer of $L(\boldsymbol{\beta}, \boldsymbol{\gamma})$ subject to $\boldsymbol{\beta} \in \mathcal{M}_{\mathcal{G}}$ and $\boldsymbol{\gamma} \in \mathcal{M}_{\mathcal{T}}$, we have that $Q\left(\breve{\boldsymbol{\beta}}, \breve{\boldsymbol{\gamma}}\right) \geqslant Q\left(\widehat{\boldsymbol{\beta}}^0, \widehat{\boldsymbol{\gamma}}^0\right)$. Hence, (1) is proved.

For any $(\boldsymbol{\beta}, \boldsymbol{\gamma}) \in \Theta \cap \Theta_n$, we have that $\left(\breve{\boldsymbol{\beta}}, \breve{\boldsymbol{\gamma}}\right) \in \Theta \cap \Theta_n$. By Taylor's expansion, let $\bar{\boldsymbol{\beta}} = \chi \boldsymbol{\beta} + (1-\chi)\breve{\boldsymbol{\beta}}$ and $\bar{\boldsymbol{\gamma}} = \chi \boldsymbol{\gamma} + (1-\chi)\breve{\boldsymbol{\gamma}}$ where $\chi \in [0,1]$, we have that $Q(\boldsymbol{\beta}, \boldsymbol{\gamma}) - Q\left(\breve{\boldsymbol{\beta}}, \breve{\boldsymbol{\gamma}}\right) = \Gamma_1 + \Gamma_2 + \Gamma_3$, where

$$\Gamma_1 = -\sum_{i=1}^{n}\left(\boldsymbol{X}_i - \bar{\boldsymbol{\beta}}_i\right)^{\mathrm{T}}\left(\boldsymbol{\beta}_i - \breve{\boldsymbol{\beta}}_i\right) - \sum_{i=1}^{n}\left(\boldsymbol{Z}_i - \bar{\boldsymbol{\gamma}}_i\right)^{\mathrm{T}}\left(\boldsymbol{\gamma}_i - \breve{\boldsymbol{\gamma}}_i\right) \triangleq \Gamma_{11} + \Gamma_{12},$$

$$\Gamma_2 = \sum_{(j,m)\in\mathcal{A}\backslash\mathcal{A}^{\mathrm{p}},(m,j)\in\mathcal{A}\backslash\mathcal{A}^{\mathrm{p}}} p'\left(\left\|\bar{\boldsymbol{\theta}}_j - \bar{\boldsymbol{\theta}}_m\right\|_2; \lambda_1\right)\left\|\bar{\boldsymbol{\theta}}_j - \bar{\boldsymbol{\theta}}_m\right\|_2^{-1}\left(\bar{\boldsymbol{\theta}}_j - \bar{\boldsymbol{\theta}}_m\right)^{\mathrm{T}}\left(\boldsymbol{\theta}_j - \breve{\boldsymbol{\theta}}_j\right),$$

$$\Gamma_3 = \sum_{(j,m)\in\mathcal{A}\backslash\mathcal{A}^{\mathrm{p}},(m,j)\in\mathcal{A}\backslash\mathcal{A}^{\mathrm{p}}} p'\left(\left\|\bar{\boldsymbol{\beta}}_j - \bar{\boldsymbol{\beta}}_m\right\|_2; \lambda_2\right)\left\|\bar{\boldsymbol{\beta}}_j - \bar{\boldsymbol{\beta}}_m\right\|_2^{-1}\left(\bar{\boldsymbol{\beta}}_j - \bar{\boldsymbol{\beta}}_m\right)^{\mathrm{T}}\left(\boldsymbol{\beta}_j - \breve{\boldsymbol{\beta}}_j\right),$$

where $\boldsymbol{\theta}_i = \left(\boldsymbol{\beta}_i^{\mathrm{T}}, \boldsymbol{\gamma}_i^{\mathrm{T}}\right)^{\mathrm{T}}$. For Γ_{11}, let $\boldsymbol{B}_{1j} = \boldsymbol{X}_j - \bar{\boldsymbol{\beta}}_j$, we have that

$$\begin{aligned}\Gamma_{11} &= -\sum_{j=1}^{n}\boldsymbol{B}_{1j}^{\mathrm{T}}\left(\boldsymbol{\beta}_j - \breve{\boldsymbol{\beta}}_j\right) = -\sum_{k_1=1}^{K_1}\sum_{j\in\mathcal{G}_{k_1}^*}\sum_{m\in\mathcal{G}_{k_1}^*}\left|\mathcal{G}_{k_1}^*\right|^{-1}\boldsymbol{B}_{1j}^{\mathrm{T}}\left(\boldsymbol{\beta}_j - \boldsymbol{\beta}_m + \boldsymbol{\beta}_m - \breve{\boldsymbol{\beta}}_j\right)\\ &= -\sum_{k_1=1}^{K_1}\sum_{j,m\in\mathcal{G}_{k_1}^*}\left|\mathcal{G}_{k_1}^*\right|^{-1}\boldsymbol{B}_{1j}^{\mathrm{T}}\left(\boldsymbol{\beta}_j - \boldsymbol{\beta}_m\right) - \sum_{k_1=1}^{K_1}\sum_{j\in\mathcal{G}_{k_1}^*}\boldsymbol{B}_{1j}^{\mathrm{T}}\left(\sum_{m\in\mathcal{G}_{k_1}^*}\left|\mathcal{G}_{k_1}^*\right|^{-1}\boldsymbol{\beta}_m - \breve{\boldsymbol{\beta}}_j\right).\end{aligned}$$

Note that for $j, m \in \mathcal{G}_{k_1}^*$, $\breve{\boldsymbol{\beta}}_j = \breve{\boldsymbol{\beta}}_m = \sum_{m\in\mathcal{G}_{k_1}^*}\left|\mathcal{G}_{k_1}^*\right|^{-1}\boldsymbol{\beta}_m$, then

$$\begin{aligned}\Gamma_{11} &= -\sum_{k_1=1}^{K_1}\sum_{j,m\in\mathcal{G}_{k_1}^*}\frac{1}{2}\left|\mathcal{G}_{k_1}^*\right|^{-1}\left[\boldsymbol{B}_{1j}^{\mathrm{T}}\left(\boldsymbol{\beta}_j - \boldsymbol{\beta}_m\right) + \boldsymbol{B}_{1m}^{\mathrm{T}}\left(\boldsymbol{\beta}_m - \boldsymbol{\beta}_j\right)\right]\\ &= -\sum_{k_1=1}^{K_1}\sum_{j,m\in\mathcal{G}_{k_1}^*,(j,m)\in\mathcal{A}\backslash\mathcal{A}^{\mathrm{p}}}\left|\mathcal{G}_{k_1}^*\right|^{-1}\left(\boldsymbol{B}_{1j} - \boldsymbol{B}_{1m}\right)^{\mathrm{T}}\left(\boldsymbol{\beta}_j - \boldsymbol{\beta}_m\right)\\ &\geqslant -2\sup_{1\leqslant i\leqslant n}\left\|\boldsymbol{B}_{1i}\right\|_2\sum_{k_1=1}^{K_1}\sum_{j,m\in\mathcal{G}_{k_1}^*,(j,m)\in\mathcal{A}\backslash\mathcal{A}^{\mathrm{p}}} G_{\min}^{-1}\left\|\boldsymbol{\beta}_j - \boldsymbol{\beta}_m\right\|_2.\end{aligned}$$

Note that $\boldsymbol{X}_i = \boldsymbol{\beta}_i^* + \boldsymbol{\epsilon}_{1i}$, then we have $\boldsymbol{B}_{1i} = \boldsymbol{\epsilon}_{1i} + \chi\left(\boldsymbol{\beta}_i^* - \boldsymbol{\beta}_i\right) + (1-\chi)\left(\boldsymbol{\beta}_i^* - \breve{\boldsymbol{\beta}}_i\right)$. Hence,

$$\left\|\boldsymbol{B}_{1i}\right\|_2 \leqslant \left\|\boldsymbol{\epsilon}_{1i}\right\|_2 + \chi\left\|\boldsymbol{\beta}_i^* - \boldsymbol{\beta}_i\right\|_2 + (1-\chi)\left\|\boldsymbol{\beta}_i^* - \breve{\boldsymbol{\beta}}_i\right\|_2 \leqslant \left\|\boldsymbol{\epsilon}_{1i}\right\|_2 + C\phi_n.$$

Note that $\boldsymbol{\epsilon}_{1i}$ follows sub-Gaussian distribution with variance proxy σ_0^2 and $\mathrm{E}\left(\boldsymbol{\epsilon}_{1i}\right) = \mathbf{0}$. By Lemma 1, for fixed i, with probability at least $1 - n^{-2}$, $\left\|\boldsymbol{\epsilon}_{1i}\right\|_2 \leqslant 4\sigma_0 q^{\frac{1}{2}} + 2\sigma_0\left(4\log n\right)^{\frac{1}{2}}$. Hence, with probability at least $1 - n^{-1}$,

$$\sup_{1\leqslant i\leqslant n}\left\|\boldsymbol{B}_{1i}\right\|_2 \leqslant C_1\left(q^{\frac{1}{2}} + (\log n)^{\frac{1}{2}}\right),$$

where C_1 is a positive constant. Then,

$$\Gamma_{11} \geqslant -\sum_{k_1=1}^{K_1}\sum_{j,m\in\mathcal{G}_{k_1}^*,(j,m)\in\mathcal{A}\setminus\mathcal{A}^{\mathrm{p}}} 2C_1 G_{\min}^{-1}\left(q^{\frac{1}{2}} + (\log n)^{\frac{1}{2}}\right)\left\|\boldsymbol{\beta}_j - \boldsymbol{\beta}_m\right\|_2.$$

Similarly, there exists a positive constant C_2, with probability at least $1 - n^{-1}$,

$$\begin{aligned}\Gamma_{12} &\geqslant -\sum_{k_2=1}^{K_2}\sum_{j,m\in\mathcal{T}_{k_2}^*,(j,m)\in\mathcal{A}\setminus\mathcal{A}^{\mathrm{p}}} 2C_2 T_{\min}^{-1}\left(p^{\frac{1}{2}} + (\log n)^{\frac{1}{2}}\right)\left\|\boldsymbol{\gamma}_j - \boldsymbol{\gamma}_m\right\|_2 \\ &\geqslant -\sum_{k_2=1}^{K_2}\sum_{j,m\in\mathcal{T}_{k_2}^*,(j,m)\in\mathcal{A}\setminus\mathcal{A}^{\mathrm{p}}} 2C_2 T_{\min}^{-1}\left(p^{\frac{1}{2}} + (\log n)^{\frac{1}{2}}\right)\left\|\boldsymbol{\theta}_j - \boldsymbol{\theta}_m\right\|_2.\end{aligned}$$

