Supplement to "Individual Data Protected Integrative Regression Analysis of High-dimensional Heterogeneous Data"

Tianxi Cai, Molei Liu, and Yin Xia

In the supplement, we first provide justifications for the random (sub-gaussian) design Compatibility Condition and introduce and verify the Irrepresentable Condition for some common correlation structures. Then the detailed proofs of Theorems 1–3 and the rate property of $(\widehat{\boldsymbol{\mu}}_{\mathsf{L\&B}}, \widehat{\boldsymbol{\alpha}}_{\mathsf{L\&B}})$ are presented. Finally we outline theoretical analyses of SHIR for various penalty functions and present additional numerical results.

A.1. JUSTIFICATION OF THE COMPATIBILITY CONDITION

We provide justification for Proposition 1 in this section.

Proof. First, we show that for any $\boldsymbol{\beta}^{(\mathsf{m})}$ satisfying $\|\boldsymbol{\beta}^{(\mathsf{m})} - \boldsymbol{\beta}_0^{(\mathsf{m})}\|_2 = o(1)$,

$$(2C_x)^{-1} \leq \Lambda_{\min}\{\bar{\mathbb{H}}_m(\boldsymbol{\beta}^{(\mathsf{m})})\} \leq \Lambda_{\max}\{\bar{\mathbb{H}}_m(\boldsymbol{\beta}^{(\mathsf{m})})\} \leq 2C_x. \tag{S1}$$

By $\max_{\mathbf{x}\in\mathscr{B}_1(\mathbf{0})} \mathsf{E}[\mathbf{x}^{\mathsf{T}}\mathbf{X}_i^{(\mathsf{m})}]^4 \leq C_x$, for any $\mathbf{x} \in \mathscr{B}_1(\mathbf{0})$ and $\boldsymbol{\beta}^{(\mathsf{m})}$ satisfying $\|\boldsymbol{\beta}^{(\mathsf{m})} - \boldsymbol{\beta}_0^{(\mathsf{m})}\|_2 = o(1)$,

$$\begin{aligned}
&\left|\mathbf{x}^{\mathsf{T}}\bar{\mathbb{H}}_m(\boldsymbol{\beta}_0^{(\mathsf{m})})\mathbf{x} - \mathbf{x}^{\mathsf{T}}\bar{\mathbb{H}}_m(\boldsymbol{\beta}^{(\mathsf{m})})\mathbf{x}\right| \\
=&\left|\mathsf{E}(\mathbf{x}^{\mathsf{T}}\mathbf{X}_i^{(\mathsf{m})})^2\left\{f_1''(\mathbf{X}_i^{(\mathsf{m})\mathsf{T}}\boldsymbol{\beta}_0^{(\mathsf{m})}, Y_i^{(\mathsf{m})}) - f_1''(\mathbf{X}_i^{(\mathsf{m})\mathsf{T}}\boldsymbol{\beta}^{(\mathsf{m})}, Y_i^{(\mathsf{m})})\right\}\right| \\
\leq&\mathsf{E}\left[(\mathbf{x}^{\mathsf{T}}\mathbf{X}_i^{(\mathsf{m})})^2 C_L|\mathbf{X}_i^{(\mathsf{m})\mathsf{T}}(\boldsymbol{\beta}_0^{(\mathsf{m})} - \boldsymbol{\beta}^{(\mathsf{m})})|\right] \leq C_L\left(\mathsf{E}[\mathbf{x}^{\mathsf{T}}\mathbf{X}_i^{(\mathsf{m})}]^4\mathsf{E}[\mathbf{X}_i^{(\mathsf{m})\mathsf{T}}(\boldsymbol{\beta}_0^{(\mathsf{m})} - \boldsymbol{\beta}^{(\mathsf{m})})]^2\right)^{1/2} \\
\leq&C_L\left(\mathsf{E}[\mathbf{x}^{\mathsf{T}}\mathbf{X}_i^{(\mathsf{m})}]^4\max_{\boldsymbol{v}\in\mathscr{B}_1(\mathbf{0})}\mathsf{E}[\boldsymbol{v}^{\mathsf{T}}\mathbf{X}_i^{(\mathsf{m})}]^2\|\boldsymbol{\beta}_0^{(\mathsf{m})} - \boldsymbol{\beta}^{(\mathsf{m})}\|_2^2\right)^{1/2} \leq C_xC_L\|\boldsymbol{\beta}_0^{(\mathsf{m})} - \boldsymbol{\beta}^{(\mathsf{m})}\|_2 = o(1).
\end{aligned}$$

So by $C_x^{-1} \leq \Lambda_{\min}(\bar{\mathbb{H}}_m) \leq \Lambda_{\max}(\bar{\mathbb{H}}_m) \leq C_x$, equation (S1) holds. For any $\delta_1 = \Theta\{(s_0M\log p/N)^{1/2}\}$

and $\boldsymbol{\beta}^{(\bullet)} = (\boldsymbol{\beta}^{(1)\mathsf{T}}, \ldots, \boldsymbol{\beta}^{(\mathsf{M})\mathsf{T}})^{\mathsf{T}}$ satisfying $\boldsymbol{\beta}^{(\mathsf{m})} \in \mathscr{B}_{\delta_1}(\boldsymbol{\beta}_0^{(\mathsf{m})})$, since $s_0 = o\{N/(M \log p)\}$, we have $\|\boldsymbol{\beta}^{(\mathsf{m})} - \boldsymbol{\beta}_0^{(\mathsf{m})}\|_2 = o(1)$ and thus $(2C_x)^{-1} \le \Lambda_{\min}\{\bar{\mathbb{H}}_m(\boldsymbol{\beta}^{(\mathsf{m})})\} \le \Lambda_{\max}\{\bar{\mathbb{H}}_m(\boldsymbol{\beta}^{(\mathsf{m})})\} \le 2C_x$ for all $m \in [M]$. Let $\widetilde{\mathbf{X}}_i^{(\mathsf{m})} = \mathbf{X}_i^{(\mathsf{m})}\{f_1''(\boldsymbol{\beta}^{(\mathsf{m})\mathsf{T}}\mathbf{X}_i, Y_i^{(\mathsf{m})})\}^{1/2}$, and by the assumption in Proposition 1, we have that $\|\widetilde{\mathbf{X}}_i^{(\mathsf{m})}\|_{\psi_2} \le \kappa_x$.

Now we follow similar procedures as the proof of Theorem 1.6 in Rudelson and Zhou (2012) to show that, for the mixture penalty in our case, $\mathbb{H}(\boldsymbol{\beta}^{(\bullet)})$ satisfies $\mathscr{C}_{\mathsf{comp}}$ with probability approaching 1. We first define the complexity measure of any set $\mathcal{V} \subseteq \mathscr{B}_1(\mathbf{0})$ as follow.

Definition A1. *For any $\mathcal{V} \subseteq \mathscr{B}_1(\mathbf{0})$, define $c_d(\mathcal{V}) = \mathsf{E}\sup_{\boldsymbol{v}\in\mathcal{V}} |\boldsymbol{g}^{\mathsf{T}}\boldsymbol{v}|$, where $\boldsymbol{g} = (g_1, g_2, \ldots, g_d)^{\mathsf{T}}$ and $g_1, g_2, \ldots, g_d$ are independent $\mathrm{N}(0,1)$ variables.*

We recall that

$$\mathcal{C}(t, \mathcal{S}) = \Big\{(\boldsymbol{u}^{\mathsf{T}}, \boldsymbol{v}^{(\bullet)\mathsf{T}})^{\mathsf{T}} = (\boldsymbol{u}^{\mathsf{T}}, \boldsymbol{v}^{(1)\mathsf{T}}, \ldots, \boldsymbol{v}^{(\mathsf{M})\mathsf{T}})^{\mathsf{T}} : \boldsymbol{v}^{(1)} + \cdots + \boldsymbol{v}^{(\mathsf{M})} = \mathbf{0},$$

$$\|\boldsymbol{u}_{\mathcal{S}^c}\|_1 + \lambda_g \|\boldsymbol{v}_{\mathcal{S}^c}^{(\bullet)}\|_{2,1} \le t(\|\boldsymbol{u}_{\mathcal{S}}\|_1 + \lambda_g \|\boldsymbol{v}_{\mathcal{S}}^{(\bullet)}\|_{2,1})\Big\},$$

as introduced in Definition 1. Denote by

$$\widetilde{\mathscr{B}}_1 = \left\{(\boldsymbol{u}^{\mathsf{T}}, \boldsymbol{v}^{(\bullet)\mathsf{T}})^{\mathsf{T}} = (\boldsymbol{u}^{\mathsf{T}}, \boldsymbol{v}^{(1)\mathsf{T}}, \ldots, \boldsymbol{v}^{(\mathsf{M})\mathsf{T}})^{\mathsf{T}} : \|\boldsymbol{u}\|_2^2 + \lambda_g^2 \|\boldsymbol{v}^{(\bullet)}\|_2^2 = 1\right\},$$

$\bar{\mathcal{C}}_t = \mathcal{C}(t, \mathcal{S}_0) \cap \widetilde{\mathscr{B}}_1$, and define that

$$\Gamma_t = \left\{\frac{1}{N^{1/2}}\left[n_1^{1/2}(\boldsymbol{\mu}_\Delta + \boldsymbol{\alpha}_\Delta^{(1)})^{\mathsf{T}}\bar{\mathbb{H}}_1^{1/2}(\boldsymbol{\beta}^{(1)}), \ldots, n_M^{1/2}(\boldsymbol{\mu}_\Delta + \boldsymbol{\alpha}_\Delta^{(\mathsf{M})})^{\mathsf{T}}\bar{\mathbb{H}}_M^{1/2}(\boldsymbol{\beta}^{(\mathsf{M})})\right]^{\mathsf{T}} : (\boldsymbol{\mu}_\Delta^{\mathsf{T}}, \boldsymbol{\alpha}_\Delta^{(1)\mathsf{T}}, \ldots, \boldsymbol{\alpha}_\Delta^{(\mathsf{M})\mathsf{T}})^{\mathsf{T}} \in \bar{\mathcal{C}}_t\right\},$$

which is a subset of $\mathbb{R}^{Mp}$. We now provides bound for $c_{Mp}(\Gamma_t)$, the complexity measure of Γ_t. Let $\boldsymbol{g}^{(\bullet)} = (\boldsymbol{g}^{(1)\mathsf{T}}, \boldsymbol{g}^{(2)\mathsf{T}}, \ldots, \boldsymbol{g}^{(\mathsf{M})\mathsf{T}})^{\mathsf{T}}$ where $\boldsymbol{g}^{(\mathsf{m})} = (g_1^{(\mathsf{m})}, g_2^{(\mathsf{m})}, \ldots, g_p^{(\mathsf{m})})^{\mathsf{T}}$ are independent gaussian vectors and $g_1^{(\mathsf{m})}, \dots, g_p^{(\mathsf{m})} \sim \mathrm{N}(0,1)$ are independent. We have

$$\begin{aligned}
c_{Mp}(\Gamma_t) \le & \mathsf{E}\sup\left\{ \frac{1}{N^{1/2}} \sum_{m=1}^{M} n_m^{1/2} (\boldsymbol{\mu}_\Delta + \boldsymbol{\alpha}_\Delta^{(\mathsf{m})})^\mathsf{T} \bar{\mathbb{H}}_m^{1/2}(\boldsymbol{\beta}^{(\mathsf{m})}) \boldsymbol{g}^{(\mathsf{m})} : (\boldsymbol{\mu}_\Delta^\mathsf{T}, \boldsymbol{\alpha}_\Delta^{(1)\mathsf{T}}, \dots, \boldsymbol{\alpha}_\Delta^{(\mathsf{M})\mathsf{T}})^\mathsf{T} \in \bar{\mathcal{C}}_t \right\} \\
\le & \mathsf{E}\sup\left\{ \|\boldsymbol{\mu}_\Delta\|_1 \left\| \frac{1}{N^{1/2}} \sum_{m=1}^{M} n_m^{1/2} \bar{\mathbb{H}}_m^{1/2}(\boldsymbol{\beta}^{(\mathsf{m})}) \boldsymbol{g}^{(\mathsf{m})} \right\|_\infty : (\boldsymbol{\mu}_\Delta^\mathsf{T}, \boldsymbol{\alpha}_\Delta^{(\bullet)\mathsf{T}})^\mathsf{T} \in \bar{\mathcal{C}}_t \right\} \\
& + \mathsf{E}\sup\left\{ \|\boldsymbol{\alpha}_\Delta^{(\bullet)}\|_{2,1} \left\| \frac{1}{N^{1/2}} \left[n_1^{1/2} \boldsymbol{g}^{(1)\mathsf{T}} \bar{\mathbb{H}}_1^{1/2}(\boldsymbol{\beta}^{(1)}), \dots, n_M^{1/2} \boldsymbol{g}^{(\mathsf{M})\mathsf{T}} \bar{\mathbb{H}}_M^{1/2}(\boldsymbol{\beta}^{(\mathsf{M})}) \right]^\mathsf{T} \right\|_{2,\infty} : (\boldsymbol{\mu}_\Delta^\mathsf{T}, \boldsymbol{\alpha}_\Delta^{(\bullet)\mathsf{T}})^\mathsf{T} \in \bar{\mathcal{C}}_t \right\},
\end{aligned}$$

where the $\|\cdot\|_{2,\infty}$ norm is defined as

$$\left\| \frac{1}{N^{1/2}} \left[n_1^{1/2} \boldsymbol{g}^{(1)\mathsf{T}} \bar{\mathbb{H}}_1^{1/2}(\boldsymbol{\beta}^{(1)}), \dots, n_M^{1/2} \boldsymbol{g}^{(\mathsf{M})\mathsf{T}} \bar{\mathbb{H}}_M^{1/2}(\boldsymbol{\beta}^{(\mathsf{M})}) \right]^\mathsf{T} \right\|_{2,\infty} = \max_{j\in[p]} \sqrt{ \frac{1}{N} \sum_{m=1}^{M} n_m \left[\bar{\mathbb{H}}_1^{1/2}(\boldsymbol{\beta}^{(\mathsf{m})}) \boldsymbol{g}^{(\mathsf{m})} \right]_j^2 }.$$

By $n_m = \Theta(N/M)$, $\Lambda_{\max}\{\bar{\mathbb{H}}_M^{1/2}(\boldsymbol{\beta}^{(\mathsf{M})})\} \le 2C_x$ for all $m \in [M]$ and that $\boldsymbol{g}^{(\bullet)}$ is gaussian, and similar to the derivation below the proof of Lemma A1, we can show there exists an absolute constant $C_g > 0$ such that

$$\begin{aligned}
& \mathsf{E}\left\| \frac{1}{N^{1/2}} \sum_{m=1}^{M} n_m^{1/2} \bar{\mathbb{H}}_m^{1/2}(\boldsymbol{\beta}^{(\mathsf{m})}) \boldsymbol{g}^{(\mathsf{m})} \right\|_\infty \le C_g \sqrt{\log p}; \\
& \mathsf{E}\left\| \frac{1}{N^{1/2}} \left[n_1^{1/2} \boldsymbol{g}^{(1)\mathsf{T}} \bar{\mathbb{H}}_1^{1/2}(\boldsymbol{\beta}^{(1)}), \dots, n_M^{1/2} \boldsymbol{g}^{(\mathsf{M})\mathsf{T}} \bar{\mathbb{H}}_M^{1/2}(\boldsymbol{\beta}^{(\mathsf{M})}) \right]^\mathsf{T} \right\|_{2,\infty} \le C_g \sqrt{\frac{M + \log p}{M}},
\end{aligned}$$

through some calculation on the order statistics of gaussian or χ^2-type (quadratic form of gaussian) variables. These combined with $\lambda_g = \Theta(M^{-1/2})$ lead to that there exists absolute constant $C > 0$ such that

$$c_{Mp}(\Gamma_t) \le C\sqrt{\log p + M} \sup\left\{ \|\boldsymbol{\mu}_\Delta\|_1 + \lambda_g \|\boldsymbol{\alpha}_\Delta^{(\bullet)}\|_{2,1} : (\boldsymbol{\mu}_\Delta^\mathsf{T}, \boldsymbol{\alpha}_\Delta^{(\bullet)\mathsf{T}})^\mathsf{T} \in \bar{\mathcal{C}}_t \right\}. \tag{S2}$$

Following that $\bar{\mathcal{C}}_t = \mathcal{C}(t, \mathcal{S}_0) \cap \widetilde{\mathscr{B}}_1$, we have

$$\sup\left\{\|\boldsymbol{\mu}_\Delta\|_1 + \lambda_g\|\boldsymbol{\alpha}_\Delta^{(\bullet)}\|_{2,1} : (\boldsymbol{\mu}_\Delta^\mathsf{T}, \boldsymbol{\alpha}_\Delta^{(\bullet)\mathsf{T}})^\mathsf{T} \in \bar{\mathcal{C}}_t\right\}$$
$$\leq \sup\left\{(t+1)^2|\mathcal{S}_0|\left(\|\boldsymbol{\mu}_\Delta\|_2^2 + \lambda_g^2\|\boldsymbol{\alpha}_\Delta^{(\bullet)}\|_2^2\right) : (\boldsymbol{\mu}_\Delta^\mathsf{T}, \boldsymbol{\alpha}_\Delta^{(\bullet)\mathsf{T}})^\mathsf{T} \in \bar{\mathcal{C}}_t\right\} = (t+1)^2 s_0$$

So by (S2), we have $c_{Mp}(\Gamma_t) \leq C(t+1)\sqrt{s_0(\log p + M)}$. Now similar to Rivasplata (2012), we introduce the following theorem from Mendelson et al. (2007, 2008) (adapted to our notation and setting), as the foundation of our proof.

Theorem A1 (Mendelson et al. (2007, 2008))**.** *Recall that*

$$\mathbb{H}(\boldsymbol{\beta}^{(\bullet)}) = N^{-1}\mathsf{bdiag}\{n_1\mathbb{H}_1(\boldsymbol{\beta}^{(1)}), \ldots, n_M\mathbb{H}_M(\boldsymbol{\beta}^{(\mathsf{M})})\}$$

where $\mathbb{H}_m(\boldsymbol{\beta}^{(\mathsf{m})}) = n_m^{-1}\sum_{i=1}^{n_m}\widetilde{\mathbf{X}}_i^{(\mathsf{m})}\widetilde{\mathbf{X}}_i^{(\mathsf{m})\mathsf{T}}$*. If there exists constants* $\kappa_x > 0$ *and* $C' > 0$ *such that* $\|\widetilde{\mathbf{X}}_i^{(\mathsf{m})}\|_{\psi_2} \leq \kappa_x$ *and* $N > C'c_{Mp}^2(\Gamma_t)$*, then there exists a constant* $\phi_0 > 0$ *depending only on* κ_x *and* C'*, such that with probability approaching* 1*,* $\mathbb{H}(\boldsymbol{\beta}^{(\bullet)})$ *and* $\mathcal{S}_0$ *satisfy the Compatibility Condition* $\mathscr{C}_{\mathsf{comp}}$ *with the compatibility constant* $\phi_0\{t, \mathcal{S}_0, \mathbb{H}(\boldsymbol{\beta}^{(\bullet)})\} \geq \phi_0$*.*

Theorem A1 could be viewed as a special case of Corollary 2.7 and Theorem 2.1 in Mendelson et al. (2008) with the complexity measure and $\mathscr{C}_{\mathsf{comp}}$ specific to our case. Because we assume that $s_0 = o\{N/(M\log p)\} \leq o\{N/(M + \log p)\}$, and it has been shown that $c_{Mp}(\Gamma_t) \leq C(t+1)\sqrt{s_0(\log p + M)}$, we have $N > C'c_{Mp}^2(\Gamma_t)$ for any constant $C' > 0$ when N is large enough. Combining this with $\|\widetilde{\mathbf{X}}_i^{(\mathsf{m})}\|_{\psi_2} \leq \kappa_x$ and Theorem A1, Proposition 1 is proved.

