Supplement to "Individualized Multi-directional Variable Selection"

This supplementary note collects auxiliary lemmas, detailed proofs for the theorems and corollaries in Section 3 (Appendix A), and additional numerical analyses (Appendix B).

A Supplementary Materials to Section 3: Theory

A.1 Some Notation and Matrix Algebra

(N1). Denote $a \wedge b = \min(a, b)$.

(N2). Define " $\circ$ " as the Hadamard product, that is, for two matrices, A and B, of the same dimension $m \times n$, then $A \circ B$ is a matrix, of the same dimension of A and B, with elements given by $(A \circ B)_{ij} = A_{ij} \cdot B_{ij}$.

(N3). Define the order between two $n \times n$ square matrices as $\boldsymbol{A} > \boldsymbol{B}$ if $\forall \boldsymbol{x} \in \mathbf{R}^n$, $\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} > \boldsymbol{x}^T\boldsymbol{B}\boldsymbol{x}$ holds. Let $\boldsymbol{A} \asymp \boldsymbol{B}$ denote $c_1\boldsymbol{A} \leq \boldsymbol{B} \leq c_2\boldsymbol{A}$ for some constants $0 < c_1 \leq c_2 < \infty$. Define a sequence of $m \times m$ matrices $\boldsymbol{A_n}$ as $\boldsymbol{A_n} = O(n)$ if $c_1 n\boldsymbol{I}_m \leq \boldsymbol{A}_n \leq c_2 n\boldsymbol{I}_m$ when n is large.

Next, we provide some useful results as well as the proofs for some matrix algebra. For two square matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ with the same dimension,

(M1). $\boldsymbol{AB}$ and $\boldsymbol{BA}$ have the same non-zero eigenvalues.

Proof: For any eigenvalue λ of $\boldsymbol{AB}$, there exists a non-zero vector $\boldsymbol{\mu}$ such that $\boldsymbol{AB\mu} = \lambda\boldsymbol{\mu}$. It implies that $\boldsymbol{BAB\mu} = \lambda\boldsymbol{B\mu}$. Let $\boldsymbol{B\mu} = \boldsymbol{\mu}^*$ and we have $\boldsymbol{BA\mu}^* = \lambda\boldsymbol{\mu}^*$ indicating that λ is also an eigenvalue of $\boldsymbol{BA}$.

(M2). If $\boldsymbol{A}$ and $\boldsymbol{B}$ are non-singular and $\boldsymbol{A} \leq \boldsymbol{B}$, for any matrix $\boldsymbol{C}$, we have $\boldsymbol{C}^T\boldsymbol{AC} \leq \boldsymbol{C}^T\boldsymbol{BC}$, and $\boldsymbol{A}^{-1} \geq \boldsymbol{B}^{-1}$.

Proof: Note that $\boldsymbol{A} \leq \boldsymbol{B}$ is equivalent to $\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{x}^T\boldsymbol{B}\boldsymbol{x}$ for any vector $\boldsymbol{x}$. $\forall$ $\boldsymbol{x}$, denote $\boldsymbol{Cx} = \boldsymbol{x}^*$ such that $\boldsymbol{x}^T\boldsymbol{C}^T\boldsymbol{ACx} = (\boldsymbol{x}^*)^T\boldsymbol{A}\boldsymbol{x}^* \leq (\boldsymbol{x}^*)^T\boldsymbol{B}\boldsymbol{x}^* = \boldsymbol{x}^T\boldsymbol{C}^T\boldsymbol{BCx}$, implies $\boldsymbol{C}^T\boldsymbol{AC} \leq \boldsymbol{C}^T\boldsymbol{BC}$.

It is trivial that if $\boldsymbol{A} \geq \boldsymbol{I}$, then we have $\boldsymbol{A}^{-1} \leq \boldsymbol{I}$. Hence, $\boldsymbol{A} \leq \boldsymbol{B} \Rightarrow \boldsymbol{B}^{-\frac{1}{2}}\boldsymbol{AB}^{-\frac{1}{2}} \leq \boldsymbol{I} \Rightarrow \boldsymbol{B}^{\frac{1}{2}}\boldsymbol{A}^{-1}\boldsymbol{B}^{\frac{1}{2}} \geq \boldsymbol{I} \Rightarrow \boldsymbol{A}^{-1} \geq \boldsymbol{B}^{-1}$.

A.2 Regularity Conditions

We require some common regularity conditions for establishing theoretical results in Section 3.

(A1) The unknown parameter $(\boldsymbol{\gamma}', \boldsymbol{\alpha}')'$ belongs to a compact subset $\mathcal{B} \subseteq \mathbf{R}^{p+q}$ and its true value lies in the interior of $\mathcal{B}$;

(A2) $\boldsymbol{D}_{N,m}$ and $\boldsymbol{H}_{N,m}$ are positive definite when N or m is large.

(A3) There exist $\nu_l > 0, \nu_l' > 0$, such that $\lambda_{min}(\boldsymbol{R}_i^0) > \nu_l$ and $\lambda_{min}(\boldsymbol{R}_i) > \nu_l'$ for all i and m; and $tr(\boldsymbol{R}_i^{-1}) = O(m)$.

(A4) $\tilde{\boldsymbol{X}}_{ij} = (\boldsymbol{X}_{ij}', \boldsymbol{Z}_{ij}')'_{(p+q)\times 1}$ belongs to a compact set $\mathcal{X} \subset \mathbf{R}^{p+q}$ for $1 \le i \le N$ and $1 \le j \le m$;

(A5) Let $\tilde{\boldsymbol{X}}_{i,k}$ denote the kth column of $\tilde{\boldsymbol{X}}_i$, assume $\|\tilde{\boldsymbol{X}}_{i\cdot,k}\|_2^2 = O_p(m)$ and $\sum_{i=1}^N m^{-1}\|\tilde{\boldsymbol{X}}_{i\cdot,k}\|_2^2 = O_p(N)$, for $1 \le k \le p+q$;

(A6) $m^{-1}\lambda_{min}(\boldsymbol{X}_i^T\boldsymbol{X}_i) > c_3$ for any i and $\frac{1}{Nm}\lambda_{min}\left(\sum_{i=1}^N \boldsymbol{Z}_i^T(\boldsymbol{I}_m - \boldsymbol{H}_{\boldsymbol{X}_i})\boldsymbol{Z}_i\right) > c_4$, where $\boldsymbol{H}_{\boldsymbol{X}_i} = \boldsymbol{X}_i(\boldsymbol{X}_i^T\boldsymbol{X}_i)^{-1}\boldsymbol{X}_i^T$, for some constants $0 < c_3 < \infty$, $0 < c_4 < \infty$.

Conditions (A4)-(A6) are regularity conditions which are typically required for the bounded regressors. However, these are less restrictive than other assumptions, e.g., $\frac{1}{m}\boldsymbol{X}_i^T\boldsymbol{X}_i$ converges to a positive constant matrix. Note that condition (A6) allows within-individual invariant covariates, and is less restrictive since it does not require $\tilde{\boldsymbol{X}}_i^T\tilde{\boldsymbol{X}}_i$ to be positive definite.

A.3 Proof of Lemma 1

For an estimator $\hat{\boldsymbol{\theta}}$ obtained by solving the estimating equation $G_{N,m}(\boldsymbol{\theta}) = 0$ in (7), under regularity condition (A2), by Taylor's expansion, we have $(\hat{\boldsymbol{\theta}}^u - \boldsymbol{\theta}^0) = -\boldsymbol{D}_{N,m}^{-1}\boldsymbol{G}_{N,m}$ and thus $\boldsymbol{H}_{N,m}^{-\frac{1}{2}}\boldsymbol{D}_{N,m}(\hat{\boldsymbol{\theta}}^u - \boldsymbol{\theta}^0) = -\boldsymbol{H}_{N,m}^{-\frac{1}{2}}\boldsymbol{G}_{N,m}$, where $\boldsymbol{G}_{N,m} = \boldsymbol{G}_{N,m}(\boldsymbol{\theta}^0)$.

By the Chebyshev inequality,

$$\begin{aligned}
P\left(p_{\boldsymbol{\theta}}^{-\frac{1}{2}}\|\boldsymbol{H}_{N,m}^{-\frac{1}{2}}\boldsymbol{D}_{N,m}(\hat{\boldsymbol{\theta}}^u - \boldsymbol{\theta}^0)\|_2 > \delta\right) &= P\left(p_{\boldsymbol{\theta}}^{-\frac{1}{2}}\|\boldsymbol{H}_{N,m}^{-\frac{1}{2}}\boldsymbol{G}_{N,m}\|_2 > \delta\right) \\
&\le p_{\boldsymbol{\theta}}^{-1}\delta^{-2}E(\|\boldsymbol{H}_{N,m}^{-\frac{1}{2}}\boldsymbol{G}_{N,m}\|_2^2) \\
&= p_{\boldsymbol{\theta}}^{-1}\delta^{-2}E(tr(\boldsymbol{H}_{N,m}^{-\frac{1}{2}}\boldsymbol{G}_{N,m}\boldsymbol{G}_{N,m}^T\boldsymbol{H}_{N,m}^{-\frac{1}{2}})) \\
&= p_{\boldsymbol{\theta}}^{-1}\delta^{-2}tr(\boldsymbol{H}_{N,m}^{-\frac{1}{2}}E(\boldsymbol{G}_{N,m}\boldsymbol{G}_{N,m}^T)\boldsymbol{H}_{N,m}^{-\frac{1}{2}}) \\
&= p_{\boldsymbol{\theta}}^{-1}\delta^{-2}tr(\boldsymbol{H}_{N,m}^{-\frac{1}{2}}\boldsymbol{H}_{N,m}\boldsymbol{H}_{N,m}^{-\frac{1}{2}}) = \delta^{-2}.
\end{aligned}$$

Furthermore, noting that $\|\boldsymbol{H}_{N,m}^{-\frac{1}{2}}\boldsymbol{D}_{N,m}(\hat{\boldsymbol{\theta}}^u - \boldsymbol{\theta}^0)\|_2 \ge \lambda_{min}(\boldsymbol{D}_{N,m}\boldsymbol{H}_{N,m}^{-1}\boldsymbol{D}_{N,m})^{\frac{1}{2}}\|(\hat{\boldsymbol{\theta}}^u - \boldsymbol{\theta}^0)\|_2$ and thus

$$\begin{aligned}
P\left(p_{\boldsymbol{\theta}}^{-\frac{1}{2}}\|(\hat{\boldsymbol{\theta}}^u - \boldsymbol{\theta}^0)\|_2 > \delta\right) &= P\left(p_{\boldsymbol{\theta}}^{-\frac{1}{2}}\lambda_{min}(\boldsymbol{D}_{N,m}\boldsymbol{H}_{N,m}^{-1}\boldsymbol{D}_{N,m})^{\frac{1}{2}}\|(\hat{\boldsymbol{\theta}}^u - \boldsymbol{\theta}^0)\|_2 > \lambda_{min}(\boldsymbol{D}_{N,m}\boldsymbol{H}_{N,m}^{-1}\boldsymbol{D}_{N,m})^{\frac{1}{2}}\delta\right) \\
&\le P\left(p_{\boldsymbol{\theta}}^{-\frac{1}{2}}\|\boldsymbol{H}_{N,m}^{-\frac{1}{2}}\boldsymbol{D}_{N,m}^{\frac{1}{2}}(\hat{\boldsymbol{\theta}}^u - \boldsymbol{\theta}^0)\|_2 > \lambda_{min}(\boldsymbol{D}_{N,m}\boldsymbol{H}_{N,m}^{-1}\boldsymbol{D}_{N,m})^{\frac{1}{2}}\delta\right) \\
&\le \lambda_{max}(\boldsymbol{D}_{N,m}\boldsymbol{H}_{N,m}^{-1}\boldsymbol{D}_{N,m})^{-1}\sigma^{-2}.
\end{aligned}$$

As $\lambda_{max}(\boldsymbol{D}_{N,m}\boldsymbol{H}_{N,m}^{-1}\boldsymbol{D}_{N,m}) \to \infty$, we have $P\left(p_{\boldsymbol{\theta}}^{-\frac{1}{2}}\|(\hat{\boldsymbol{\theta}}^u - \boldsymbol{\theta}^0)\|_2 > \delta\right) \to 0$. □

A.4 Proof of Theorem 1

Let $\tilde{\boldsymbol{X}}_i = (\boldsymbol{X}_i, \boldsymbol{Z}_i)$ and $\tilde{\boldsymbol{\omega}}_i = (\boldsymbol{\omega}_i', \mathbf{1}_q')'$, and $\tilde{\boldsymbol{X}}_i^{or} = \tilde{\boldsymbol{X}}_i\tilde{\boldsymbol{\Omega}}_i$ where $\tilde{\boldsymbol{\Omega}}_i = \text{diag}(\tilde{\boldsymbol{\omega}}_i)$.
We denote $\boldsymbol{H}_{N,m}^{or} = \sum_{i=1}^N (\tilde{\boldsymbol{X}}_i^{or})^T \boldsymbol{V}_i^{-1}\boldsymbol{\Sigma}_i\boldsymbol{V}_i^{-1}\tilde{\boldsymbol{X}}_i^{or}$, $\boldsymbol{D}_{N,m}^{or} = \sum_{i=1}^N (\tilde{\boldsymbol{X}}_i^{or})^T \boldsymbol{V}_i^{-1}\tilde{\boldsymbol{X}}_i^{or}$, and Lemma 1 directly applies for the oracle estimator by replacing $\boldsymbol{H}_{N,m}$ and $\boldsymbol{D}_{N,m}$ with $\boldsymbol{H}_{N,m}^{or}$ and $\boldsymbol{D}_{N,m}^{or}$, respectively. Let $\hat{\boldsymbol{\theta}}^{or} = \text{vec}(\hat{\boldsymbol{\gamma}}^{or}, \hat{\boldsymbol{\alpha}}^{or})$ and $\tilde{\boldsymbol{\theta}}^0 = \text{vec}(\boldsymbol{\gamma}^0, \boldsymbol{\alpha}^0)$, according to Lemma 1 we have

$$(\boldsymbol{H}_{N,m}^{or})^{-\frac{1}{2}}(\boldsymbol{D}_{N,m}^{or})(\hat{\boldsymbol{\theta}}^{or} - \tilde{\boldsymbol{\theta}}^0) = O_p(1). \tag{A.1}$$

Note that the divergence rates of $\boldsymbol{H}_{N,m}^{or}$ and $\boldsymbol{D}_{N,m}^{or}$ are associated with the subpopulation size $|\mathcal{G}_k|$'s as N goes to infinity. However, in contrast to other clustering approaches based on an entire set of coefficient vector $\boldsymbol{\beta}_i$ (e.g., [9, 7]), the proposed model allows the subgroup partitions corresponding to different individualized predictors to be different. Therefore the design matrix for the oracle estimator here cannot be formulated as a block diagonal form, which leads to non-trivial subgroup effects on divergence rates.