Hence, with probability at least $1 - 2n^{-1}$, the following holds,

$$\begin{aligned}\Gamma_1 \geqslant &-\sum_{k_1=1}^{K_1}\sum_{j,m\in\mathcal{G}_{k_1}^*,(j,m)\in\mathcal{A}\setminus\mathcal{A}^{\mathrm{p}}} 2C_1 G_{\min}^{-1}\left(q^{\frac{1}{2}} + (\log n)^{\frac{1}{2}}\right)\left\|\boldsymbol{\beta}_j - \boldsymbol{\beta}_m\right\|_2 \\ &-\sum_{k_2=1}^{K_2}\sum_{j,m\in\mathcal{T}_{k_2}^*,(j,m)\in\mathcal{A}\setminus\mathcal{A}^{\mathrm{p}}} 2C_2 T_{\min}^{-1}\left(p^{\frac{1}{2}} + (\log n)^{\frac{1}{2}}\right)\left\|\boldsymbol{\theta}_j - \boldsymbol{\theta}_m\right\|_2.\end{aligned}$$

For Γ_2, we have that

$$\begin{aligned}
\Gamma_2 =& \sum_{(j,m)\in\mathcal{A}\backslash\mathcal{A}^{\mathrm{p}}} p'\left(\left\|\bar{\boldsymbol{\theta}}_j-\bar{\boldsymbol{\theta}}_m\right\|_2;\lambda_1\right)\left\|\bar{\boldsymbol{\theta}}_j-\bar{\boldsymbol{\theta}}_m\right\|_2^{-1}\left(\bar{\boldsymbol{\theta}}_j-\bar{\boldsymbol{\theta}}_m\right)^{\mathrm{T}}\left(\boldsymbol{\theta}_j-\breve{\boldsymbol{\theta}}_j\right)\\
&+\sum_{(m,j)\in\mathcal{A}\backslash\mathcal{A}^{\mathrm{p}}} p'\left(\left\|\bar{\boldsymbol{\theta}}_j-\bar{\boldsymbol{\theta}}_m\right\|_2;\lambda_1\right)\left\|\bar{\boldsymbol{\theta}}_j-\bar{\boldsymbol{\theta}}_m\right\|_2^{-1}\left(\bar{\boldsymbol{\theta}}_j-\bar{\boldsymbol{\theta}}_m\right)^{\mathrm{T}}\left(\boldsymbol{\theta}_j-\breve{\boldsymbol{\theta}}_j\right)\\
=& \sum_{(j,m)\in\mathcal{A}\backslash\mathcal{A}^{\mathrm{p}}} p'\left(\left\|\bar{\boldsymbol{\theta}}_j-\bar{\boldsymbol{\theta}}_m\right\|_2;\lambda_1\right)\left\|\bar{\boldsymbol{\theta}}_j-\bar{\boldsymbol{\theta}}_m\right\|_2^{-1}\left(\bar{\boldsymbol{\theta}}_j-\bar{\boldsymbol{\theta}}_m\right)^{\mathrm{T}}\left(\left(\boldsymbol{\theta}_j-\breve{\boldsymbol{\theta}}_j\right)-\left(\boldsymbol{\theta}_m-\breve{\boldsymbol{\theta}}_m\right)\right).
\end{aligned}$$

Note that for $j,m\in\mathcal{T}_{k_2}^*$, we have that $\breve{\boldsymbol{\theta}}_j=\breve{\boldsymbol{\theta}}_m$, and

$$\left\|\bar{\boldsymbol{\theta}}_j-\bar{\boldsymbol{\theta}}_m\right\|_2^{-1}\left(\bar{\boldsymbol{\theta}}_j-\bar{\boldsymbol{\theta}}_m\right)^{\mathrm{T}}\left(\boldsymbol{\theta}_j-\boldsymbol{\theta}_m\right)=\left\|\boldsymbol{\theta}_j-\boldsymbol{\theta}_m\right\|_2.$$

Moreover,

$$\begin{aligned}
\left\|\bar{\boldsymbol{\theta}}_j-\bar{\boldsymbol{\theta}}_m\right\|_2 &\leqslant \left\|\bar{\boldsymbol{\theta}}_j-\breve{\boldsymbol{\theta}}_j\right\|_2+\left\|\breve{\boldsymbol{\theta}}_m-\bar{\boldsymbol{\theta}}_m\right\|_2\leqslant\left\|\boldsymbol{\theta}_j-\breve{\boldsymbol{\theta}}_j\right\|_2+\left\|\boldsymbol{\theta}_m-\breve{\boldsymbol{\theta}}_m\right\|_2\\
&\leqslant \left\|\boldsymbol{\theta}_j-\widehat{\boldsymbol{\theta}}_j^0\right\|_2+\left\|\widehat{\boldsymbol{\theta}}_j^0-\breve{\boldsymbol{\theta}}_j\right\|_2+\left\|\boldsymbol{\theta}_m-\widehat{\boldsymbol{\theta}}_m^0\right\|_2+\left\|\widehat{\boldsymbol{\theta}}_m^0-\breve{\boldsymbol{\theta}}_m\right\|_2\leqslant 8\psi_n.
\end{aligned}$$

Hence, for $j,m\in\mathcal{T}_{k_2}^*$,

$$p'\left(\left\|\bar{\boldsymbol{\theta}}_j-\bar{\boldsymbol{\theta}}_m\right\|_2;\lambda_1\right)\left\|\bar{\boldsymbol{\theta}}_j-\bar{\boldsymbol{\theta}}_m\right\|_2^{-1}\left(\bar{\boldsymbol{\theta}}_j-\bar{\boldsymbol{\theta}}_m\right)^{\mathrm{T}}\left(\boldsymbol{\theta}_j-\boldsymbol{\theta}_m\right)\geqslant p'\left(8\psi_n;\lambda_1\right)\left\|\boldsymbol{\theta}_j-\boldsymbol{\theta}_m\right\|_2.$$

Note that for $j\in\mathcal{T}_{k_2}^*$ and $m\in\mathcal{T}_{k_2'}^*$, we have that

$$\left\|\bar{\boldsymbol{\theta}}_j-\bar{\boldsymbol{\theta}}_m\right\|_2\geqslant\left\|\boldsymbol{\theta}_j^*-\boldsymbol{\theta}_m^*\right\|_2-\left\|\boldsymbol{\theta}_j^*-\bar{\boldsymbol{\theta}}_j\right\|_2-\left\|\bar{\boldsymbol{\theta}}_m-\boldsymbol{\theta}_m^*\right\|_2\geqslant d_n-2C\phi_n>a\lambda_1,$$

hence, $p'\left(\left\|\bar{\boldsymbol{\theta}}_j-\bar{\boldsymbol{\theta}}_m\right\|_2;\lambda_1\right)=0$. Then we have that

$$\Gamma_2\geqslant\sum_{k_2=1}^{K_2}\sum_{j,m\in\mathcal{T}_{k_2}^*,(j,m)\in\mathcal{A}\backslash\mathcal{A}^{\mathrm{p}}} p'\left(8\psi_n;\lambda_1\right)\left\|\boldsymbol{\theta}_j-\boldsymbol{\theta}_m\right\|_2.$$

In a similar way, we have that

$$\Gamma_3 \geqslant \sum_{k_1=1}^{K_1} \sum_{j,m \in \mathcal{G}_{k_1}^*, (j,m) \in \mathcal{A} \setminus \mathcal{A}^{\mathrm{p}}} p'\left(4\psi_n; \lambda_2\right) \left\|\boldsymbol{\beta}_j - \boldsymbol{\beta}_m\right\|_2 .$$

In conclusion,

$$\begin{aligned}
& Q\left(\boldsymbol{\beta}, \boldsymbol{\gamma}\right) - Q\left(\breve{\boldsymbol{\beta}}, \breve{\boldsymbol{\gamma}}\right) \\
\geqslant & \sum_{k_1=1}^{K_1} \sum_{j,m \in \mathcal{G}_{k_1}^*, (j,m) \in \mathcal{A} \setminus \mathcal{A}^{\mathrm{p}}} \left\{p'\left(4\psi_n; \lambda_2\right) - 2C_1 G_{\min}^{-1}\left(q^{\frac{1}{2}} + \left(\log n\right)^{\frac{1}{2}}\right)\right\} \left\|\boldsymbol{\beta}_j - \boldsymbol{\beta}_m\right\|_2 \\
& + \sum_{k_2=1}^{K_2} \sum_{j,m \in \mathcal{T}_{k_2}^*, (j,m) \in \mathcal{A} \setminus \mathcal{A}^{\mathrm{p}}} \left\{p'\left(8\psi_n; \lambda_1\right) - 2C_2 T_{\min}^{-1}\left(p^{\frac{1}{2}} + \left(\log n\right)^{\frac{1}{2}}\right)\right\} \left\|\boldsymbol{\theta}_j - \boldsymbol{\theta}_m\right\|_2 .
\end{aligned}$$

Note that $\psi_n \to 0$, then we have $p'\left(4\psi_n; \lambda_2\right) \to \lambda_2$ and $p'\left(8\psi_n; \lambda_1\right) \to \lambda_1$. Moreover, since $\min\left\{\lambda_1, \lambda_2\right\} \gg \phi_n = (q+p)^{\frac{1}{2}} T_{\min}^{-\frac{1}{2}} \left(\log n\right)^{\frac{1}{2}} \gg (q+p)^{\frac{1}{2}} T_{\min}^{-1} \left(\log n\right)^{\frac{1}{2}} > (q+p)^{\frac{1}{2}} G_{\min}^{-1} \left(\log n\right)^{\frac{1}{2}}$. Hence, with probability at least $1 - 2n^{-1}$, for sufficiently large n, $Q\left(\boldsymbol{\beta}, \boldsymbol{\gamma}\right) \geqslant Q\left(\breve{\boldsymbol{\beta}}, \breve{\boldsymbol{\gamma}}\right)$ and (2) is proved. By (1) and (2), we have that with probability at least $1 - 4n^{-1}$, (*) holds and for any $\left(\boldsymbol{\beta}, \boldsymbol{\gamma}\right) \in \Theta \cap \Theta_n$, we have that $Q\left(\boldsymbol{\beta}, \boldsymbol{\gamma}\right) \geqslant Q\left(\breve{\boldsymbol{\beta}}, \breve{\boldsymbol{\gamma}}\right) \geqslant Q\left(\widehat{\boldsymbol{\beta}}^0, \widehat{\boldsymbol{\gamma}}^0\right)$, resulting in $\left(\widehat{\boldsymbol{\beta}}^0, \widehat{\boldsymbol{\gamma}}^0\right)$ as a local minimizer of $Q\left(\boldsymbol{\beta}, \boldsymbol{\gamma}\right)$ subject to $\boldsymbol{\beta} \in \mathcal{M}_1$ and $\boldsymbol{\gamma} \in \mathcal{M}_2$. Hence, Lemma 3 is proved.

References

[1] Phillippe Rigollet and Jan-Christian Hütter. High dimensional statistics. *Lecture notes for course 18S997*, 813(814):46, 2015.

B Details of the computational algorithm

Recall $\mathcal{L}(\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\omega},\boldsymbol{\eta},\boldsymbol{v},\boldsymbol{u})$ and (4.1)-(4.3) in the main text. In this section, we give the details of the deveriation of (4.4) and (4.5).