□

A.2. THE IRREPRESENTABLE CONDITION AND ITS JUSTIFICATION

We first introduce the Irrepresentable Condition used in Condition 6. For any matrix $\mathbb{A} = [\mathbf{A}_1, \ldots, \mathbf{A}_d] \in \mathbb{R}^{n\times d}$ and index set $\mathcal{S}_1, \mathcal{S}_2 \subseteq [d]$, let $\mathbb{A}_{j\bullet}$ and $\mathbb{A}_{\bullet j}$ respectively denote the j^{th} row and column of $\mathbb{A}$, $\mathbb{A}_{\mathcal{S}_1\mathcal{S}_2}$ denote the submatrix corresponding to rows in $\mathcal{S}_1$ and columns in $\mathcal{S}_2$, $\mathbb{A}_{\bullet\mathcal{S}} = [\mathbb{A}_{\bullet j_1}, \ldots, \mathbb{A}_{\bullet j_k}]$ for $\mathcal{S} = \{j_1, \ldots, j_k : j_1 < \cdots < j_k\} \subseteq [d]$. The weighted design matrix corresponding to $\widehat{\mathcal{L}}_{\mathsf{SHIR}}(\boldsymbol{\mu}, \boldsymbol{\alpha}^{(\bullet)})$ with respect to $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\alpha}^{(2)\mathsf{T}}, \ldots, \boldsymbol{\alpha}^{(\mathsf{M})\mathsf{T}})^\mathsf{T}$ after setting $\boldsymbol{\alpha}^{(1)} = -\sum_{m=2}^M \boldsymbol{\alpha}^{(\mathsf{m})}$

can be expressed as

$$\mathbb{W}(\boldsymbol{\beta}^{(\bullet)}) = \mathsf{bdiag}\{\boldsymbol{\Omega}_1^{1/2}(\boldsymbol{\beta}^{(1)}), \dots, \boldsymbol{\Omega}_M^{1/2}(\boldsymbol{\beta}^{(\mathsf{M})})\}\mathbb{Z},$$

where "bdiag" is the block diagonal operator, $\boldsymbol{\Omega}_m(\boldsymbol{\beta}) = \mathsf{diag}\{f_1''(\boldsymbol{\beta}^{\mathsf{T}}\mathbf{X}_1^{(\mathsf{m})}, Y_1^{(\mathsf{m})}), \dots, f_1''(\boldsymbol{\beta}^{\mathsf{T}}\mathbf{X}_{n_m}^{(\mathsf{m})}, Y_{n_m}^{(\mathsf{m})})\}$ is a $n_m \times n_m$ dimensional matrix, $\mathbb{Z} = \mathbb{Z}_{[p],[p]}$, and for any $\mathcal{S}_1, \mathcal{S}_2 \subseteq [p]$,

$$\mathbb{Z}_{\mathcal{S}_1,\mathcal{S}_2} = \begin{pmatrix} \mathbb{X}_{\bullet\mathcal{S}_1}^{(1)} & -\mathbb{X}_{\bullet\mathcal{S}_2}^{(1)} & -\mathbb{X}_{\bullet\mathcal{S}_2}^{(1)} & \cdots & -\mathbb{X}_{\bullet\mathcal{S}_2}^{(1)} \\ \mathbb{X}_{\bullet\mathcal{S}_1}^{(2)} & \mathbb{X}_{\bullet\mathcal{S}_2}^{(2)} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbb{X}_{\bullet\mathcal{S}_1}^{(3)} & \mathbf{0} & \mathbb{X}_{\bullet\mathcal{S}_2}^{(3)} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbb{X}_{\bullet\mathcal{S}_1}^{(\mathsf{M})} & \mathbf{0} & \mathbf{0} & \cdots & \mathbb{X}_{\bullet\mathcal{S}_2}^{(\mathsf{M})} \end{pmatrix}.$$

For any $\mathcal{S}_1, \mathcal{S}_2 \subseteq [p]$, let $\mathbb{H}_{m,\mathcal{S}_1}(\boldsymbol{\beta}^{(\mathsf{m})})$ represent the sub-matrix of $\mathbb{H}_m(\boldsymbol{\beta}^{(\mathsf{m})}) := \nabla^2\widehat{\mathcal{L}}_m(\boldsymbol{\beta}^{(\mathsf{m})})$ with its rows and columns corresponding to $\mathcal{S}_1$, and $\mathbb{W}_{\mathcal{S}_1,\mathcal{S}_2}(\boldsymbol{\beta}^{(\bullet)})$ denote the sub-matrix of $\mathbb{W}(\boldsymbol{\beta}^{(\bullet)})$ corresponding to $\mathbb{Z}_{\mathcal{S}_1,\mathcal{S}_2}$ and $(\boldsymbol{\mu}_{\mathcal{S}_1}^{\mathsf{T}}, \boldsymbol{\alpha}_{\mathcal{S}_2}^{(2)\mathsf{T}}, \dots, \boldsymbol{\alpha}_{\mathcal{S}_2}^{(\mathsf{M})\mathsf{T}})^{\mathsf{T}}$. Let $\mathcal{S}_{\mathsf{full}} = \{\mathcal{S}_\mu, \mathcal{S}_\alpha\}$ and $\mathbb{W}_{\mathcal{S}_{\mathsf{full}}}(\boldsymbol{\beta}^{(\bullet)}) = \mathbb{W}_{\mathcal{S}_\mu,\mathcal{S}_\alpha}(\boldsymbol{\beta}^{(\bullet)})$. Also, denote by $\mathbb{T} = (\mathbf{1}_{(M-1)\times 1}, \mathbb{I}_{(M-1)\times(M-1)})^{\mathsf{T}}$ and define $\|\mathbf{x}\|_{\mathbb{T}} := \|\mathbb{T}\mathbf{x}\|_2$ for $\mathbf{x} \in \mathbb{R}^{M-1}$ and its conjugate norm as $\|\mathbf{x}\|_{\widetilde{\mathbb{T}}} := \|\mathbb{T}(\mathbb{T}^{\mathsf{T}}\mathbb{T})^{-1}\mathbf{x}\|_2$.

Definition A2. Irrepresentable Condition ($\mathscr{C}_{\mathsf{Irrep}}$): *The design matrix $\mathbb{W}(\boldsymbol{\beta}^{(\bullet)})$ satisfies the Irrepresentable Condition on $\mathcal{S}_{\mathsf{full}} = (\mathcal{S}_\mu, \mathcal{S}_\alpha)$ with parameter $\epsilon > 0$, if for all $j \in \mathcal{S}_\mu^c$ and $j' \in \mathcal{S}_\alpha^c$,*

$$\sup_{\boldsymbol{u}\in\mathscr{G}_{\mathcal{S}_\mu}, \boldsymbol{v}^{(\bullet)}\in\mathscr{G}_{\mathcal{S}_\alpha}} \left\{ \left| (\boldsymbol{u}^{\mathsf{T}}, \lambda_g \boldsymbol{v}^{(\bullet)\mathsf{T}}) \left[\mathbb{W}_{\mathcal{S}_{\mathsf{full}}}^{\mathsf{T}}(\boldsymbol{\beta}^{(\bullet)})\mathbb{W}_{\mathcal{S}_{\mathsf{full}}}(\boldsymbol{\beta}^{(\bullet)})\right]^{-1} \mathbb{W}_{\mathcal{S}_{\mathsf{full}}}^{\mathsf{T}}(\boldsymbol{\beta}^{(\bullet)})\mathbb{W}_{j,\emptyset}(\boldsymbol{\beta}^{(\bullet)}) \right| \right\} \leq 1 - \epsilon;$$

$$\sup_{\boldsymbol{u}\in\mathscr{G}_{\mathcal{S}_\mu}, \boldsymbol{v}^{(\bullet)}\in\mathscr{G}_{\mathcal{S}_\alpha}} \left\{ \left\| (\boldsymbol{u}^{\mathsf{T}}, \lambda_g \boldsymbol{v}^{(\bullet)\mathsf{T}}) \left[\mathbb{W}_{\mathcal{S}_{\mathsf{full}}}^{\mathsf{T}}(\boldsymbol{\beta}^{(\bullet)})\mathbb{W}_{\mathcal{S}_{\mathsf{full}}}(\boldsymbol{\beta}^{(\bullet)})\right]^{-1} \mathbb{W}_{\mathcal{S}_{\mathsf{full}}}^{\mathsf{T}}(\boldsymbol{\beta}^{(\bullet)})\mathbb{W}_{\emptyset,j'}(\boldsymbol{\beta}^{(\bullet)}) \right\|_{\widetilde{\mathbb{T}}} \right\} \leq \lambda_g(1 - \epsilon),$$

where

$$\mathscr{G}_{\mathcal{S}_\mu} = \left\{ \boldsymbol{u} = (u_1, \cdots, u_{|\mathcal{S}_\mu|})^{\mathsf{T}} \in \mathbb{R}^{|\mathcal{S}_\mu|} : \max_{j\in[|\mathcal{S}_\mu|]} |u_j| \leq 1 \right\},$$
$$\mathscr{G}_{\mathcal{S}_\alpha} = \left\{ \boldsymbol{v}^{(\bullet)} = (\boldsymbol{v}^{(2)\mathsf{T}}, \dots, \boldsymbol{v}^{(\mathsf{M})\mathsf{T}})^{\mathsf{T}} \in \mathbb{R}^{(M-1)|\mathcal{S}_\alpha|} : \max_{j\in[|\mathcal{S}_\alpha|]} \|\boldsymbol{v}_j\|_{\widetilde{\mathbb{T}}} \leq 1,\ \boldsymbol{v}_j = (v_j^{(2)}, \dots, v_j^{(\mathsf{M})})^{\mathsf{T}} \right\}$$

represent the sub-gradient space corresponding to $\mathcal{S}_\mu$ and $\mathcal{S}_\alpha$ of the mixture penalty.

Next, we demonstrate that Condition 6 is a reasonable assumption and is similar to those required for the sparsistency of LASSO and group LASSO (Zhao and Yu, 2006; Nardi et al., 2008). Specifically, we present detailed justifications for the Irrepresentable Condition $\mathscr{C}_{\mathsf{Irrep}}$ of the weighted design matrix $\mathbb{W}(\boldsymbol{\beta}^{(\bullet)})$ when the local Hessian matrix satisfies two commonly seen correlation structures, the constant positive correlation and auto-regressive correlation defined respectively by

$$\mathsf{Cons}(r) = \{r^{\mathbf{I}(i\neq j)}\}_{p\times p} \quad \text{and} \quad \mathsf{AR}(\rho) = \{\rho^{|i-j|}\}_{p\times p}.$$

To see the design matrix associated with $\boldsymbol{\theta} = (\boldsymbol{\mu}^{\mathsf{T}}, \boldsymbol{\alpha}^{(2)\mathsf{T}}, \ldots, \boldsymbol{\alpha}^{(\mathsf{M})\mathsf{T}})^{\mathsf{T}}$, let $\mathbf{A}$ be the transformation operator between $\boldsymbol{\beta}^{(\bullet)}$ and $\boldsymbol{\theta}$ such that $\boldsymbol{\beta}^{(\bullet)}_{\mathcal{S}} = \mathbf{A}_{\mathcal{S},\mathcal{S}}\boldsymbol{\theta}_{\mathcal{S},\mathcal{S}}$, where $\boldsymbol{\beta}^{(\bullet)}_{\mathcal{S}} = (\boldsymbol{\beta}^{(1)\mathsf{T}}_{\mathcal{S}}, \ldots, \boldsymbol{\beta}^{(\mathsf{M})\mathsf{T}}_{\mathcal{S}})^{\mathsf{T}}$. For any $\mathcal{S}_1, \mathcal{S}_2 \subseteq [p]$, let $\boldsymbol{\theta}_{\mathcal{S}_1,\mathcal{S}_2} = (\boldsymbol{\mu}^{\mathsf{T}}_{\mathcal{S}_1}, \boldsymbol{\alpha}^{(-1)\mathsf{T}}_{\mathcal{S}_2})^{\mathsf{T}}$, and $\boldsymbol{\alpha}^{(-1)}_{\mathcal{S}_2} = (\boldsymbol{\alpha}^{(2)\mathsf{T}}_{\mathcal{S}_2}, \ldots, \boldsymbol{\alpha}^{(\mathsf{M})\mathsf{T}}_{\mathcal{S}_2})^{\mathsf{T}}$. Then it follows that $\mathbb{Z}_{\mathcal{S},\mathcal{S}} = \mathbb{X}_{\mathcal{S}}\mathbf{A}_{\mathcal{S},\mathcal{S}}$, where $\mathbb{X}_{\mathcal{S}} = \mathsf{bdiag}\{\mathbb{X}^{(\mathsf{m})}_{\bullet\mathcal{S}}\}_{m=1}^{M}$. For simplicity, we take $\mathcal{S}_\mu = \mathcal{S}_\alpha = \mathcal{S}_0$, $s = |\mathcal{S}_0|$ and $n_1 = n_2 = \ldots = n_M = n$ in our following analysis. Denote by $h = \lambda_g/(1/M^{1/2})$.

A.2.1 Constant correlation structure

First, we consider the scenario that the local Hessian matrices satisfy $\mathbb{H}_m(\boldsymbol{\beta}^{(\mathsf{m})}) = \mathbf{D}^{(\mathsf{m})}\mathsf{Cons}(r_m)\mathbf{D}^{(\mathsf{m})}$, where $r_m \in (0,1)$ and $\mathbf{D}^{(\mathsf{m})} = \mathsf{diag}\{d_{m1}, \ldots, d_{mp}\}$ with $d_{mj} > 0$, for $m \in [M]$, in analog to Corollary 1 of Zhao and Yu (2006). Without loss of generality, we assume $\mathcal{S}_0 = \{1, 2, \ldots, |\mathcal{S}_0|\}$.

Proposition A1. *Let* $\mathbb{H}_m(\boldsymbol{\beta}^{(\mathsf{m})}) = \mathbf{D}^{(\mathsf{m})}\mathsf{Cons}(r_m)\mathbf{D}^{(\mathsf{m})}$ *with* $0 \le r_m \le r$ *and* $\mathbf{D}^{(\mathsf{m})} = \mathsf{diag}\{d_{m1}, \ldots, d_{mp}\}$ *for all* $m \in [M]$. *Define that* $\delta = \max_{m\in[M], j\in\mathcal{S}_0^c, k\in\mathcal{S}_0} d_{mj}/d_{mk}$. *Then Condition 6 holds with constant* $\epsilon \in (0,1)$ *if*

$$\frac{\delta r s(1+h)}{1+(s-1)r} \le 1-\epsilon \quad \textit{and} \quad \frac{\delta r s\{2(1+h^{-2})\}^{\frac{1}{2}}}{1+(s-1)r} \le 1-\epsilon.$$

Remark A1. *If we further simplify Proposition A1 by setting* $\delta = 1$ *and* $h = 1$, *i.e.* $\lambda_g = 1/M^{1/2}$, *then the condition on* r *can be relaxed and simplified to* $r \le (1-\epsilon)/(1+s)$.

Proof. Let $\mathbf{d}^{(\mathsf{m})} = (d_{m1}, \ldots, d_{mp})^\top$ and $\breve{\mathbf{d}}^{(\mathsf{m})} = (d_{m1}^{-1}, \ldots, d_{mp}^{-1})^\top$. First, for any $j \in \mathcal{S}_0^c$,

$$\begin{aligned}
&\Big[\mathbb{W}^\top_{\mathcal{S}_{\mathsf{full}}}(\boldsymbol{\beta}^{(\bullet)})\mathbb{W}_{\mathcal{S}_{\mathsf{full}}}(\boldsymbol{\beta}^{(\bullet)})\Big]^{-1}\mathbb{W}^\top_{\mathcal{S}_{\mathsf{full}}}(\boldsymbol{\beta}^{(\bullet)})\mathbb{W}_{j,\emptyset}(\boldsymbol{\beta}^{(\bullet)}) \\
=&\Big[\mathbf{A}^\top_{\mathcal{S}_{\mathsf{full}}}\mathsf{bdiag}\{\mathbf{D}^{(\mathsf{m})}_{\mathcal{S}_0,\mathcal{S}_0}[\mathsf{Cons}(r_m)]_{\mathcal{S}_0,\mathcal{S}_0}\mathbf{D}^{(\mathsf{m})}_{\mathcal{S}_0,\mathcal{S}_0}\}_{m=1}^M\mathbf{A}_{\mathcal{S}_{\mathsf{full}}}\Big]^{-1} \\
&\cdot\mathbf{A}^\top_{\mathcal{S}_{\mathsf{full}}}\Big\{d_{1j}[\mathsf{Cons}(r_1)]^\top_{\mathcal{S}_0,j}\mathbf{D}^{(1)}_{\mathcal{S}_0,\mathcal{S}_0}, \ldots, d_{Mj}[\mathsf{Cons}(r_M)]^\top_{\mathcal{S}_0,j}\mathbf{D}^{(\mathsf{M})}_{\mathcal{S}_0,\mathcal{S}_0}\Big\}^\top \\
=&[\mathbf{A}_{\mathcal{S}_{\mathsf{full}}}]^{-1}\mathsf{bdiag}\Big\{[\mathbf{D}^{(\mathsf{m})}_{\mathcal{S}_0,\mathcal{S}_0}]^{-1}[\mathsf{Cons}(r_m)]^{-1}_{\mathcal{S}_0,\mathcal{S}_0}\Big\}_{m=1}^M\Big\{d_{1j}[\mathsf{Cons}(r_1)]^\top_{\mathcal{S}_0,j}, \ldots, d_{Mj}[\mathsf{Cons}(r_M)]^\top_{\mathcal{S}_0,j}\Big\}^\top.
\end{aligned}\tag{S3}$$

Then recall $\mathbb{T} = (\mathbf{1}_{(M-1)\times 1}, \mathbb{I}_{(M-1)\times(M-1)})^\top$, $\|\mathbf{x}\|_{\mathbb{T}} := \|\mathbb{T}\mathbf{x}\|_2$ and $\|\mathbf{x}\|_{\widetilde{\mathbb{T}}} := \|\mathbb{T}(\mathbb{T}^\top\mathbb{T})^{-1}\mathbf{x}\|_2$, it follows that for any $\boldsymbol{u} \in \mathscr{G}_{\mathcal{S}_\mu}, \boldsymbol{v}^{(\bullet)} \in \mathscr{G}_{\mathcal{S}_\alpha}$:

$$\begin{aligned}
&\left|(\boldsymbol{u}^\top, \lambda_g\boldsymbol{v}^{(\bullet)\top})\Big[\mathbb{W}^\top_{\mathcal{S}_{\mathsf{full}}}(\boldsymbol{\beta}^{(\bullet)})\mathbb{W}_{\mathcal{S}_{\mathsf{full}}}(\boldsymbol{\beta}^{(\bullet)})\Big]^{-1}\mathbb{W}^\top_{\mathcal{S}_{\mathsf{full}}}(\boldsymbol{\beta}^{(\bullet)})\mathbb{W}_{j,\emptyset}(\boldsymbol{\beta}^{(\bullet)})\right| \\
=&\left|(\boldsymbol{u}^\top, \lambda_g\boldsymbol{v}^{(\bullet)\top})[\mathbf{A}_{\mathcal{S}_{\mathsf{full}}}]^{-1}\left(\frac{r_1 d_{1j}\breve{\mathbf{d}}^{(1)}_{\mathcal{S}_0}}{1+(s-1)r_1}, \ldots, \frac{r_M d_{Mj}\breve{\mathbf{d}}^{(\mathsf{M})}_{\mathcal{S}_0}}{1+(s-1)r_M}\right)^\top\right| \\
\leq&\sum_{k=1}^{|\mathcal{S}_0|}\left|(u_k, \lambda_g\boldsymbol{v}_k^\top)\Big[\mathbf{A}_{\mathcal{S}_0[k],\mathcal{S}_0[k]}\Big]^{-1}\left(\frac{r_1 d_{1j}/d_{1k}}{1+(s-1)r_1}, \ldots, \frac{r_M d_{Mj}/d_{Mk}}{1+(s-1)r_M}\right)^\top\right|,
\end{aligned}\tag{S4}$$

where $\boldsymbol{v}_k = (v_k^{(2)}, \ldots, v_k^{(\mathsf{M})})^\top$, $\mathcal{S}_0[k]$ represents the k-th element in $\mathcal{S}_0$ and the "$\leq$" follows from the fact that $\mathbf{A}_{\mathcal{S}_{\mathsf{full}}}$ is blocked-diagonal in $\mathbf{A}_{\mathcal{S}_0[k],\mathcal{S}_0[k]}$. Note that

$$\Big[\mathbf{A}_{\mathcal{S}_0[k],\mathcal{S}_0[k]}\Big]^{-1} = \begin{pmatrix}
M^{-1} & M^{-1} & M^{-1} & \ldots & M^{-1} \\
-M^{-1} & 1-M^{-1} & -M^{-1} & \ldots & -M^{-1} \\
-M^{-1} & -M^{-1} & 1-M^{-1} & \ldots & -M^{-1} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
-M^{-1} & -M^{-1} & -M^{-1} & \ldots & 1-M^{-1}
\end{pmatrix}.$$