To get a better understanding of the group effects on the oracle estimator, we reformulate $\boldsymbol{D}_{N,m}^{or} = \sum_{i=1}^N \tilde{\boldsymbol{\Omega}}_i^T\tilde{\boldsymbol{X}}_i^T\boldsymbol{V}_i^{-1}\tilde{\boldsymbol{X}}_i\tilde{\boldsymbol{\Omega}}_i = \sum_{i=1}^N(\tilde{\boldsymbol{\Omega}}_i\tilde{\boldsymbol{\Omega}}_i^T) \circ (\boldsymbol{X}_i^T\boldsymbol{V}_i^{-1}\boldsymbol{X}_i)$, where $\tilde{\boldsymbol{\Omega}}_i^T\tilde{\boldsymbol{\Omega}}_i^T$ is a symmetric square matrix with entries to be zero or one. Suppose there are some positive constant sequences $\{\kappa_m^l\}_{m=1}^\infty$ and $\{\kappa_m^u\}_{m=1}^\infty$ such that

$$\kappa_m^l \le \lambda_{min}(\tilde{\boldsymbol{X}}_i^T\boldsymbol{V}_i^{-1}\tilde{\boldsymbol{X}}_i) \le \lambda_{max}(\tilde{\boldsymbol{X}}_i^T\boldsymbol{V}_i^{-1}\tilde{\boldsymbol{X}}_i) \le \kappa_m^u, \quad 1 \le i \le N,$$

then we have $\kappa_m^l\sum_{i=1}^N\tilde{\boldsymbol{\Omega}}_i \le \boldsymbol{D}_{N,m}^{or} \le \kappa_m^u\sum_{i=1}^N\tilde{\boldsymbol{\Omega}}_i$ by noting $\tilde{\boldsymbol{\Omega}}_i^2 = \tilde{\boldsymbol{\Omega}}_i$. Similarly, we could also show that $\phi_m^l\sum_{i=1}^N\tilde{\boldsymbol{\Omega}}_i \le \boldsymbol{H}_{N,m}^{or} \le \phi_m^u\sum_{i=1}^N\tilde{\boldsymbol{\Omega}}_i$ for some positive constant sequences $\{\phi_m^l\}_{m=1}^\infty$ and $\{\phi_m^u\}_{m=1}^\infty$. Let $\boldsymbol{\Lambda}_N = \sum_{i=1}^N\tilde{\boldsymbol{\Omega}}_i$ and note that $\boldsymbol{\Lambda}_N = \text{diag}(N\mathbf{1}_q', |\mathcal{G}_1|, \ldots, |\mathcal{G}_p|)$ is a diagonal matrix, where $|\mathcal{G}_k|$'s ($1 \le k \le p$) are signal-subgroup sizes corresponding to p individualized predictors, respectively. Since $\sum_{i=1}^N\tilde{\boldsymbol{\Omega}}_i$ is non-singular, then

$$(\phi_m^u)^{-1}(\kappa_m^l)^2\boldsymbol{\Lambda}_N \le \boldsymbol{D}_{N,m}^{or}(\boldsymbol{H}_{N,m}^{or})^{-1}\boldsymbol{D}_{N,m}^{or} \le (\phi_m^l)^{-1}(\kappa_m^u)^2\boldsymbol{\Lambda}_N. \tag{A.2}$$

The bounds in (A.2) provide the convergence rate for the oracle estimator. It is clear that $\boldsymbol{\Lambda}_N$ contains the subgroup effects on estimation, while $(\phi_m^u)^{-1}(\kappa_m^l)^2$ and $(\phi_m^l)^{-1}(\kappa_m^u)^2$ reflect the information accumulated from the increasing individual-wise measurements. For example, in the independent-error model, it is straightforward that κ_m^l, κ_m^u, ϕ_m^l and ϕ_m^u are all of order $O_p(m)$ under the regularity conditions.

Let $N_k = \sum_{i\in\mathcal{G}_k} m_i = m|\mathcal{G}_k|$ denote the number of observations in group $\mathcal{G}_k$ and $N_a = \sum_{i=1}^N m_i = mN$ denote the total number of observations. For the independent-error model, we establish asymptotic normality for the oracle estimators with convergence rates associated to the sample size N and the individual measurement size m.

Following the matrix algebra in Section A.1, we have

$$(\phi_m^u)^{-1}(\sum_{i=1}^N\tilde{\boldsymbol{\Omega}}_i)^{-1} \le (\boldsymbol{H}_{N,m}^{or})^{-1} \le (\phi_m^l)^{-1}(\sum_{i=1}^N\tilde{\boldsymbol{\Omega}}_i)^{-1},$$

and therefore (A.2) holds.

Recall that $\hat{\boldsymbol{\theta}}^{or} = \text{vec}((\hat{\boldsymbol{\gamma}}^{or}, \hat{\boldsymbol{\alpha}}^{or})$, by Taylor's expansion, we note that $(\tilde{\boldsymbol{\theta}}^{or} - \tilde{\boldsymbol{\theta}}^0) = -(\boldsymbol{D}_{N,m}^{or})^{-1}\boldsymbol{G}_{N,m}^{or} = -(\boldsymbol{H}_{N,m}^{or})^{-1}\boldsymbol{G}_{N,m}^{or}$, where

$$\boldsymbol{G}_{N,m}^{or} = \sum_{i=1}^N \tilde{\boldsymbol{X}}_i^T\boldsymbol{V}_i^{-1}(\boldsymbol{y}_i - \tilde{\boldsymbol{X}}_i\tilde{\boldsymbol{\theta}}^0),$$

since $\boldsymbol{R}_i^0 = \boldsymbol{R}_i = \boldsymbol{I}_m$ holds for any i. By the standard central limit theorem, we have $(\boldsymbol{H}_{N,m}^{or})^{-1/2}\boldsymbol{G}_{N,m}^{or} \to N(\mathbf{0}, \boldsymbol{I}_{p+q})$, implying that $(\boldsymbol{H}_{N,m}^{or})^{1/2}(\hat{\boldsymbol{\theta}}^{or} - \tilde{\boldsymbol{\theta}}^0) \to N(\mathbf{0}, \boldsymbol{I}_{p+q})$, as either $m \to \infty$ or $\min_{1\le k\le p}(|\mathcal{G}_k|) \to \infty$. In addition, if $\boldsymbol{R}_i^0 \ne \boldsymbol{I}_m$ but m is bounded, then the asymptotic normality still holds when N goes to infinity regardless of the choice of working correlation matrix $\boldsymbol{R}_i$.

Moreover, under regularity conditions (A5)-(A6), we have $\lambda_{min}(\sum_i^N \tilde{\boldsymbol{X}}_i^T\tilde{\boldsymbol{X}}_i) = O(mN)$ and $\lambda_{max}(\sum_i^N \tilde{\boldsymbol{X}}_i^T\tilde{\boldsymbol{X}}_i) = O(mN)$. When $\boldsymbol{R}_i^0 = \boldsymbol{R}_i = \boldsymbol{I}_m$, it is trivial that $\boldsymbol{H}_{N,m}^{or} = \boldsymbol{D}_{N,m}^{or} \asymp m\boldsymbol{\Lambda}_{\boldsymbol{N},\boldsymbol{m}} = \boldsymbol{M}_{N,m}$. □

A.5 Proof of Theorem 2

Following Lemma 1, we have

$$P\left((p+q)^{-\frac{1}{2}}\|(\boldsymbol{H}_{N,m}^{or})^{-\frac{1}{2}}\boldsymbol{D}_{N,m}^{or}(\hat{\boldsymbol{\theta}}^{or}-\tilde{\boldsymbol{\theta}}^0)\|_2 > \delta\right) < \tfrac{1}{\delta^2}.$$

Note that

$$\begin{aligned}
\boldsymbol{H}_{N,m} &= \sum_{i=1}^N \boldsymbol{U}_i^T\boldsymbol{V}_i^{-1}\boldsymbol{\Sigma}_i\boldsymbol{V}_i^{-1}\boldsymbol{U}_i \\
&= \sum_{i=1}^N \boldsymbol{U}_i^T\boldsymbol{V}_i^{-1/2}\boldsymbol{V}_i^{-1/2}\boldsymbol{\Sigma}_i\boldsymbol{V}_i^{-1/2}\boldsymbol{V}_i^{-1/2}\boldsymbol{U}_i \\
&\le \lambda_{max}(\boldsymbol{R}_i^{-1/2}\boldsymbol{R}_i^0\boldsymbol{R}_i^{-1/2})\sum_{i=1}^N \boldsymbol{U}_i^T\boldsymbol{V}_i^{-1/2}\boldsymbol{V}_i^{-1/2}\boldsymbol{U}_i \\
&= \lambda_{max}(\boldsymbol{R}_i^{-1}\boldsymbol{R}_i^0)\boldsymbol{D}_{N,m} = \eta_m\boldsymbol{D}_{N,m}.
\end{aligned}$$

Therefore we have $\boldsymbol{D}_{N,m}\boldsymbol{H}_{N,m}^{-1}\boldsymbol{D}_{N,m} \ge \eta_m^{-1}\boldsymbol{D}_{N,m}$, which implies that

$$\|(\boldsymbol{H}_{N,m}^{or})^{-\frac{1}{2}}\boldsymbol{D}_{N,m}^{or}(\hat{\boldsymbol{\theta}}^{or}-\tilde{\boldsymbol{\theta}}^0)\|_2 \ge \eta_m^{-\frac{1}{2}}\|(\boldsymbol{D}_{N,m}^{or})^{\frac{1}{2}}(\hat{\boldsymbol{\theta}}^{or}-\tilde{\boldsymbol{\theta}}^0)\|_2,$$

and thus

$$P\left(\eta_m^{-\frac{1}{2}}\|(\boldsymbol{D}_{N,m}^{or})^{\frac{1}{2}}(\hat{\boldsymbol{\theta}}^{or}-\tilde{\boldsymbol{\theta}}^0)\|_2 > \delta\right) < c_0\frac{1}{\delta^2}$$

for some $c_0 > 0$. The proof of Theorem 2 is completed. □

A.6 A Few Remarks and Conclusions on Divergent Correlation Structure

Remark A.1. For any N and m, according to regularity condition (A3), note that

$$\eta_m \le (\min_{1\le i\le N}\{\lambda_{min}(\boldsymbol{R}_i)\})^{-1}\max_{1\le i\le N}\{\lambda_{max}(\boldsymbol{R}_i^0)\} \le (\nu_l')^{-1}\text{tr}(\boldsymbol{R}_1^0) \le (\nu_l')^{-1}m.$$

If m is bounded, then η_m is bounded, which implies that the condition $\mathcal{C}_a^*$ does not depend on unknown true correlation structure $\boldsymbol{R}_i^0$. As $N\to\infty$, we have $\lambda_{min}(\boldsymbol{D}_{N,m}^{or})\to\infty$ regardless of the choice of working correlation $\boldsymbol{R}_i$. Hence, similar to standard results for the GEE estimator, the oracle estimator $\hat{\boldsymbol{\theta}}^{or}$ has asymptotic normality, although it may not achieve optimal efficiency if $\boldsymbol{R}_i \ne \boldsymbol{R}_i^0$.

Remark A.2. If $m\to\infty$, η_m is not always bounded. For example, if $\boldsymbol{R}_i^0$ admits an Exchangeable correlation structure and we choose working correlation $\boldsymbol{R}_i$ as an identity matrix, we have $\eta_m = O(m)$. For any bounded N, $\boldsymbol{D}_{N,m}^{or} = O(m)$, which implies that the condition ($\mathcal{C}_a^*$) fails. Although the condition ($\mathcal{C}_a$) may still hold with some constraints on the design matrix to ensure consistency (see following Example A.1),

We use the following example of a simple linear regression to illustrate some details about the conditions $\mathcal{C}_a$ and $\mathcal{C}_a^*$ with specific covariates design.