Define $\mathrm{vec}(\cdot)$ as the vectorization of matrices, specifically, for a $r_1 \times r_2$ matrix $\boldsymbol{A} = (\boldsymbol{A}_1, \cdots, \boldsymbol{A}_{r_2})$ of each column being a r_1-dimensional vector, the vectorization of $\boldsymbol{A}$ is $\mathrm{vec}(\boldsymbol{A}) = \left(\boldsymbol{A}_1^{\mathrm{T}}, \cdots, \boldsymbol{A}_{r_2}^{\mathrm{T}}\right)^{\mathrm{T}}$ which is a $r_1 r_2$-dimensional vector. For sample data $\boldsymbol{X} = (\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n)$ and $\boldsymbol{Z} = (\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n)$, the vectorization $\mathrm{vec}\left(\left(\boldsymbol{X}^{\mathrm{T}}, \boldsymbol{Z}^{\mathrm{T}}\right)^{\mathrm{T}}\right)$ is a $n(q+p)$-dimensional vector. Similarly, the parameters matrix and Lagrange multipliers can be also expressed as the form of vectorization. By discarding terms independent of $(\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\omega},\boldsymbol{\eta})$, the augmented Lagrangian function $\mathcal{L}(\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\omega},\boldsymbol{\eta},\boldsymbol{v},\boldsymbol{u})$ can be reformulated as

$$
\begin{aligned}
\widetilde{\mathcal{L}} = & \frac{1}{2}\sum_{i=1}^{n}\left(\left\|\left(\boldsymbol{X}_i^{\mathrm{T}}, \boldsymbol{Z}_i^{\mathrm{T}}\right)^{\mathrm{T}} - \left(\boldsymbol{\beta}_i^{\mathrm{T}}, \boldsymbol{\gamma}_i^{\mathrm{T}}\right)^{\mathrm{T}}\right\|_2^2\right) \\
& + \sum_{(j,m)\in\mathcal{A}} p\left(\left(\|\boldsymbol{\omega}_{jm}\|_2^2 + \|\boldsymbol{\eta}_{jm}\|_2^2\right)^{\frac{1}{2}}; \lambda_1\right) + \sum_{(j,m)\in\mathcal{A}} p\left(\|\boldsymbol{\omega}_{jm}\|_2; \lambda_2\right) \\
& + \frac{\vartheta}{2}\sum_{(j,m)\in\mathcal{A}}\left\|\left(\boldsymbol{\beta}_j^{\mathrm{T}} - \boldsymbol{\beta}_m^{\mathrm{T}}, \boldsymbol{\gamma}_j^{\mathrm{T}} - \boldsymbol{\gamma}_m^{\mathrm{T}}\right)^{\mathrm{T}} - \left(\boldsymbol{\omega}_{jm}^{\mathrm{T}} - \vartheta^{-1}\boldsymbol{v}_{jm}^{\mathrm{T}}, \boldsymbol{\eta}_{jm}^{\mathrm{T}} - \vartheta^{-1}\boldsymbol{u}_{jm}^{\mathrm{T}}\right)^{\mathrm{T}}\right\|_2^2.
\end{aligned}
$$

Note that $\boldsymbol{\beta} \in \mathcal{M}_1$ and $\boldsymbol{\gamma} \in \mathcal{M}_2$. There exists unique matrices $\widetilde{\boldsymbol{\beta}} \in \mathbb{R}^q \times \mathbb{R}^K$ and $\widetilde{\boldsymbol{\gamma}} \in \mathbb{R}^p \times \mathbb{R}^K$ such that $\mathrm{vec}\left(\left(\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\gamma}^{\mathrm{T}}\right)^{\mathrm{T}}\right) = \boldsymbol{J}\mathrm{vec}\left(\left(\widetilde{\boldsymbol{\beta}}^{\mathrm{T}}, \widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\right)^{\mathrm{T}}\right)$. By discarding terms independent of $(\boldsymbol{\beta},\boldsymbol{\gamma})$, the optimization problem (4.1) is equivalent to minimizing that

$$
\begin{aligned}
f\left(\widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\gamma}}\right) = & \frac{1}{2}\left\|\mathrm{vec}\left(\left(\boldsymbol{X}^{\mathrm{T}}, \boldsymbol{Z}^{\mathrm{T}}\right)^{\mathrm{T}}\right) - \boldsymbol{J}\mathrm{vec}\left(\left(\widetilde{\boldsymbol{\beta}}^{\mathrm{T}}, \widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\right)^{\mathrm{T}}\right)\right\|_2^2 \\
& + \frac{\vartheta}{2}\left\|\boldsymbol{H}\boldsymbol{J}\mathrm{vec}\left(\left(\widetilde{\boldsymbol{\beta}}^{\mathrm{T}}, \widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\right)^{\mathrm{T}}\right) - \mathrm{vec}\left(\left(\boldsymbol{\omega}^{(t)\mathrm{T}} - \vartheta^{-1}\boldsymbol{v}^{(t)\mathrm{T}}, \boldsymbol{\eta}^{(t)\mathrm{T}} - \vartheta^{-1}\boldsymbol{u}^{(t)\mathrm{T}}\right)^{\mathrm{T}}\right)\right\|_2^2.
\end{aligned}
$$

Minimizing $f\left(\widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\gamma}}\right)$ is a standard quadratic convex optimization problem and the global solution is unique. Hence, the updates for $\left(\widetilde{\boldsymbol{\beta}}^{(t+1)}, \widetilde{\boldsymbol{\gamma}}^{(t+1)}\right)$ satisfy

$$\begin{aligned}
&\left(\boldsymbol{J}^{\mathrm{T}}\boldsymbol{J}+\vartheta \boldsymbol{J}^{\mathrm{T}}\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{J}\right)\operatorname{vec}\left(\left(\widetilde{\boldsymbol{\beta}}^{(t+1)\mathrm{T}}, \widetilde{\boldsymbol{\gamma}}^{(t+1)\mathrm{T}}\right)^{\mathrm{T}}\right)\\
=&\boldsymbol{J}^{\mathrm{T}}\operatorname{vec}\left(\left(\boldsymbol{X}^{\mathrm{T}}, \boldsymbol{Z}^{\mathrm{T}}\right)^{\mathrm{T}}\right)+\boldsymbol{J}^{\mathrm{T}}\boldsymbol{H}^{\mathrm{T}}\operatorname{vec}\left(\left(\vartheta\boldsymbol{\omega}^{(t)\mathrm{T}}-\boldsymbol{v}^{(t)\mathrm{T}}, \vartheta\boldsymbol{\eta}^{(t)\mathrm{T}}-\boldsymbol{u}^{(t)\mathrm{T}}\right)^{\mathrm{T}}\right).
\end{aligned}$$

It is noted that $\boldsymbol{J}^{\mathrm{T}}\boldsymbol{J}+\vartheta \boldsymbol{J}^{\mathrm{T}}\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{J}$ is a $K(q+p)\times K(q+p)$ matrix, calculation of its inverse matrix costs huge computational expenses in high-dimensional situation. By some linear algebra knowledge, we aim to find out the special structure of such defined matrix and obtain its inverse simply. By properties of Kronecker product, we have that

$$\begin{aligned}
&\boldsymbol{J}=\boldsymbol{L}\otimes\boldsymbol{I}_{q+p}, \quad \boldsymbol{J}^{\mathrm{T}}\boldsymbol{J}=\left(\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L}\right)\otimes\boldsymbol{I}_{q+p},\\
&\boldsymbol{H}\boldsymbol{J}=(\boldsymbol{D}\boldsymbol{L})\otimes\boldsymbol{I}_{q+p}, \quad \boldsymbol{J}^{\mathrm{T}}\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{J}=\left(\boldsymbol{L}^{\mathrm{T}}\boldsymbol{D}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{L}\right)\otimes\boldsymbol{I}_{q+p},\\
&\boldsymbol{J}^{\mathrm{T}}\boldsymbol{J}+\vartheta\boldsymbol{J}^{\mathrm{T}}\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{J}=\left(\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L}+\vartheta\boldsymbol{L}^{\mathrm{T}}\boldsymbol{D}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{L}\right)\otimes\boldsymbol{I}_{q+p}.
\end{aligned}$$

By properties of Kronecker product with vectorization ($\operatorname{vec}(\boldsymbol{A}\boldsymbol{B})=\left(\boldsymbol{B}^{\mathrm{T}}\otimes\boldsymbol{I}\right)\operatorname{vec}(\boldsymbol{A})$), the updates for $\left(\widetilde{\boldsymbol{\beta}}^{(t+1)}, \widetilde{\boldsymbol{\gamma}}^{(t+1)}\right)$ satisfy

$$\begin{aligned}
&\left(\widetilde{\boldsymbol{\beta}}^{(t+1)\mathrm{T}}, \widetilde{\boldsymbol{\gamma}}^{(t+1)\mathrm{T}}\right)^{\mathrm{T}}\left(\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L}+\vartheta\boldsymbol{L}^{\mathrm{T}}\boldsymbol{D}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{L}\right)\\
=&\left(\boldsymbol{X}^{\mathrm{T}}, \boldsymbol{Z}^{\mathrm{T}}\right)^{\mathrm{T}}\boldsymbol{L}+\left(\vartheta\boldsymbol{\omega}^{(t)\mathrm{T}}-\boldsymbol{v}^{(t)\mathrm{T}}, \vartheta\boldsymbol{\eta}^{(t)\mathrm{T}}-\boldsymbol{u}^{(t)\mathrm{T}}\right)^{\mathrm{T}}\boldsymbol{D}\boldsymbol{L}.
\end{aligned}$$

Since $\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L}+\vartheta\boldsymbol{L}^{\mathrm{T}}\boldsymbol{D}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{L}$ is a $K\times K$ matrix, and K is much smaller than $K(q+p)$ to some extent, hence its inverse matrix is efficient to calculate. The updates for $\left(\widetilde{\boldsymbol{\beta}}^{(t+1)}, \widetilde{\boldsymbol{\gamma}}^{(t+1)}\right)$ are

$$\begin{aligned}
&\left(\widetilde{\boldsymbol{\beta}}^{(t+1)\mathrm{T}}, \widetilde{\boldsymbol{\gamma}}^{(t+1)\mathrm{T}}\right)^{\mathrm{T}}\\
=&\left\{\left(\boldsymbol{X}^{\mathrm{T}}, \boldsymbol{Z}^{\mathrm{T}}\right)^{\mathrm{T}}+\left(\vartheta\boldsymbol{\omega}^{(t)\mathrm{T}}-\boldsymbol{v}^{(t)\mathrm{T}}, \vartheta\boldsymbol{\eta}^{(t)\mathrm{T}}-\boldsymbol{u}^{(t)\mathrm{T}}\right)^{\mathrm{T}}\boldsymbol{D}\right\}\boldsymbol{L}\left(\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L}+\vartheta\boldsymbol{L}^{\mathrm{T}}\boldsymbol{D}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{L}\right)^{-1}.
\end{aligned}$$