Let $\Big[\mathbf{A}_{\mathcal{S}_0[k],\mathcal{S}_0[k]}\Big]^{-1}_{-1,\bullet}$ denote the second to the M-th rows of $\Big[\mathbf{A}_{\mathcal{S}_0[k],\mathcal{S}_0[k]}\Big]^{-1}$ and

$$\widetilde{\boldsymbol{r}}_k = (\widetilde{r}_{k1}, \ldots, \widetilde{r}_{kM})^\top = \left(\frac{r_1 d_{1j}/d_{1k}}{1+(s-1)r_1}, \ldots, \frac{r_M d_{Mj}/d_{Mk}}{1+(s-1)r_M}\right)^\top.$$

Recall that $\lambda_g = h/M^{1/2}$ and $d_{mj}/d_{mk} \le \delta$ for $j \in \mathcal{S}_0^c$ and $k \in \mathcal{S}_0$, we have that

$$\begin{aligned}
&\left|(\boldsymbol{u}^{\mathsf{T}}, \lambda_g \boldsymbol{v}^{(\bullet)\mathsf{T}}) \left[\mathbb{W}^{\mathsf{T}}_{\mathcal{S}_{\mathsf{full}}}(\boldsymbol{\beta}^{(\bullet)})\mathbb{W}_{\mathcal{S}_{\mathsf{full}}}(\boldsymbol{\beta}^{(\bullet)})\right]^{-1} \mathbb{W}^{\mathsf{T}}_{\mathcal{S}_{\mathsf{full}}}(\boldsymbol{\beta}^{(\bullet)})\mathbb{W}_{j,\emptyset}(\boldsymbol{\beta}^{(\bullet)})\right| \\
\le& \sum_{k=1}^{|\mathcal{S}_0|} |u_k| \left\{ M^{-1} \sum_{m=1}^{M} \widetilde{r}_{km} \right\} + \lambda_g \sum_{k=1}^{|\mathcal{S}_0|} \|\boldsymbol{v}_k\|_{\widetilde{\mathbb{T}}} \left\| \left[\mathbf{A}_{\mathcal{S}_0[k],\mathcal{S}_0[k]}\right]^{-1}_{-1,\bullet} \widetilde{\boldsymbol{r}}_k \right\|_{\mathbb{T}} \qquad \text{(S5)} \\
\le& s M^{-1} \sum_{m=1}^{M} \widetilde{r}_{km} + s\lambda_g \left\| \widetilde{\boldsymbol{r}}^{\mathsf{T}}_{k,-1} \right\|_2 \le \frac{\delta r s(1 + \lambda_g \sqrt{M-1})}{1 + (s-1)r} \le \frac{\delta r s (1+h)}{1+(s-1)r} \le 1 - \epsilon,
\end{aligned}$$

where we use the fact $\mathbb{T}\left[\mathbf{A}_{\mathcal{S}_0[k],\mathcal{S}_0[k]}\right]^{-1}_{-1,\bullet} = (\mathbf{0}, \mathbb{I}_{M-1})^{\mathsf{T}}$ for the second "$\le$".

While for $j' \in \mathcal{S}_\alpha^c$ and $\boldsymbol{u} \in \mathscr{G}_{\mathcal{S}_\mu}, \boldsymbol{v}^{(\bullet)} \in \mathscr{G}_{\mathcal{S}_\alpha}$, define that $\widetilde{\boldsymbol{v}}_k = (\widetilde{v}_k^{(1)}, \ldots, \widetilde{v}_k^{(\mathsf{M})})^{\mathsf{T}} = \lambda_g \mathbb{T}(\mathbb{T}^{\mathsf{T}}\mathbb{T})^{-1}\boldsymbol{v}_k$ and similar to (S3) and (S4),

$$\begin{aligned}
&\left\|(\boldsymbol{u}^{\mathsf{T}}, \lambda_g \boldsymbol{v}^{(\bullet)\mathsf{T}}) \left[\mathbb{W}^{\mathsf{T}}_{\mathcal{S}_{\mathsf{full}}}(\boldsymbol{\beta}^{(\bullet)})\mathbb{W}_{\mathcal{S}_{\mathsf{full}}}(\boldsymbol{\beta}^{(\bullet)})\right]^{-1} \mathbb{W}^{\mathsf{T}}_{\mathcal{S}_{\mathsf{full}}}(\boldsymbol{\beta}^{(\bullet)})\mathbb{W}_{\emptyset,j'}(\boldsymbol{\beta}^{(\bullet)})\right\|_{\widetilde{\mathbb{T}}} \\
\le& \sum_{k=1}^{|\mathcal{S}_0|} \left\| (u_k, \lambda_g \boldsymbol{v}_k^{\mathsf{T}}) \left[\mathbf{A}_{\mathcal{S}_0[k],\mathcal{S}_0[k]}\right]^{-1} \left(\widetilde{r}_{k1}\mathbf{1}_{M-1}, \mathsf{diag}\{\widetilde{r}_{k2}, \ldots, \widetilde{r}_{kM}\}\right)^{\mathsf{T}} \right\|_{\widetilde{\mathbb{T}}} \\
=& \sum_{k=1}^{|\mathcal{S}_0|} \left\| (u_k, \widetilde{\boldsymbol{v}}_k^{\mathsf{T}})(M^{-1}\mathbf{1}_M, \mathbb{I}_M)^{\mathsf{T}} \left(\widetilde{r}_{k1}\mathbf{1}_{M-1}, \mathsf{diag}\{\widetilde{r}_{k2}, \ldots, \widetilde{r}_{kM}\}\right)^{\mathsf{T}} \right\|_{\widetilde{\mathbb{T}}}.
\end{aligned}$$

Due to the fact that $|u_k| \le 1$, $\|\widetilde{\boldsymbol{v}}_k\|_2 \le \lambda_g$, $\mathbf{1}^{\mathsf{T}}\widetilde{\boldsymbol{v}}_k = 0$, and note that $\mathbf{x}^{\mathsf{T}}(\mathbb{T}^{\mathsf{T}}\mathbb{T})^{-1}\mathbf{x}$ is the sample variance of $\mathbf{x}$, which is smaller or equal to $\|\mathbf{x} - c\|_2^2$ for any constant c, we have that

$$\begin{aligned}
&\left\|(\boldsymbol{u}^{\mathsf{T}}, \lambda_g \boldsymbol{v}^{(\bullet)\mathsf{T}}) \left[\mathbb{W}^{\mathsf{T}}_{\mathcal{S}_{\mathsf{full}}}(\boldsymbol{\beta}^{(\bullet)})\mathbb{W}_{\mathcal{S}_{\mathsf{full}}}(\boldsymbol{\beta}^{(\bullet)})\right]^{-1} \mathbb{W}^{\mathsf{T}}_{\mathcal{S}_{\mathsf{full}}}(\boldsymbol{\beta}^{(\bullet)})\mathbb{W}_{\emptyset,j'}(\boldsymbol{\beta}^{(\bullet)})\right\|_{\widetilde{\mathbb{T}}} \\
\le& \sum_{k=1}^{|\mathcal{S}_0|} \inf_{c \in \mathbb{R}, c \perp t} \left[\sum_{t \ne 1} (M^{-1}u_1\widetilde{r}_{k1} + M^{-1}u_1\widetilde{r}_{kt} + M^{-1}\widetilde{v}_k^{(1)}\widetilde{r}_{k1} + \widetilde{v}_k^{(\mathsf{t})}\widetilde{r}_{kt} - c)^2 \right]^{\frac{1}{2}} \\
\le& \sum_{k=1}^{|\mathcal{S}_0|} \left[\sum_{t \ne 1} \widetilde{r}_{kt}^2 (M^{-1}u_1 + \widetilde{v}_k^{(\mathsf{t})})^2 \right]^{\frac{1}{2}} = \sum_{k=1}^{|\mathcal{S}_0|} \frac{\delta r}{1 + (s-1)r} \left[\sum_{t \ne 1} 2M^{-2}u_1^2 + 2(\widetilde{v}_k^{(\mathsf{t})})^2 \right]^{\frac{1}{2}} \\
\le& \frac{s\delta r}{1 + (s-1)r} \left(2M^{-1} + 2\lambda_g^2 \right)^{\frac{1}{2}} = \frac{\{2(1 + h^{-2})\}^{\frac{1}{2}} \lambda_g s \delta r}{1 + (s-1)r} \le \lambda_g(1 - \epsilon).
\end{aligned}$$

□

A.2.2 Auto-regressive correlation structure

Now we turn to the auto-regressive correlation structure, i.e., $\mathbb{H}_m(\boldsymbol{\beta}^{(\mathsf{m})}) = \mathbf{D}^{(\mathsf{m})}\mathsf{AR}(\rho_m)\mathbf{D}^{(\mathsf{m})}$, where $\rho_m \in (-1, 1)$ and $\mathbf{D}^{(\mathsf{m})} = \mathsf{diag}\{d_{m1}, \ldots, d_{mp}\}$ with $d_{mj} > 0$, for $m \in [M]$, in analog to Corollary 3 of Zhao and Yu (2006).

Proposition A2. *Let $\mathbb{H}_m(\boldsymbol{\beta}^{(\mathsf{m})}) = \mathbf{D}^{(\mathsf{m})}\mathsf{AR}(\rho_m)\mathbf{D}^{(\mathsf{m})}$ with $\mathbf{D}^{(\mathsf{m})} = \mathsf{diag}\{d_{m1}, \ldots, d_{mp}\}$ and $0 \leq \rho_m \leq \rho$ for all $m \in [M]$. Again denote by $\delta = \max_{m\in[M], j\in\mathcal{S}_0^c, k\in\mathcal{S}_0} d_{mj}/d_{mk}$. Then Condition 6 holds with constant $\epsilon \in (0, 1)$ if*

$$\frac{2\delta\rho(1+h)}{1+\rho^2} \leq 1-\epsilon \quad \textit{and} \quad \frac{2\delta\rho\{2(1+h^{-2})\}^{\frac{1}{2}}}{1+\rho^2} \leq 1-\epsilon.$$

Remark A2. *If we again simplify Proposition A2 by setting $\delta = 1$ and $h = 1$, i.e. $\lambda_g = 1/M^{1/2}$, then the condition on ρ can be simplified to*

$$\rho \leq \frac{1}{2 + \sqrt{4 - (1-\epsilon)^2}},$$

which can be approximated by $\rho \leq 2 - \sqrt{3} \approx 0.27$ if we set $\epsilon \approx 0$.

Proof. Again denote by $\mathbf{d}^{(\mathsf{m})} = (d_{m1}, \ldots, d_{mp})^\mathsf{T}$. Let $\mathcal{S}_0 = \{k_1, \ldots, k_s\}$ where $k_1 < \ldots < k_s$. Without loss of generality, we let $k_{s+1} = p$ if $k_s < p$. For $j \in \mathcal{S}_0^c$ satisfying $k_\ell < j < k_{\ell+1}$, similar to the proof of Corollary 3 in Zhao and Yu (2006), we have that the $k_{\ell+1}$-th element of $[\mathbf{D}^{(\mathsf{m})}_{\mathcal{S}_0,\mathcal{S}_0}]^{-1}[\mathsf{AR}(\rho_m)]^{-1}_{\mathcal{S}_0,\mathcal{S}_0} d_{mj}[\mathsf{AR}(\rho_m)]_{\mathcal{S}_0,j}$ is $d_{mj}/d_{mk_{\ell+1}} \cdot (\rho_m^{k_{\ell+1}-j} - \rho_m^{j-k_{\ell+1}})/(\rho_m^{k_{\ell+1}-k_\ell} - \rho_m^{k_\ell-k_{\ell+1}})$, and the k_ℓ-th element is $d_{mj}/d_{mk_\ell} \cdot (\rho_m^{j-k_\ell} - \rho_m^{k_\ell-j})/(\rho_m^{k_{\ell+1}-k_\ell} - \rho_m^{k_\ell-k_{\ell+1}})$, while the remaining elements are all 0. Then similar to (S5) as shown in the proof of Proposition A1, for any $\boldsymbol{u} \in \mathscr{G}_{\mathcal{S}_\mu}, \boldsymbol{v}^{(\bullet)} \in \mathscr{G}_{\mathcal{S}_\alpha}$, we have

$$
\begin{aligned}
&\left|(\boldsymbol{u}^{\mathsf{T}}, \lambda_g \boldsymbol{v}^{(\bullet)\mathsf{T}}) \left[\mathbb{W}_{\mathcal{S}_{\mathsf{full}}}^{\mathsf{T}}(\boldsymbol{\beta}^{(\bullet)})\mathbb{W}_{\mathcal{S}_{\mathsf{full}}}(\boldsymbol{\beta}^{(\bullet)})\right]^{-1} \mathbb{W}_{\mathcal{S}_{\mathsf{full}}}^{\mathsf{T}}(\boldsymbol{\beta}^{(\bullet)})\mathbb{W}_{j,\emptyset}(\boldsymbol{\beta}^{(\bullet)})\right| \\
\leq & \sum_{t\in\{\ell,\ell+1\}} |u_t| M^{-1} \sum_{m=1}^{M} \frac{d_{mj}}{d_{mk_t}} \cdot \left| \frac{\rho_m^{k_t-j} - \rho_m^{j-k_t}}{\rho_m^{k_{\ell+1}-k_\ell} - \rho_m^{k_\ell - k_{\ell+1}}} \right| + \sum_{t\in\{\ell,\ell+1\}} \lambda_g \|\boldsymbol{v}_j\|_{\widetilde{\mathbb{T}}} \left\| \left[\mathbf{A}^{(1)}_{\mathcal{S}_0[j],\mathcal{S}_0[j]}\right]^{-1}_{-1,\bullet} \widetilde{\boldsymbol{\rho}}_t \right\|_{\mathbb{T}} \\
\leq & \frac{2\delta\rho}{1+\rho^2} + \sum_{t\in\{\ell,\ell+1\}} \lambda_g \|\widetilde{\boldsymbol{\rho}}_{t,-1}^{\mathsf{T}}\|_2 \leq \frac{2\delta\rho}{1+\rho^2} + \lambda_g \sqrt{2(\|\widetilde{\boldsymbol{\rho}}_{\ell,-1}^{\mathsf{T}}\|_2^2 + \|\widetilde{\boldsymbol{\rho}}_{\ell+1,-1}^{\mathsf{T}}\|_2^2)} \\
\leq & \frac{2\delta\rho(1+\lambda_g M^{\frac{1}{2}})}{1+\rho^2} = \frac{2\delta\rho(1+h)}{1+\rho^2} \leq 1-\epsilon,
\end{aligned}
$$

where $\widetilde{\boldsymbol{\rho}}_t = \mathbf{0}$ if $t \notin \{\ell, \ell+1\}$,

$$
\widetilde{\boldsymbol{\rho}}_t = (\widetilde{\rho}_{t1}, \ldots, \widetilde{\rho}_{tM})^{\mathsf{T}} = \left(\frac{d_{1j}}{d_{1k_t}} \left| \frac{\rho_1^{k_t-j} - \rho_1^{j-k_t}}{\rho_1^{k_{\ell+1}-k_\ell} - \rho_1^{k_\ell-k_{\ell+1}}} \right|, \ldots, \frac{d_{Mj}}{d_{Mk_t}} \left| \frac{\rho_M^{k_t-j} - \rho_M^{j-k_t}}{\rho_M^{k_{\ell+1}-k_\ell} - \rho_M^{k_\ell-k_{\ell+1}}} \right| \right)^{\mathsf{T}},
$$

when $t \in \{\ell, \ell+1\}$ and we use the fact that $\widetilde{\rho}_{t1}, \ldots, \widetilde{\rho}_{tM} \leq \delta\rho/(1+\rho^2)$.

While for $j' \in \mathcal{S}_\alpha^c$ and $\boldsymbol{u} \in \mathscr{G}_{\mathcal{S}_\mu}, \boldsymbol{v}^{(\bullet)} \in \mathscr{G}_{\mathcal{S}_\alpha}$, we again define that $\widetilde{\boldsymbol{v}}_k = (\widetilde{v}_k^{(1)}, \ldots, \widetilde{v}_k^{(\mathsf{M})})^{\mathsf{T}} = \lambda_g \mathbb{T}(\mathbb{T}^{\mathsf{T}}\mathbb{T})^{-1}\boldsymbol{v}_k$ and similar to the proof of Proposition A1, we have

$$
\begin{aligned}
&\left\|(\boldsymbol{u}^{\mathsf{T}}, \lambda_g \boldsymbol{v}^{(\bullet)\mathsf{T}}) \left[\mathbb{W}_{\mathcal{S}_{\mathsf{full}}}^{\mathsf{T}}(\boldsymbol{\beta}^{(\bullet)})\mathbb{W}_{\mathcal{S}_{\mathsf{full}}}(\boldsymbol{\beta}^{(\bullet)})\right]^{-1} \mathbb{W}_{\mathcal{S}_{\mathsf{full}}}^{\mathsf{T}}(\boldsymbol{\beta}^{(\bullet)})\mathbb{W}_{\emptyset,j'}(\boldsymbol{\beta}^{(\bullet)})\right\|_{\widetilde{\mathbb{T}}} \\
\leq & \sum_{k\in\{\ell,\ell+1\}} \inf_{c\in\mathbb{R}, c\perp t} \left[\sum_{t\neq 1} (M^{-1}u_1\widetilde{\rho}_1 + M^{-1}u_1\widetilde{\rho}_t + M^{-1}\widetilde{v}_k^{(1)}\widetilde{\rho}_1 + \widetilde{v}_k^{(\mathsf{t})}\widetilde{\rho}_t - c)^2 \right]^{\frac{1}{2}} \\
\leq & \sum_{k\in\{\ell,\ell+1\}} \left[\sum_{t\neq 1} \widetilde{\rho}_t^2 (M^{-1}u_1 + \widetilde{v}_k^{(\mathsf{t})})^2 \right]^{\frac{1}{2}} \leq \frac{2\delta\rho}{1+\rho^2} \left(2M^{-1} + 2\lambda_g^2\right)^{\frac{1}{2}} \\
\leq & \lambda_g \{2(1+h^{-2})\}^{\frac{1}{2}} \frac{2\delta\rho}{1+\rho^2} \leq \lambda_g(1-\epsilon),
\end{aligned}
$$

which finishes the proof. □

A.2.3 Conclusion

For both constant correlation structure and auto-regressive correlation structure, our Irrepresentable Condition $\mathscr{C}_{\mathsf{Irrep}}$ is comparable to that of the LASSO estimator as in Corollaries 1 and

3 of Zhao and Yu (2006). Specifically, we both have the upper bound for r in the $\mathsf{Cons}(r)$ structure decaying with a rate of s^{-1}, and both have constant rate for ρ in the $\mathsf{AR}(\rho)$ structure. Note that in terms of the multiplicative constants for the rates on r or ρ, our assumptions seem to be stronger. This is due to the fact that the supports of $\boldsymbol{\mu}_0$ and $\boldsymbol{\alpha}_0^{(\bullet)}$ are set to be the same for the simplicity of construction, and as a result it produces more regularization bias than the simple LASSO case.

A.3. PROOF OF THE MAIN THEOREMS

Throughout, we define the *model complexity adjusted effective* sample size for each study as $n_m^{\mathsf{eff}} = n_m/(s_0 \log p)$ and $n^{\mathsf{eff}} = N/[s_0(\log p + M)]$, which are the main drivers for the rates of the proposed estimators.

A.3.1 Outline of the proof

Due to the lengthy proof, we begin with the outline of the main steps as below.