Example A.1. Consider an individual-wise model with homogeneous effect,

$$y_{it} = x_{it}\beta + \varepsilon_{it}, \quad i = 1, \ldots, N; t = 1, \ldots, m,$$

where $\boldsymbol{\varepsilon_i} = (\varepsilon_{i1}, \ldots, \varepsilon_{im})' \sim N(\mathbf{0}, \sigma^2 \boldsymbol{R^0})$ and $\boldsymbol{R^0}$ admits an exchangeable structure with parameter $\rho > 0$, x_{ij}'s are iid $N(\mu, 1)$. For the case of bounded N, without loss of generality, we assume $N = 1$. By using an independent working correlation $\boldsymbol{R}_i = \boldsymbol{I}_m$, we have $\boldsymbol{D}_m = \boldsymbol{x}_1^T \boldsymbol{x}_1 = O(m)$ and $\eta_m = \lambda_{max}(\boldsymbol{R}^0) = m\rho + 1 - \rho$, where $\boldsymbol{x}_1 = (x_{11}, \ldots, x_{1m})'$. Thus condition $\mathcal{C}_a^*$ fails. However, note that $\boldsymbol{R}^0(\rho) = (1-\rho)\boldsymbol{I}_m + \rho \mathbf{1}_m \mathbf{1}_m^T$. We have $\boldsymbol{H}_m = \sigma^2 \boldsymbol{x}_1^T \boldsymbol{R}^0 \boldsymbol{x}_1 = \sigma^2 \boldsymbol{x}_1^T((1-\rho)\boldsymbol{I}_m + \rho \mathbf{1}_m \mathbf{1}_m^T)\boldsymbol{x}_1^T = \sigma^2(1-\rho)\boldsymbol{x}_1^T \boldsymbol{x}_1 + m\rho(m^{-\frac{1}{2}} \sum_{i=1}^m x_{1t})^2 = O(m) + O(m)$ if $\mu = 0$, and thus $\lambda_{min}(\boldsymbol{D}_m \boldsymbol{H}_m^{-1} \boldsymbol{D}_m) = O(m) \to \infty$ as $m \to \infty$. But if $\mu > 0$, it is clear that $m\rho(m^{-\frac{1}{2}} \sum_{i=1}^m x_{1t})^2 = O(m^2)$ and thus $\lambda_{min}(\boldsymbol{D}_m \boldsymbol{H}_m^{-1} \boldsymbol{D}_m) = O(1)$.

Under mild conditions on correlation structures, we have a simplified result for the oracle estimator with subgrouping effects with correlated data as follows.

Corollary A.1. *Suppose $\eta_m \leq C_1$ holds uniformly for some constant $0 < C_1 < \infty$, under regularity conditions, we have*

$$\|\boldsymbol{M}_{N,m}^{\frac{1}{2}}(\hat{\boldsymbol{\theta}}^{or} - \tilde{\boldsymbol{\theta}}^0)\|_2 \leq O_p(1),$$

where $\boldsymbol{M}_{N,m}$ is defined in Theorem 1.

Proof: Following the proof of Theorem 2, if $\eta_m \leq C_1$ holds uniformly for some positive constant C_1, it is straightforward that $\|(\boldsymbol{D}_{N,m}^{or})^{\frac{1}{2}}(\hat{\boldsymbol{\theta}}^{or} - \tilde{\boldsymbol{\theta}}^0)\|_2 = O_p(1)$. Note that $\boldsymbol{D}_{N,m}^{or} = \sum_i^N \tilde{\boldsymbol{\Omega}}_i \tilde{\boldsymbol{X}}_i^T \boldsymbol{R}_i^{-1} \tilde{\boldsymbol{X}}_i \tilde{\boldsymbol{\Omega}}_i$ and $\tilde{\boldsymbol{X}}_i^T \boldsymbol{R}_i^{-1} \tilde{\boldsymbol{X}}_i \asymp \tilde{\boldsymbol{X}}_i^T \tilde{\boldsymbol{X}}_i$ according to the regularity conditions. Following the similar argument in Section A.4, we have $\boldsymbol{D}_{N,m}^{or} \asymp \boldsymbol{M}_{N,m}$ and thus $\|(\boldsymbol{M}_{N,m})^{\frac{1}{2}}(\hat{\boldsymbol{\theta}}^{or} - \tilde{\boldsymbol{\theta}}^0)\|_2 \leq O_p(1)$. The proof of Corollary A.1 is completed. □

The condition of uniformly bounded η_m in Corollary A.1 naturally holds when m is bounded. However, when m goes to infinity, it implies that either we choose a working correlation matrix $\boldsymbol{R}_i$ close to the true one, or the true correlation is not too strong. The first case involves a consistent and efficient estimator of the correlation structure, which has been discussed in [1], [5] and [4]. For the second case, a variety of conditions can be imposed on the correlation structures to ensure a "weak" dependency. In the following, we provide a sufficient condition which can be verified easily in practice.

Proposition A.1. *Under regularity condition (A3), for any arbitrary true correlation matrix $\boldsymbol{R}^0(\rho_{ij})$, if $|\rho_{ij}| \leq \rho_{|i-j|}$ for $i \neq j$ and $\sum_{k=1}^\infty \rho_k < \infty$, then $\eta_m = \max_{1 \leq i \leq N} \{\lambda_{max}(\boldsymbol{R}_i^{-1} \boldsymbol{R}^0)\}$ is bounded uniformly for any working correlation structures $\boldsymbol{R}_i$'s.*

This indicates that $\boldsymbol{R}^0$ is bounded if the within-individual correlation decays rapidly as m increases. In practice, a wide family of correlation structures satisfy the conditions in Proposition A.1 including the AR-1 and the M-dependent correlation matrices.

Proof: The correlation matrix $\boldsymbol{R}(\rho_{ij})$ is symmetric, which implies that $\|\boldsymbol{R}_{m\times m}\|_1 = \|\boldsymbol{R}_{m\times m}\|_\infty \leq \sum_{k=0}^{m-1} |\rho_k| < \sum_{k=0}^\infty |\rho_k| < \infty$. By noting that $\|\boldsymbol{R}_{m\times m}\|_2^2 \leq \|\boldsymbol{R}_{m\times m}\|_1 \|\boldsymbol{R}_{m\times m}\|_\infty$, we have $\lambda_{max}(\boldsymbol{R}) = \|\boldsymbol{R}\|_2$ uniformly bounded, and thus $\eta_m \leq (\nu_l')^{-1} \sum_{k=0}^\infty |\rho_k| < \infty$. The proof of Proposition A.1 is completed. □

A.7 Preparation to Theorem 3

A.7.1 Convergence Rate of Individual-wise Unpenalized Estimator

In general, the unpenalized heterogeneous estimator plays an important intermediate role in investigating the large sample theory of the penalized estimator. Hence, prior to presenting the theoretical results for the proposed estimator, we discuss the asymptotic behavior of the divergent-dimensional individual-wise least squares estimator $\hat{\boldsymbol{\theta}}^u_{(N)} = \text{vec}(\hat{\boldsymbol{\beta}}^u_{(N)}, \hat{\boldsymbol{\alpha}}^u)$ obtained by minimizing $L_{N,m}(\boldsymbol{\theta})$ in (4).

Note that, for the proposed estimator and the individual-wise heterogeneous estimator, each term of $\boldsymbol{U}_i^T\boldsymbol{V}_i^{-1}\boldsymbol{U}_i$ in $\boldsymbol{D}_{N,m}$ does not equal to $\boldsymbol{X}_i^T\boldsymbol{V}_i^{-1}\boldsymbol{X}_i$, but is a block sparse matrix as $\boldsymbol{\mu}_i(\boldsymbol{\alpha}, \boldsymbol{\beta}_i)$ does not contain any other individualized parameter $\boldsymbol{\beta}_j$ for $j \neq i$. We denote

$$\boldsymbol{D}_{N,m} = \begin{pmatrix} \boldsymbol{D}_{xx}(Np \times Np) & \boldsymbol{D}_{xz}(Np \times q) \\ \boldsymbol{D}_{zx}(q \times Np) & \boldsymbol{D}_{zz}(q \times q) \end{pmatrix},$$

for the individual-wise estimator, where $\boldsymbol{D}_{xx} = \text{bdiag}\Big(\{\boldsymbol{X}_i^T\boldsymbol{V}_i^{-1}\boldsymbol{X}_i\}_{i=1}^N\Big)$ and $\text{bdiag}(\cdot)$ denotes a block-diagonal matrix. Similarly, we have $\boldsymbol{H}_{xx} = \text{bdiag}\Big(\{\boldsymbol{X}_i^T\boldsymbol{V}_i^{-1}\boldsymbol{\Sigma}_i\boldsymbol{V}_i^{-1}\boldsymbol{X}_i\}_{i=1}^N\Big)$ in $\boldsymbol{H}_{N,m}$, and both $\boldsymbol{D}_{xx}$ and $\boldsymbol{H}_{xx}$ will expand as N increases. Following Lemma 1, we obtain the following result:

Lemma A.1. *Under regularity conditions, for any $\delta > 0$ and $\boldsymbol{a} \in \mathbf{R}^{Np+q}$, we have*

$$P\Big(|\boldsymbol{a}^T(\hat{\boldsymbol{\theta}}^u_{(N)} - \boldsymbol{\theta}^0_{(N)})|^2 > \delta\Big) \leq \delta^{-2}\boldsymbol{a}^T(\boldsymbol{D}^s_{N,m}(\boldsymbol{H}_{N,m})^{-1}\boldsymbol{D}^s_{N,m})^{-1}\boldsymbol{a}.$$

If we choose $\boldsymbol{a}$ as a coordinate indicator for $\boldsymbol{\beta}_i$ in $\boldsymbol{\theta}_{(N)}$, that is, $\boldsymbol{a} = (\mathbf{0}_q', \boldsymbol{a}_1', \ldots, \boldsymbol{a}_N')'$, where $\boldsymbol{a}_j \in \mathbf{R}^p$, $1 \leq j \leq N$, $\boldsymbol{a}_j = \mathbf{1}_p$ if $j = i$ or $\boldsymbol{a}_j = \mathbf{0}_p$ if $j \neq i$, Lemma A.1 implies the following corollary, which provides a detailed view of the convergence property for each individual-wise estimator $\hat{\boldsymbol{\beta}}^u_i$ and the population-shared estimator $\hat{\boldsymbol{\alpha}}^u$.

Corollary A.2. *Under regularity conditions, for any $\delta > 0$ and individualized estimator $\hat{\boldsymbol{\beta}}^u_i$,*

$$P\Big(\|\hat{\boldsymbol{\beta}}^u_i - \boldsymbol{\beta}^0_i\|_2 > \delta\Big) \leq p\delta^{-2}\eta_m\lambda_{min}(\boldsymbol{D}_{\boldsymbol{X}_i})^{-1},$$

where $\boldsymbol{D}_{\boldsymbol{X}_i} = \boldsymbol{X}_i^T\boldsymbol{V}_i^{-1}\boldsymbol{X}_i$, $i = 1, \ldots, N$, and for the population-shared estimator $\hat{\boldsymbol{\alpha}}^u$,

$$P\Big(\|(\hat{\boldsymbol{\alpha}}^u - \boldsymbol{\alpha}^0)\|_2 > \delta\Big) \leq q\delta^{-2}\eta_m\lambda_{min}(\boldsymbol{D}_{\boldsymbol{Z}})^{-1},$$

where $\boldsymbol{D}_{\boldsymbol{Z}} = \sum_{i=1}^N \boldsymbol{Z}_i^T\boldsymbol{V}_i^{-1}\boldsymbol{Z}_i$.

Note that the condition ($\mathcal{C}_a$) requires that $m \to \infty$. In the case of bounded m and diverging N, it is straightforward that the consistency of any individualized parameter cannot be achieved since $\lambda_{min}(\boldsymbol{D}_{\boldsymbol{X}_i})$ does not diverge. Intuitively, the increasing number of individuals does not accumulate additional information for the individual-wise parameters. However, the estimator of population-shared parameter $\hat{\boldsymbol{\alpha}}$ could still be consistent as $N \to \infty$ by noting that η_m is bounded and $\lambda_{min}(\boldsymbol{D}_{\boldsymbol{Z}}) \to \infty$.

A.7.2 Proof of Lemma A.1 and Corollary A.2

Note that, for the proposed estimator and the individual-wise least squares estimator, each term of $\boldsymbol{U}_i^T\boldsymbol{V}_i^{-1}\boldsymbol{U}_i$ in $\boldsymbol{D}_{N,m}$ does not equal to $\boldsymbol{X}_i^T\boldsymbol{V}_i^{-1}\boldsymbol{X}_i$, but is a block sparse matrix as $\boldsymbol{\mu}_i$ does not contain any other individualized parameter $\boldsymbol{\beta}_j$ for $j \neq i$. We denote

$$\boldsymbol{D}_{N,m} = \begin{pmatrix} \tau_{xx}(Np \times Np) & \boldsymbol{D}_{xz}(Np \times q) \\ \boldsymbol{D}_{zx}(q \times Np) & \boldsymbol{D}_{zz}(q \times q) \end{pmatrix},$$

for the individual-wise estimator. Specifically,

$$\boldsymbol{D}_{N,m} = \begin{pmatrix} \boldsymbol{X}_1^T\boldsymbol{V}_1^{-1}\boldsymbol{X}_1 & 0 & \dots & 0 & \boldsymbol{X}_1^T\boldsymbol{V}_1^{-1}\boldsymbol{Z}_1 \\ 0 & \boldsymbol{X}_2^T\boldsymbol{V}_2^{-1}\boldsymbol{X}_2 & \dots & 0 & \boldsymbol{X}_2^T\boldsymbol{V}_2^{-1}\boldsymbol{Z}_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & \boldsymbol{X}_N^T\boldsymbol{V}_N^{-1}\boldsymbol{X}_N & \boldsymbol{X}_N^T\boldsymbol{V}_N^{-1}\boldsymbol{Z}_N \\ \boldsymbol{Z}_1^T\boldsymbol{V}_1^{-1}\boldsymbol{X}_1 & \boldsymbol{Z}_2^T\boldsymbol{V}_2^{-1}\boldsymbol{X}_2 & \dots & \boldsymbol{Z}_N^T\boldsymbol{V}_N^{-1}\boldsymbol{X}_N & \sum_{i=1}^N \boldsymbol{Z}_i^T\boldsymbol{V}_i^{-1}\boldsymbol{Z}_i \end{pmatrix},$$