Then, the updates for $\left(\boldsymbol{\beta}^{(t+1)}, \boldsymbol{\gamma}^{(t+1)}\right)$ are

$$\operatorname{vec}\left(\left(\boldsymbol{\beta}^{(t+1)\mathrm{T}}, \boldsymbol{\gamma}^{(t+1)\mathrm{T}}\right)^{\mathrm{T}}\right) = \boldsymbol{J}\operatorname{vec}\left(\left(\widetilde{\boldsymbol{\beta}}^{(t+1)\mathrm{T}}, \widetilde{\boldsymbol{\gamma}}^{(t+1)\mathrm{T}}\right)^{\mathrm{T}}\right),$$

When there is no prior information, that is, $\mathcal{A}^{\mathrm{P}} = \emptyset$. Without any information, the matrix $\boldsymbol{L} = \boldsymbol{I}_n$ and $K = n$. By the definition of matrix $\boldsymbol{D}$, it holds that $\boldsymbol{I}_n + \vartheta \boldsymbol{D}^{\mathrm{T}}\boldsymbol{D} = (1 + n\vartheta)\,\boldsymbol{I}_n - \vartheta \mathbf{1}_n \mathbf{1}_n^{\mathrm{T}}$, where $\mathbf{1}_n$ is a n-dimensional vector whose all elements are 1. By Sherman-Morrison formula, it holds that $\left((1 + n\vartheta)\,\boldsymbol{I}_n - \vartheta \mathbf{1}_n \mathbf{1}_n^{\mathrm{T}}\right)^{-1} = (1 + n\vartheta)^{-1}\left(\boldsymbol{I}_n + \vartheta \mathbf{1}_n \mathbf{1}_n^{\mathrm{T}}\right)$. Hence,

$$\begin{aligned}
&\left(\boldsymbol{\beta}^{(t+1)\mathrm{T}}, \boldsymbol{\gamma}^{(t+1)\mathrm{T}}\right)^{\mathrm{T}} \\
=&\frac{1}{1 + n\vartheta}\left\{\left(\boldsymbol{X}^{\mathrm{T}}, \boldsymbol{Z}^{\mathrm{T}}\right)^{\mathrm{T}} + \left(\vartheta \boldsymbol{\omega}^{(t)\mathrm{T}} - \boldsymbol{v}^{(t)\mathrm{T}}, \vartheta \boldsymbol{\eta}^{(t)\mathrm{T}} - \boldsymbol{u}^{(t)\mathrm{T}}\right)^{\mathrm{T}} \boldsymbol{D}\right\}\left(\boldsymbol{I}_n + \vartheta \mathbf{1}_n \mathbf{1}_n^{\mathrm{T}}\right).
\end{aligned}$$

Let $\boldsymbol{\omega}_{jm}^{*(t)} = \boldsymbol{\beta}_j^{(t+1)} - \boldsymbol{\beta}_m^{(t+1)} + \vartheta^{-1}\boldsymbol{v}_{jm}^{(t)}$ and $\boldsymbol{\eta}_{jm}^{*(t)} = \boldsymbol{\gamma}_j^{(t+1)} - \boldsymbol{\gamma}_m^{(t+1)} + \vartheta^{-1}\boldsymbol{u}_{jm}^{(t)}$. By discarding the terms independent of $(\boldsymbol{\omega}, \boldsymbol{\eta})$, the optimization problem (4.2) is equivalent to minimizing

$$\frac{\vartheta}{2}\left\|\left(\boldsymbol{\omega}_{jm}^{\mathrm{T}}, \boldsymbol{\eta}_{jm}^{\mathrm{T}}\right)^{\mathrm{T}} - \left(\boldsymbol{\omega}_{jm}^{*(t)\mathrm{T}}, \boldsymbol{\eta}_{jm}^{*(t)\mathrm{T}}\right)^{\mathrm{T}}\right\|_2^2 + p\left(\left(\|\boldsymbol{\omega}_{jm}\|_2^2 + \|\boldsymbol{\eta}_{jm}\|_2^2\right)^{\frac{1}{2}}; \lambda_1\right) + p\left(\|\boldsymbol{\omega}_{jm}\|_2; \lambda_2\right),$$

with respect to $\left(\boldsymbol{\omega}_{jm}^{\mathrm{T}}, \boldsymbol{\eta}_{jm}^{\mathrm{T}}\right)^{\mathrm{T}}$. The penalty $p(\cdot;\lambda)$ is MCP with a fixed parameter $a > \vartheta^{-1}$. Denote $\boldsymbol{\zeta}_{jm} = \left(\boldsymbol{\omega}_{jm}^{\mathrm{T}}, \boldsymbol{\eta}_{jm}^{\mathrm{T}}\right)^{\mathrm{T}}$ and $(s)_+ = s\mathbb{I}(s > 0)$. The solution of (4.2) is $S\left(\boldsymbol{\omega}_{jm}^{*(t)}, \boldsymbol{\eta}_{jm}^{*(t)}\right)$, which can be expressed by the following discussion.

Case 1. If $\left\|\boldsymbol{\zeta}_{jm}^{*(t)}\right\|_2 > a\lambda_1$ and $\left\|\boldsymbol{\omega}_{jm}^{*(t)}\right\|_2 > a\lambda_2$, then $\boldsymbol{\omega}_{jm}^{(t+1)} = \boldsymbol{\omega}_{jm}^{*(t)}$ and $\boldsymbol{\eta}_{jm}^{(t+1)} = \boldsymbol{\eta}_{jm}^{*(t)}$.

Case 2. If $\left\|\boldsymbol{\zeta}_{jm}^{*(t)}\right\|_2 \leqslant a\lambda_1$ and $\frac{\left(1-\lambda_1\left(\vartheta\left\|\boldsymbol{\zeta}_{jm}^{*(t)}\right\|_2\right)^{-1}\right)_+}{1-(a\vartheta)^{-1}}\left\|\boldsymbol{\omega}_{jm}^{*(t)}\right\|_2 > a\lambda_2$, then,

$$\boldsymbol{\omega}_{jm}^{(t+1)} = \frac{\left(1 - \lambda_1\left(\vartheta\left\|\boldsymbol{\zeta}_{jm}^{*(t)}\right\|_2\right)^{-1}\right)_+}{1 - (a\vartheta)^{-1}}\boldsymbol{\omega}_{jm}^{*(t)}, \quad \boldsymbol{\eta}_{jm}^{(t+1)} = \frac{\left(1 - \lambda_1\left(\vartheta\left\|\boldsymbol{\zeta}_{jm}^{*(t)}\right\|_2\right)^{-1}\right)_+}{1 - (a\vartheta)^{-1}}\boldsymbol{\eta}_{jm}^{*(t)}.$$

Case 3. If $\left\|\boldsymbol{\eta}_{jm}^{*(t)}\right\|_2^2 + \left(\frac{\left(1-\lambda_2\left(\vartheta\left\|\boldsymbol{\omega}_{jm}^{*(t)}\right\|_2\right)^{-1}\right)_+}{1-(a\vartheta)^{-1}}\right)^2 \left\|\boldsymbol{\omega}_{jm}^{*(t)}\right\|_2^2 > (a\lambda_1)^2$ and $\left\|\boldsymbol{\omega}_{jm}^{*(t)}\right\|_2 \leqslant a\lambda_2$, then,

$$\boldsymbol{\omega}_{jm}^{(t+1)} = \frac{\left(1-\lambda_2\left(\vartheta\left\|\boldsymbol{\omega}_{jm}^{*(t)}\right\|_2\right)^{-1}\right)_+}{1-(a\vartheta)^{-1}}\boldsymbol{\omega}_{jm}^{*(t)}, \quad \boldsymbol{\eta}_{jm}^{(t+1)} = \boldsymbol{\eta}_{jm}^{*(t)}.$$

Other cases. Using the local quadratic approximation technique to obtain an explicit solution (Here, to avoid 0 appearing on the denominator, we can make a very small disturbance on $\left\|\boldsymbol{\zeta}_{jm}^{(t)}\right\|_2$ and $\left\|\boldsymbol{\omega}_{jm}^{(t)}\right\|_2$ in computation, such as 10^{-7}), then,

$$\begin{aligned}
\boldsymbol{\omega}_{jm}^{(t+1)} &= \left(1+\frac{p'\left(\left\|\boldsymbol{\zeta}_{jm}^{(t)}\right\|_2;\lambda_1\right)}{\vartheta\left\|\boldsymbol{\zeta}_{jm}^{(t)}\right\|_2}+\frac{p'\left(\left\|\boldsymbol{\omega}_{jm}^{(t)}\right\|_2;\lambda_2\right)}{\vartheta\left\|\boldsymbol{\omega}_{jm}^{(t)}\right\|_2}\right)^{-1}\boldsymbol{\omega}_{jm}^{*(t)},\\
\boldsymbol{\eta}_{jm}^{(t+1)} &= \left(1+\frac{p'\left(\left\|\boldsymbol{\zeta}_{jm}^{(t)}\right\|_2;\lambda_1\right)}{\vartheta\left\|\boldsymbol{\zeta}_{jm}^{(t)}\right\|_2}\right)^{-1}\boldsymbol{\eta}_{jm}^{*(t)}.
\end{aligned}$$