1) To account for the randomness of $\nabla\widehat{\mathcal{L}}_\bullet(\boldsymbol{\beta}_0^{(\bullet)}) = (\nabla\widehat{\mathcal{L}}_1(\boldsymbol{\beta}_0^{(\bullet)})^\mathsf{T}, \ldots, \nabla\widehat{\mathcal{L}}_M(\boldsymbol{\beta}_0^{(\bullet)})^\mathsf{T})^\mathsf{T}$, bound

$$\|\nabla\widehat{\mathcal{L}}_\bullet(\boldsymbol{\beta}_0^{(\bullet)})\|_{2,\infty} := \max_{j\in[p]}\left\{N^{-1}\sqrt{\sum_{m=1}^{M}\left[n_m\nabla_j\widehat{\mathcal{L}}_m(\boldsymbol{\beta}_0^{(\bullet)})\right]^2}\right\} \quad \text{and} \quad \left\|N^{-1}\sum_{m=1}^{M} n_m\nabla\widehat{\mathcal{L}}_m(\boldsymbol{\beta}_0^{(\mathsf{m})})\right\|_\infty$$

using Condition 2 and Lemma A1, where $\nabla_j\widehat{\mathcal{L}}_m(\boldsymbol{\beta}_0^{(\bullet)})$ is the jth element of $\nabla\widehat{\mathcal{L}}_m(\boldsymbol{\beta}_0^{(\bullet)})$. This is a crucial step to control the empirical process $\nabla\widehat{\mathcal{L}}_\bullet(\boldsymbol{\beta}_0^{(\bullet)})(\widehat{\boldsymbol{\beta}}_{\mathsf{SHIR}}^{(\bullet)} - \boldsymbol{\beta}_0^{(\bullet)})$ by the terms $\|\nabla\widehat{\mathcal{L}}_\bullet(\boldsymbol{\beta}_0^{(\bullet)})\|_{2,\infty}$, $\|N^{-1}\sum_{m=1}^{M} n_m\nabla\widehat{\mathcal{L}}_m(\boldsymbol{\beta}_0^{(\mathsf{m})})\|_\infty$, and $\|\widehat{\boldsymbol{\mu}}_{\mathsf{SHIR}} - \boldsymbol{\mu}_0\|_1 + \lambda_g\|\widehat{\boldsymbol{\alpha}}_{\mathsf{SHIR}}^{(\bullet)} - \boldsymbol{\alpha}_0^{(\bullet)}\|_{2,1}$.

2) Bound the additional noise terms from the integrating process using Conditions 2, 3 and 4.

3) Start from the basic inequality $\widehat{Q}_{\mathsf{SHIR}}(\widehat{\boldsymbol{\beta}}_{\mathsf{SHIR}}^{(\bullet)}) \le \widehat{Q}_{\mathsf{SHIR}}(\boldsymbol{\beta}_0^{(\bullet)})$, use the Condition $\mathscr{C}_{\mathsf{comp}}$ and the results of Steps 1) and 2) to prove Theorem 1.

4) To prove Theorem 2, base on the inequality $\widehat{Q}_{\mathsf{SHIR}}(\widehat{\boldsymbol{\beta}}_{\mathsf{SHIR}}^{(\bullet)}) \le \widehat{Q}_{\mathsf{SHIR}}(\widehat{\boldsymbol{\beta}}_{\mathsf{IPDpool}}^{(\bullet)})$ to compare $\widehat{\boldsymbol{\beta}}_{\mathsf{SHIR}}^{(\bullet)}$ and $\widehat{\boldsymbol{\beta}}_{\mathsf{IPDpool}}^{(\bullet)}$ directly and use the fact that $\widehat{\boldsymbol{\beta}}_{\mathsf{IPDpool}}^{(\bullet)}$ minimizes the individual level objective function to simplify the inequality $\widehat{Q}_{\mathsf{SHIR}}(\widehat{\boldsymbol{\beta}}_{\mathsf{SHIR}}^{(\bullet)}) \le \widehat{Q}_{\mathsf{SHIR}}(\widehat{\boldsymbol{\beta}}_{\mathsf{IPDpool}}^{(\bullet)})$.

5) To prove Theorem 3, follow the similar strategy used in Zhao and Yu (2006) and Nardi et al.

(2008). In specific, verify the KarushKuhnTucker (KKT) conditions corresponding to the true $\mathcal{S}_\mu$ and $\mathcal{S}_\alpha$, separately for the zero and non-zero parts of $(\widehat{\boldsymbol{\mu}}_{\mathsf{IPDpool}}^{\mathsf{T}}, \widehat{\boldsymbol{\alpha}}_{\mathsf{IPDpool}}^{(\bullet)\mathsf{T}})$.

A.3.2 Proofs of Theorem 1

Proof. First, we expand $\nabla\widehat{\mathcal{L}}_m(\widehat{\boldsymbol{\beta}}_{\mathsf{LASSO}}^{(\mathsf{m})})$ around $\nabla\widehat{\mathcal{L}}_m(\boldsymbol{\beta}_0^{(\mathsf{m})})$ inspired by (Feng et al., 2014). For a vector or matrix $\mathbf{A}(t)$ whose (i,j)-entry being $A_{ij}(t)$, a function of the scalar $t \in [0,1]$, define $\int_0^1 \mathbf{A}(t)dt$ as the vector or matrix with its (i,j)-entry being $\int_0^1 A_{ij}(t)dt$. We then have

$$\nabla\widehat{\mathcal{L}}_m(\widehat{\boldsymbol{\beta}}_{\mathsf{LASSO}}^{(\mathsf{m})}) = \nabla\widehat{\mathcal{L}}_m(\boldsymbol{\beta}_0^{(\mathsf{m})}) + \int_0^1 \nabla^2\widehat{\mathcal{L}}_m\left(\boldsymbol{\beta}_0^{(\mathsf{m})} + t[\widehat{\boldsymbol{\beta}}_{\mathsf{LASSO}}^{(\mathsf{m})} - \boldsymbol{\beta}_0^{(\mathsf{m})}]\right)(\widehat{\boldsymbol{\beta}}_{\mathsf{LASSO}}^{(\mathsf{m})} - \boldsymbol{\beta}_0^{(\mathsf{m})})dt, \tag{S6}$$

Thus, the gradient term $\widehat{\mathbf{g}}_m$ in equation (3) can be expressed as

$$\begin{aligned}\nabla\widehat{\mathcal{L}}_m(\widehat{\boldsymbol{\beta}}_{\mathsf{LASSO}}^{(\mathsf{m})}) - \widehat{\mathbb{H}}_m\widehat{\boldsymbol{\beta}}_{\mathsf{LASSO}}^{(\mathsf{m})} =& \nabla\widehat{\mathcal{L}}_m(\boldsymbol{\beta}_0^{(\mathsf{m})}) - \widehat{\mathbb{H}}_m\boldsymbol{\beta}_0^{(\mathsf{m})} \\ &+ \int_0^1 \left\{\nabla^2\widehat{\mathcal{L}}_m\left(\boldsymbol{\beta}_0^{(\mathsf{m})} + t[\widehat{\boldsymbol{\beta}}_{\mathsf{LASSO}}^{(\mathsf{m})} - \boldsymbol{\beta}_0^{(\mathsf{m})}]\right) - \widehat{\mathbb{H}}_m\right\}(\widehat{\boldsymbol{\beta}}_{\mathsf{LASSO}}^{(\mathsf{m})} - \boldsymbol{\beta}_0^{(\mathsf{m})})dt.\end{aligned} \tag{S7}$$

The third term of (S7)'s right hand side can be seen as the noise term introduced by our integrating procedure. Now we bound this term using Conditions 2, 3 and 4. For $t \in [0,1]$, Conditions 2 and 3 lead to

$$\begin{aligned}&\left\|\left\{\nabla^2\widehat{\mathcal{L}}_m\left(\boldsymbol{\beta}_0^{(\mathsf{m})} + t[\widehat{\boldsymbol{\beta}}_{\mathsf{LASSO}}^{(\mathsf{m})} - \boldsymbol{\beta}_0^{(\mathsf{m})}]\right) - \widehat{\mathbb{H}}_m\right\}(\widehat{\boldsymbol{\beta}}_{\mathsf{LASSO}}^{(\mathsf{m})} - \boldsymbol{\beta}_0^{(\mathsf{m})})\right\|_\infty \\ =& n_m^{-1}\left\|\mathbb{X}^{(\mathsf{m})\mathsf{T}}\left[\boldsymbol{\Omega}_m\left(\boldsymbol{\beta}_0^{(\mathsf{m})} + t[\widehat{\boldsymbol{\beta}}_{\mathsf{LASSO}}^{(\mathsf{m})} - \boldsymbol{\beta}_0^{(\mathsf{m})}]\right) - \boldsymbol{\Omega}_m(\widehat{\boldsymbol{\beta}}_{\mathsf{LASSO}}^{(\mathsf{m})})\right]\mathbb{X}^{(\mathsf{m})}(\widehat{\boldsymbol{\beta}}_{\mathsf{LASSO}}^{(\mathsf{m})} - \boldsymbol{\beta}_0^{(\mathsf{m})})\right\|_\infty \\ \leq& \frac{\max_{i,j,m}\left|X_{ij}^{(\mathsf{m})}\right|}{n_m}\sum_{i=1}^{n_m}\left|\mathbf{X}_i^{(\mathsf{m})\mathsf{T}}(\widehat{\boldsymbol{\beta}}_{\mathsf{LASSO}}^{(\mathsf{m})} - \boldsymbol{\beta}_0^{(\mathsf{m})})\right| \cdot C_L\left|(1-t)\mathbf{X}_i^{(\mathsf{m})\mathsf{T}}(\widehat{\boldsymbol{\beta}}_{\mathsf{LASSO}}^{(\mathsf{m})} - \boldsymbol{\beta}_0^{(\mathsf{m})})\right| \leq \frac{BC_L}{n_m}\left\|\mathbb{X}^{(\mathsf{m})}(\widehat{\boldsymbol{\beta}}_{\mathsf{LASSO}}^{(\mathsf{m})} - \boldsymbol{\beta}_0^{(\mathsf{m})})\right\|_2^2,\end{aligned}$$

which implies that

$$\left\|\int_0^1 \left\{\nabla^2\widehat{\mathcal{L}}_m\left(\boldsymbol{\beta}_0^{(\mathsf{m})} + t[\widehat{\boldsymbol{\beta}}_{\mathsf{LASSO}}^{(\mathsf{m})} - \boldsymbol{\beta}_0^{(\mathsf{m})}]\right) - \widehat{\mathbb{H}}_m\right\}(\widehat{\boldsymbol{\beta}}_{\mathsf{LASSO}}^{(\mathsf{m})} - \boldsymbol{\beta}_0^{(\mathsf{m})})dt\right\|_\infty \leq \frac{BC_L}{n_m}\left\|\mathbb{X}^{(\mathsf{m})}(\widehat{\boldsymbol{\beta}}_{\mathsf{LASSO}}^{(\mathsf{m})} - \boldsymbol{\beta}_0^{(\mathsf{m})})\right\|_2^2. \tag{S8}$$

Then by the fact that $\widehat{Q}_{\mathsf{SHIR}}(\widehat{\boldsymbol{\beta}}^{(\bullet)}_{\mathsf{SHIR}}) \leq \widehat{Q}_{\mathsf{SHIR}}(\boldsymbol{\beta}^{(\bullet)}_0)$, we have

$$\begin{aligned}
&N^{-1}\sum_{m=1}^{M} n_m(\widehat{\boldsymbol{\beta}}^{(\mathsf{m})}_{\mathsf{SHIR}} - \boldsymbol{\beta}^{(\mathsf{m})}_0)^{\mathsf{T}}\widehat{\mathbb{H}}_m(\widehat{\boldsymbol{\beta}}^{(\mathsf{m})}_{\mathsf{SHIR}} - \boldsymbol{\beta}^{(\mathsf{m})}_0) + \lambda\rho(\widehat{\boldsymbol{\beta}}^{(\bullet)}_{\mathsf{SHIR}}) \\
&\leq -2N^{-1}\sum_{m=1}^{M} n_m(\widehat{\boldsymbol{\beta}}^{(\mathsf{m})}_{\mathsf{SHIR}} - \boldsymbol{\beta}^{(\mathsf{m})}_0)^{\mathsf{T}}\nabla\widehat{\mathcal{L}}_m(\boldsymbol{\beta}^{(\mathsf{m})}_0) \\
&+ 2N^{-1}\sum_{m=1}^{M} n_m(\widehat{\boldsymbol{\beta}}^{(\mathsf{m})}_{\mathsf{SHIR}} - \boldsymbol{\beta}^{(\mathsf{m})}_0)^{\mathsf{T}}\textstyle\int_0^1 \nabla^2\widehat{\mathcal{L}}_m\Big(\boldsymbol{\beta}^{(\mathsf{m})}_0 + t[\widehat{\boldsymbol{\beta}}^{(\mathsf{m})}_{\mathsf{LASSO}} - \boldsymbol{\beta}^{(\mathsf{m})}_0]\Big)(\widehat{\boldsymbol{\beta}}^{(\mathsf{m})}_{\mathsf{LASSO}} - \boldsymbol{\beta}^{(\mathsf{m})}_0)dt + \lambda\rho(\boldsymbol{\beta}^{(\bullet)}_0) \\
&=: \xi_1 + \xi_2 + \lambda\rho(\boldsymbol{\beta}^{(\bullet)}_0).
\end{aligned} \tag{S9}$$

Now we bound ξ_1 and ξ_2 using Lemma A1, in terms of $\|\widehat{\boldsymbol{\mu}}_{\mathsf{SHIR}} - \boldsymbol{\mu}_0\|_1 + \lambda_g\|\widehat{\boldsymbol{\alpha}}^{(\bullet)}_{\mathsf{SHIR}} - \boldsymbol{\alpha}^{(\bullet)}_0\|_{2,1}$. Let $\lambda_1 \geq 2\max\Big\{\lambda_{01}, \lambda_{02}/(\lambda_g M^{1/2})\Big\}$, we have that with probability approaching 1,

$$\begin{aligned}
|\xi_1| \leq& 2\left\|N^{-1}\sum_{m=1}^{M} n_m\nabla\widehat{\mathcal{L}}_m(\boldsymbol{\beta}^{(\mathsf{m})}_0)\right\|_\infty \|\widehat{\boldsymbol{\mu}}_{\mathsf{SHIR}} - \boldsymbol{\mu}_0\|_1 + 2\|\nabla\widehat{\mathcal{L}}_\bullet(\boldsymbol{\beta}^{(\bullet)}_0)\|_{2,\infty}\|\widehat{\boldsymbol{\alpha}}^{(\bullet)}_{\mathsf{SHIR}} - \boldsymbol{\alpha}^{(\bullet)}_0\|_{2,1} \\
\leq& \frac{\lambda_1}{2}(\|\widehat{\boldsymbol{\mu}}_{\mathsf{SHIR}} - \boldsymbol{\mu}_0\|_1 + \lambda_g\|\widehat{\boldsymbol{\alpha}}^{(\bullet)}_{\mathsf{SHIR}} - \boldsymbol{\alpha}^{(\bullet)}_0\|_{2,1})
\end{aligned}$$

We let $\lambda_2 = 4\max(1, \lambda_g M^{1/2})C_{\mathsf{loc}}C_L B s_0 \log p/\min_{m\in[M]} n_m$, where the constant C_{loc} satisfies $\max_{m\in[M]}\|\mathbb{X}^{(\mathsf{m})}(\widehat{\boldsymbol{\beta}}^{(\mathsf{m})}_{\mathsf{LASSO}} - \boldsymbol{\beta}^{(\mathsf{m})}_0)\|_2 \leq (C_{\mathsf{loc}} n_m/n^{\mathsf{eff}}_m)^{1/2}$ with probability approaching 1 by Condition 4. Then we have

$$\begin{aligned}
|\xi_2| \leq& 2N^{-1}\sum_{m=1}^{M} BC_L\|\mathbb{X}^{(\mathsf{m})}(\widehat{\boldsymbol{\beta}}^{(\mathsf{m})}_{\mathsf{LASSO}} - \boldsymbol{\beta}^{(\mathsf{m})}_0)\|_2^2\|\widehat{\boldsymbol{\mu}}_{\mathsf{SHIR}} - \boldsymbol{\mu}_0\|_1 \\
&+ \max_{m\in M}\|\mathbb{X}^{(\mathsf{m})}(\widehat{\boldsymbol{\beta}}^{(\mathsf{m})}_{\mathsf{LASSO}} - \boldsymbol{\beta}^{(\mathsf{m})}_0)\|_2^2 \cdot \frac{2M^{\frac{1}{2}}BC_L\|\widehat{\boldsymbol{\alpha}}^{(\bullet)}_{\mathsf{SHIR}} - \boldsymbol{\alpha}^{(\bullet)}_0\|_{2,1}}{N} \\
\leq& \frac{\lambda_2}{2}(\|\widehat{\boldsymbol{\mu}}_{\mathsf{SHIR}} - \boldsymbol{\mu}_0\|_1 + \lambda_g\|\widehat{\boldsymbol{\alpha}}^{(\bullet)}_{\mathsf{SHIR}} - \boldsymbol{\alpha}^{(\bullet)}_0\|_{2,1}).
\end{aligned}$$

Then we let $\lambda = \lambda_1 + \lambda_2$ in (S9) and see that

$$\|\widehat{\boldsymbol{\mu}}_{\mathsf{SHIR},-1}\|_1 + \lambda_g\sum_{j=2}^{p}\|\widehat{\boldsymbol{\alpha}}_{\mathsf{SHIR},j}\|_2 \leq \frac{1}{2}(\|\widehat{\boldsymbol{\mu}}_{\mathsf{SHIR}} - \boldsymbol{\mu}_0\|_1 + \lambda_g\|\widehat{\boldsymbol{\alpha}}^{(\bullet)}_{\mathsf{SHIR}} - \boldsymbol{\alpha}^{(\bullet)}_0\|_{2,1}) + \|\boldsymbol{\mu}_0\|_1 + \lambda_g\|\boldsymbol{\alpha}^{(\bullet)}_0\|_{2,1}.$$

This and $1 \in \mathcal{S}_0$ yield that

$$\|\widehat{\boldsymbol{\mu}}_{\mathsf{SHIR},\mathcal{S}_0^c}\|_1 + \lambda_g \|\widehat{\boldsymbol{\alpha}}^{(\bullet)}_{\mathsf{SHIR},\mathcal{S}_0^c}\|_{2,1} \le 3(\|\widehat{\boldsymbol{\mu}}_{\mathsf{SHIR},\mathcal{S}_0} - \boldsymbol{\mu}_{0,\mathcal{S}_0}\|_1 + \lambda_g\|\widehat{\boldsymbol{\alpha}}^{(\bullet)}_{\mathsf{SHIR},\mathcal{S}_0} - \boldsymbol{\alpha}^{(\bullet)}_{0,\mathcal{S}_0}\|_{2,1}). \tag{S10}$$

Note that $\widehat{\boldsymbol{\alpha}}^{(1)}_{\mathsf{SHIR}} - \boldsymbol{\alpha}^{(1)}_0 + \cdots + \widehat{\boldsymbol{\alpha}}^{(\mathsf{M})}_{\mathsf{SHIR}} - \boldsymbol{\alpha}^{(\mathsf{M})}_0 = \mathbf{0}$, we have $(\widehat{\boldsymbol{\mu}}^{\mathsf{T}}_{\mathsf{SHIR}} - \boldsymbol{\mu}_0^{\mathsf{T}}, \widehat{\boldsymbol{\alpha}}^{(\bullet)\mathsf{T}}_{\mathsf{SHIR}} - \boldsymbol{\alpha}_0^{(\bullet)\mathsf{T}})^{\mathsf{T}} \in \mathcal{C}_2(3,\mathcal{S}_0)$. Combining Condition 4: $\|\widehat{\boldsymbol{\beta}}^{(\mathsf{m})}_{\mathsf{LASSO}} - \boldsymbol{\beta}^{(\mathsf{m})}_0\|_2 = O_{\mathsf{P}}\{(1/n_m^{\mathsf{eff}})^{1/2}\}$ with Condition 1 yields that $\mathcal{S}_0$ and $\widehat{\mathbb{H}}$ satisfy $\mathscr{C}_{\mathsf{comp}}$. Then we have

$$\begin{aligned}\|\widehat{\mathbb{H}}^{\frac{1}{2}}(\widehat{\boldsymbol{\beta}}^{(\bullet)}_{\mathsf{SHIR}} - \boldsymbol{\beta}^{(\bullet)}_0)\|_2^2 &\le \frac{3\lambda}{2}(\|\widehat{\boldsymbol{\mu}}_{\mathsf{SHIR}} - \boldsymbol{\mu}_0\|_1 + \lambda_g\|\widehat{\boldsymbol{\alpha}}^{(\bullet)}_{\mathsf{SHIR}} - \boldsymbol{\alpha}^{(\bullet)}_0\|_{2,1})\\ &\le \frac{3\lambda}{2}\sqrt{s_0\|\widehat{\mathbb{H}}^{\frac{1}{2}}(\widehat{\boldsymbol{\beta}}^{(\bullet)}_{\mathsf{SHIR}} - \boldsymbol{\beta}^{(\bullet)}_0)\|_2^2/\phi_0}.\end{aligned}$$

Since $\lambda_g = \Theta(M^{-1/2})$ and $n_m = \Theta(N/M)$ for all $m \in [M]$, we have $\lambda = \lambda_1 + \lambda_2 = \Theta(1/(s_0 n^{\mathsf{eff}})^{1/2} + B/n_m^{\mathsf{eff}})$. Then we conclude that $\|\widehat{\mathbb{H}}^{\frac{1}{2}}(\widehat{\boldsymbol{\beta}}^{(\bullet)}_{\mathsf{SHIR}} - \boldsymbol{\beta}^{(\bullet)}_0)\|_2 = O_{\mathsf{P}}\{(1/n^{\mathsf{eff}})^{\frac{1}{2}} + Bs_0^{\frac{1}{2}}/n_m^{\mathsf{eff}}\}$. For estimation error, again by Condition 1 and using the fact that $M^{-1}\|\widehat{\boldsymbol{\beta}}^{(\bullet)}_{\mathsf{SHIR}} - \boldsymbol{\beta}^{(\bullet)}_0\|_1 = O(\|\widehat{\boldsymbol{\mu}}_{\mathsf{SHIR}} - \boldsymbol{\mu}_0\|_1 + \lambda_g\|\widehat{\boldsymbol{\alpha}}^{(\bullet)}_{\mathsf{SHIR}} - \boldsymbol{\alpha}^{(\bullet)}_0\|_{2,1})$,we have $\|\widehat{\boldsymbol{\mu}}_{\mathsf{SHIR}} - \boldsymbol{\mu}_0\|_1 + \lambda_g\|\widehat{\boldsymbol{\alpha}}^{(\bullet)}_{\mathsf{SHIR}} - \boldsymbol{\alpha}^{(\bullet)}_0\|_{2,1} = O_{\mathsf{P}}\{(s_0/n^{\mathsf{eff}})^{\frac{1}{2}} + Bs_0/n_m^{\mathsf{eff}}\}$ and $M^{-1}\|\widehat{\boldsymbol{\beta}}^{(\bullet)}_{\mathsf{SHIR}} - \boldsymbol{\beta}^{(\bullet)}_0\|_1 = O_{\mathsf{P}}\{(s_0/n^{\mathsf{eff}})^{\frac{1}{2}} + Bs_0/n_m^{\mathsf{eff}}\}$.