Similarly we have

$$\boldsymbol{H}_{N,m} = \begin{pmatrix} \boldsymbol{X}_1^T\boldsymbol{V}_1^{-1}\boldsymbol{\Sigma}_1\boldsymbol{V}_1^{-1}\boldsymbol{X}_1 & 0 & \dots & 0 & \boldsymbol{X}_1^T\boldsymbol{V}_1^{-1}\boldsymbol{\Sigma}_1\boldsymbol{V}_1^{-1}\boldsymbol{Z}_1 \\ 0 & \boldsymbol{X}_2^T\boldsymbol{V}_2^{-1}\boldsymbol{\Sigma}_2\boldsymbol{V}_2^{-1}\boldsymbol{X}_2 & \dots & 0 & \boldsymbol{X}_2^T\boldsymbol{V}_2^{-1}\boldsymbol{\Sigma}_2\boldsymbol{V}_2^{-1}\boldsymbol{Z}_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & \boldsymbol{X}_N^T\boldsymbol{V}_N^{-1}\boldsymbol{\Sigma}_N\boldsymbol{V}_N^{-1}\boldsymbol{X}_N & \boldsymbol{X}_N^T\boldsymbol{V}_N^{-1}\boldsymbol{\Sigma}_N\boldsymbol{V}_N^{-1}\boldsymbol{Z}_N \\ \boldsymbol{Z}_1^T\boldsymbol{V}_1^{-1}\boldsymbol{\Sigma}_1\boldsymbol{V}_1^{-1}\boldsymbol{X}_1 & \boldsymbol{Z}_2^T\boldsymbol{V}_2^{-1}\boldsymbol{\Sigma}_2\boldsymbol{V}_2^{-1}\boldsymbol{X}_2 & \dots & \boldsymbol{Z}_N^T\boldsymbol{V}_N^{-1}\boldsymbol{\Sigma}_N\boldsymbol{V}_N^{-1}\boldsymbol{X}_N & \sum_{i=1}^N \boldsymbol{Z}_i^T\boldsymbol{V}_i^{-1}\boldsymbol{\Sigma}_i\boldsymbol{V}_i^{-1}\boldsymbol{Z}_i \end{pmatrix},$$

Since $\boldsymbol{H}_{N,m} \leq \eta_m \boldsymbol{D}_{N,m}$, we have $\boldsymbol{a}^T(\boldsymbol{D}_{N,m}(\boldsymbol{H}_{N,m})^{-1}\boldsymbol{D}_{N,m})^{-1} \leq \eta_m \boldsymbol{a}^T(\boldsymbol{D}_{N,m})^{-1}\boldsymbol{a}$. Note that $\boldsymbol{D}_{N,m}$ can be decomposed as

$$\boldsymbol{D}_{N,m} = \begin{pmatrix} \boldsymbol{I}_{Np} & \boldsymbol{0} \\ \boldsymbol{D}_{zx}(\boldsymbol{D}_{xx})^{-1} & \boldsymbol{I}_q \end{pmatrix} \begin{pmatrix} \boldsymbol{D}_{xx} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{D}_{zz} - \boldsymbol{D}_{zx}(\boldsymbol{D}_{xx})^{-1}\boldsymbol{D}_{xz} \end{pmatrix} \begin{pmatrix} \boldsymbol{I}_{Np} & (\boldsymbol{D}_{xx})^{-1}\boldsymbol{D}_{xz} \\ \boldsymbol{0} & \boldsymbol{I}_q \end{pmatrix},$$

and hence

$$(\boldsymbol{D}_{N,m})^{-1} = \begin{pmatrix} \boldsymbol{I}_{Np} & \boldsymbol{0} \\ -\boldsymbol{D}_{zx}(\boldsymbol{D}_{xx})^{-1} & \boldsymbol{I}_q \end{pmatrix} \begin{pmatrix} (\boldsymbol{D}_{xx})^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & (\boldsymbol{D}_{zz} - \boldsymbol{D}_{zx}(\boldsymbol{D}_{xx})^{-1}\boldsymbol{D}_{xz})^{-1} \end{pmatrix} \begin{pmatrix} \boldsymbol{I}_{Np} & -(\boldsymbol{D}_{xx})^{-1}\boldsymbol{D}_{xz} \\ \boldsymbol{0} & \boldsymbol{I}_q \end{pmatrix}.$$

Therefore, for any coordinate indicator $\boldsymbol{a}$ of $\boldsymbol{\beta}_i$, $\boldsymbol{a}^T(\boldsymbol{D}_{N,m})^{-1}\boldsymbol{a} = \boldsymbol{1}_p^T(\boldsymbol{X}_i^T\boldsymbol{V}_i^{-1}\boldsymbol{X}_i)^{-1}\boldsymbol{1}_p \leq p\lambda_{min}(\boldsymbol{X}_i^T\boldsymbol{V}_i^{-1}\boldsymbol{X}_i)^{-1}$. The result for the population-shared parameter $\hat{\boldsymbol{\alpha}}$ could be obtained following the same argument. □

A.7.3 Uniform Consistency with Divergent N

Next, based on the result in Lemma A.1 and Corollary A.2, under condition $(\mathcal{I}_a)$ or $(\mathcal{I}_b)$, we provide a stronger uniform consistency regarding the divergent-dimensional parameters when both $N \to \infty$ and $m \to \infty$.

Lemma A.2. *Under regularity conditions (A1)-(A6), given $\tau_m = \lambda_{min}(\boldsymbol{D}_{N,m}\boldsymbol{H}_{N,m}^{-1}\boldsymbol{D}_{N,m})$, if either condition $(\mathcal{I}_a)$ holds with $N = O(\tau_m)$ or condition $(\mathcal{I}_b)$ holds with $\log(N) = O(\tau_m)$, for any $\delta > 0$, as $\tau_m \to \infty$, we have*

$$P\left(\|\hat{\boldsymbol{\theta}}^u_{(N)} - \boldsymbol{\theta}^0_{(N)}\|_\infty > \delta\right) \to 0.$$

Lemma A.2 indicates that if N diverges at a limited rate compared to m, we are able to achieve a stronger uniform consistency in terms of the L_∞ norm. The allowed divergence rate of N depends on the tail property of the random error's distribution. Note that the τ_m in conditions $(\mathcal{I}_a)$ and $(\mathcal{I}_b)$ could also be replaced with $\eta_m^{-1}\lambda_{min}(\boldsymbol{D}_{N,m})$ analogous to the above discussion, which leads to a sufficient condition.

Proof: We denote

$$\tilde{\boldsymbol{X}} = (\boldsymbol{X}, \boldsymbol{Z}) = \begin{pmatrix} \boldsymbol{X}_1 & \dots & \boldsymbol{0} & \boldsymbol{Z}_1 \\ \vdots & \ddots & \vdots & \vdots \\ \boldsymbol{0} & \dots & \boldsymbol{X}_N & \boldsymbol{Z}_N \end{pmatrix},$$

and $\tilde{\boldsymbol{V}} = \text{diag}(\boldsymbol{V}_1, \ldots, \boldsymbol{V}_N)$, $\tilde{\boldsymbol{\Sigma}} = \text{diag}(\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_N)$, $\tilde{\boldsymbol{\varepsilon}} = (\tilde{\boldsymbol{\varepsilon}}_1', \ldots, \tilde{\boldsymbol{\varepsilon}}_N')'$.

Denote $\hat{\boldsymbol{\theta}}^u_{(N)} = \left((\hat{\boldsymbol{\beta}}^u_{(N)})', (\hat{\boldsymbol{\alpha}}^u)' \right)'$, we have the least squares estimator

$$\begin{aligned}
(\hat{\boldsymbol{\theta}}^u_{(N)} - \boldsymbol{\theta}^0_{(N)}) &= \left(\tilde{\boldsymbol{X}}^T \tilde{\boldsymbol{V}}^{-1} \tilde{\boldsymbol{X}} \right)^{-1} \tilde{\boldsymbol{X}}^T \tilde{\boldsymbol{V}}^{-1} \tilde{\boldsymbol{\varepsilon}} \\
&= \left(\tilde{\boldsymbol{X}}^T \tilde{\boldsymbol{V}}^{-1} \tilde{\boldsymbol{X}} \right)^{-1} \tilde{\boldsymbol{X}}^T \tilde{\boldsymbol{V}}^{-1} \tilde{\boldsymbol{\Sigma}}^{1/2} \tilde{\boldsymbol{\Sigma}}^{-1/2} \tilde{\boldsymbol{\varepsilon}} \\
&= \left(\tilde{\boldsymbol{X}}^T \tilde{\boldsymbol{V}}^{-1} \tilde{\boldsymbol{X}} \right)^{-1} \tilde{\boldsymbol{X}}^T \tilde{\boldsymbol{V}}^{-1} \tilde{\boldsymbol{\Sigma}}^{1/2} \tilde{\boldsymbol{\varepsilon}}^*.
\end{aligned}$$

Under condition ($\mathcal{I}_a$) that $N = o(\tau_m)$, by Chebyshev's inequality,

$$\begin{aligned}
\text{P}(\|\hat{\boldsymbol{\theta}}^u_{(N)} - \boldsymbol{\theta}^0_{(N)}\|_\infty > \delta) &= \text{P}(\|(\tilde{\boldsymbol{X}}^T \tilde{\boldsymbol{V}}^{-1} \tilde{\boldsymbol{X}})^{-1} \tilde{\boldsymbol{X}}^T \tilde{\boldsymbol{V}}^{-1} \tilde{\boldsymbol{\varepsilon}}\|_\infty > \delta) \\
&\leq \delta^{-2} tr\left((\tilde{\boldsymbol{X}}^T \tilde{\boldsymbol{V}}^{-1} \tilde{\boldsymbol{X}})^{-1} \tilde{\boldsymbol{X}}^T \tilde{\boldsymbol{V}}^{-1} \tilde{\boldsymbol{\Sigma}} \tilde{\boldsymbol{V}}^{-1} \tilde{\boldsymbol{X}} (\tilde{\boldsymbol{X}}^T \tilde{\boldsymbol{V}}^{-1} \tilde{\boldsymbol{X}})^{-1} \right) \\
&= \delta^{-2} tr\left((\boldsymbol{D}_{N,m})^{-1} \boldsymbol{H}_{N,m} (\boldsymbol{D}_{N,m})^{-1} \right) \\
&\leq \delta^{-2} (Np + q) \lambda_{max}\left((\boldsymbol{D}_{N,m})^{-1} \boldsymbol{H}_{N,m} (\boldsymbol{D}_{N,m})^{-1} \right) \\
&\leq \delta^{-2} (Np + q) (\tau_m)^{-1} \to 0
\end{aligned}$$

as $\tau_m \to \infty$.

Moreover, let $\boldsymbol{a}_t = ((\tilde{\boldsymbol{X}}^T \tilde{\boldsymbol{V}}^{-1} \tilde{\boldsymbol{X}})^{-1} \tilde{\boldsymbol{X}}^T \tilde{\boldsymbol{V}}^{-1} \tilde{\boldsymbol{\Sigma}}^{1/2})_{t\cdot}$ denote the tth row of $(\tilde{\boldsymbol{X}}^T \tilde{\boldsymbol{V}}^{-1} \tilde{\boldsymbol{X}})^{-1} \tilde{\boldsymbol{X}}^T \tilde{\boldsymbol{V}}^{-1} \tilde{\boldsymbol{\Sigma}}^{1/2}$, $t = 1, \ldots, (Np + q)$. By condition (i) in ($\mathcal{I}_b$), we have

$$\text{P}(|\boldsymbol{a}_t^T \boldsymbol{\varepsilon}_{\boldsymbol{i}}^*| > \delta) < 2\text{exp}(-\frac{\delta^2}{c_\sigma^2 \|\boldsymbol{a}_t\|_2^2}).$$

Hence

$$\begin{aligned}
\text{P}(\|\hat{\boldsymbol{\theta}}^u_{(N)} - \boldsymbol{\theta}^0_{(N)}\|_\infty > \delta) &= \text{P}(\|(\tilde{\boldsymbol{X}}^T \tilde{\boldsymbol{V}}^{-1} \tilde{\boldsymbol{X}})^{-1} \tilde{\boldsymbol{X}}^T \tilde{\boldsymbol{V}}^{-1} \tilde{\boldsymbol{\Sigma}}^{1/2} \tilde{\boldsymbol{\varepsilon}}^*\|_\infty > \delta) \\
&\leq \sum_{t=1}^{Np+q} \text{P}(|\boldsymbol{a}_t^T \boldsymbol{\varepsilon}^*| > \delta) \\
&\leq \sum_{t=1}^{Np+q} 2\text{exp}(-\frac{\delta^2}{c_\sigma^2 \|\boldsymbol{a}_t\|_2^2}) \\
&\leq (Np + q) \max_{1 \leq t \leq Np+q} (2\text{exp}(-\frac{\delta^2}{c_\sigma^2 \|\boldsymbol{a}_t\|_2^2})) \\
&= 2(Np + q)\text{exp}(-\frac{\delta^2}{c_\sigma^2 \max_{1 \leq t \leq Np+q}(\|\boldsymbol{a}_t\|_2^2)}).
\end{aligned}$$

Note that

$$\begin{aligned}
\max_{1 \leq t \leq Np+q} (\|\boldsymbol{a}_t\|_2^2) &\leq \lambda_{max}\left((\tilde{\boldsymbol{X}}^T \tilde{\boldsymbol{V}}^{-1} \tilde{\boldsymbol{X}})^{-1} \tilde{\boldsymbol{X}}^T \tilde{\boldsymbol{V}}^{-1} \tilde{\boldsymbol{\Sigma}} \tilde{\boldsymbol{V}}^{-1} \tilde{\boldsymbol{X}} (\tilde{\boldsymbol{X}}^T \tilde{\boldsymbol{V}}^{-1} \tilde{\boldsymbol{X}})^{-1} \right) \\
&= \lambda_{max}\left((\boldsymbol{D}_{N,m})^{-1} \boldsymbol{H}_{N,m} (\boldsymbol{D}_{N,m})^{-1} \right) = (\tau_m)^{-1}.
\end{aligned}$$

By condition (ii) in ($\mathcal{I}_b$) that $\log(N) = o(\tau_m)$,

$$\mathrm{P}(\|\hat{\boldsymbol{\theta}}^u_{(N)} - \boldsymbol{\theta}^0_{(N)}\|_\infty > \delta) \leq 2(Np+q)\exp(-\frac{\delta^2\tau_m}{c_\sigma^2}) \to 0$$

as $\tau_m \to \infty$. □

A.8 Proofs of Theorem 3, Corollaries 1-3

We first establish the following result.