C Additional numerical results

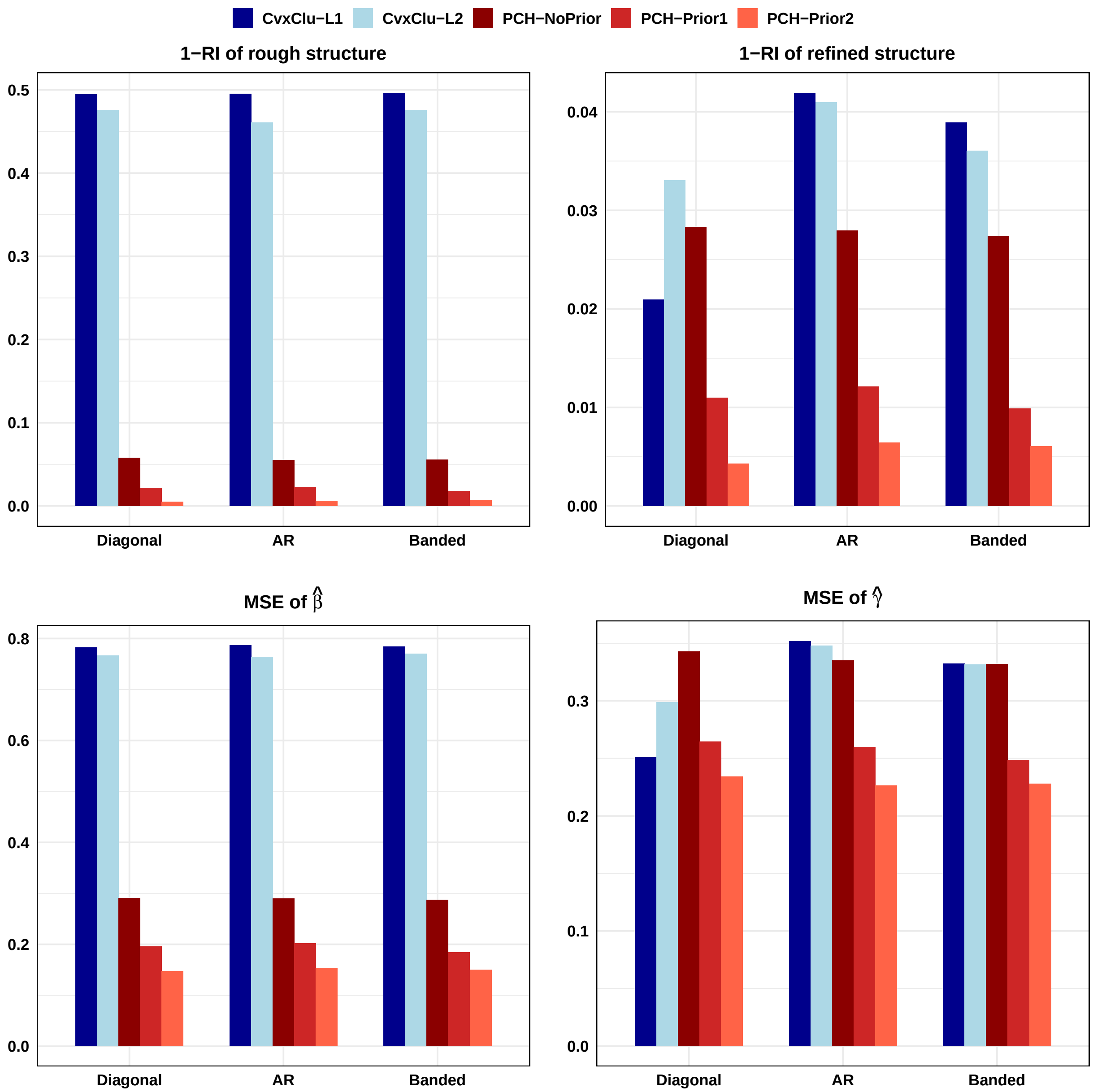

Figure S1: Simulation results with Simulation 1 and $\mu_1 = 1.2$. In each subfigure, horizontal axis displays our proposed methods and alternatives with three different covariance matrices, and longitudinal axis displays the mean of corresponding measurement values under 100 simulated replicates. The top-left subfigure displays $1 - \text{RI}$ of rough clustering structure, the top-right subfigure displays $1 - \text{RI}$ of refined clustering structure, the bottom-left subfigure displays MSE of $\hat{\boldsymbol{\beta}}$, and the bottom-right subfigure displays MSE of $\hat{\boldsymbol{\gamma}}$.

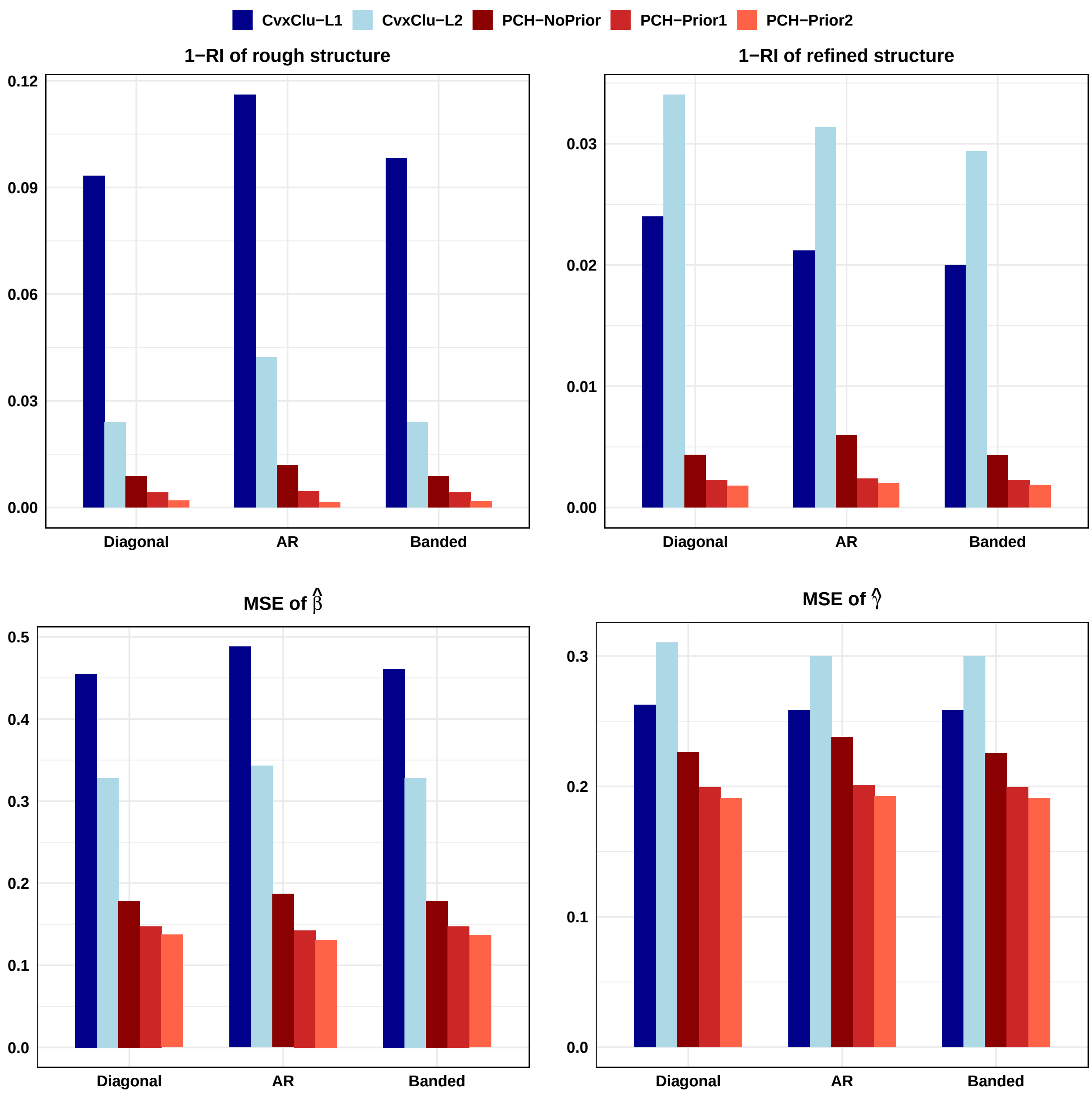

Figure S2: Simulation results with Simulation 1 and $\mu_2 = 1.6$. In each subfigure, horizontal axis displays our proposed methods and alternatives with three different covariance matrices, and longitudinal axis displays the mean of corresponding measurement values under 100 simulated replicates. The top-left subfigure displays $1 - \mathrm{RI}$ of rough clustering structure, the top-right subfigure displays $1 - \mathrm{RI}$ of refined clustering structure, the bottom-left subfigure displays MSE of $\hat{\boldsymbol{\beta}}$, and the bottom-right subfigure displays MSE of $\hat{\boldsymbol{\gamma}}$.

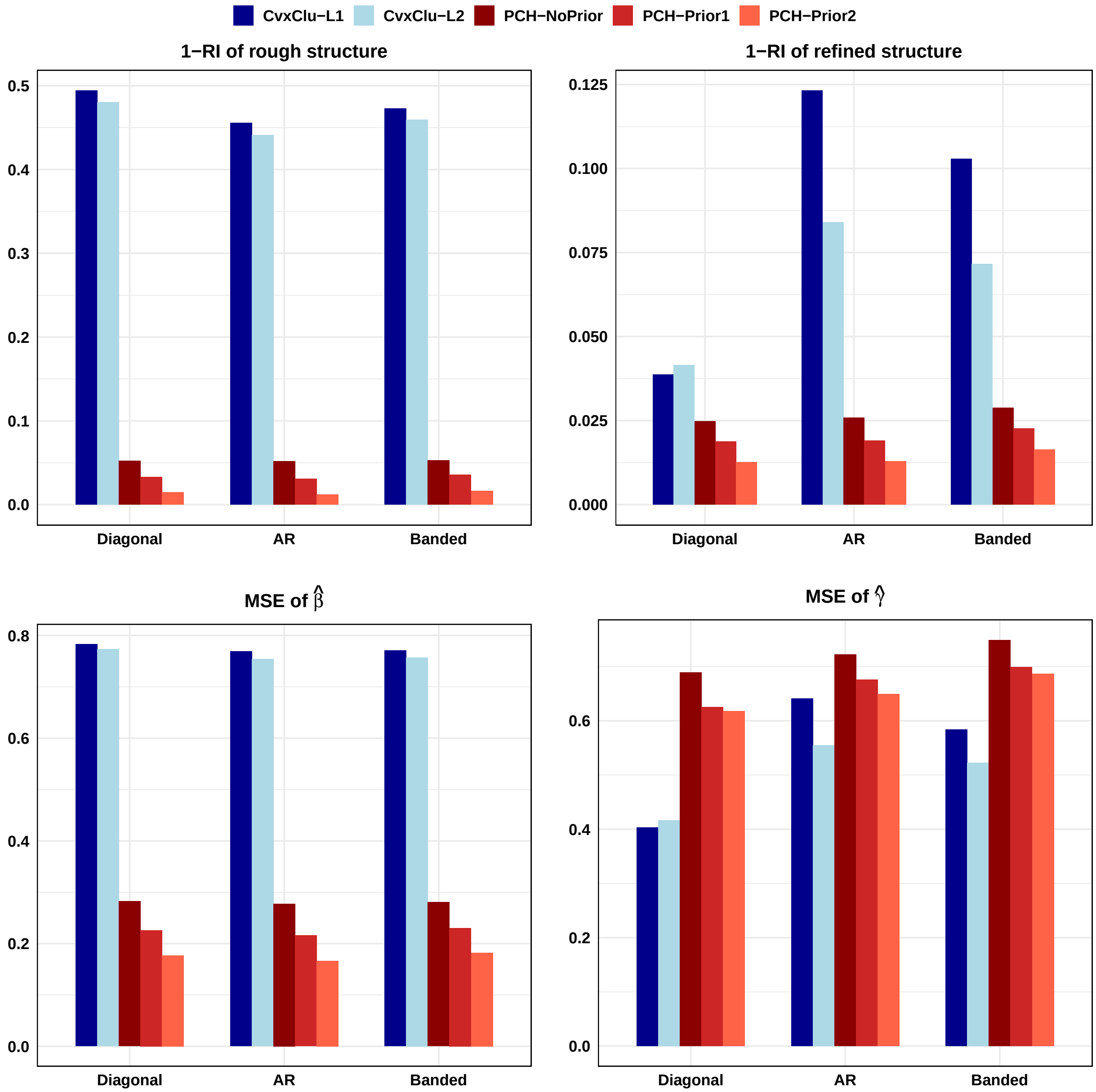

Figure S3: Simulation results with Simulation 2 and $\mu_1 = 1.2$. In each subfigure, horizontal axis displays our proposed methods and alternatives with three different covariance matrices, and longitudinal axis displays the mean of corresponding measurement values under 100 simulated replicates. The top-left subfigure displays $1-\text{RI}$ of rough clustering structure, the top-right subfigure displays $1-\text{RI}$ of refined clustering structure, the bottom-left subfigure displays MSE of $\hat{\boldsymbol{\beta}}$, and the bottom-right subfigure displays MSE of $\hat{\boldsymbol{\gamma}}$.