□

A.3.3 Proof of Theorem 2

To establish the equivalence between $\widehat{\boldsymbol{\beta}}^{(\bullet)}_{\mathsf{SHIR}}$ and $\widehat{\boldsymbol{\beta}}^{(\bullet)}_{\mathsf{IPDpool}}$, we need to compare these two estimators directly via an inequality similar to (S9), which is shown in (S13) in the following proof. The way we utilize (S13) to prove Theorem 2 is similar to (S9) in Theorem 1 but this is more elaborative since the two estimators are not necessarily as sparse as $\boldsymbol{\beta}^{(\bullet)}_0$. Specifically, based on the results and proof procedures of Theorem 1, we prove Theorem 2 as follows.

Proof. Let λ_1 and λ_2 be as defined in the proof of Theorem 1. First, using the conclusion of Negahban et al. (2012), proof of which actually implements similar steps as in the proofs of Theorem 1, we have that there exists $\tilde{\lambda} = \Theta(\lambda_1)$ as defined in the proof of Theorem 1, the $\mathsf{IPDpool}$ estimator $\widehat{\boldsymbol{\beta}}^{(\bullet)}_{\mathsf{IPDpool}}$ satisfies that

$$\|\widehat{\mathbb{H}}^{\frac{1}{2}}(\widehat{\boldsymbol{\beta}}^{(\bullet)}_{\mathsf{IPDpool}} - \boldsymbol{\beta}^{(\bullet)}_0)\|_2 = O_{\mathsf{P}}\{(1/n^{\mathsf{eff}})^{\frac{1}{2}}\}; \quad \|\widehat{\boldsymbol{\mu}}_{\mathsf{IPDpool}} - \boldsymbol{\mu}_0\|_1 + \lambda_g\|\widehat{\boldsymbol{\alpha}}^{(\bullet)}_{\mathsf{IPDpool}} - \boldsymbol{\alpha}^{(\bullet)}_0\|_{2,1} = O_{\mathsf{P}}\{(s_0/n^{\mathsf{eff}})^{\frac{1}{2}}\}.$$

To control the additional noise introduced by integrating the summarized statistics, which is characterized by λ_2 as defined in the proof of Theorem 1, λ need to be larger than $\tilde{\lambda}$ by some $\lambda_\Delta = \lambda - \tilde{\lambda} > 0$. Under the assumptions in Theorem 2, such λ_Δ can be selected to have smaller order than $\tilde{\lambda}$ but still control the aggregation noise. Thus the difference between the prediction and estimation risks of the two estimators is also of smaller order than the risks themselves. Now we demonstrate this intuition by the rigorous proofs as below.

Since $s_0 = o\{(n_m^{\mathsf{eff}})^2/(B^2 n^{\mathsf{eff}})\}$, $\lambda_2 = \Theta(B/n_m^{\mathsf{eff}})$, and $\tilde{\lambda} = \Theta\{1/(s_0 n^{\mathsf{eff}})^{1/2}\}$, we have $\lambda_2 = o(\tilde{\lambda})$. So there exists λ_Δ satisfying $\lambda_\Delta = \omega(\lambda_2)$ and $\lambda_\Delta = o(\tilde{\lambda})$. Then as N is large enough, $\lambda = \tilde{\lambda} + \lambda_\Delta \geq \lambda_1 + \lambda_2$. So by Theorem 1, we have $\|\widehat{\boldsymbol{\mu}}_{\mathsf{SHIR}} - \boldsymbol{\mu}_0\|_1 + \lambda_g \|\widehat{\boldsymbol{\alpha}}^{(\bullet)}_{\mathsf{SHIR}} - \boldsymbol{\alpha}^{(\bullet)}_0\|_{2,1} = O_{\mathsf{P}}\{(s_0/n^{\mathsf{eff}})^{\frac{1}{2}} + Bs_0/n_m^{\mathsf{eff}}\}$ and $M^{-1}\|\widehat{\boldsymbol{\beta}}^{(\bullet)}_{\mathsf{SHIR}} - \boldsymbol{\beta}^{(\bullet)}_0\|_1 = O_{\mathsf{P}}\{(s_0/n^{\mathsf{eff}})^{\frac{1}{2}} + Bs_0/n_m^{\mathsf{eff}}\}$.

Similar to Theorem 1, Taylor expansion on $\nabla\widehat{\mathcal{L}}_m(\widehat{\boldsymbol{\beta}}^{(\mathsf{m})}_{\mathsf{LASSO}})$ around the $\mathsf{IPDpool}$ $\widehat{\boldsymbol{\beta}}^{(\mathsf{m})}_{\mathsf{IPDpool}}$ yields that

$$
\begin{aligned}
\nabla\widehat{\mathcal{L}}_m(\widehat{\boldsymbol{\beta}}^{(\mathsf{m})}_{\mathsf{LASSO}}) - \widehat{\mathbb{H}}_m\widehat{\boldsymbol{\beta}}^{(\mathsf{m})}_{\mathsf{LASSO}} =& \nabla\widehat{\mathcal{L}}_m(\widehat{\boldsymbol{\beta}}^{(\mathsf{m})}) - \widehat{\mathbb{H}}_m\widehat{\boldsymbol{\beta}}^{(\mathsf{m})} \\
&+ \int_0^1 \left\{\nabla^2\widehat{\mathcal{L}}_m\left(\widehat{\boldsymbol{\beta}}^{(\mathsf{m})} + t[\widehat{\boldsymbol{\beta}}^{(\mathsf{m})}_{\mathsf{LASSO}} - \widehat{\boldsymbol{\beta}}^{(\mathsf{m})}]\right) - \widehat{\mathbb{H}}_m\right\}(\widehat{\boldsymbol{\beta}}^{(\mathsf{m})}_{\mathsf{LASSO}} - \widehat{\boldsymbol{\beta}}^{(\mathsf{m})})dt.
\end{aligned} \tag{S11}
$$

Similar to (S8) in proof of Theorem 1 and by $\lambda_2 = o(\lambda_\Delta)$, we then have

$$
\begin{aligned}
\xi_3 :=& \frac{2}{N}\sum_{m=1}^M n_m(\widehat{\boldsymbol{\beta}}^{(\mathsf{m})}_{\mathsf{SHIR}} - \widehat{\boldsymbol{\beta}}^{(\mathsf{m})}_{\mathsf{IPDpool}})^{\mathsf{T}} \int_0^1 \left\{\nabla^2\widehat{\mathcal{L}}_m\left(\widehat{\boldsymbol{\beta}}^{(\mathsf{m})} + t[\widehat{\boldsymbol{\beta}}^{(\mathsf{m})}_{\mathsf{LASSO}} - \widehat{\boldsymbol{\beta}}^{(\mathsf{m})}]\right) - \widehat{\mathbb{H}}_m\right\}(\widehat{\boldsymbol{\beta}}^{(\mathsf{m})}_{\mathsf{LASSO}} - \widehat{\boldsymbol{\beta}}^{(\mathsf{m})})dt \\
\leq& N^{-1}C_L B\left(\max_{m\in[M]}\|\mathbb{X}^{(\mathsf{m})}(\widehat{\boldsymbol{\beta}}^{(\mathsf{m})}_{\mathsf{LASSO}} - \widehat{\boldsymbol{\beta}}^{(\mathsf{m})}_{\mathsf{IPDpool}})\|_2^2\right)\|\widehat{\boldsymbol{\beta}}^{(\bullet)}_{\mathsf{SHIR}} - \widehat{\boldsymbol{\beta}}^{(\bullet)}_{\mathsf{IPDpool}}\|_1 \\
=& O_{\mathsf{P}}\left(Bs_0\log p/N\right)O_{\mathsf{P}}\{M(s_0/n^{\mathsf{eff}})^{1/2}\} = o_{\mathsf{P}}\{\lambda_\Delta(s_0/n^{\mathsf{eff}})^{1/2}\}.
\end{aligned} \tag{S12}
$$

Then by $\widehat{Q}_{\mathsf{SHIR}}(\widehat{\boldsymbol{\beta}}^{(\bullet)}_{\mathsf{SHIR}}) \leq \widehat{Q}_{\mathsf{SHIR}}(\widehat{\boldsymbol{\beta}}^{(\bullet)}_{\mathsf{IPDpool}})$, (S11) and (S12), we have

$$
\begin{aligned}
&N^{-1}\sum_{m=1}^M n_m(\widehat{\boldsymbol{\beta}}^{(\mathsf{m})}_{\mathsf{SHIR}} - \widehat{\boldsymbol{\beta}}^{(\mathsf{m})}_{\mathsf{IPDpool}})^{\mathsf{T}}\widehat{\mathbb{H}}_m(\widehat{\boldsymbol{\beta}}^{(\mathsf{m})}_{\mathsf{SHIR}} - \widehat{\boldsymbol{\beta}}^{(\mathsf{m})}_{\mathsf{IPDpool}}) + \lambda\rho_2(\widehat{\boldsymbol{\beta}}^{(\mathsf{m})}_{\mathsf{SHIR}}) \\
\leq& 2N^{-1}\sum_{m=1}^M n_m(\widehat{\boldsymbol{\beta}}^{(\mathsf{m})}_{\mathsf{IPDpool}} - \widehat{\boldsymbol{\beta}}^{(\mathsf{m})}_{\mathsf{SHIR}})^{\mathsf{T}}\nabla\widehat{\mathcal{L}}_m(\widehat{\boldsymbol{\beta}}^{(\mathsf{m})}_{\mathsf{IPDpool}}) + \xi_3 + \lambda\rho_2(\widehat{\boldsymbol{\beta}}^{(\bullet)}_{\mathsf{IPDpool}}),
\end{aligned} \tag{S13}
$$

which enables us to compare the two estimators. Note that

$$(\widehat{\boldsymbol{\mu}}_{\mathsf{IPDpool}}, \widehat{\boldsymbol{\alpha}}^{(\bullet)}_{\mathsf{IPDpool}}, \widehat{\boldsymbol{\zeta}}_{\mathsf{IPDpool}}) = \underset{\boldsymbol{\mu}, \boldsymbol{\alpha}^{(\bullet)}, \boldsymbol{\zeta}}{\operatorname{argmin}} \widehat{\mathcal{L}}(\boldsymbol{\beta}^{(\bullet)}) + \tilde{\lambda}\rho_2(\boldsymbol{\mu}, \boldsymbol{\alpha}^{(\bullet)}; \lambda_g) + \boldsymbol{\zeta}^{\mathsf{T}}(\boldsymbol{\alpha}^{(1)} + \cdots + \boldsymbol{\alpha}^{(\mathsf{M})}),$$

where $\boldsymbol{\zeta} \in \mathbb{R}^p$ is the Lagrangian multiplier for the constraint: $\boldsymbol{\alpha}^{(1)} + \cdots + \boldsymbol{\alpha}^{(\mathsf{M})} = \mathbf{0}$. By KKT condition for the above optimization problem, we have

$$\begin{pmatrix} 2\nabla_{\boldsymbol{\mu}}\widehat{\mathcal{L}}(\widehat{\boldsymbol{\beta}}^{(\bullet)}_{\mathsf{IPDpool}}) \\ 2\nabla_{\boldsymbol{\alpha}}\widehat{\mathcal{L}}(\widehat{\boldsymbol{\beta}}^{(\bullet)}_{\mathsf{IPDpool}}) \end{pmatrix} + (\lambda - \lambda_\Delta) \begin{pmatrix} \nabla_{\boldsymbol{\mu}}\rho_2(\widehat{\boldsymbol{\mu}}_{\mathsf{IPDpool}}, \widehat{\boldsymbol{\alpha}}^{(\bullet)}_{\mathsf{IPDpool}}; \lambda_g) \\ \nabla_{\boldsymbol{\alpha}}\rho_2(\widehat{\boldsymbol{\mu}}_{\mathsf{IPDpool}}, \widehat{\boldsymbol{\alpha}}^{(\bullet)}_{\mathsf{IPDpool}}; \lambda_g) \end{pmatrix} + \begin{pmatrix} \mathbf{0}_{p\times 1} \\ \widehat{\boldsymbol{\zeta}}^{(\bullet)}_{\mathsf{IPDpool}} \end{pmatrix} = \mathbf{0},$$

where $\nabla_{\boldsymbol{\mu}}\widehat{\mathcal{L}}(\boldsymbol{\beta}^{(\bullet)}) = \partial\widehat{\mathcal{L}}(\boldsymbol{\beta}^{(\bullet)})/\partial\boldsymbol{\mu}$, $\nabla_{\boldsymbol{\alpha}}\widehat{\mathcal{L}}(\boldsymbol{\beta}^{(\bullet)}) = \partial\widehat{\mathcal{L}}(\boldsymbol{\beta}^{(\bullet)})/\partial\boldsymbol{\alpha}$, $\nabla_{\boldsymbol{\mu}}\rho_2$ and $\nabla_{\boldsymbol{\alpha}}\rho_2$ are the sub-gradients of ρ_2 on $\widehat{\boldsymbol{\mu}}_{\mathsf{IPDpool}}$ and $\widehat{\boldsymbol{\alpha}}^{(\bullet)}_{\mathsf{IPDpool}}$, and $\widehat{\boldsymbol{\zeta}}^{(\bullet)}_{\mathsf{IPDpool}} = (\widehat{\boldsymbol{\zeta}}^{\mathsf{T}}_{\mathsf{IPDpool}}, \ldots, \widehat{\boldsymbol{\zeta}}^{\mathsf{T}}_{\mathsf{IPDpool}})^{\mathsf{T}}$ is the M-time replication of the Lagrangian multiplier $\widehat{\boldsymbol{\zeta}}_{\mathsf{IPDpool}}$. We note that for $j = 1$, the sub-gradient equals to 0 and for $j \in \{2, 3, \ldots, p\}$,

- $|\nabla_{\mu_j}\rho_2(\widehat{\boldsymbol{\mu}}_{\mathsf{IPDpool}}, \widehat{\boldsymbol{\alpha}}^{(\bullet)}_{\mathsf{IPDpool}}; \lambda_g)| \leq 1$, $\nabla_{\mu_j}\rho_2(\widehat{\boldsymbol{\mu}}_{\mathsf{IPDpool}}, \widehat{\boldsymbol{\alpha}}^{(\bullet)}_{\mathsf{IPDpool}}; \lambda_g) = \operatorname{sign}(\widehat{\mu}_{\mathsf{SHIR},j})$ when $\widehat{\mu}_{\mathsf{SHIR},j} \neq 0$;
- $\|\nabla_{\alpha_j}\rho_2(\widehat{\boldsymbol{\mu}}_{\mathsf{IPDpool}}, \widehat{\boldsymbol{\alpha}}^{(\bullet)}_{\mathsf{IPDpool}}; \lambda_g)\|_2 \leq \lambda_g$, $\nabla_{\boldsymbol{\alpha}_j}\rho_2(\widehat{\boldsymbol{\mu}}_{\mathsf{IPDpool}}, \widehat{\boldsymbol{\alpha}}^{(\bullet)}_{\mathsf{IPDpool}}; \lambda_g) = \lambda_g\widehat{\boldsymbol{\alpha}}_{\mathsf{IPD},j}/\|\widehat{\boldsymbol{\alpha}}_{\mathsf{IPD},j}\|_2$ when $\|\widehat{\boldsymbol{\alpha}}_{\mathsf{IPD},j}\|_2 \neq 0$.