Lemma A.3. *Suppose there is a sequence of numbers $\{a_i\}_{i=1,\ldots,N}$ associated with a partition of index sets $\mathcal{G}_l$ ($l = 1,\ldots,L$), such that $|a_i - b_l| \leq \epsilon$ for any $i \in \mathcal{G}_l$, where ϵ is a small positive value. Then there is a local minimizer $\hat{\boldsymbol{b}}$ of following objective function*

$$S(\boldsymbol{b}|\boldsymbol{a}) = \sum_{i=1}^{N}\Big(\underset{1\leq l\leq L}{\wedge}|a_i - b_l|\Big),$$

such that $\|\hat{\boldsymbol{b}} - \boldsymbol{b}\|_\infty \leq 2\epsilon$, where $\underset{1\leq l\leq L}{\wedge}|a_i - b_l| = \min_{1\leq l\leq L}(|a_i - b_l|)$.

Proof: Without loss of generality, assume $b_1 = 0$, we have $|a_i| \leq \epsilon$ for any $i \in \mathcal{G}_1$ and hence $\sum_{i\in\mathcal{G}_1}|a_i| \leq |\mathcal{G}_1|\epsilon$. Moreover, note that $\sum_{i\in\mathcal{G}_1}|a_i - 2\epsilon| = \sum_{i\in\mathcal{G}_1}(2\epsilon - a_i) \geq \sum_{i\in\mathcal{G}_1}\epsilon = |\mathcal{G}_1|\epsilon$ and $\sum_{i\in\mathcal{G}_1}|a_i + 2\epsilon| = \sum_{i\in\mathcal{G}_1}(2\epsilon - a_i) \geq \sum_{i\in\mathcal{G}_1}\epsilon = |\mathcal{G}_1|\epsilon$. Therefore there is a minimizer $|\hat{b}_1| \leq 2\epsilon$ and the proof of Lemma A.3 is completed.

Next we establish another lemma regarding subgrouping on heterogeneous parameters. Denote $\boldsymbol{B}_{\boldsymbol{\beta}^0_i}(r)$ as a ball in $\mathbf{R}^p$ centered at $\boldsymbol{\beta}^0_i$ with a radius $r > 0$.

Lemma A.4. *Suppose either condition ($\mathcal{I}_a$) holds with $N = O(\tau_m)$ or condition ($\mathcal{I}_b$) holds with $\log(N) = O(\tau_m)$, for any constant $r > 0$, as $\tau_m \to \infty$, there exists a local minimizer $(\hat{\boldsymbol{\alpha}}^T, \hat{\boldsymbol{\beta}}^T, \hat{\boldsymbol{\gamma}}^T)^T$ of $Q_{N,m}$ in (3) such that*

$$P\Big(\bigcap_{1\leq i\leq N}\{\hat{\boldsymbol{\beta}}_i \in \boldsymbol{B}_{\boldsymbol{\beta}^0_i}(r)\}\bigcap\{\hat{\boldsymbol{\alpha}} \in \boldsymbol{B}_{\boldsymbol{\alpha}^0}(r)\}\bigcap\{\hat{\boldsymbol{\gamma}} \in \boldsymbol{B}_{\boldsymbol{\gamma}^0}(r)\}\Big) \to 1.$$

Proof: The proposed objective function is

$$\begin{aligned} Q_{N,m}(\boldsymbol{\beta}_{(N)}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) &= \frac{1}{2}\sum_{i=1}^{N}\| \boldsymbol{y}_i - \boldsymbol{X_i}\boldsymbol{\beta_i} - \boldsymbol{Z_i}\boldsymbol{\alpha} \|_2^2 + \lambda_{N,m}\sum_{i=1}^{N}\sum_{k=1}^{p}s(\beta_{ik}, \gamma_k) \\ &= L_{N,m}(\boldsymbol{\theta}_{(N)}) + S_{\lambda_{N,m}}(\boldsymbol{\beta}_{(N)}, \boldsymbol{\gamma}). \end{aligned}$$

Let $\boldsymbol{\theta}^*_{(N)} = \boldsymbol{\theta}^0_{(N)} + (\tau_m)^{-1/2}\boldsymbol{u}$, $\boldsymbol{\gamma}^* \in \mathbb{R}^p$, where $\| \boldsymbol{u} \|_2 = d$. Note that $S_{\lambda_{N,m}}(\boldsymbol{\beta}^0_{(N)}, \boldsymbol{\gamma}^0) = 0$, by Taylor's expansion, we have

$$\begin{aligned} D_{N,m}(\boldsymbol{u}) &= Q_{N,m}(\boldsymbol{\theta}^*_{(N)}, \boldsymbol{\gamma}^*) - Q_{N,m}(\boldsymbol{\theta}^0_{(N)}, \boldsymbol{\gamma}^0) \\ &= L_{N,m}(\boldsymbol{\theta}^*_{(N)}) - L_{N,m}(\boldsymbol{\theta}^0_{(N)}) + S_{\lambda_{N,m}}(\boldsymbol{\beta}^*_{(N)}, \boldsymbol{\gamma}^*) \\ &= (\tau_m)^{-1/2}\dot{L}^T_{N,m}(\boldsymbol{\theta}^0_{(N)})\boldsymbol{u} + \frac{1}{2}(\tau_m)^{-1}\boldsymbol{u}^T\ddot{L}_{N,m}(\boldsymbol{\theta}^0_{(N)})\boldsymbol{u} + S_{\lambda_{N,m}}(\boldsymbol{\beta}^*_{(N)}, \boldsymbol{\gamma}^*), \\ &= (\tau_m)^{-1/2}(\boldsymbol{G}_{N,m})^T\boldsymbol{u} + \frac{1}{2}(\tau_m)^{-1}\boldsymbol{u}^T\boldsymbol{D}_{N,m}\boldsymbol{u} + S_{\lambda_{N,m}}(\boldsymbol{\beta}^*_{(N)}, \boldsymbol{\gamma}^*), \end{aligned}$$

where $\dot{L}_{N,m}$ is the gradient vector of $L_{N,m}(\boldsymbol{\theta})$ and $\ddot{L}_{N,m}$ is the Jacobian matrix. Note that

$$\begin{aligned} P(\boldsymbol{u}^T(\boldsymbol{H}_{N,m})^{-1/2}\boldsymbol{G}_{N,m}| > \delta) &\leq \delta^{-2}\boldsymbol{u}^T\mathbf{E}((\boldsymbol{H}_{N,m})^{-1/2}\boldsymbol{G}_{N,m}(\boldsymbol{G}_{N,m})^T(\boldsymbol{H}_{N,m})^{-1/2})\boldsymbol{u} \\ &\leq \delta^{-2}d^2, \end{aligned}$$

implying that $\boldsymbol{u}^T(\boldsymbol{H}_{N,m})^{-1/2}\boldsymbol{G}_{N,m} = O_p(d)$. Moreover, we have

$$\begin{aligned} (\boldsymbol{H}_{N,m})^{1/2} &= (\boldsymbol{D}_{N,m})^{1/2}(\boldsymbol{D}_{N,m})^{-1/2}(\boldsymbol{H}_{N,m})^{1/2}(\boldsymbol{D}_{N,m})^{-1/2}(\boldsymbol{D}_{N,m})^{1/2} \\ &\leq (\boldsymbol{D}_{N,m})^{1/2}\lambda_{max}\left((\boldsymbol{D}_{N,m})^{-1/2}(\boldsymbol{H}_{N,m})^{1/2}(\boldsymbol{D}_{N,m})^{-1/2}\right)(\boldsymbol{D}_{N,m})^{1/2} \\ &= \lambda_{min}\left((\boldsymbol{D}_{N,m})^{1/2}(\boldsymbol{H}_{N,m})^{-1/2}(\boldsymbol{D}_{N,m})^{1/2}\right)^{-1}\boldsymbol{D}_{N,m} \\ &= (\tau_m)^{-1/2}\boldsymbol{D}_{N,m}, \end{aligned}$$

and thus $(\tau_m)^{-1/2}(\boldsymbol{H}_{N,m})^{1/2} \leq (\tau_m)^{-1}\boldsymbol{D}_{N,m}$. Consequently, if d is sufficiently large, then the second term in $D_{N,m}(\boldsymbol{u})$ dominates the first term, which implies that, with probability tending to 1, $D_{N,m}(\boldsymbol{u}) > 0$ at $\| \boldsymbol{u} \|_2 = d$. Hence we have

$$P\left\{ \inf_{\|\boldsymbol{u}\|_2=d} D_{N,m}(\boldsymbol{u}) > 0 \right\} \to 1.$$

This implies that, with probability tending to 1, there exists a local minimizer $\hat{\boldsymbol{\theta}}_{(N)}$ in the ball $B(\boldsymbol{\theta}^0_{(N)}, (\tau_m)^{-1/2}d)$. In particular, this indicates that the convergence rate for estimator of any individualized parameter $\hat{\boldsymbol{\beta}}_i$ is $(\tau_m)^{1/2}$. Following the proof of Lemma A.2, under condition $\mathcal{I}_a$ or $\mathcal{I}_b$, we have $P(\|\hat{\boldsymbol{\beta}}_{(N)} - \boldsymbol{\beta}^0_{(N)}\|_\infty > p^{-1}r) \to 0$ for any positive constant r. By Lemma A.3, given $\|\hat{\boldsymbol{\beta}}_{(N)} - \boldsymbol{\beta}^0_{(N)}\|_\infty \leq p^{-1}r$, there exists a minimizer $\hat{\boldsymbol{\gamma}}$ of $S_{\lambda_{N,m}}(\boldsymbol{\gamma}|\hat{\boldsymbol{\beta}}_{(N)})$, such that $\hat{\boldsymbol{\gamma}} \in B(\boldsymbol{\gamma}^0, r)$. The proof of Lemma A.4 is completed.

Next we show that the objective function $Q_{N,m}(\boldsymbol{\theta}^*_{(N)}, \boldsymbol{\gamma}^*)$ is convex at $\{\boldsymbol{\theta}^*_{(N)} \in B(\boldsymbol{\theta}^0_{(N)}, (\tau_m)^{-1/2}d)\} \cap \{\boldsymbol{\gamma}^* \in B(\boldsymbol{\gamma}^0, (\tau_m)^{-1/2}d)\}$ when m is sufficiently large. Note that, if $\beta^0_{ik} = \gamma^0_k$, we have

$$\begin{aligned} \sup_{\beta^*_{ik}\in B(\beta^0_{ik}), \gamma^*_k \in B(\gamma^0_k)} |\beta^*_{ik} - \gamma^*_k| &\leq \sup_{\beta^*_{ik}\in B(\beta^0_{ik})} |\beta^*_{ik} - \beta^0_{ik}| + \sup_{\gamma^*_k \in B(\gamma^0_k)} |\gamma^*_k - \gamma^0_k| + |\beta^0_{ik} - \gamma^0_k| \\ &\leq 2(\tau_m)^{-1/2}d + |\beta^0_{ik} - \gamma^0_k| \to 0, \end{aligned}$$

and $\inf_{\beta^*_{ik}\in B(\beta^0_{ik})} |\beta^*_{ik}| \geq (|\gamma^0_k| - (\tau_m)^{-1/2}d)_+ \to |\gamma^0_k|$. It follows

$$P\left(\sup_{\beta^*_{ik}\in B(\beta^0_{ik}), \gamma^*_k \in B(\gamma^0_k)} |\beta^*_{ik} - \gamma^*_k| \leq \inf_{\beta^*_{ik}\in B(\beta^0_{ik})} |\beta^*_{ik}| \right) \to 1.$$

Define

$$\tilde{S}_{\lambda_{N,m}}(\boldsymbol{\beta}^*_{(N)}, \boldsymbol{\gamma}^*) = \lambda_{N,m} \sum_{k=1}^{p} \left\{ \sum_{i\in\mathcal{G}^c_k} |\beta^*_{ik}| + \sum_{i\in\mathcal{G}_k} |\beta^*_{ik} - \gamma^*_k| \right\},$$

and $\tilde{Q}_{N,m} = L_{N,m} + \tilde{S}_{\lambda_{N,m}}$. We have $Q_{N,m}(\boldsymbol{\theta}^*_{(N)}, \boldsymbol{\gamma}^*) = \tilde{Q}_{N,m}(\boldsymbol{\theta}^*_{(N)}, \boldsymbol{\gamma}^*)$ at $\{\boldsymbol{\theta}^*_{(N)} \in B(\boldsymbol{\theta}^0_{(N)}, (\tau_m)^{-1/2}d)\} \cap \{\boldsymbol{\gamma}^* \in B(\boldsymbol{\gamma}^0, (\tau_m)^{-1/2}d)\}$ when τ_m is sufficiently large, and thus $\operatorname{argmin} Q_{N,m} = \operatorname{argmin} \tilde{Q}_{N,m}$.