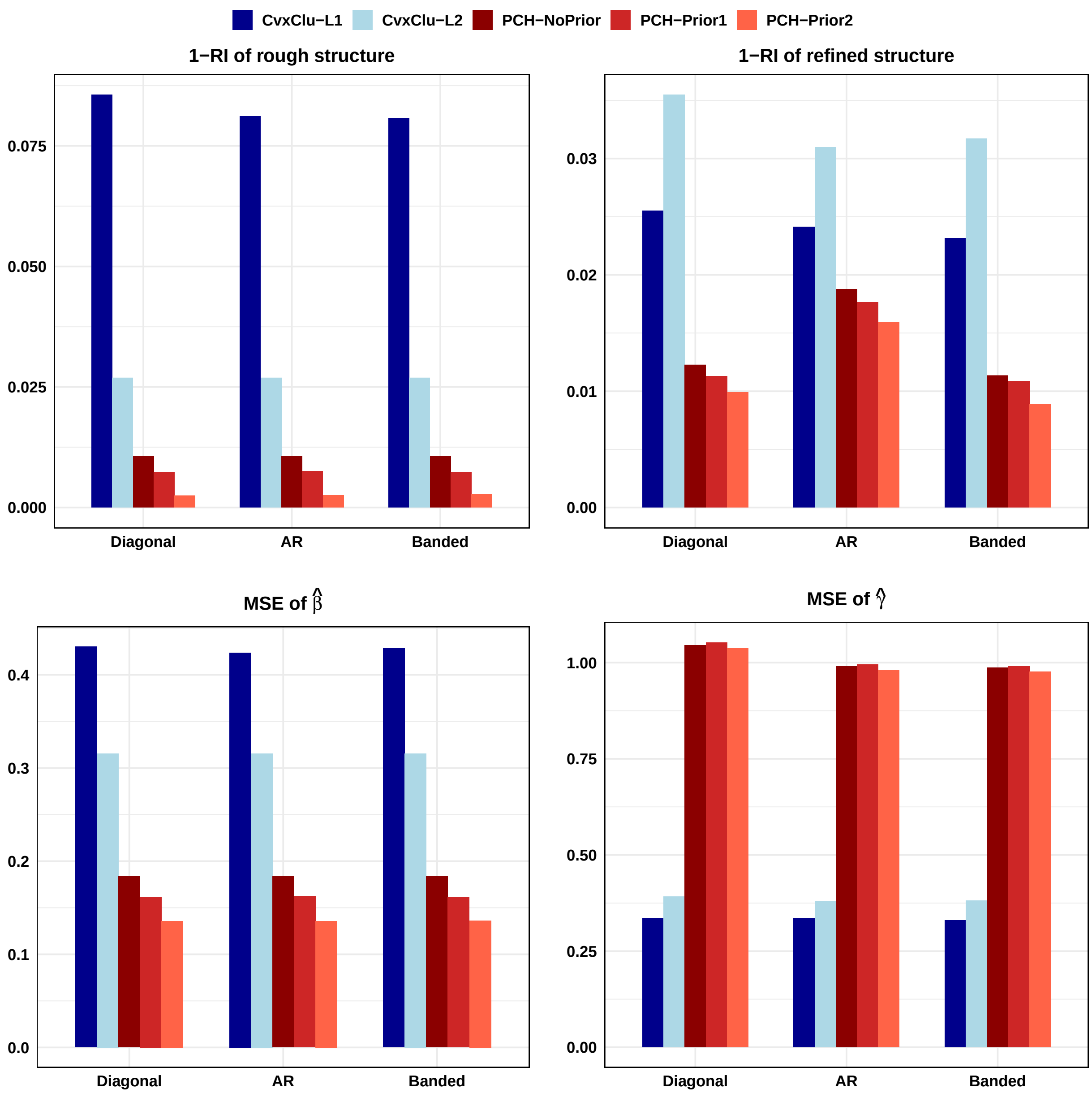

Figure S4: Simulation results with Simulation 2 and $\mu_2 = 1.6$. In each subfigure, horizontal axis displays our proposed methods and alternatives with three different covariance matrices, and longitudinal axis displays the mean of corresponding measurement values under 100 simulated replicates. The top-left subfigure displays $1-\mathrm{RI}$ of rough clustering structure, the top-right subfigure displays $1-\mathrm{RI}$ of refined clustering structure, the bottom-left subfigure displays MSE of $\hat{\boldsymbol{\beta}}$, and the bottom-right subfigure displays MSE of $\hat{\boldsymbol{\gamma}}$.

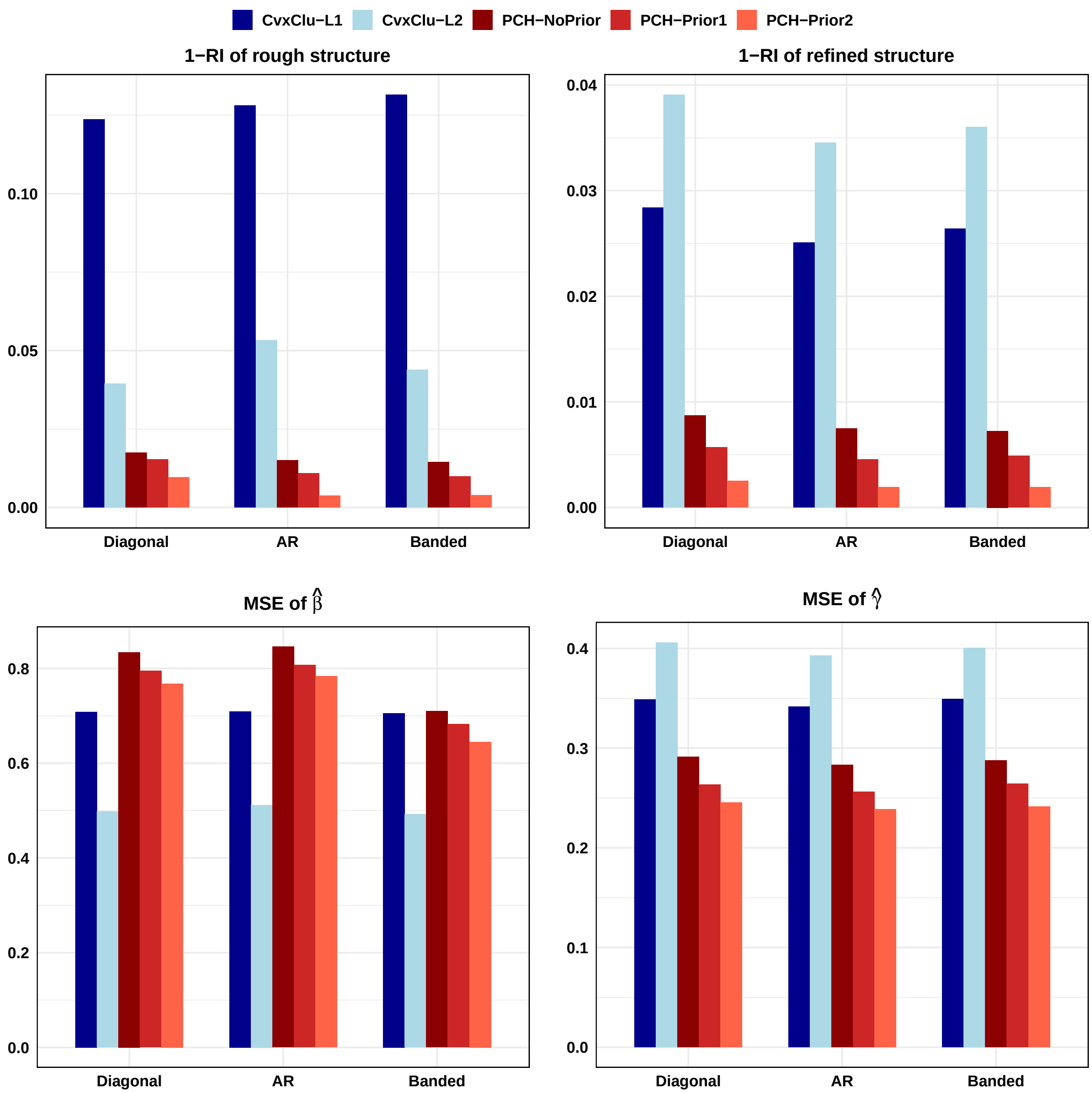

Figure S5: Simulation results with Simulation 3 and $\mu_2 = 1.6$. In each subfigure, horizontal axis displays our proposed methods and alternatives with three different covariance matrices, and longitudinal axis displays the mean of corresponding measurement values under 100 simulated replicates. The top-left subfigure displays $1-\mathrm{RI}$ of rough clustering structure, the top-right subfigure displays $1-\mathrm{RI}$ of refined clustering structure, the bottom-left subfigure displays MSE of $\hat{\boldsymbol{\beta}}$, and the bottom-right subfigure displays MSE of $\hat{\boldsymbol{\gamma}}$.

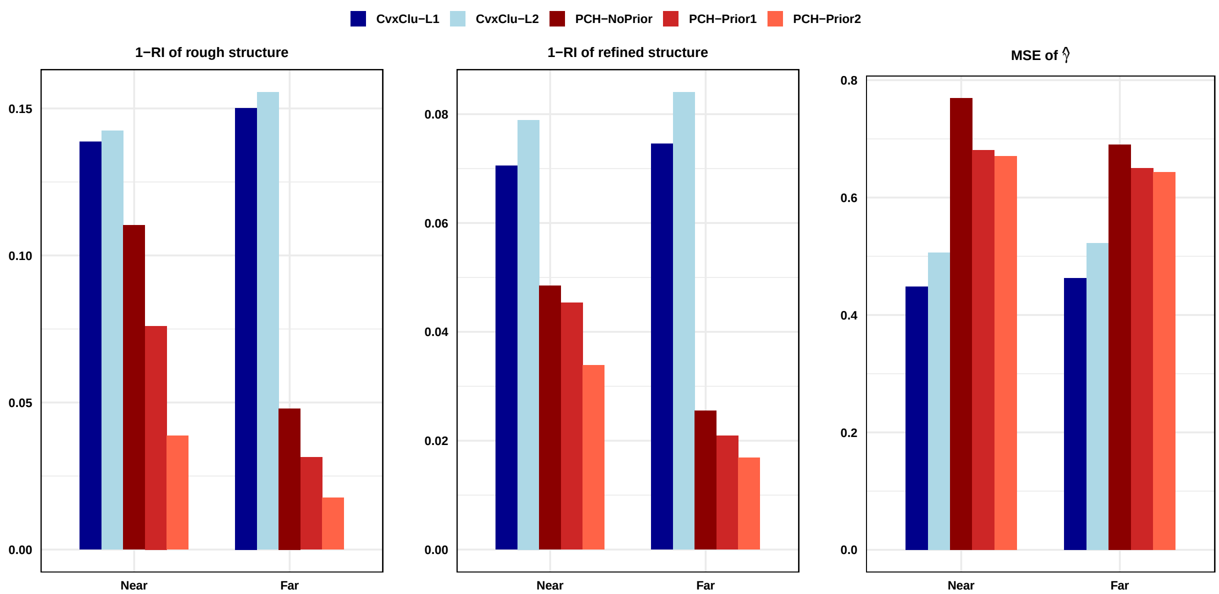

Figure S6: Simulation results with Simulation 5 in two half-moon clusters case. In each subfigure, horizontal axis displays our proposed methods and alternatives with near and far centers, and longitudinal axis displays the mean of corresponding measurement values under 100 simulated replicates. The left subfigure displays $1-\text{RI}$ of rough clustering structure, the middle subfigure displays $1-\text{RI}$ of refined clustering structure, and the right subfigure displays MSE of $\hat{\gamma}$.