From $\widehat{\boldsymbol{\alpha}}^{(1)}_{\mathsf{SHIR}} - \widehat{\boldsymbol{\alpha}}^{(1)}_{\mathsf{IPDpool}} + \cdots + \widehat{\boldsymbol{\alpha}}^{(\mathsf{M})}_{\mathsf{SHIR}} - \widehat{\boldsymbol{\alpha}}^{(\mathsf{M})}_{\mathsf{IPDpool}} = \mathbf{0}$, we have $(\widehat{\boldsymbol{\alpha}}^{(\bullet)\mathsf{T}}_{\mathsf{SHIR}} - \widehat{\boldsymbol{\alpha}}^{(\bullet)\mathsf{T}}_{\mathsf{IPDpool}})\widehat{\boldsymbol{\zeta}}^{(\bullet)}_{\mathsf{IPDpool}}$=0. By the sub-gradient condition and Cauchy-Schwarz inequality,

$$\begin{aligned} &\widehat{\boldsymbol{\mu}}^{\mathsf{T}}_{\mathsf{SHIR}}\nabla_{\boldsymbol{\mu}}\rho_2(\widehat{\boldsymbol{\mu}}_{\mathsf{IPDpool}}, \widehat{\boldsymbol{\alpha}}^{(\bullet)}_{\mathsf{IPDpool}}; \lambda_g) + \widehat{\boldsymbol{\alpha}}^{(\bullet)\mathsf{T}}_{\mathsf{SHIR}}\nabla_{\boldsymbol{\alpha}}\rho_2(\widehat{\boldsymbol{\mu}}_{\mathsf{IPDpool}}, \widehat{\boldsymbol{\alpha}}^{(\bullet)}_{\mathsf{IPDpool}}; \lambda_g) \\ \leq &\|\widehat{\boldsymbol{\mu}}_{\mathsf{SHIR}}\|_1 + \|\widehat{\boldsymbol{\alpha}}^{(\bullet)}_{\mathsf{SHIR}}\|_{2,1} = \rho_2(\widehat{\boldsymbol{\mu}}_{\mathsf{SHIR}}, \widehat{\boldsymbol{\alpha}}^{(\bullet)}_{\mathsf{SHIR}}; \lambda_g). \end{aligned}$$

Thus, we have

$$\begin{aligned} &-2N^{-1}\sum_{m=1}^{M} n_m(\widehat{\boldsymbol{\beta}}^{(\mathsf{m})}_{\mathsf{SHIR}} - \widehat{\boldsymbol{\beta}}^{(\mathsf{m})}_{\mathsf{IPDpool}})^{\mathsf{T}}\nabla\widehat{\mathcal{L}}_m(\widehat{\boldsymbol{\beta}}^{(\mathsf{m})}_{\mathsf{IPDpool}}) \\ =&(\lambda - \lambda_\Delta)(\widehat{\boldsymbol{\mu}}^{\mathsf{T}}_{\mathsf{SHIR}} - \widehat{\boldsymbol{\mu}}^{\mathsf{T}}_{\mathsf{IPDpool}})\nabla_{\boldsymbol{\mu}}\rho_2(\widehat{\boldsymbol{\mu}}_{\mathsf{IPDpool}}, \widehat{\boldsymbol{\alpha}}^{(\bullet)}_{\mathsf{IPDpool}}; \lambda_g) \\ &+ (\lambda - \lambda_\Delta)(\widehat{\boldsymbol{\alpha}}^{(\bullet)\mathsf{T}}_{\mathsf{SHIR}} - \widehat{\boldsymbol{\alpha}}^{(\bullet)\mathsf{T}}_{\mathsf{IPDpool}})[\nabla_{\boldsymbol{\alpha}}\rho_2(\widehat{\boldsymbol{\mu}}_{\mathsf{IPDpool}}, \widehat{\boldsymbol{\alpha}}^{(\bullet)}_{\mathsf{IPDpool}}; \lambda_g) + \widehat{\boldsymbol{\zeta}}^{(\bullet)}_{\mathsf{IPDpool}}] \\ \leq&(\lambda - \lambda_\Delta)[\rho_2(\widehat{\boldsymbol{\mu}}_{\mathsf{SHIR}}, \widehat{\boldsymbol{\alpha}}^{(\bullet)}_{\mathsf{SHIR}}; \lambda_g) - \rho_2(\widehat{\boldsymbol{\mu}}_{\mathsf{IPDpool}}, \widehat{\boldsymbol{\alpha}}^{(\bullet)}_{\mathsf{IPDpool}}; \lambda_g)]. \end{aligned}$$

Substituting this into (S13), we have

$$N^{-1}\|\widehat{\mathbb{H}}^{\frac{1}{2}}(\widehat{\boldsymbol{\beta}}^{(\bullet)}_{\mathsf{SHIR}} - \widehat{\boldsymbol{\beta}}^{(\bullet)}_{\mathsf{IPDpool}})\|_2^2 + \lambda_\Delta \rho_2(\widehat{\boldsymbol{\mu}}_{\mathsf{SHIR}}, \widehat{\boldsymbol{\alpha}}^{(\bullet)}_{\mathsf{SHIR}}; \lambda_g) \le \xi_3 + \lambda_\Delta \rho_2(\widehat{\boldsymbol{\mu}}_{\mathsf{IPDpool}}, \widehat{\boldsymbol{\alpha}}^{(\bullet)}_{\mathsf{IPDpool}}; \lambda_g). \tag{S14}$$

Consequently, by (S12), Theorem 1 and $\lambda_\Delta = o(\tilde{\lambda})$, we have

$$\begin{aligned} & N^{-1}\|\widehat{\mathbb{H}}^{\frac{1}{2}}(\widehat{\boldsymbol{\beta}}^{(\bullet)}_{\mathsf{SHIR}} - \widehat{\boldsymbol{\beta}}^{(\bullet)}_{\mathsf{IPDpool}})\|_2^2 \le \xi_3 + \lambda_\Delta \left(\|\widehat{\boldsymbol{\mu}}_{\mathsf{SHIR}} - \widehat{\boldsymbol{\mu}}_{\mathsf{IPDpool}}\|_1 + \lambda_g \|\widehat{\boldsymbol{\alpha}}^{(\bullet)}_{\mathsf{SHIR}} - \widehat{\boldsymbol{\alpha}}^{(\bullet)}_{\mathsf{IPDpool}}\|_{2,1}\right) \\ \le & o_{\mathsf{P}}\{\lambda_\Delta (s_0/n^{\mathsf{eff}})^{\frac{1}{2}}\} + \lambda_\Delta \left(\|\widehat{\boldsymbol{\mu}}_{\mathsf{SHIR}} - \boldsymbol{\mu}_0\|_1 + \lambda_g \|\widehat{\boldsymbol{\alpha}}^{(\bullet)}_{\mathsf{SHIR}} - \boldsymbol{\alpha}^{(\bullet)}_0\|_{2,1} + \|\widehat{\boldsymbol{\mu}}_{\mathsf{IPDpool}} - \boldsymbol{\mu}_0\|_1 + \lambda_g \|\widehat{\boldsymbol{\alpha}}^{(\bullet)}_{\mathsf{IPDpool}} - \boldsymbol{\alpha}^{(\bullet)}_0\|_{2,1}\right) \\ = & o_{\mathsf{P}}\{\tilde{\lambda}(s_0/n^{\mathsf{eff}})^{\frac{1}{2}}\} = o_{\mathsf{P}}(1/n^{\mathsf{eff}}). \end{aligned}$$

Thus, we finish proving the equivalence of prediction risk:

$$N^{-\frac{1}{2}}\|\widehat{\mathbb{H}}^{\frac{1}{2}}(\widehat{\boldsymbol{\beta}}^{(\bullet)}_{\mathsf{SHIR}} - \boldsymbol{\beta}^{(\bullet)}_0)\|_2 \le N^{-1}\|\widehat{\mathbb{H}}^{\frac{1}{2}}(\widehat{\boldsymbol{\beta}}^{(\bullet)}_{\mathsf{IPDpool}} - \boldsymbol{\beta}^{(\bullet)}_0)\|_2 + o_{\mathsf{P}}\{(1/n^{\mathsf{eff}})^{\frac{1}{2}}\}.$$

For estimation equivalence, we will first show by contradiction that

$$\begin{aligned} & \rho_2(\widehat{\boldsymbol{\mu}}_{\mathsf{IPD},\mathcal{S}_0} - \widehat{\boldsymbol{\mu}}_{\mathsf{SHIR},\mathcal{S}_0}, \widehat{\boldsymbol{\alpha}}^{(\bullet)}_{\mathsf{IPD},\mathcal{S}_0} - \widehat{\boldsymbol{\alpha}}_{\mathsf{SHIR},\mathcal{S}_0}; \lambda_g) \\ \le & \|\widehat{\boldsymbol{\mu}}_{\mathsf{IPD},\mathcal{S}_0} - \widehat{\boldsymbol{\mu}}_{\mathsf{SHIR},\mathcal{S}_0}\|_1 + \lambda_g \|\widehat{\boldsymbol{\alpha}}^{(\bullet)}_{\mathsf{IPD},\mathcal{S}_0} - \widehat{\boldsymbol{\alpha}}^{(\bullet)}_{\mathsf{SHIR},\mathcal{S}_0}\|_{2,1} = o_{\mathsf{P}}\{(s_0/n^{\mathsf{eff}})^{\frac{1}{2}}\}. \end{aligned}$$

We assume that there exists a subsequence of N (for simplicity, we still denote it as N) and constants $C_1 > 0$ and $0 < q < 1$ that with probability at least q,

$$\|\widehat{\boldsymbol{\mu}}_{\mathsf{SHIR},\mathcal{S}_0} - \widehat{\boldsymbol{\mu}}_{\mathsf{IPD},\mathcal{S}_0}\|_1 + \lambda_g \|\widehat{\boldsymbol{\alpha}}^{(\bullet)}_{\mathsf{SHIR},\mathcal{S}_0} - \widehat{\boldsymbol{\alpha}}^{(\bullet)}_{\mathsf{IPD},\mathcal{S}_0}\|_{2,1} \ge C_1 (s_0/n^{\mathsf{eff}})^{\frac{1}{2}}. \tag{S15}$$

Then using the error rates of the $\mathsf{IPDpool}$ and SHIR estimators, we have that there exists constant C_2 that with probability at least q,

$$\begin{aligned} & \|\widehat{\boldsymbol{\mu}}_{\mathsf{SHIR},\mathcal{S}_0^c} - \widehat{\boldsymbol{\mu}}_{\mathsf{IPD},\mathcal{S}_0^c}\|_1 + \lambda_g \|\widehat{\boldsymbol{\alpha}}^{(\bullet)}_{\mathsf{SHIR},\mathcal{S}_0^c} - \widehat{\boldsymbol{\alpha}}^{(\bullet)}_{\mathsf{IPD},\mathcal{S}_0^c}\|_{2,1} \le C_2 (s_0/n^{\mathsf{eff}})^{\frac{1}{2}} \\ \le & \frac{C_2}{C_1}(\|\widehat{\boldsymbol{\mu}}_{\mathsf{SHIR},\mathcal{S}_0} - \widehat{\boldsymbol{\mu}}_{\mathsf{IPD},\mathcal{S}_0}\|_1 + \lambda_g \|\widehat{\boldsymbol{\alpha}}^{(\bullet)}_{\mathsf{SHIR},\mathcal{S}_0} - \widehat{\boldsymbol{\alpha}}^{(\bullet)}_{\mathsf{IPD},\mathcal{S}_0}\|_{2,1}). \end{aligned}$$

Since $\widehat{\boldsymbol{\alpha}}^{(1)}_{\mathsf{SHIR}} - \widehat{\boldsymbol{\alpha}}^{(1)}_{\mathsf{IPDpool}} + \cdots + \widehat{\boldsymbol{\alpha}}^{(\mathsf{M})}_{\mathsf{SHIR}} - \widehat{\boldsymbol{\alpha}}^{(\mathsf{M})}_{\mathsf{IPDpool}} = \mathbf{0}$, $(\widehat{\boldsymbol{\mu}}^{\mathsf{T}}_{\mathsf{SHIR}} - \widehat{\boldsymbol{\mu}}^{\mathsf{T}}_{\mathsf{IPDpool}}, \widehat{\boldsymbol{\alpha}}^{(\bullet)\mathsf{T}}_{\mathsf{SHIR}} - \widehat{\boldsymbol{\alpha}}^{(\bullet)\mathsf{T}}_{\mathsf{IPDpool}})^{\mathsf{T}} \in \mathcal{C}_2(t_1, \mathcal{S}_0)$, where

$t_1 = C_2/C_1$. So using Condition 1, there exists constant $C_3 > 0$,

$$\begin{aligned} &\|\widehat{\boldsymbol{\mu}}_{\mathsf{SHIR},\mathcal{S}_0} - \widehat{\boldsymbol{\mu}}_{\mathsf{IPD},\mathcal{S}_0}\|_1 + \lambda_g \|\widehat{\boldsymbol{\alpha}}^{(\bullet)}_{\mathsf{SHIR},\mathcal{S}_0} - \widehat{\boldsymbol{\alpha}}^{(\bullet)}_{\mathsf{IPD},\mathcal{S}_0}\|_{2,1} \\ \le &\|\widehat{\boldsymbol{\mu}}_{\mathsf{SHIR}} - \widehat{\boldsymbol{\mu}}_{\mathsf{IPDpool}}\|_1 + \lambda_g \|\widehat{\boldsymbol{\alpha}}^{(\bullet)}_{\mathsf{SHIR}} - \widehat{\boldsymbol{\alpha}}^{(\bullet)}_{\mathsf{IPDpool}}\|_{2,1} \\ \le &C_3 (s_0/N)^{\frac{1}{2}} \|\widehat{\mathbb{H}}^{\frac{1}{2}}(\widehat{\boldsymbol{\beta}}^{(\bullet)}_{\mathsf{SHIR}} - \widehat{\boldsymbol{\beta}}^{(\bullet)}_{\mathsf{IPDpool}})\|_2 = o_{\mathsf{P}}\{(s_0/n^{\mathsf{eff}})^{\frac{1}{2}}\}, \end{aligned}$$

which contradicts what we assumed in (S15), as N is large enough. Thus,

$$\|\widehat{\boldsymbol{\mu}}_{\mathsf{SHIR},\mathcal{S}_0} - \widehat{\boldsymbol{\mu}}_{\mathsf{IPD},\mathcal{S}_0}\|_1 + \lambda_g \|\widehat{\boldsymbol{\alpha}}^{(\bullet)}_{\mathsf{SHIR},\mathcal{S}_0} - \widehat{\boldsymbol{\alpha}}^{(\bullet)}_{\mathsf{IPD},\mathcal{S}_0}\|_{2,1} = o_{\mathsf{P}}\{(s_0/n^{\mathsf{eff}})^{\frac{1}{2}}\}.$$

It follows that

$$\begin{aligned} &\|\widehat{\boldsymbol{\mu}}_{\mathsf{SHIR},\mathcal{S}_0} - \boldsymbol{\mu}_{0,\mathcal{S}_0}\|_1 + \lambda_g \|\widehat{\boldsymbol{\alpha}}^{(\bullet)}_{\mathsf{SHIR},\mathcal{S}_0} - \boldsymbol{\alpha}^{(\bullet)}_{0,\mathcal{S}_0}\|_{2,1} \\ \le &\|\widehat{\boldsymbol{\mu}}_{\mathsf{IPD},\mathcal{S}_0} - \boldsymbol{\mu}_{0,\mathcal{S}_0}\|_1 + \lambda_g \|\widehat{\boldsymbol{\alpha}}^{(\bullet)}_{\mathsf{IPD},\mathcal{S}_0} - \boldsymbol{\alpha}^{(\bullet)}_{0,\mathcal{S}_0}\|_{2,1} + o_{\mathsf{P}}\{(s_0/n^{\mathsf{eff}})^{\frac{1}{2}}\}. \end{aligned} \tag{S16}$$

By (S14) we have

$$\begin{aligned} &\lambda_\Delta \rho_2(\widehat{\boldsymbol{\mu}}_{\mathsf{SHIR},\mathcal{S}_0^c}, \widehat{\boldsymbol{\alpha}}^{(\bullet)}_{\mathsf{SHIR},\mathcal{S}_0^c}; \lambda_g) \\ \le &|\xi_3| + \lambda_\Delta \rho_2(\widehat{\boldsymbol{\mu}}_{\mathsf{IPD},\mathcal{S}_0^c}, \widehat{\boldsymbol{\alpha}}^{(\bullet)}_{\mathsf{IPD},\mathcal{S}_0^c}; \lambda_g) + \lambda_\Delta \rho_2(\widehat{\boldsymbol{\mu}}_{\mathsf{SHIR},\mathcal{S}_0} - \widehat{\boldsymbol{\mu}}_{\mathsf{IPD},\mathcal{S}_0}, \widehat{\boldsymbol{\alpha}}^{(\bullet)}_{\mathsf{SHIR},\mathcal{S}_0} - \widehat{\boldsymbol{\alpha}}^{(\bullet)}_{\mathsf{IPD},\mathcal{S}_0}; \lambda_g). \end{aligned}$$

Combine this with (S12) and adding the difference of intercept term to the right hand side, we have

$$\begin{aligned} &\|\widehat{\boldsymbol{\mu}}_{\mathsf{SHIR},\mathcal{S}_0^c}\|_1 + \lambda_g \|\widehat{\boldsymbol{\alpha}}^{(\bullet)}_{\mathsf{SHIR},\mathcal{S}_0^c}\|_{2,1} \\ \le &\xi_3/\lambda_\Delta + \|\widehat{\boldsymbol{\mu}}_{\mathsf{SHIR},\mathcal{S}_0} - \widehat{\boldsymbol{\mu}}_{\mathsf{IPD},\mathcal{S}_0}\|_1 + \lambda_g \|\widehat{\boldsymbol{\alpha}}^{(\bullet)}_{\mathsf{SHIR},\mathcal{S}_0} - \widehat{\boldsymbol{\alpha}}^{(\bullet)}_{\mathsf{IPD},\mathcal{S}_0}\|_{2,1} + \|\widehat{\boldsymbol{\mu}}_{\mathsf{IPD},\mathcal{S}_0^c}\|_1 + \lambda_g \|\widehat{\boldsymbol{\alpha}}^{(\bullet)}_{\mathsf{IPD},\mathcal{S}_0^c}\|_{2,1} \\ \le &o_{\mathsf{P}}\{(s_0/n^{\mathsf{eff}})^{\frac{1}{2}}\} + \|\widehat{\boldsymbol{\mu}}_{\mathsf{IPD},\mathcal{S}_0^c}\|_1 + \lambda_g \|\widehat{\boldsymbol{\alpha}}^{(\bullet)}_{\mathsf{IPD},\mathcal{S}_0^c}\|_{2,1}. \end{aligned}$$

Since $\boldsymbol{\mu}_{0,\mathcal{S}_0^c} = \mathbf{0}$ and $\boldsymbol{\alpha}_{0,\mathcal{S}_0^c} = \mathbf{0}$, we combine this with (S16) and obtain that

$$\|\widehat{\boldsymbol{\mu}}_{\mathsf{SHIR}} - \boldsymbol{\mu}_0\|_1 + \lambda_g \|\widehat{\boldsymbol{\alpha}}^{(\bullet)}_{\mathsf{SHIR}} - \boldsymbol{\alpha}^{(\bullet)}_0\|_{2,1} \le \|\widehat{\boldsymbol{\mu}}_{\mathsf{IPDpool}} - \boldsymbol{\mu}_0\|_1 + \lambda_g \|\widehat{\boldsymbol{\alpha}}^{(\bullet)}_{\mathsf{IPDpool}} - \boldsymbol{\alpha}^{(\bullet)}_0\|_{2,1} + o_{\mathsf{P}}\{(s_0/n^{\mathsf{eff}})^{\frac{1}{2}}\},$$

which finishes the proof. □

A.3.4 Excessive risk for the debiased LASSO based approaches

We outline below the key steps to derive the error rate for the debiased LASSO based estimators (Lee et al., 2017; Battey et al., 2018) introduced in Section 4.4. First, by Lee et al. (2017) and Battey et al. (2018), we have

$$\widehat{\boldsymbol{\beta}}^{(\mathsf{m})}_{\mathsf{dLASSO}} - \boldsymbol{\beta}^{(\mathsf{m})}_0 = \boldsymbol{\varphi}^{(\mathsf{m})}/\sqrt{n_m} + O_{\mathsf{P}}\{B(s_0+s_1)\log p/n_m\},$$

where $\boldsymbol{\varphi}^{(\mathsf{m})}$ is a sub-gaussian vector of mean $\mathbf{0}$ satisfying $\|\boldsymbol{\varphi}^{(\mathsf{m})}\|_{\psi_2} = \Theta(1)$. Then using the concentration results similar to Lemma A1, for $\lambda_g = \Theta(1/M^{1/2})$, we have

$$\|\widehat{\boldsymbol{\mu}}_{\mathsf{dLASSO}} - \boldsymbol{\mu}_0\|_\infty \le O_{\mathsf{P}}\{(\log p/N)^{\frac{1}{2}}\} + O_{\mathsf{P}}\{B(s_0+s_1)\log p/n_m\}$$
$$\lambda_g\|\widehat{\boldsymbol{\alpha}}^{(\bullet)}_{\mathsf{dLASSO}} - \boldsymbol{\alpha}^{(\bullet)}_0\|_{2,\infty} \le O_{\mathsf{P}}\{[(\log p + M)/N]^{\frac{1}{2}}\} + O_{\mathsf{P}}\{B(s_0+s_1)\log p/n_m\},$$

where $\widehat{\boldsymbol{\alpha}}^{(\bullet)}_{\mathsf{dLASSO}} = (\widehat{\boldsymbol{\alpha}}^{(1)\mathsf{T}}_{\mathsf{dLASSO}}, \ldots, \widehat{\boldsymbol{\alpha}}^{(\mathsf{M})\mathsf{T}}_{\mathsf{dLASSO}})^{\mathsf{T}}$. Then following a similar procedure as Theorem 4.3 of Battey et al. (2018) and Theorem 22 of Lee et al. (2017), one can obtain the following bound for both hard and soft thresholding estimators:

$$\|\widehat{\boldsymbol{\mu}}_{\mathsf{L\&B}} - \boldsymbol{\mu}_0\|_1 + \lambda_g\|\widehat{\boldsymbol{\alpha}}^{(\bullet)}_{\mathsf{L\&B}} - \boldsymbol{\alpha}^{(\bullet)}_0\|_{2,1} = O_{\mathsf{P}}\{(s_0/n^{\mathsf{eff}})^{\frac{1}{2}} + B(s_0+s_1)/n^{\mathsf{eff}}_m\}.$$

A.3.5 Proof of Theorem 3

Selection consistency (or sparsistency) of the linear model with LASSO and group LASSO penalty has been established by Zhao and Yu (2006) and Nardi et al. (2008), respectively. Compared with their proof procedures, our theoretical analysis takes into consideration of the additional aggregation noise terms bounded in (S8) and the techniques for handling the mixture penalty ρ_2. We prove Theorem 3 as follows.