Let $\boldsymbol{\theta}^{**}_{(N)} = \boldsymbol{\theta}^{0}_{(N)} + \lambda^{-1}_{N,m}\boldsymbol{u}$ and $\boldsymbol{\gamma}^{**} = \boldsymbol{\gamma}^{0} + \lambda^{-1}_{N,m}\boldsymbol{v}$, similarly it follows that

$$\begin{aligned} D_{N,m}(\boldsymbol{u},\boldsymbol{v}) &= \tilde{Q}_{N,m}(\boldsymbol{\theta}^{**}_{(N)},\boldsymbol{\gamma}^{**}) - \tilde{Q}_{N,m}(\boldsymbol{\theta}^{0}_{(N)},\boldsymbol{\gamma}^{0}) = L_{N,m}(\boldsymbol{\theta}^{**}_{(N)}) - L_{N,m}(\boldsymbol{\theta}^{0}_{(N)}) + \tilde{S}_{\lambda_{N,m}}(\boldsymbol{\beta}^{**}_{(N)},\boldsymbol{\gamma}^{**}) \\ &= \frac{(\tau_m)^{1/2}}{\lambda_{N,m}}(\tau_m)^{-1/2}\dot{L}^{T}_{N,m}(\boldsymbol{\theta}^{0}_{(N)})\boldsymbol{u} + \frac{\tau_m}{\lambda^{2}_{N,m}}\frac{1}{2}(\tau_m)^{-1}\boldsymbol{u}^{T}\ddot{L}_{N,m}(\boldsymbol{\theta}^{0}_{(N)})\boldsymbol{u} \\ &\quad + \lambda_{N,m}\sum_{k=1}^{p}\Big\{\sum_{i\in\mathcal{G}^{c}_{k}}\lambda^{-1}_{N,m}|u_{ik}| + \sum_{i\in\mathcal{G}_{k}}\lambda^{-1}_{N,m}|u_{ik}-v_{k}|\Big\}. \end{aligned}$$

Since $\frac{\lambda_{N,m}}{(\tau_m)^{1/2}} \to \infty$, hence $D_{N,m}(\boldsymbol{u},\boldsymbol{v}) \to_p D(\boldsymbol{u},\boldsymbol{v})$, where

$$D(\boldsymbol{u},\boldsymbol{v}) = \sum_{k=1}^{p}\Big\{\sum_{i\in\mathcal{G}^{c}_{k}}|u_{ik}| + \sum_{i\in\mathcal{G}_{k}}|u_{ik}-v_{k}|\Big\},$$

which is minimized at $\{u_{ik}=0 | i\in\mathcal{G}^{c}_{k};\quad u_{ik}=v_{k} | i\in\mathcal{G}_{k}\}$. Because $D_{N,m}(\boldsymbol{u},\boldsymbol{v})$ is a convex function, it follows [3] that $\operatorname{argmin} D_{N,m} \to \operatorname{argmin} D$, and thus $\operatorname{argmin} Q_{N,m} \to \operatorname{argmin} D$. This implies that $P(\hat{\beta}_{ik}=0 | i\in\mathcal{G}^{c}_{k}) \to 1$ and $P(\hat{\beta}_{ik}=\hat{\gamma}_{k} | i\in\mathcal{G}_{k}) \to 1$. The proof of Theorem 3 is completed and the proof of Corollaries 1, 2 and 3 follow immediately. □

A.9 Proof of Theorem 4

First, we prove the estimation consistency as $\lambda_{m^*} = o(m^*)$. Recall that $\boldsymbol{\theta}^{*}_{\boldsymbol{i}} = \text{vec}(\boldsymbol{\beta}^{*}_{\boldsymbol{i}},\boldsymbol{\alpha}^{*})$ and $\tilde{\boldsymbol{X}}_i^{\,*} = (\boldsymbol{X}^{*}_{i},\boldsymbol{Z}^{*}_{i})$. Given $\hat{\boldsymbol{\gamma}}$, let

$$\begin{aligned} Q_{i,m^*}(\boldsymbol{\theta}^{*}_{\boldsymbol{i}}|\hat{\boldsymbol{\gamma}}) &= \| \boldsymbol{y}^{*}_{\boldsymbol{i}} - \tilde{\boldsymbol{X}}^{*}_{i}\boldsymbol{\theta}^{*}_{\boldsymbol{i}} \|^{2}_{2} + (\lambda_{m^*})\sum_{k=1}^{p} s(\beta^{*}_{ik},\hat{\gamma}_{k}) \\ &= L_{i,m^*}(\boldsymbol{\theta}^{*}_{\boldsymbol{i}}) + S_{\lambda_{m^*}}(\boldsymbol{\beta}^{*}_{\boldsymbol{i}}|\hat{\boldsymbol{\gamma}}), \end{aligned}$$

which is minimized at $\hat{\boldsymbol{\theta}}^{*}_{\boldsymbol{i}}$, where $L_{i,m^*}(\cdot)$ is the squared loss function and $S_{\lambda_{m^*}}(\cdot)$ is the MDSP function.

Suppose $\frac{1}{m^*}\tilde{\boldsymbol{X}}_i^{*\,T}\tilde{\boldsymbol{X}}^{*}_{i} \to \boldsymbol{C}_i$ where C_i is a positive definite matrix. Following [6], we define another function not related to m^*

$$Q_i(\boldsymbol{\theta}^{*}_{\boldsymbol{i}}|\boldsymbol{\gamma}^{0}) = (\boldsymbol{\theta}^{*}_{\boldsymbol{i}} - \boldsymbol{\theta}^{\boldsymbol{0}}_{\boldsymbol{i}})^{T}C_i(\boldsymbol{\theta}^{*}_{\boldsymbol{i}} - \boldsymbol{\theta}^{\boldsymbol{0}}_{\boldsymbol{i}}) + \lambda_0\sum_{k=1}^{p} s(\beta^{*}_{ik},\gamma^{0}_{k}),$$

and $\lambda_{m^*}/m^* \to \lambda_0$. Since C_i is not singular, if $\lambda_0 = 0$, then Q_i has a unique minimizer $\boldsymbol{\theta}^{\boldsymbol{0}}_{\boldsymbol{i}}$. Following [6], we need to show

$$\sup_{\boldsymbol{\theta}^{*}_{\boldsymbol{i}}\in\boldsymbol{\Theta}}\left|\frac{1}{m^*}Q_{i,m^*}(\boldsymbol{\theta}^{*}_{\boldsymbol{i}}|\hat{\boldsymbol{\gamma}}) - Q_i(\boldsymbol{\theta}^{*}_{\boldsymbol{i}}|\boldsymbol{\gamma}^{0}) - \sigma^{2}\right| \to_p 0, \tag{A.3}$$

for any compact set $\boldsymbol{\Theta}$ and also that

$$\hat{\boldsymbol{\theta}}^{*}_{\boldsymbol{i}} = O_p(1). \tag{A.4}$$

The result in (A.3) follows

$$\frac{1}{m*}\| \boldsymbol{y}^{*}_{\boldsymbol{i}} - \tilde{\boldsymbol{X}}^{*}_{i}\boldsymbol{\theta}^{*}_{\boldsymbol{i}} \|^{2}_{2} \longrightarrow_p (\boldsymbol{\theta}^{*}_{\boldsymbol{i}} - \boldsymbol{\theta}^{\boldsymbol{0}}_{\boldsymbol{i}})^{T}C_i(\boldsymbol{\theta}^{*}_{\boldsymbol{i}} - \boldsymbol{\theta}^{\boldsymbol{0}}_{\boldsymbol{i}}) + \sigma^{2}$$

according to standard results [10] and also

$$\sup_{\boldsymbol{\theta}_i^* \in \boldsymbol{\Theta}} \left| \frac{1}{m} S_{\lambda_{m^*}}(\boldsymbol{\beta}_i^* | \hat{\boldsymbol{\gamma}}) - S_0(\boldsymbol{\beta}_i^* | \boldsymbol{\gamma}^0) \right| \leq \sup_{\boldsymbol{\theta}_i^* \in \boldsymbol{\Theta}} \frac{1}{m^*} \left| S_{\lambda_{m^*}}(\boldsymbol{\beta}_i^* | \hat{\boldsymbol{\gamma}}) - S_{\lambda_{m^*}}(\boldsymbol{\beta}_i^* | \boldsymbol{\gamma}^0) \right| + \sup_{\boldsymbol{\theta}_i^* \in \boldsymbol{\Theta}} \left| \frac{1}{m^*} S_{\lambda_{m^*}}(\boldsymbol{\beta}_i^* | \boldsymbol{\gamma}^0) - S_0(\boldsymbol{\beta}_i^* | \boldsymbol{\gamma}^0) \right|$$
$$\leq \frac{\lambda_{m^*} p}{m^*} \parallel \hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^0 \parallel_2 + c | \frac{\lambda_{m^*}}{m^*} - \lambda_0 | \to 0,$$

where $c > 0$ is a constant. Although Q_{i,m^*} is not convex, we note that $\operatorname{argmin}(L_{i,m^*}) = O_p(1)$ and $\operatorname{argmin}(S_{\lambda_{m^*}}) = O_p(1)$. It follows that $\hat{\boldsymbol{\theta}}_i^* = \operatorname{argmin}(Q_{i,m^*}) = O_p(1)$. Under (A.3) and (A.4), we have

$$\operatorname{argmin}(Q_{i,m^*}) \to_p \operatorname{argmin}(Q_i).$$

Next, we prove the selection consistency as $\lambda_{m^*}/\sqrt{m^*} \to \infty$. Let $\boldsymbol{\beta}_i^* = \boldsymbol{\beta}_i^0 + \frac{\boldsymbol{u}}{\lambda_{m^*}}$ and $\boldsymbol{\alpha}^* = \boldsymbol{\alpha}^0 + \frac{\boldsymbol{v}}{\lambda_{m^*}}$, where $\boldsymbol{u} = O_p(1)$ and $\boldsymbol{v} = O_p(1)$. Let

$$\begin{aligned}
D_{i,m^*}(\boldsymbol{u}, \boldsymbol{v}) &= Q_{i,m^*}(\boldsymbol{\beta}_i^*, \boldsymbol{\alpha}^* | \hat{\boldsymbol{\gamma}}) - Q_{i,m^*}(\boldsymbol{\beta}_i^0, \boldsymbol{\alpha}^0 | \hat{\boldsymbol{\gamma}}) \\
&= L_{i,m^*}(\boldsymbol{\beta}_i^*, \boldsymbol{\alpha}^*) - L_{i,m^*}(\boldsymbol{\beta}_i^0, \boldsymbol{\alpha}^0) + S_{\lambda_{m^*}}(\boldsymbol{\beta}_i^* | \hat{\boldsymbol{\gamma}}) - S_{\lambda_{m^*}}(\boldsymbol{\beta}_i^0 | \hat{\boldsymbol{\gamma}}) \\
&= \| \boldsymbol{\varepsilon}_i - \boldsymbol{X}_i^* \frac{\boldsymbol{u}}{\lambda_{m^*}} - \boldsymbol{Z}_i^* \frac{\boldsymbol{v}}{\lambda_{m^*}} \|_2^2 - \| \boldsymbol{\varepsilon}_i \|_2^2 + \lambda_{m^*} \sum_{k=1}^{p} [s(\beta_{ik}^0 + \frac{u_k}{\lambda_{m^*}}, \hat{\gamma}_k) - s(\beta_{ik}^0, \hat{\gamma}_k)] \\
&= \frac{\sqrt{m^*}}{\lambda_{m^*}} \frac{1}{\sqrt{m^*}} \boldsymbol{\varepsilon}_i^T (\boldsymbol{X}_i^* \boldsymbol{u} + \boldsymbol{Z}_i^* \boldsymbol{v}) + \frac{m^*}{\lambda_{m^*}^2} (\boldsymbol{u}^T, \boldsymbol{v}^T) \left(\frac{1}{m^*} (\boldsymbol{X}_i^*, \boldsymbol{Z}_i^*)^T (\boldsymbol{X}_i^*, \boldsymbol{Z}_i^*) \right) (\boldsymbol{u}^T, \boldsymbol{v}^T)^T \\
&\quad + \lambda_{m^*} \sum_{k=1}^{p} [s(\beta_{ik}^0 + \frac{u_k}{\lambda_{m^*}}, \hat{\gamma}_k) - s(\beta_{ik}^0, \hat{\gamma}_k)].
\end{aligned}$$

The first two terms vanish as $\lambda_{m^*}/\sqrt{m^*} \to \infty$. Let $\hat{\boldsymbol{\gamma}} \to \boldsymbol{\gamma}^{(0)}$, it follows that

$$D_{i,m^*}(\boldsymbol{u}, \boldsymbol{v}) \to \sum_{k \in \mathcal{A}_i^c} |u_k| + \sum_{k \in \mathcal{A}_i} u_k \operatorname{sign}(\gamma_k^{(0)} - \gamma_k^0).$$

Since $\sqrt{m^*}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^0) \leq O_p(1)$, that is, $\boldsymbol{\gamma}^{(0)} = \boldsymbol{\gamma}^0$, then the second term above also vanishes, therefore $D_{i,m^*}(\boldsymbol{u}, \boldsymbol{v})$ is minimized at $u_k = 0, k \in \mathcal{A}_i^c$. Note that $\boldsymbol{u} = \lambda_{m^*}(\boldsymbol{\beta}_i^* - \boldsymbol{\beta}_i^0)$ and thus $\operatorname{argmin}(Q_{i,m^*}) = \operatorname{argmin}(D_{i,m^*})$, the proof is hence completed.