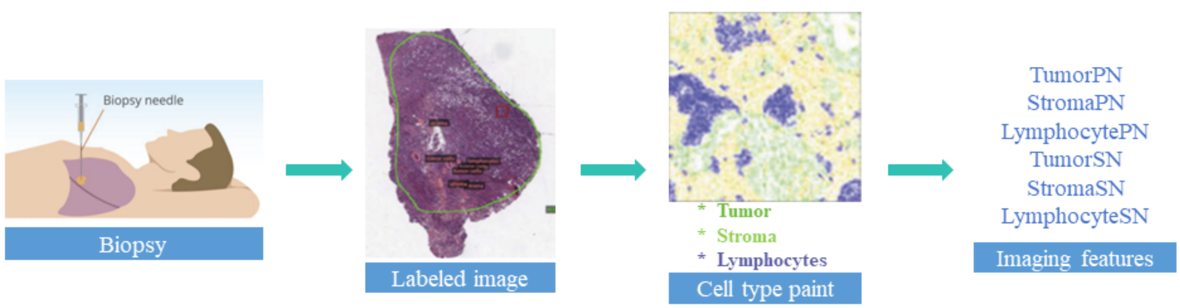

Figure S7: Analysis of LUAD data. Pipelines for extracting six clinical imaging features. LymphocytesPN: Perimeter of lymphocyte cell region/square root of image size. StromaPN: Perimeter of stromal cell region/square root of image size. TumorPN: Perimeter of tumor cell region/square root of image size. LymphocytesSN: Size of lymphocyte cell region/image size. StromaSN: Size of stromal cell region/image size. TumorSN: Size of tumor cell region/image size.

Table S1: Simulation results with $(K_1, K_2) = (2, 4)$ and diagonal covariance structure in Gaussian clusters case. Simulation results include the mean and standard deviation (SD) of RI of rough and refined clustering structure, MSE of $\hat{\boldsymbol{\beta}}$, and MSE of $\hat{\boldsymbol{\gamma}}$ under 100 simulated replicates with $\mu_1 = 1.2$ and $\mu_2 = 1.6$.

		Rough Structure				**Refined Structure**			
		Rand Index		**MSE of $\hat{\boldsymbol{\beta}}$**		**Rand Index**		**MSE of $\hat{\boldsymbol{\gamma}}$**	
	Methods	Mean	SD	Mean	SD	Mean	SD	Mean	SD
$\mu_1 = 1.2$	CvxClu-L_1	0.5053	0.0472	0.7824	0.0445	0.9791	0.0100	0.2511	0.0415
	CvxClu-L_2	0.5238	0.1003	0.7669	0.0901	0.9669	0.0125	0.2988	0.0452
	PCH-NoPrior	0.9423	0.0298	0.2909	0.0696	0.9717	0.0150	0.3430	0.1380
	PCH-Prior1	0.9783	0.0199	0.1954	0.0589	0.9890	0.0094	0.2648	0.1287
	PCH-Prior2	0.9947	0.0097	0.1476	0.0368	0.9957	0.0063	0.2342	0.1298
$\mu_2 = 1.6$	CvxClu-L_1	0.9068	0.1790	0.4544	0.3023	0.9760	0.0103	0.2626	0.0381
	CvxClu-L_2	0.9760	0.0503	0.3281	0.1815	0.9660	0.0120	0.3104	0.0440
	PCH-NoPrior	0.9912	0.0109	0.1782	0.0674	0.9956	0.0055	0.2263	0.0558
	PCH-Prior1	0.9958	0.0079	0.1475	0.0486	0.9977	0.0043	0.1995	0.0365
	PCH-Prior2	0.9981	0.0059	0.1374	0.0414	0.9982	0.0039	0.1910	0.0282

Table S2: Simulation results with $(K_1, K_2) = (2, 4)$ and AR covariance structure in Gaussian clusters case. Simulation results include the mean and standard deviation (SD) of RI of rough and refined clustering structure, MSE of $\hat{\boldsymbol{\beta}}$, and MSE of $\hat{\boldsymbol{\gamma}}$ under 100 simulated replicates with $\mu_1 = 1.2$ and $\mu_2 = 1.6$.

		Rough Structure				**Refined Structure**			
		Rand Index		**MSE of $\hat{\boldsymbol{\beta}}$**		**Rand Index**		**MSE of $\hat{\boldsymbol{\gamma}}$**	
	Methods	Mean	SD	Mean	SD	Mean	SD	Mean	SD
$\mu_1 = 1.2$	CvxClu-L_1	0.5048	0.0455	0.7871	0.0084	0.9581	0.0517	0.3521	0.1702
	CvxClu-L_2	0.5393	0.1245	0.7641	0.0805	0.9590	0.0424	0.3481	0.1305
	PCH-NoPrior	0.9448	0.0278	0.2901	0.0669	0.9721	0.0141	0.3352	0.1101
	PCH-Prior1	0.9778	0.0204	0.2020	0.0567	0.9879	0.0105	0.2595	0.1000
	PCH-Prior2	0.9936	0.0109	0.1536	0.0449	0.9936	0.0087	0.2266	0.0972
$\mu_2 = 1.6$	CvxClu-L_1	0.8840	0.1944	0.4881	0.3221	0.9788	0.0092	0.2585	0.0390
	CvxClu-L_2	0.9578	0.0965	0.3430	0.2068	0.9687	0.0111	0.3000	0.0434
	PCH-NoPrior	0.9881	0.0146	0.1870	0.0791	0.9940	0.0072	0.2379	0.0683
	PCH-Prior1	0.9955	0.0089	0.1426	0.0482	0.9976	0.0046	0.2009	0.0399
	PCH-Prior2	0.9984	0.0055	0.1311	0.0401	0.9980	0.0044	0.1925	0.0308

Table S3: Simulation results with $(K_1, K_2) = (2, 4)$ and banded covariance structure in Gaussian clusters case. Simulation results include the mean and standard deviation (SD) of RI of rough and refined clustering structure, MSE of $\hat{\boldsymbol{\beta}}$, and MSE of $\hat{\boldsymbol{\gamma}}$ under 100 simulated replicates with $\mu_1 = 1.2$ and $\mu_2 = 1.6$.

		Rough Structure				**Refined Structure**			
		Rand Index		**MSE of $\hat{\boldsymbol{\beta}}$**		**Rand Index**		**MSE of $\hat{\boldsymbol{\gamma}}$**	
	Methods	Mean	SD	Mean	SD	Mean	SD	Mean	SD
$\mu_1 = 1.2$	CvxClu-L_1	0.5038	0.0487	0.7841	0.0393	0.9611	0.0507	0.3325	0.1640
	CvxClu-L_2	0.5247	0.1020	0.7700	0.0757	0.9639	0.0348	0.3318	0.1139
	PCH-NoPrior	0.9444	0.0262	0.2871	0.0611	0.9726	0.0134	0.3319	0.1109
	PCH-Prior1	0.9820	0.0159	0.1842	0.0529	0.9901	0.0079	0.2488	0.0922
	PCH-Prior2	0.9930	0.0103	0.1497	0.0457	0.9939	0.0075	0.2281	0.0980
$\mu_2 = 1.6$	CvxClu-L_1	0.9018	0.1834	0.4610	0.3079	0.9800	0.0090	0.2584	0.0381
	CvxClu-L_2	0.9760	0.0503	0.3281	0.1815	0.9706	0.0108	0.2999	0.0417
	PCH-NoPrior	0.9912	0.0109	0.1782	0.0674	0.9957	0.0055	0.2255	0.0554
	PCH-Prior1	0.9958	0.0079	0.1475	0.0486	0.9977	0.0043	0.1995	0.0376
	PCH-Prior2	0.9983	0.0058	0.1369	0.0410	0.9981	0.0039	0.1912	0.0280

Table S4: Simulation results with $(K_1, K_2) = (2, 6)$ and diagonal covariance structure in Gaussian clusters case. Simulation results include the mean and standard deviation (SD) of RI of rough and refined clustering structure, MSE of $\hat{\boldsymbol{\beta}}$, and MSE of $\hat{\boldsymbol{\gamma}}$ under 100 simulated replicates with $\mu_1 = 1.2$ and $\mu_2 = 1.6$.

		Rough Structure				**Refined Structure**			
		Rand Index		**MSE of $\hat{\boldsymbol{\beta}}$**		**Rand Index**		**MSE of $\hat{\boldsymbol{\gamma}}$**	
	Methods	Mean	SD	Mean	SD	Mean	SD	Mean	SD
$\mu_1 = 1.2$	CvxClu-L_1	0.5060	0.0472	0.7830	0.0374	0.9613	0.0270	0.4034	0.1029
	CvxClu-L_2	0.5198	0.0901	0.7735	0.0681	0.9584	0.0217	0.4168	0.0804
	PCH-NoPrior	0.9477	0.0262	0.2825	0.0610	0.9752	0.0194	0.6890	0.5049
	PCH-Prior1	0.9669	0.0229	0.2258	0.0644	0.9812	0.0193	0.6250	0.4537
	PCH-Prior2	0.9852	0.0173	0.1768	0.0578	0.9873	0.0188	0.6176	0.4754
$\mu_2 = 1.6$	CvxClu-L_1	0.9144	0.1682	0.4303	0.2964	0.9745	0.0093	0.3362	0.0476
	CvxClu-L_2	0.9731	0.0509	0.3156	0.1680	0.9645	0.0104	0.3915	0.0517
	PCH-NoPrior	0.9894	0.0131	0.1840	0.0720	0.9877	0.0209	1.0453	0.7595
	PCH-Prior1	0.9927	0.0104	0.1617	0.0579	0.9887	0.0207	1.0528	0.7854
	PCH-Prior2	0.9975	0.0064	0.1358	0.0392	0.9901	0.0206	1.0387	0.7758

Table S5: Simulation results with $(K_1, K_2) = (2, 6)$ and AR covariance structure in Gaussian clusters case. Simulation results include the mean and standard deviation (SD) of RI of rough and refined clustering structure, MSE of $\hat{\boldsymbol{\beta}}$, and MSE of $\hat{\boldsymbol{\gamma}}$ under 100 simulated replicates with $\mu_1 = 1.2$ and $\mu_2 = 1.6$.