Proof. For any m and $\mathcal{S}_1, \mathcal{S}_2 \subseteq [p]$, let $\boldsymbol{\alpha}^{(-1)}_{\mathcal{S}_2} = (\boldsymbol{\alpha}^{(2)\mathsf{T}}_{\mathcal{S}_2}, \ldots, \boldsymbol{\alpha}^{(\mathsf{M})\mathsf{T}}_{\mathcal{S}_2})^{\mathsf{T}}$, $\boldsymbol{\theta}_{\mathcal{S}_1,\mathcal{S}_2} = (\boldsymbol{\mu}^{\mathsf{T}}_{\mathcal{S}_1}, \boldsymbol{\alpha}^{(-1)\mathsf{T}}_{\mathcal{S}_2})^{\mathsf{T}}$, $\boldsymbol{\theta} = \boldsymbol{\theta}_{[p],[p]}$ and similarly we define $\widehat{\boldsymbol{\theta}}_{\mathsf{SHIR}}$ and $\boldsymbol{\theta}_0$. For any m and $\widehat{\boldsymbol{\theta}}_{\mathsf{SHIR}}$, after substituting $\boldsymbol{\alpha}^{(1)}$ with the remaining

$\boldsymbol{\alpha}^{(\mathsf{m})}$'s, by (S7), we can express the corresponding KKT condition as

$$2N^{-1}\mathbb{W}^{\mathsf{T}}(\widehat{\boldsymbol{\beta}}^{(\bullet)}_{\mathsf{LASSO}})\mathbb{W}(\widehat{\boldsymbol{\beta}}^{(\bullet)}_{\mathsf{LASSO}})\begin{pmatrix}\widehat{\boldsymbol{\mu}}_{\mathsf{SHIR}}-\boldsymbol{\mu}_0\\ \widehat{\boldsymbol{\alpha}}^{(-1)}_{\mathsf{SHIR}}-\boldsymbol{\alpha}^{(-1)}_0\end{pmatrix}-2\begin{pmatrix}\boldsymbol{\Upsilon}_{[p],\emptyset}\\ \boldsymbol{\Upsilon}_{\emptyset,[p]}\end{pmatrix}-2\begin{pmatrix}\boldsymbol{\Xi}_{[p],\emptyset}\\ \boldsymbol{\Xi}_{\emptyset,[p]}\end{pmatrix}+\lambda\begin{pmatrix}\boldsymbol{\eta}_{[p],\emptyset}\\ \boldsymbol{\eta}_{\emptyset,[p]}\end{pmatrix}=0, \quad \text{(S17)}$$

where the sub-gradient $\boldsymbol{\eta}=(\boldsymbol{\eta}^{\mathsf{T}}_{[p],\emptyset},\boldsymbol{\eta}^{\mathsf{T}}_{\emptyset,[p]})^{\mathsf{T}}$ and the gradients $\boldsymbol{\Upsilon}=(\boldsymbol{\Upsilon}^{\mathsf{T}}_{[p],\emptyset},\boldsymbol{\Upsilon}^{\mathsf{T}}_{\emptyset,[p]})^{\mathsf{T}}$ and $\boldsymbol{\Xi}=(\boldsymbol{\Xi}^{\mathsf{T}}_{[p],\emptyset},\boldsymbol{\Xi}^{\mathsf{T}}_{\emptyset,[p]})^{\mathsf{T}}$ are defined as follow: (i) For any $\mathcal{S}_1,\mathcal{S}_2\subseteq[p]$, denote by $\boldsymbol{\eta}_{\mathcal{S}_1,\emptyset}$ and $\boldsymbol{\eta}_{\emptyset,\mathcal{S}_2}$ the sub-gradient corresponding to $\boldsymbol{\mu}_{\mathcal{S}_1}$ and $\boldsymbol{\alpha}^{(-1)}_{\mathcal{S}_2}$, satisfying the sub-gradient condition: $\boldsymbol{\eta}_{j,\emptyset}=\text{sign}(\mu_j)$ if $\mu_j\neq 0$ and $|\boldsymbol{\eta}_{j,\emptyset}|\leq 1$ for all $j\in[p]$; $\boldsymbol{\eta}_{\emptyset,j}=\lambda_g\mathbb{T}^{\mathsf{T}}\mathbb{T}\boldsymbol{\alpha}_j/\|\boldsymbol{\alpha}_j\|_{\mathbb{T}}$ if $\boldsymbol{\alpha}_j\neq 0$ and $\|\boldsymbol{\eta}_{\emptyset,j}\|_{\widetilde{\mathbb{T}}}\leq\lambda_g$ for all $j\in[p]$. (ii) Let $\mathbf{A}$ be the transformation matrix between $\boldsymbol{\beta}^{(\bullet)}$ and $\boldsymbol{\theta}$ such that $\boldsymbol{\beta}^{(\bullet)}=\mathbf{A}\boldsymbol{\theta}$.

Then $\boldsymbol{\Upsilon}$ and $\boldsymbol{\Xi}$ defined in above equation could be written as:

$$\boldsymbol{\Upsilon}=N^{-1}\mathbf{A}^{\mathsf{T}}\begin{pmatrix}n_1\nabla\widehat{\mathcal{L}}_1(\boldsymbol{\beta}^{(1)}_0)\\ \vdots\\ n_M\nabla\widehat{\mathcal{L}}_M(\boldsymbol{\beta}^{(\mathsf{M})}_0)\end{pmatrix}\quad\text{and}\quad\boldsymbol{\Xi}=\mathbf{A}^{\mathsf{T}}\begin{pmatrix}\boldsymbol{\Psi}_1\\ \vdots\\ \boldsymbol{\Psi}_M\end{pmatrix},$$

where we denote by

$$\boldsymbol{\Psi}_m=\frac{n_m}{N}\int_0^1\{\nabla^2\widehat{\mathcal{L}}_m(\boldsymbol{\beta}^{(1)}_0+t[\widehat{\boldsymbol{\beta}}^{(\mathsf{m})}_{\mathsf{LASSO}}-\boldsymbol{\beta}^{(\mathsf{m})}_0])-\widehat{\mathbb{H}}_m\}(\widehat{\boldsymbol{\beta}}^{(\mathsf{m})}_{\mathsf{LASSO}}-\boldsymbol{\beta}^{(\mathsf{m})}_0)dt.$$

For any $\mathcal{S}_1,\mathcal{S}_2\subseteq[p]$, let $\boldsymbol{\Upsilon}_{\mathcal{S}_1,\emptyset}$ and $\boldsymbol{\Xi}_{\mathcal{S}_1,\emptyset}$ be the sub-vector of the gradients $\boldsymbol{\Upsilon}$ and $\boldsymbol{\Xi}$ corresponding to $\boldsymbol{\mu}_{\mathcal{S}_1}$ while $\boldsymbol{\Upsilon}_{\emptyset,\mathcal{S}_2}$ and $\boldsymbol{\Xi}_{\emptyset,\mathcal{S}_2}$ corresponds to $\boldsymbol{\alpha}^{(-1)}_{\mathcal{S}_2}$. Denote by $\boldsymbol{\Psi}=(\boldsymbol{\Psi}_1,\ldots,\boldsymbol{\Psi}_M)^{\mathsf{T}}$, $\boldsymbol{\Phi}_m=\left\{f_1'(\mathbf{X}^{(\mathsf{m})\mathsf{T}}_1\boldsymbol{\beta}^{(\mathsf{m})}_0,Y^{(\mathsf{m})}_1),\ldots,f_1'(\mathbf{X}^{(\mathsf{m})\mathsf{T}}_{n_m}\boldsymbol{\beta}^{(\mathsf{m})}_0,Y^{(\mathsf{m})}_{n_m})\right\}^{\mathsf{T}}$ and $\boldsymbol{\Phi}=\left(\boldsymbol{\Phi}^{\mathsf{T}}_1,\boldsymbol{\Phi}^{\mathsf{T}}_2,\ldots,\boldsymbol{\Phi}^{\mathsf{T}}_M\right)^{\mathsf{T}}$, then

$$\boldsymbol{\Upsilon}=N^{-1}\mathbf{A}^{\mathsf{T}}\mathbb{X}^{\mathsf{T}}\boldsymbol{\Phi}\quad\text{and}\quad\boldsymbol{\Xi}=\mathbf{A}^{\mathsf{T}}\boldsymbol{\Psi}.$$

Recall that $\mathcal{S}_{\mathsf{full}}=\{\mathcal{S}_\mu,\mathcal{S}_\alpha\}$. By the KKT condition in (S17) and note the fact that we can reparameterize $\boldsymbol{\beta}^{(\bullet)}$ with $\boldsymbol{\theta}$ for arbitrary $m\in[M]$ and the KKT equations are essentially equivalent with different $m\in[M]$, the event $\mathscr{O}_\mu\cap\mathscr{O}_\alpha$ holds if and only if the following events hold:

- The estimator $\widehat{\boldsymbol{\theta}}_{\mathsf{SHIR},\mathcal{S}_{\mathsf{full}}}$ obtained from

$$\widehat{\boldsymbol{\theta}}_{\mathsf{SHIR},\mathcal{S}_{\mathsf{full}}} = \boldsymbol{\theta}_{0,\mathcal{S}_{\mathsf{full}}} + N\left[\mathbb{W}^{\mathsf{T}}_{\mathcal{S}_{\mathsf{full}}}(\widehat{\boldsymbol{\beta}}^{(\bullet)}_{\mathsf{LASSO}})\mathbb{W}_{\mathcal{S}_{\mathsf{full}}}(\widehat{\boldsymbol{\beta}}^{(\bullet)}_{\mathsf{LASSO}})\right]^{-1}\left(\boldsymbol{\Upsilon}_{\mathcal{S}_{\mathsf{full}}} + \boldsymbol{\Xi}_{\mathcal{S}_{\mathsf{full}}} - \frac{\lambda}{2}\boldsymbol{\eta}_{\mathcal{S}_{\mathsf{full}}}\right), \tag{S18}$$

  satisfies that $\max\{\|\widehat{\boldsymbol{\mu}}_{\mathsf{SHIR},\mathcal{S}_\mu} - \boldsymbol{\mu}_{0,\mathcal{S}_\mu}\|_\infty, \|\widehat{\boldsymbol{\alpha}}^{(\bullet)}_{\mathsf{SHIR},\mathcal{S}_\alpha} - \boldsymbol{\alpha}^{(\bullet)}_{0,\mathcal{S}_\alpha}\|_{2,\infty}\} < \nu$.

- For any $j \in \mathcal{S}_\mu^c$, the sub-gradient $\boldsymbol{\eta}_{j,\emptyset}$ obtained from

$$\begin{aligned}\lambda\boldsymbol{\eta}_{j,\emptyset} =& 2\boldsymbol{\Upsilon}_{j,\emptyset} + 2\boldsymbol{\Xi}_{j,\emptyset}\\ &- \mathbb{W}^{\mathsf{T}}_{j,\emptyset}(\widehat{\boldsymbol{\beta}}^{(\bullet)}_{\mathsf{LASSO}})\mathbb{W}_{\mathcal{S}_{\mathsf{full}}}(\widehat{\boldsymbol{\beta}}^{(\bullet)}_{\mathsf{LASSO}})\left[\mathbb{W}^{\mathsf{T}}_{\mathcal{S}_{\mathsf{full}}}(\widehat{\boldsymbol{\beta}}^{(\bullet)}_{\mathsf{LASSO}})\mathbb{W}_{\mathcal{S}_{\mathsf{full}}}(\widehat{\boldsymbol{\beta}}^{(\bullet)}_{\mathsf{LASSO}})\right]^{-1}\left(2\boldsymbol{\Upsilon}_{\mathcal{S}_{\mathsf{full}}} + 2\boldsymbol{\Xi}_{\mathcal{S}_{\mathsf{full}}} - \lambda\boldsymbol{\eta}_{\mathcal{S}_{\mathsf{full}}}\right),\end{aligned} \tag{S19}$$

  satisfies that $|\boldsymbol{\eta}_{j,\emptyset}| < 1$.

- For any $j \in \mathcal{S}_\alpha^c$, the term $\boldsymbol{\eta}_{\emptyset,j}$ obtained from

$$\begin{aligned}\lambda\boldsymbol{\eta}_{\emptyset,j} =& 2\boldsymbol{\Upsilon}_{\emptyset,j} + 2\boldsymbol{\Xi}_{\emptyset,j}\\ &- \mathbb{W}^{\mathsf{T}}_{\emptyset,j}(\widehat{\boldsymbol{\beta}}^{(\bullet)}_{\mathsf{LASSO}})\mathbb{W}_{\mathcal{S}_{\mathsf{full}}}(\widehat{\boldsymbol{\beta}}^{(\bullet)}_{\mathsf{LASSO}})\left[\mathbb{W}^{\mathsf{T}}_{\mathcal{S}_{\mathsf{full}}}(\widehat{\boldsymbol{\beta}}^{(\bullet)}_{\mathsf{LASSO}})\mathbb{W}_{\mathcal{S}_{\mathsf{full}}}(\widehat{\boldsymbol{\beta}}^{(\bullet)}_{\mathsf{LASSO}})\right]^{-1}\left(2\boldsymbol{\Upsilon}_{\mathcal{S}_{\mathsf{full}}} + 2\boldsymbol{\Xi}_{\mathcal{S}_{\mathsf{full}}} - \lambda\boldsymbol{\eta}_{\mathcal{S}_{\mathsf{full}}}\right),\end{aligned} \tag{S20}$$

  satisfies that $\|\boldsymbol{\eta}_{\emptyset,j}\|_{\widetilde{\mathbb{T}}} < \lambda_g$.

Note that $\widehat{\boldsymbol{\theta}}_{\mathsf{SHIR},\mathcal{S}_{\mathsf{full}}}$ is the unique solution to (S17) and is the minimizer of $\widehat{Q}_{\mathsf{SHIR}}(\boldsymbol{\beta}^{(\bullet)})$ whenever (S18), (S19) and (S20) are satisfied for all j, with $\boldsymbol{\eta}$ satisfying the subgradient condition. So we only need to show that

$$\mathsf{P}(\|\widehat{\boldsymbol{\mu}}_{\mathsf{SHIR},\mathcal{S}_\mu} - \boldsymbol{\mu}_{0,\mathcal{S}_\mu}\|_\infty < \nu;\ M^{-\frac{1}{2}}\|\widehat{\boldsymbol{\alpha}}^{(\bullet)}_{\mathsf{SHIR},\mathcal{S}_\alpha} - \boldsymbol{\alpha}^{(\bullet)}_{0,\mathcal{S}_\alpha}\|_{2,\infty} < \nu) \to 1, \tag{S21}$$

and that as $N \to \infty$,

$$\mathsf{P}(\forall\ j \in \mathcal{S}_\mu^c,\ |\boldsymbol{\eta}_{j,\emptyset}| < 1;\ \forall j \in \mathcal{S}_\alpha^c,\ \|\boldsymbol{\eta}_{\emptyset,j}\|_{\widetilde{\mathbb{T}}} < \lambda_g) \to 1. \tag{S22}$$

Similar to the proof of Theorem 1, there exists constant C_Ψ such that

$$M^{-1}\sum_{m=1}^{M}\|\boldsymbol{\Psi}_m\|_\infty \le C_\Psi B/n_m^{\mathsf{eff}}; \quad \|\boldsymbol{\Psi}_m\|_\infty \le C_\Psi B/n_m^{\mathsf{eff}}. \tag{S23}$$

And in the following deductions, we base on (S18), (S19) and its corresponding sub-gradient condition of $\boldsymbol{\eta}_{\mathcal{S}_{\mathsf{full}}}$, to define $\widehat{\boldsymbol{\theta}}_{\mathsf{SHIR},\mathcal{S}_{\mathsf{full}}}$ and $\boldsymbol{\eta}$ to show (S21) and (S22). Here note that $\mathcal{S}_0 = \mathcal{S}_\mu \cup \mathcal{S}_\alpha$. For (S21), we will prove its sufficient condition:

$$\mathsf{P}(\|\widehat{\boldsymbol{\mu}}_{\mathsf{SHIR},\mathcal{S}_0} - \boldsymbol{\mu}_{0,\mathcal{S}_0}\|_\infty < \nu;\ M^{-\frac{1}{2}}\|\widehat{\boldsymbol{\alpha}}^{(\bullet)}_{\mathsf{SHIR},\mathcal{S}_0} - \boldsymbol{\alpha}^{(\bullet)}_{0,\mathcal{S}_0}\|_{2,\infty} < \nu) \to 1 \tag{S24}$$

To prove this, denote by $\widetilde{\mathcal{S}_0} = \{\mathcal{S}_0, \mathcal{S}_0\}$ and let

$$\widehat{\boldsymbol{\theta}}_{\mathsf{SHIR},\widetilde{\mathcal{S}}_0} = \boldsymbol{\theta}_{0,\widetilde{\mathcal{S}}_0} + N\left[\mathbb{W}^{\mathsf{T}}_{\widetilde{\mathcal{S}}_0}(\widehat{\boldsymbol{\beta}}^{(\bullet)}_{\mathsf{LASSO}})\mathbb{W}_{\widetilde{\mathcal{S}}_0}(\widehat{\boldsymbol{\beta}}^{(\bullet)}_{\mathsf{LASSO}})\right]^{-1}\left(\boldsymbol{\Upsilon}_{\widetilde{\mathcal{S}}_0} + \boldsymbol{\Xi}_{\widetilde{\mathcal{S}}_0} - \frac{\lambda}{2}\boldsymbol{\eta}_{\widetilde{\mathcal{S}}_0}\right).$$

Recall $\widehat{\mathbb{H}}_{m,\mathcal{S}_0} = n_m^{-1}\mathbb{X}^{(\mathsf{m})\mathsf{T}}_{\bullet\mathcal{S}_0}\boldsymbol{\Omega}_m(\widehat{\boldsymbol{\beta}}^{(\mathsf{m})}_{\mathsf{LASSO}})\mathbb{X}^{(\mathsf{m})}_{\bullet\mathcal{S}_0}$, $\boldsymbol{\eta}_\mu = \nabla_{\boldsymbol{\mu}}\rho_2(\widehat{\boldsymbol{\mu}}_{\mathsf{SHIR}}, \widehat{\boldsymbol{\alpha}}^{(\bullet)}_{\mathsf{SHIR}}; \lambda_g)$ and $\boldsymbol{\eta}_{\alpha^{(\mathsf{m})}} = \nabla_{\alpha^{(\mathsf{m})}}\rho_2(\widehat{\boldsymbol{\mu}}_{\mathsf{SHIR}}, \widehat{\boldsymbol{\alpha}}^{(\bullet)}_{\mathsf{SHIR}}; \lambda_g)$. We first get back to the KKT condition for $\widehat{\boldsymbol{\beta}}^{(\mathsf{m})}_{\mathsf{SHIR},\mathcal{S}_0}$:

$$\widehat{\boldsymbol{\beta}}^{(\mathsf{m})}_{\mathsf{SHIR},\mathcal{S}_0} = \boldsymbol{\beta}^{(\mathsf{m})}_{0,\mathcal{S}_0} + \widehat{\mathbb{H}}^{-1}_{m,\mathcal{S}_0}\left[2MN^{-1}\mathbb{X}^{(\mathsf{m})\mathsf{T}}_{\mathcal{S}_0\bullet}\boldsymbol{\Phi}_m + 2\boldsymbol{\Psi}_{m,\mathcal{S}_0} + \lambda(\boldsymbol{\eta}_{\mu,\mathcal{S}_0} + \boldsymbol{\eta}_{\alpha^{(\mathsf{m})},\mathcal{S}_0})\right]$$