In general, the regularity condition (A6) only guarantees that $\frac{1}{m^*}(\boldsymbol{X}_i^*)^T \boldsymbol{X}_i^*$ is positive definite, but not for $\frac{1}{m^*} \tilde{\boldsymbol{X}}_i^{*T} \tilde{\boldsymbol{X}}_i^*$ since there could be invariant population-shared covariates $\boldsymbol{Z}_i^*$ within the individual. However, the above argument still holds by taking a transformation $\bar{\boldsymbol{Z}}_i^* = \boldsymbol{Z}_i^* \boldsymbol{T}_i$ such that $\frac{1}{m^*}(\bar{\boldsymbol{Z}}_i^*)^T \bar{\boldsymbol{Z}}_i^*$ is positive definite. □

B Additional Numerical Studies and Algorithm Implementation

B.1 Subgroup Number Selection

In this simulation study, we first investigate the performance of the data-driven method discussed in Section 4 to select the number of shrinkage centers (subgroups). We compared the proposed method (MDSP) based on BIC-type criterion with a two-stage approach (OLSK) which employs the gap statistic [11] to choose the number of subgroups for the K-means algorithm based on the least squares estimators of individualized coefficients. The OLSK method is implemented by R package *cluster* (version 2.0.5) [8]. The number of bootstrap samples in calculating the gap statistic is set as 100.

We generate the data following (15) in Section 5.1 under various scenarios. Scenario 1 has only a noise individualized variable ($\beta_i = 0, i = 1, \ldots, N$), while Scenarios 2 and 3 have two ($\beta_i = 0, 1$) or three subgroups ($\beta_i = 0, 2, 5$) for one individualized predictor, respectively, and Scenario 4 assumes a model of two individualized predictors with two ($\beta_{i1} = 0, 2$) or three ($\beta_{i2} = 0, -2, 1$) subgroups, respectively. The subgroup size in each scenario is balanced.

Table 1 provides the mean estimated number of subgroups and proportion of selecting the correct number of subgroups based on 100 replications. Overall, the proposed method is able to select the correct number of subgroup with more than 85% probability over all scenarios with different sample sizes ($N = 60, 120$) and individual measurement sizes ($m = 5, 10, 20$). The chance of selecting the correct number of subgroups increases as the individual measurement size increases. In addition, the proposed method consistently outperforms the two-stage OLSK method, especially when the individual measurement size is small ($m = 5$).

Table 1: The mean of identified subgroup numbers of the proposed model compared with the two-stage OLSK method based on 100 simulations, with sample size $N = 60, 120$, individual measurement size $m = 5, 10, 20$. The first three scenarios contain one individualized predictor ($p = 1$) of one, two and three groups, respectively. The last scenario contains two individualized predictors ($p = 2$), one with two groups and the other with three groups. The subgroup sizes are equal in each scenario. The subgroup homogeneous effects are listed as possible values for β_i in the table.

Number of individualized variables		$p = 1$						$p = 2$			
Sample	Cluster	$\beta_i = 0$		$\beta_i = 0, 1$		$\beta_i = 0, 2, 5$		$\beta_{1i} = 0, 2$		$\beta_{2i} = -2, 0, 1$	
Size (N)	Size(m)	MDSP	OLSK	MDSP	OLSK	MDSP	OLSK	MDSP	OLSK	MDSP	OLSK
	5	1.0(100)	1.0(100)	2.0(95)	1.0(2)	2.9(88)	2.5(68)	2.0(100)	1.5(52)	3.2(85)	1.2(0)
60	10	1.0(100)	1.0(100)	2.0(100)	1.3(26)	3.1(90)	2.7(74)	2.0(100)	2.0(100)	3.1(90)	2.4(44)
	20	1.0(100)	1.0(100)	2.0(100)	2.0(100)	3.1(92)	2.8(78)	2.0(100)	2.0(100)	3.0(100)	2.8(80)
	5	1.0(100)	1.0(100)	2.0(96)	1.0(2)	3.2(86)	2.8(82)	2.0(100)	1.7(72)	3.1(90)	1.4(0)
120	10	1.0(100)	1.0(100)	2.0(100)	1.2(24)	3.1(92)	2.9(86)	2.0(100)	2.0(100)	3.1(90)	2.6(64)
	20	1.0(100)	1.0(100)	2.0(100)	2.0(100)	3.0(98)	2.9(96)	2.0(100)	2.0(100)	3.1(92)	2.78(78)

B.2 ACTG Data Analysis

In this section, we illustrate the proposed individualized variable selection method using the Harvard longitudinal AIDS clinical trial group (ACTG) data. One of the goals from this study is to test the treatment effect of Zidovudine on CD4 cell counts, e.g., [2].

The 140 patients from this study are repeated measured over 14 time points with a missing rate of 8.5% and maintain CD4 counts above 50 at the baseline measures. The demographic information includes age and gender for each patient. We denote ZDV=1 if the patient receives the treatment and ZDV=0 if the patient is in the control group. Let y_{it} be the CD4 counts for the ith patient at time t. Each individuals' CD4 measurements are standardized by within-individual standard deviation to achieve a uniform scale. A marginal model to incorporate time, treatment, interaction of time and treatment, age and gender is provided as follows:

$$y_{it} = \beta_0 + \beta_t * Time + \beta_z * ZDV + \beta_{zt} * ZDV * Time + \beta_a * Age + \beta_g * Gender + \varepsilon_{it}. \tag{A.5}$$

We are particularly interested in the treatment effect of Zidovudine over time. The standard analysis concludes that the marginal treatment effect over time $\hat{\beta}_{zt}$ is not significant with p-value$= 0.113$.

However, if we examine the time trend of CD4 counts from individuals, there exist subgroups for the treatment group. Given the treatment ZDV, some individuals' CD4 counts are more stable over time while some patients' CD4 counts decrease more rapidly than the average of the control group over time. This could be interpreted that some patients respond more positively, while some respond more negatively, and the remaining patients have no effects from receiving ZDV treatment compared to the average effect of the control group.

Clearly, the subgroup differences are washed out if we apply the above marginal model in (A.5). Therefore, we employ an individualized regression model which accommodates the personalized treatment effects ZDV over time as the following:

$$y_{it} = \beta_0 + \beta_t * Time + \beta_z * ZDV + \beta_{izt} * ZDV * Time + \beta_a * Age + \beta_g * Gender + \varepsilon_{it}.$$

We assume for the β_{izt} coefficient, that it falls into three subgroups ($\beta_{izt} = \gamma^+ > 0$, $\beta_{izt} = \gamma^- < 0$ or $\beta_{izt} = 0$). Note that for patients in the control group, we set $\beta_{izt} = 0$ since their personalized effects corresponding to the treatment are unobserved. Since the treatment variable is constant over time, we compare our proposed method with the individual-wise Lasso model, the standard population homogeneous model, the random-effects model assuming a random slope of ZDV and time interaction and the fused Lasso model.

We choose observations at times $t = 1, \ldots, 12$ as the training set and the remaining observations at $t = 13, 14$ as the testing set. On the testing set, we calculate the root mean square prediction error for each individual at $t = 13, 14$, where the median of the individuals' prediction errors is reported. Table 2 shows that the proposed method has the smallest median prediction error among all methods. For example, the proposed method has 16.0%, 13.9% and 18.1% improvement in prediction accuracy compared to the marginal model, the random-effects model and the Lasso model, respectively.

Furthermore, Figure 1 shows the individuals corresponding to no effect, positive effect and negative effect in the treatment group identified by the Lasso method and the proposed method respectively. The proposed method is able to detect more individuals with significant responses to the treatment than the Lasso method does, as the proposed separation penalty enables us to shrink the estimated coefficients in multiple directions.

To examine whether subgrouping provides more informative treatment effect over time, we refit a marginal regression model in (A.5) for each subgroup, where each subgroup consists of the corresponding individuals identified in the treatment group and all individuals in the control group. Table 3 illustrates that the treatment effect over time from the positive-effect subgroup selected by the Lasso method is still not significant, while the negative-effect subgroup is significant with p-value of 0.02. In contrast, the proposed method identifies both positive and negative subgroups with significant p-values of 0.02 and 0.00 respectively.

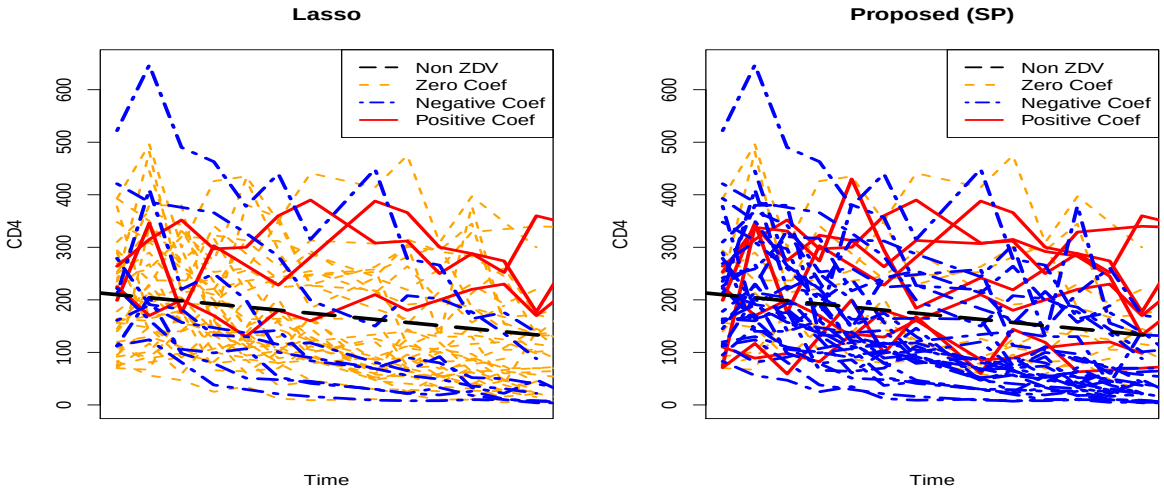

Figure 1: The different individuals corresponding to no effect, positive effect and negative effect in the treatment group selected by the Lasso model and the proposed method.

Table 2: The estimated coefficients of the population model, the random-effects model, the L_1-penalty model and the proposed model with corresponding median prediction errors (MPE) for the ACTG data. The individualized coefficient estimators $\hat{\beta}_{izt}$'s in the Lasso model, the fused Lasso (fusedL) model and the proposed (MDSP) model are not listed.

Model	$\hat{\beta}_0$	$\hat{\beta}_t$	$\hat{\beta}_z$	$\hat{\beta}_a$	$\hat{\beta}_g$	$\hat{\beta}_{zt}$	$\hat{\gamma}^+$	$\hat{\gamma}^-$	MPE
Population	3.09	-0.68	-0.54	0.01	-0.01	-0.24	-	-	1.67
Random-effects	2.56	-0.68	-0.57	0.02	-0.01	-0.29	-	-	1.70
Lasso	3.09	-0.76	-0.54	0.01	-0.01	-	-	-	1.64
fusedL	3.05	-0.72	-0.52	0.01	-0.01	-	-	-	1.62
MDSP	3.10	-0.68	-0.56	0.01	-0.01	-	0.62	-0.60	**1.44**

Table 3: The treatment effect estimators within each subgroup model (zero-effect group: β^0_{zt}, negative-effect group: β^-_{zt} and positive-effect group β^+_{zt}) as well as the standard errors (s.e.) and the p-values. Each subgroup consists of the corresponding individuals in the treatment group identified by the Lasso model or the proposed model (MDSP) as well as all the individuals in the control group. The proportion of individuals with the treatment classified into each subgroup is provided.

Model		Estimates	s.e.	p-value	Proportion
Lasso	$\hat{\beta}^0_{zt}$	-0.24	0.17	0.14	0.75
	$\hat{\beta}^-_{zt}$	-0.73	0.31	0.02	0.18
	$\hat{\beta}^+_{zt}$	0.82	0.48	0.10	0.07
MDSP	$\hat{\beta}^0_{zt}$	-0.04	0.30	0.89	0.20
	$\hat{\beta}^-_{zt}$	-0.68	0.08	**0.00**	0.64
	$\hat{\beta}^+_{zt}$	0.72	0.33	**0.02**	0.16

B.3 Supplementary Results to Simulation Study in Section 5.1

This section collects the supplementary numerical results to the simulation study in Section 5.1. Specifically, Table 4 summarizes the individualized variable selection results; Table 5 presents the estimated sub-homogeneous effects from the MDSP model; Figures 2 and 3 provide the boxplots of the variable selection evaluations with a sample size $N = 40$.

Table 4: The average correct variable selection rate (CVSR), sensitivity and specificity of the proposed MDSP model compared with other approaches based on 100 simulations, with sample size $N = 40, 100$, individual measurement size $m = 10, 20$, and subgroup homogeneous effect $\gamma = 1, 2$, where Sub, Homo, FusedL, Lasso, AdapL, SCAD and MCP stand for individual-wise model, homogeneous model, the fused Lasso, the Lasso, the adaptive Lasso, the SCAD and the MCP regularization models, respectively. The number of subgroups (two) is correctly specified in the proposed model.