		Rough Structure				**Refined Structure**			
		Rand Index		**MSE of $\hat{\boldsymbol{\beta}}$**		**Rand Index**		**MSE of $\hat{\boldsymbol{\gamma}}$**	
	Methods	Mean	SD	Mean	SD	Mean	SD	Mean	SD
$\mu_1 = 1.2$	CvxClu-L_1	0.5446	0.1368	0.7690	0.0761	0.8769	0.0903	0.6413	0.2023
	CvxClu-L_2	0.5587	0.1543	0.7540	0.1057	0.9159	0.0582	0.5548	0.1394
	PCH-NoPrior	0.9484	0.0275	0.2777	0.0628	0.9741	0.0214	0.7220	0.5300
	PCH-Prior1	0.9690	0.0236	0.2161	0.0651	0.9810	0.0202	0.6756	0.5135
	PCH-Prior2	0.9881	0.0145	0.1667	0.0489	0.9872	0.0198	0.6494	0.5024
$\mu_2 = 1.6$	CvxClu-L_1	0.9189	0.1632	0.4237	0.2901	0.9759	0.0116	0.3357	0.0611
	CvxClu-L_2	0.9731	0.0509	0.3156	0.1680	0.9690	0.0100	0.3800	0.0492
	PCH-NoPrior	0.9894	0.0131	0.1840	0.0720	0.9812	0.0239	0.9900	0.7252
	PCH-Prior1	0.9926	0.0110	0.1625	0.0599	0.9823	0.0235	0.9951	0.7426
	PCH-Prior2	0.9975	0.0066	0.1359	0.0392	0.9841	0.0240	0.9804	0.7448

Table S6: Simulation results with $(K_1, K_2) = (2, 6)$ and banded covariance structure in Gaussian clusters case. Simulation results include the mean and standard deviation (SD) of RI of rough and refined clustering structure, MSE of $\hat{\boldsymbol{\beta}}$, and MSE of $\hat{\boldsymbol{\gamma}}$ under 100 simulated replicates with $\mu_1 = 1.2$ and $\mu_2 = 1.6$.

		Rough Structure				**Refined Structure**			
		Rand Index		**MSE of $\hat{\boldsymbol{\beta}}$**		**Rand Index**		**MSE of $\hat{\boldsymbol{\gamma}}$**	
	Methods	Mean	SD	Mean	SD	Mean	SD	Mean	SD
$\mu_1 = 1.2$	CvxClu-L_1	0.5273	0.1098	0.7708	0.0792	0.8971	0.1072	0.5842	0.1845
	CvxClu-L_2	0.5409	0.1315	0.7570	0.1113	0.9284	0.0477	0.5228	0.1305
	PCH-NoPrior	0.9470	0.0324	0.2809	0.0737	0.9712	0.0230	0.7492	0.4854
	PCH-Prior1	0.9644	0.0315	0.2301	0.0725	0.9773	0.0220	0.6989	0.4554
	PCH-Prior2	0.9834	0.0189	0.1822	0.0584	0.9837	0.0209	0.6868	0.4687
$\mu_2 = 1.6$	CvxClu-L_1	0.9192	0.1629	0.4284	0.2930	0.9769	0.0093	0.3303	0.0449
	CvxClu-L_2	0.9731	0.0509	0.3156	0.1680	0.9683	0.0100	0.3817	0.0491
	PCH-NoPrior	0.9894	0.0131	0.1840	0.0720	0.9887	0.0190	0.9871	0.7812
	PCH-Prior1	0.9927	0.0109	0.1614	0.0586	0.9891	0.0192	0.9907	0.8018
	PCH-Prior2	0.9972	0.0069	0.1362	0.0393	0.9911	0.0193	0.9769	0.7951

Table S7: Simulation results with $(K_1, K_2) = (3, 6)$ and diagonal covariance structure in Gaussian clusters case. Simulation results include the mean and standard deviation (SD) of RI of rough and refined clustering structure, MSE of $\hat{\boldsymbol{\beta}}$, and MSE of $\hat{\boldsymbol{\gamma}}$ under 100 simulated replicates with $\mu_1 = 1.2$ and $\mu_2 = 1.6$.

		Rough Structure				**Refined Structure**			
		Rand Index		**MSE of $\hat{\boldsymbol{\beta}}$**		**Rand Index**		**MSE of $\hat{\boldsymbol{\gamma}}$**	
	Methods	Mean	SD	Mean	SD	Mean	SD	Mean	SD
$\mu_1 = 1.2$	CvxClu-L_1	0.3461	0.0594	0.9058	0.0151	0.9540	0.0379	0.4208	0.1228
	CvxClu-L_2	0.3609	0.0958	0.8992	0.0459	0.9528	0.0244	0.4351	0.0898
	PCH-NoPrior	0.9313	0.0325	0.7463	0.5241	0.9657	0.0162	0.3754	0.0572
	PCH-Prior1	0.9580	0.0345	0.6208	0.4629	0.9796	0.0166	0.3218	0.0628
	PCH-Prior2	0.9809	0.0293	0.5722	0.5060	0.9904	0.0142	0.2762	0.0548
$\mu_2 = 1.6$	CvxClu-L_1	0.8763	0.1703	0.7085	0.3164	0.9716	0.0095	0.3487	0.0478
	CvxClu-L_2	0.9605	0.0533	0.4981	0.2194	0.9609	0.0103	0.4061	0.0520
	PCH-NoPrior	0.9826	0.0244	0.8339	0.8184	0.9913	0.0121	0.2912	0.0720
	PCH-Prior1	0.9847	0.0377	0.7949	0.8371	0.9943	0.0100	0.2634	0.0618
	PCH-Prior2	0.9904	0.0378	0.7681	0.8297	0.9975	0.0085	0.2454	0.0561

Table S8: Simulation results with $(K_1, K_2) = (3, 6)$ and banded covariance structure in Gaussian clusters case. Simulation results include the mean and standard deviation (SD) of RI of rough and refined clustering structure, MSE of $\hat{\boldsymbol{\beta}}$, and MSE of $\hat{\boldsymbol{\gamma}}$ under 100 simulated replicates with $\mu_1 = 1.2$ and $\mu_2 = 1.6$.

		Rough Structure				**Refined Structure**			
		Rand Index		**MSE of $\hat{\boldsymbol{\beta}}$**		**Rand Index**		**MSE of $\hat{\boldsymbol{\gamma}}$**	
	Methods	Mean	SD	Mean	SD	Mean	SD	Mean	SD
$\mu_1 = 1.2$	CvxClu-L_1	0.3542	0.0830	0.9035	0.0266	0.9058	0.0727	0.5849	0.1636
	CvxClu-L_2	0.3574	0.0917	0.9002	0.0431	0.9220	0.0515	0.5335	0.1282
	PCH-NoPrior	0.9278	0.0360	0.6875	0.4896	0.9645	0.0164	0.3929	0.1173
	PCH-Prior1	0.9541	0.0380	0.5711	0.4405	0.9784	0.0154	0.3386	0.1106
	PCH-Prior2	0.9788	0.0327	0.5132	0.4692	0.9902	0.0129	0.2915	0.1118
$\mu_2 = 1.6$	CvxClu-L_1	0.8685	0.1907	0.7056	0.3250	0.9736	0.0113	0.3493	0.0530
	CvxClu-L_2	0.9561	0.0596	0.4925	0.2108	0.9640	0.0137	0.4005	0.0602
	PCH-NoPrior	0.9855	0.0124	0.7099	0.7607	0.9927	0.0063	0.2876	0.0537
	PCH-Prior1	0.9901	0.0101	0.6825	0.7767	0.9951	0.0052	0.2645	0.0430
	PCH-Prior2	0.9961	0.0066	0.6453	0.7944	0.9981	0.0031	0.2416	0.0314

Table S9: Simulation results with $(K_1, K_2) = (2, 6)$ in two half-moon clusters case. Simulation results include the mean and standard deviation (SD) of RI of rough and refined clustering structure, and MSE of $\hat{\gamma}$ under 100 simulated replicates with near and far centers.

		Rough Structure		**Refined Structure**			
		Rand Index		**Rand Index**		**MSE of $\hat{\gamma}$**	
	Methods	Mean	SD	Mean	SD	Mean	SD
Near centers	CvxClu-L_1	0.8613	0.0811	0.9295	0.0289	0.4482	0.0793
	CvxClu-L_2	0.8576	0.0825	0.9211	0.0285	0.5066	0.0832
	PCH-NoPrior	0.8898	0.0720	0.9515	0.0319	0.7692	0.3951
	PCH-Prior1	0.9240	0.0998	0.9546	0.0506	0.6807	0.3384
	PCH-Prior2	0.9614	0.0921	0.9661	0.0547	0.6704	0.3502
Far centers	CvxClu-L_1	0.8499	0.0799	0.9254	0.0288	0.4626	0.0771
	CvxClu-L_2	0.8445	0.0802	0.9160	0.0286	0.5228	0.0834
	PCH-NoPrior	0.9520	0.0460	0.9745	0.0257	0.6901	0.4390
	PCH-Prior1	0.9686	0.0320	0.9791	0.0272	0.6501	0.4175
	PCH-Prior2	0.9824	0.0232	0.9832	0.0289	0.6432	0.4158

Table S10: Analysis of LUAD data. The clustering results by the proposed method. There are identified two rough clusters with sizes 203 and 152, respectively. There are four identified refined clusters nested in rough clusters, with sizes 158, 45, 118, and 34, respectively. The estimated centers of distinct rough and refined clusters are displayed with respect to imaging features and principal components.

Imaging Features	**Centers of Rough Clusters** 1 (203)	2 (152)
LymphocytesPN	-0.1120	0.1496
StromaPN	0.6094	-0.8139
TumorPN	0.3466	-0.4629
LymphocytesSN	-0.0679	0.0907
StromaSN	0.5772	-0.7709
TumorSN	-0.5648	0.7542

Principal Components	**Centers of Refined Clusters** 1 (158)	2 (45)	3 (118)	4 (34)
PC1	0.0750	0.0059	-0.0340	-0.2383
PC2	-0.0581	0.1493	-0.0107	0.1095
PC3	0.0153	-0.2853	-0.0674	0.5402
PC4	-0.2604	1.2218	-0.0022	-0.3990
PC5	-0.0097	-0.0149	0.0300	-0.0394
PC6	-0.1900	0.3570	0.1758	-0.1998
PC7	-0.0561	-0.0829	-0.0209	0.4432
PC8	0.0456	-0.2065	0.0262	-0.0298
PC9	0.1149	-0.1638	0.0430	-0.4660
PC10	0.0185	-0.4697	0.0596	0.3290
PC11	-0.1984	0.6676	-0.0349	0.1595
PC12	0.0091	0.0778	0.1021	-0.4996
PC13	-0.3119	0.6533	-0.0222	0.6618
PC14	0.0318	0.0174	-0.0956	0.1610
PC15	0.2588	0.1175	-0.2509	-0.4873
PC16	-0.0459	0.2430	-0.1693	0.4793
PC17	-0.2216	0.0919	0.2842	-0.0785
PC18	0.0018	0.3206	-0.2260	0.3518
PC19	0.0905	-0.1138	-0.1858	0.3748
PC20	-0.1526	0.1862	0.3404	-0.7188