Variable Selection	Sample Size (N)	Cluster Size(m)	Methods					
			MDSP	FusedL	Lasso	AdapL	SCAD	MCP
					$\gamma = 1$			
CVSR	40	10	0.916	0.692	0.876	0.820	0.717	0.741
		20	0.970	0.678	0.924	0.869	0.778	0.829
	100	10	0.909	0.673	0.862	0.840	0.718	0.754
		20	0.963	0.682	0.890	0.888	0.773	0.833
Sensitivity	40	10	0.942	0.978	0.898	0.943	0.975	0.966
		20	0.985	1.000	0.990	0.997	0.999	0.999
	100	10	0.946	0.986	0.917	0.941	0.974	0.967
		20	0.990	0.999	0.993	0.994	0.999	0.997
Specificity	40	10	0.909	0.406	0.853	0.696	0.460	0.517
		20	0.956	0.356	0.857	0.742	0.557	0.659
	100	10	0.886	0.360	0.807	0.739	0.462	0.542
		20	0.942	0.364	0.787	0.782	0.547	0.669
					$\gamma = 2$			
CVSR	40	10	0.959	0.639	0.886	0.884	0.800	0.852
		20	0.972	0.670	0.928	0.940	0.908	0.953
	100	10	0.940	0.648	0.868	0.898	0.809	0.871
		20	0.965	0.682	0.890	0.888	0.773	0.832
Sensitivity	40	10	0.997	0.996	0.997	0.998	1.000	0.998
		20	1.000	1.000	1.000	1.000	1.000	1.000
	100	10	0.998	0.997	0.998	0.998	0.999	0.999
		20	1.000	0.999	0.993	0.994	0.999	0.997
Specificity	40	10	0.922	0.282	0.774	0.771	0.602	0.705
		20	0.945	0.340	0.856	0.880	0.816	0.906
	100	10	0.882	0.299	0.738	0.797	0.620	0.744
		20	0.930	0.365	0.787	0.782	0.546	0.668

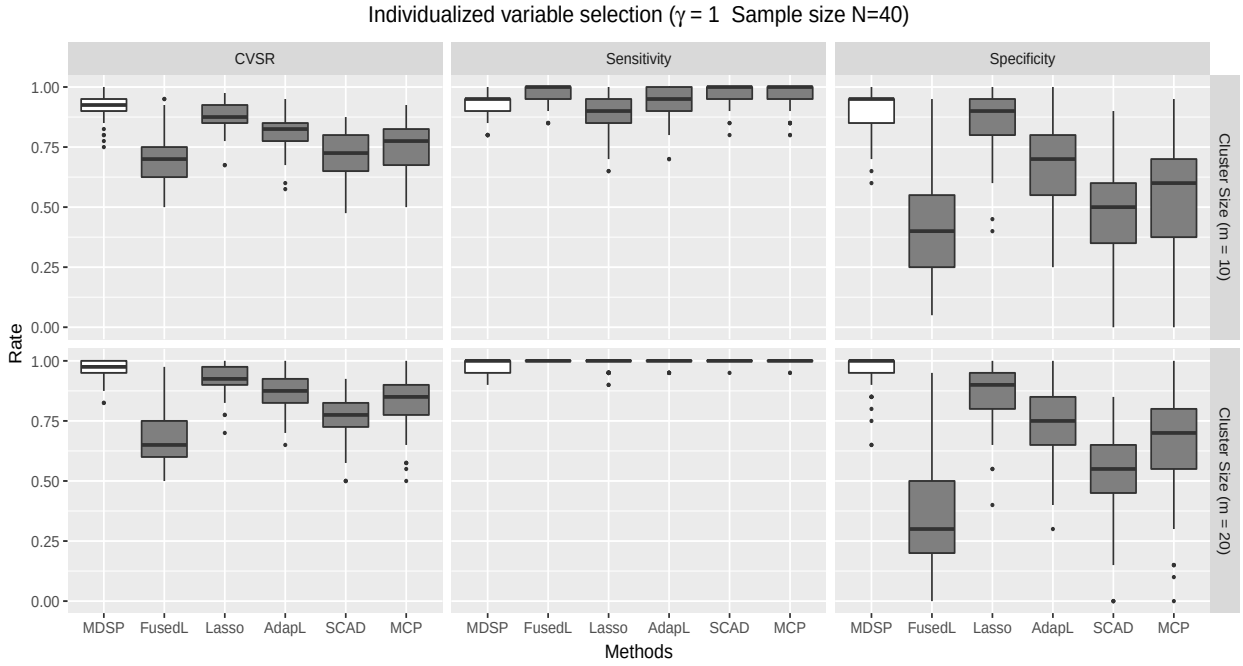

Figure 2: The boxplots of CVSR, sensitivity and specificity for all regularization approaches based on 100 simulations, with individual measurement size $m = 10, 20$, where homogeneous effect $\gamma = 1$ and sample size $N = 40$.

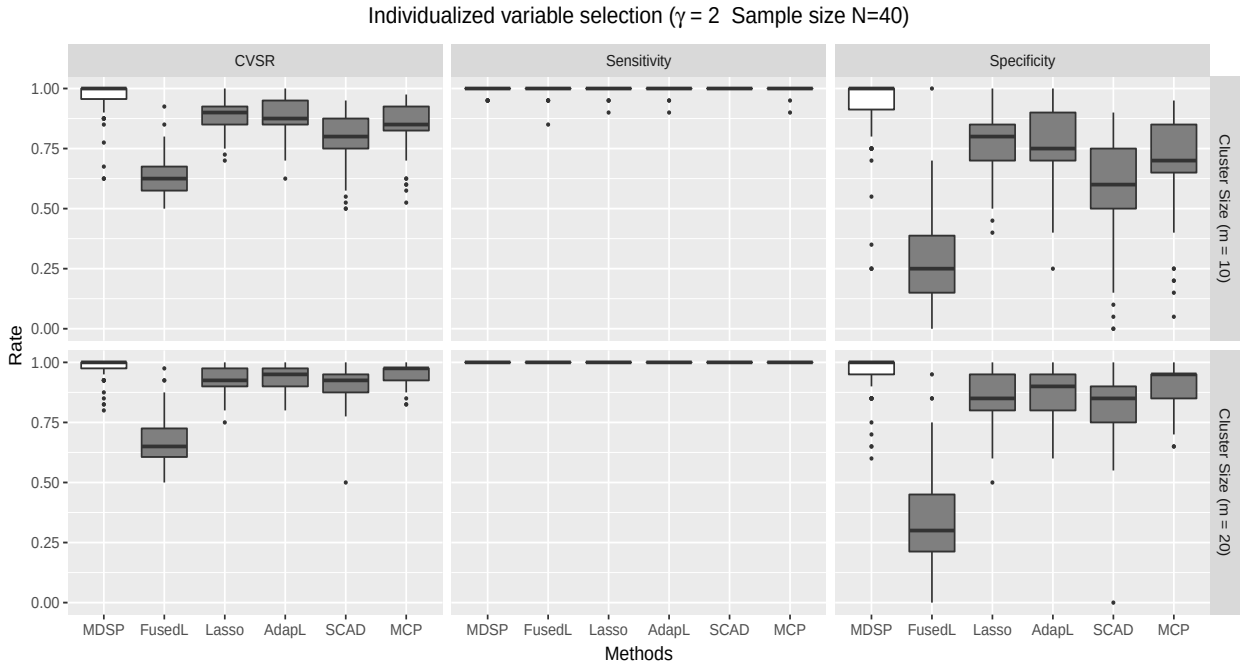

Figure 3: The boxplots of CVSR, sensitivity and specificity for all regularization approaches based on 100 simulations, with individual measurement size $m = 10, 20$, where homogeneous effect $\gamma = 2$ and sample size $N = 40$.

Table 5: The average RMSE of the estimated subgroup homogeneous effect $\hat{\gamma}$ from the proposed model based on 100 simulations (empirical standard errors in parenthesis), with sample size $N = 40, 100$, individual measurement size $m = 10, 20$.

Homogeneous	N=40		N=100	
Effect	$T = 10$	$T = 20$	$T = 10$	$T = 20$
$\gamma = 1$	1.03(0.08)	1.00(0.05)	1.02(0.05)	1.00(0.03)
$\gamma = 2$	2.01(0.07)	2.00(0.05)	2.00(0.05)	2.00(0.03)

B.4 ADMM Algorithm Implementation

In this section, we provide some implementation details for the proposed ADMM algorithm in Section 4.1 with an independent model. In the proposed algorithm, we update $\{\boldsymbol{\alpha}, \boldsymbol{\beta}\}$, $\{\boldsymbol{\nu}, \boldsymbol{\gamma}\}$ and $\boldsymbol{\Lambda}$ alternately at the $(l+1)$th iteration as follows:

$$\{\boldsymbol{\alpha}^{(l+1)}, \boldsymbol{\beta}^{(l+1)}\} = \underset{\boldsymbol{\alpha},\boldsymbol{\beta}}{\operatorname{argmin}} L_{N,m}(\boldsymbol{\alpha}, \boldsymbol{\beta}) + \frac{\kappa}{2}\|\boldsymbol{\beta} - \boldsymbol{\nu}^{(l)} + \kappa^{-1}\boldsymbol{\Lambda}^{(l)}\|_2^2, \tag{A.6}$$

$$\{\boldsymbol{\nu}^{(l+1)}, \boldsymbol{\gamma}^{(l+1)}\} = \underset{\boldsymbol{\nu},\boldsymbol{\gamma}}{\operatorname{argmin}} S_{\lambda_{N,m}}(\boldsymbol{\nu}, \boldsymbol{\gamma}) + \frac{\kappa}{2}\|\boldsymbol{\beta}^{(l+1)} - \boldsymbol{\nu} + \kappa^{-1}\boldsymbol{\Lambda}^{(l)}\|_2^2, \tag{A.7}$$

$$\boldsymbol{\Lambda}^{(l+1)} = \boldsymbol{\Lambda}^{(l)} + \kappa(\boldsymbol{\beta}^{(l+1)} - \boldsymbol{\nu}^{(l+1)}).$$

The optimization in (A.6) has an explicit solution as

$$\{\boldsymbol{\alpha}^{(l+1)}, \boldsymbol{\beta}^{(l+1)}\} = \left(\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}} + \text{bdiag}\{\mathbf{0}_{q\times q}, \kappa\boldsymbol{I}_{Np+q}\}\right)^{-1}\left(\tilde{\boldsymbol{X}}^T\boldsymbol{Y} + \text{vec}\{\mathbf{0}_{q\times 1}, \kappa\boldsymbol{\nu}^{(l)} - \lambda_{N,m}\boldsymbol{\Lambda}^{(l)}\}\right).$$

The optimization in (A.7) is achieved by iteratively updating $\boldsymbol{\nu}$ and $\boldsymbol{\gamma}$ in

$$\underset{\boldsymbol{\nu}_{\cdot j}}{\operatorname{argmin}} \sum_{i=1}^{N}\left\{\frac{\kappa}{2}(\nu_{ij} - \beta_{ij}^{(l+1)} - \kappa^{-1}\Lambda_{ij}^{(l)})^2 + \lambda_{N,m}\min(|\nu_{ij}|, |\nu_{ij} - \gamma_j|)\right\}. \tag{A.8}$$

Given $\boldsymbol{\gamma}$, let $\tilde{\beta}_{ik}^{(l+1)} = \beta_{ik}^{(l+1)} + \kappa^{-1}\Lambda_{ik}^{(l)}$, the updates of $\boldsymbol{\nu}$ are obtained as

$$\nu_{ik}^{(new)} = \begin{cases} \text{sign}(\tilde{\beta}_{ik}^{(l+1)}) \cdot \max(0, |\tilde{\beta}_{ik}^{(l+1)}| - \frac{\lambda_{N,m}}{\kappa}), & \text{if } |\tilde{\beta}_{ik}^{(l+1)}| \le |\tilde{\beta}_{ik}^{(l+1)} - \gamma_k^{(old)}| \\ \gamma_k^{(old)} + \text{sign}(\tilde{\beta}_{ik}^{(l+1)} - \gamma_k^{(old)}) \cdot \max(0, |\tilde{\beta}_{ik}^{(l+1)} - \gamma_k^{(old)}| - \frac{\lambda_{N,m}}{\kappa}), & \text{if } |\tilde{\beta}_{ik}^{(l+1)}| > |\tilde{\beta}_{ik}^{(l+1)} - \gamma_k^{(old)}| \end{cases},$$

for $k = 1, \ldots, p$, $i = 1, \ldots, N$. And given $\boldsymbol{\nu}$, the $\boldsymbol{\gamma}$ is estimated via a one-dimensional exhaustive search along each covariate for $k = 1, \ldots, p$.

References

[1] Bickel, P. J. and Levina, E. (2008). Regularized estimation of large covariance matrices. *The Annals of Statistics 36*, 199-227.

[2] Dolin, R., Amato, D. A., Fischl, M. A. et al. (1995). Zidovudine compared with Didanosine in patients with advanced HIV type 1 infection and little or no previous experience with Zidovudine. *Archives of Internal Medicine 155*, 961-74.

[3] Geyer, C. J. (1996). On the asymptotics of convex stochastic optimization. Unpublished manuscript.

[4] Han, F. and Liu, H. (2017). Statistical analysis of latent generalized correlation matrix estimation in transelliptical distribution. *Bernoulli 23(1)*, 23-57.

[5] Jung, S. and Marron, J. S. (2009). PCA consistency in high dimension, low sample size context. *The Annals of Statistics 37*, 4104-4130.

[6] Knight, K. and Fu, W. (2000). Asymptotics for Lasso-type estimators. *The Annals of Statistics 28*, 1356-1378.

[7] Ma, S., and Huang, J. (2016) A concave pairwise fusion approach to subgroup analysis. *Journal of the American Statistical Association 112(517)*, 410-432.

[8] Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K., Studer, M., Roudier, P. and Gonzalez, J. (2016). Cluster: Finding groups in data. R package version 2.0.5.

[9] Pan, W., Shen, X. and Liu, B. (2013). Cluster analysis: Unsupervised learning via supervised learning with a non-convex penalty. *Journal of Machine Learning Research 14*, 1865-1889.

[10] Pollard, D. (1991). Asymptotics for least absolute deviation regression estimators. *Econometric Theory 7*, 186-199.

[11] Tibshirani, R., Walther, G. and Hastie, T. (2001). Estimating the number of data clusters via the gap statistic. *Journal of the Royal Statistical Society: Ser. B 63*, 411-423.