Combining this with $\boldsymbol{\beta}^{(\mathsf{m})} = \boldsymbol{\mu} + \boldsymbol{\alpha}^{(\mathsf{m})}$ and $\boldsymbol{\alpha}^{(1)} + \cdots + \boldsymbol{\alpha}^{(\mathsf{M})} = \mathbf{0}$, we then have

$$\begin{aligned}
\widehat{\boldsymbol{\mu}}_{\mathsf{SHIR},\mathcal{S}_0} &= \boldsymbol{\mu}_{0,\mathcal{S}_0} + M^{-1}\sum_{m=1}^{M}\widehat{\mathbb{H}}^{-1}_{m,\mathcal{S}_0}\left[2MN^{-1}\mathbb{X}^{(\mathsf{m})\mathsf{T}}_{\mathcal{S}_0\bullet}\boldsymbol{\Phi}_m + 2\boldsymbol{\Psi}_{m,\mathcal{S}_0} + \lambda(\boldsymbol{\eta}_{\mu,\mathcal{S}_0} + \boldsymbol{\eta}_{\alpha^{(\mathsf{m})},\mathcal{S}_0})\right];\\
\widehat{\boldsymbol{\alpha}}^{(\mathsf{m})}_{\mathsf{SHIR},\mathcal{S}_0} &= \boldsymbol{\alpha}^{(\mathsf{m})}_{0,\mathcal{S}_0} + (\boldsymbol{\mu}_{0,\mathcal{S}_0} - \widehat{\boldsymbol{\mu}}_{\mathsf{SHIR},\mathcal{S}_0}) + \widehat{\mathbb{H}}^{-1}_{m,\mathcal{S}_0}\left[2MN^{-1}\mathbb{X}^{(\mathsf{m})\mathsf{T}}_{\mathcal{S}_0\bullet}\boldsymbol{\Phi}_m + 2\boldsymbol{\Psi}_{m,\mathcal{S}_0} + \lambda(\boldsymbol{\eta}_{\mu,\mathcal{S}_0} + \boldsymbol{\eta}_{\alpha^{(\mathsf{m})},\mathcal{S}_0})\right].
\end{aligned} \tag{S25}$$

Now, we base on (S25) to prove (S24). Combining Condition 5 and Condition 4 that $\|\widehat{\boldsymbol{\beta}}^{(\mathsf{m})}_{\mathsf{LASSO}} - \boldsymbol{\beta}^{(\mathsf{m})}_0\|_2 = O_{\mathsf{P}}\{(1/n_m^{\mathsf{eff}})^{1/2}\}$, we have $\Lambda_{\max}\left(\widehat{\mathbb{H}}^{-1}_{m,\mathcal{S}_0}\right) \le (C_{\min})^{-1}$ with probability approaching 1. Also, by Condition 6, $\mathbb{W}(\widehat{\boldsymbol{\beta}}^{(\bullet)}_{\mathsf{LASSO}})$ satisfies the Irrepresentable Condition $\mathscr{C}_{\mathsf{Irrep}}$ (Definition A2). Then it follows from (S8) and $\lambda_g = \Theta(M^{-1/2}) < 1$ that for $m \in [M]$,

$$\begin{aligned}
&\left\|\widehat{\mathbb{H}}_{m,\mathcal{S}_0}^{-1}\left[2\boldsymbol{\Psi}_{m,\mathcal{S}_0} + \lambda\boldsymbol{\eta}_{\mu,\mathcal{S}_0} + \lambda\boldsymbol{\eta}_{\alpha^{(\mathsf{m})},\mathcal{S}_0}\right]\right\|_\infty \leq \left\|\widehat{\mathbb{H}}_{m,\mathcal{S}_0}^{-1}\right\|_2\left(2\left\|\boldsymbol{\Psi}_{m,\mathcal{S}_0}\right\|_2 + \lambda\left\|\boldsymbol{\eta}_{\mu,\mathcal{S}_0} + \boldsymbol{\eta}_{\alpha^{(\mathsf{m})},\mathcal{S}_0}\right\|_2\right)\\
\leq &(C_{\min})^{-1}\sqrt{s_0}\left(2\left\|\boldsymbol{\Psi}_{m,\mathcal{S}_0}\right\|_\infty + \lambda\left\|\boldsymbol{\eta}_{\mu,\mathcal{S}_0} + \boldsymbol{\eta}_{\alpha^{(\mathsf{m})},\mathcal{S}_0}\right\|_\infty\right) \leq 2(C_{\min})^{-1}\sqrt{s_0}\left(\left\|\boldsymbol{\Psi}_{m,\mathcal{S}_0}\right\|_\infty + \lambda\right).
\end{aligned}\tag{S26}$$

By Condition 2 and similar to Lemma A1, we can prove the concentration result: there exists positive constant C_4 that with probability approaching 1,

$$\begin{aligned}
&\left\|M^{-1}\sum_{m=1}^{M}\widehat{\mathbb{H}}_{m,\mathcal{S}_0}^{-1}N^{-1}M\mathbb{X}_{\mathcal{S}_0\bullet}^{(\mathsf{m})\mathsf{T}}\boldsymbol{\Phi}_m\right\|_\infty \leq \frac{C_4\sqrt{s_0}}{C_{\min}}\cdot\sqrt{\frac{\log s_0}{N}} \leq \frac{C_4\sqrt{s_0}}{C_{\min}}\cdot\sqrt{\frac{\log p}{N}};\\
&\max_{j\in[s_0]} M^{-\frac{1}{2}}\sqrt{\sum_{m=1}^{M}\left(2MN^{-1}\left[\widehat{\mathbb{H}}_{m,\mathcal{S}_0}^{-1}\mathbb{X}_{\mathcal{S}_0\bullet}^{(\mathsf{m})\mathsf{T}}\boldsymbol{\Phi}_m\right]_j\right)^2} \leq \frac{C_4\sqrt{s_0}}{C_{\min}}\sqrt{\frac{M+\log s_0}{N}} \leq \frac{C_4\sqrt{s_0}}{C_{\min}}\sqrt{\frac{M+\log p}{N}}.
\end{aligned}\tag{S27}$$

By Condition 7 and combining (S23), the first equation of (S25), (S26) and the first row of (S27),

$$\begin{aligned}
&\frac{1}{\nu}\left\|\widehat{\boldsymbol{\mu}}_{\mathsf{SHIR},\mathcal{S}_0} - \boldsymbol{\mu}_{0,\mathcal{S}_0}\right\|_\infty\\
\leq&\frac{1}{\nu}\left(\left\|M^{-1}\sum_{m=1}^{M}\widehat{\mathbb{H}}_{m,\mathcal{S}_0}^{-1}N^{-1}M\mathbb{X}_{\mathcal{S}_0\bullet}^{(\mathsf{m})\mathsf{T}}\boldsymbol{\Phi}_m\right\|_\infty + M^{-1}\sum_{m=1}^{M}\left\|\widehat{\mathbb{H}}_{m,\mathcal{S}_0}^{-1}\left[2\boldsymbol{\Psi}_{m,\mathcal{S}_0} + \lambda\boldsymbol{\eta}_{\mu,\mathcal{S}_0} + \lambda\boldsymbol{\eta}_{\alpha^{(\mathsf{m})},\mathcal{S}_0}\right]\right\|_\infty\right)\\
\leq&\frac{(C_{\min})^{-1}\sqrt{s_0}}{\nu}\left[C_4\sqrt{\frac{\log p}{N}} + \frac{C_\Phi B}{n_m^{\mathsf{eff}}} + 2\lambda\right] = \frac{\sqrt{s_0}}{\nu}\Theta\left(\sqrt{\frac{\log p}{N}} + \frac{Bs_0M(\log p)}{N} + \lambda\right) \to 0,
\end{aligned}$$

with probability tending to 1. For $\widehat{\boldsymbol{\alpha}}_{\mathsf{SHIR},\mathcal{S}_0}^{(\bullet)}$, again by Condition 7 and combining (S23), the second equation of (S25), (S26) and the second row of (S27), we have that with probability tending to 1,

$$
\begin{aligned}
&\frac{1}{\sqrt{M}\nu}\|\widehat{\boldsymbol{\alpha}}^{(\bullet)}_{\mathsf{SHIR},\mathsf{IPD},\mathcal{S}_0}-\boldsymbol{\alpha}^{(\bullet)}_{0,\mathsf{IPD},\mathcal{S}_0}\|_{2,\infty}\\
\le&\frac{1}{\nu}\left\|\widehat{\boldsymbol{\mu}}_{\mathsf{SHIR},\mathcal{S}_0}-\boldsymbol{\mu}_{0,\mathcal{S}_0}\right\|_\infty+\frac{1}{\nu}\max_{j\in[s_0]}M^{-\frac{1}{2}}\sqrt{\sum_{m=1}^M\left(2MN^{-1}\left[\widehat{\mathbb{H}}^{-1}_{m,\mathcal{S}_0}\mathbb{X}^{(\mathsf{m})\mathsf{T}}_{\mathcal{S}_0\bullet}\boldsymbol{\Phi}_m\right]_j\right)^2}\\
&+\frac{1}{\sqrt{M}\nu}\sqrt{\sum_{m=1}^M\left\|\widehat{\mathbb{H}}^{-1}_{m,\mathcal{S}_0}\left[2\boldsymbol{\Psi}_{m,\mathcal{S}_0}+\lambda\boldsymbol{\eta}_{\mu,\mathcal{S}_0}+\lambda\boldsymbol{\eta}_{\alpha^{(\mathsf{m})},\mathcal{S}_0}\right]\right\|_\infty^2}\\
\le&\frac{(C_{\min})^{-1}\sqrt{s_0}}{\nu}\left[C_4\sqrt{\frac{M+\log p}{N}}+\frac{C_\Phi B}{n_m^{\mathsf{eff}}}+2\lambda\right]=\frac{\sqrt{s_0}}{\nu}\Theta\left(\sqrt{\frac{M+\log p}{N}}+\frac{Bs_0M(\log p)}{N}+\lambda\right)\to 0.
\end{aligned}
$$

Given $\mathcal{S}_0=\mathcal{S}_\mu\cup\mathcal{S}_\alpha$, these yield that

$$
\mathsf{P}(\|\widehat{\boldsymbol{\mu}}_{\mathsf{SHIR},\mathcal{S}_\mu}-\boldsymbol{\mu}_{0,\mathcal{S}_\mu}\|_\infty<\nu;\ M^{-\frac{1}{2}}\|\widehat{\boldsymbol{\alpha}}^{(\bullet)}_{\mathsf{SHIR},\mathcal{S}_\alpha}-\boldsymbol{\alpha}^{(\bullet)}_{0,\mathcal{S}_\alpha}\|_{2,\infty}<\nu)\to 1,\ \text{as } N\to\infty.
$$

Then we adopt similar approaches in Zhao and Yu (2006); Nardi et al. (2008) to bound the terms on the right hand side of (S19). Note that for any $\mathbf{x}\in\mathbb{R}^{M-1}$,

$$
\|\mathbf{x}\|^2_{\widetilde{\mathbb{T}}}=\mathbf{x}^{\mathsf{T}}(\mathbb{T}^{\mathsf{T}}\mathbb{T})^{-1}\mathbf{x}\le\|\mathbf{x}\|_2^2/\Lambda_{\min}(\mathbb{T}^{\mathsf{T}}\mathbb{T})=\|\mathbf{x}\|_2^2.
$$

Then by Lemma A1 and that $n_m=\Theta(N/M)$, there exists some constant $C_5>0$ that with probability approaching 1,

$$
\begin{aligned}
|\boldsymbol{\Upsilon}_{j,\emptyset}|\le\|N^{-1}\sum_{m=1}^M n_m\nabla\widehat{\mathcal{L}}_m(\boldsymbol{\beta}^{(\mathsf{m})}_0)\|_\infty\le C_5\lambda_{01};\\
\|\boldsymbol{\Upsilon}_{\emptyset,j}\|_{\widetilde{\mathbb{T}}}=\|\mathbb{T}(\mathbb{T}^{\mathsf{T}}\mathbb{T})^{-1}\boldsymbol{\Upsilon}_{\emptyset,j}\|_2\le 2\|\nabla\widehat{\mathcal{L}}_\bullet(\boldsymbol{\beta}^{(\bullet)}_0)\|_{2,\infty}\le C_5M^{-\frac{1}{2}}\lambda_{02}.
\end{aligned}
\tag{S28}
$$

And again using (S23), we have that for $j\in[p]$,

$$
|\boldsymbol{\Xi}_{j,\emptyset}|\le C_\Psi B/n_m^{\mathsf{eff}};\quad\|\boldsymbol{\Xi}_{\emptyset,j}\|_{\widetilde{\mathbb{T}}}\le\|\boldsymbol{\Xi}_{\emptyset,j}\|_2\le C_\Psi B/(\sqrt{M}n_m^{\mathsf{eff}}).
\tag{S29}
$$

We let $\mathbf{U} = 2\boldsymbol{\Upsilon}_{\mathcal{S}_{\mathsf{full}}^c} + 2\boldsymbol{\Xi}_{\mathcal{S}_{\mathsf{full}}^c}$ and

$$\mathbf{V} = N^{-1}\mathbb{W}^{\mathsf{T}}_{\mathcal{S}_{\mathsf{full}}^c}(\widehat{\boldsymbol{\beta}}^{(\bullet)}_{\mathsf{LASSO}})\mathbb{W}_{\mathcal{S}_{\mathsf{full}}}(\widehat{\boldsymbol{\beta}}^{(\bullet)}_{\mathsf{LASSO}})\left[N^{-1}\mathbb{W}^{\mathsf{T}}_{\mathcal{S}_{\mathsf{full}}}(\widehat{\boldsymbol{\beta}}^{(\bullet)}_{\mathsf{LASSO}})\mathbb{W}_{\mathcal{S}_{\mathsf{full}}}(\widehat{\boldsymbol{\beta}}^{(\bullet)}_{\mathsf{LASSO}})\right]^{-1}\left(2\boldsymbol{\Upsilon}_{\mathcal{S}_{\mathsf{full}}} + 2\boldsymbol{\Xi}_{\mathcal{S}_{\mathsf{full}}}\right)$$

Note that by (S28) and (S29),

$$\begin{aligned}
&(C_5\lambda_{01})^{-1}\boldsymbol{\Upsilon}_{\mathcal{S}_\mu,\emptyset} \in \mathscr{G}_{\mathcal{S}_\mu}; \quad \left[C_\Psi B/n_m^{\mathsf{eff}}\right]^{-1}\boldsymbol{\Xi}_{\mathcal{S}_\mu,\emptyset} \in \mathscr{G}_{\mathcal{S}_\mu};\\
&(C_5 M^{-\frac{1}{2}}\lambda_{02}\lambda_g^{-1})^{-1}\lambda_g^{-1}\boldsymbol{\Xi}_{\emptyset,\mathcal{S}_\alpha} \in \mathscr{G}_{\mathcal{S}_\alpha}; \quad \lambda_g^{-1}\left[C_\Psi B/(\lambda_g\sqrt{M}n_m^{\mathsf{eff}})\right]^{-1}\boldsymbol{\Xi}_{\emptyset,\mathcal{S}_\alpha} \in \mathscr{G}_{\mathcal{S}_\alpha}.
\end{aligned}$$

Then using Condition 6, we have that with probability approaching 1, for each $j \in \mathcal{S}_\mu^c$,

$$\begin{aligned}
|\mathbf{U}_{j,\emptyset}| \leq& 2C_5\lambda_{01} + 2C_\Psi B/n_m^{\mathsf{eff}};\\
|\mathbf{V}_{j,\emptyset}| \leq& 2(1-\epsilon)\max\left\{C_5\lambda_{01},\ C_5M^{-\frac{1}{2}}\lambda_{02}\lambda_g^{-1},\ C_\Psi B/n_m^{\mathsf{eff}},\ C_\Psi B/(\lambda_g\sqrt{M}n_m^{\mathsf{eff}})\right\}
\end{aligned}$$

Since $\lambda_g = \Theta(M^{-1/2})$, $\lambda_{01} = \Theta(\{\log p/N\}^{1/2})$ and $n_m = \Theta(N/M)$, we then have

$$|\mathbf{U}_{j,\emptyset}| + |\mathbf{V}_{j,\emptyset}| = O_{\mathsf{P}}\left(\sqrt{\frac{\log p + M}{N}} + \frac{Bs_0M\log p}{N}\right). \tag{S30}$$

And for $j \in [p]$, we have

$$\begin{aligned}
&\|\mathbf{U}_{\emptyset,j}\|_{\widetilde{\mathbb{T}}} \leq 2C_5M^{-\frac{1}{2}}\lambda_{02} + 2C_\Psi B/(\sqrt{M}n_m^{\mathsf{eff}});\\
&\|\mathbf{V}_{\emptyset,j}\|_{\widetilde{\mathbb{T}}} \leq 2\lambda_g(1-\epsilon)\max\left\{C_5\lambda_{01},\ C_5M^{-\frac{1}{2}}\lambda_{02}\lambda_g^{-1},\ C_\Psi B/n_m^{\mathsf{eff}},\ C_\Psi B/(\lambda_g\sqrt{M}n_m^{\mathsf{eff}})\right\}
\end{aligned}$$

with probability converging to 1. Given $\lambda_g = \Theta(M^{-1/2})$, this yields that

$$\|\mathbf{U}_{\emptyset,j}\|_{\widetilde{\mathbb{T}}} + \|\mathbf{V}_{\emptyset,j}\|_{\widetilde{\mathbb{T}}} = \lambda_g \cdot O_{\mathsf{P}}\left(\sqrt{\frac{\log p + M}{N}} + \frac{Bs_0M\log p}{N}\right). \tag{S31}$$

Then combining (S19) and (S30) and using Condition 6, $\boldsymbol{\eta}_{\mathcal{S}_\mu,\emptyset} \in \mathscr{G}_{\mathcal{S}_\mu}$, $\lambda_g^{-1}\boldsymbol{\eta}_{\emptyset,\mathcal{S}_\alpha} \in \mathscr{G}_{\mathcal{S}_\alpha}$ and

$$\frac{1}{\lambda\epsilon}\left(\sqrt{\frac{\log p + M}{N}} + \frac{Bs_0M\log p}{N}\right) \to 0,$$

we have that as N is large enough, for any $j \in \mathcal{S}_\mu^c$

$$
\begin{aligned}
|\boldsymbol{\eta}_{j,\emptyset}| =& \lambda^{-1} O_{\mathsf{P}}\left(\sqrt{\frac{\log p + M}{N}} + \frac{Bs_0 M \log p}{N}\right) \\
&+ \left| \mathbb{W}^{\mathsf{T}}_{j,\emptyset}(\widehat{\boldsymbol{\beta}}^{(\bullet)}_{\mathsf{LASSO}}) \mathbb{W}_{\mathcal{S}_{\mathsf{full}}}(\widehat{\boldsymbol{\beta}}^{(\bullet)}_{\mathsf{LASSO}}) \left[\mathbb{W}^{\mathsf{T}}_{\mathcal{S}_{\mathsf{full}}}(\widehat{\boldsymbol{\beta}}^{(\bullet)}_{\mathsf{LASSO}}) \mathbb{W}_{\mathcal{S}_{\mathsf{full}}}(\widehat{\boldsymbol{\beta}}^{(\bullet)}_{\mathsf{LASSO}})\right]^{-1} \boldsymbol{\eta}_{\mathcal{S}_{\mathsf{full}}} \right| \\
\leq& \frac{\epsilon}{2} + 1 - \epsilon = 1 - \frac{\epsilon}{2} < 1,
\end{aligned}
$$

with probability converging to 1. For any $j' \in \mathcal{S}_\alpha^c$, since $\lambda_g = \Theta(M^{-1/2})$, by (S20) and again by Condition 6, we have that for any $j \in \mathcal{S}_\mu^c$,

$$
\begin{aligned}
\lambda_g^{-1} \|\boldsymbol{\eta}_{\emptyset,j}\|_{\widetilde{\mathbb{T}}} =& \lambda^{-1} O_{\mathsf{P}}\left(\sqrt{\frac{\log p + M}{N}} + \frac{Bs_0 M \log p}{N}\right) \\
&+ \lambda_g^{-1} \left\| \mathbb{W}^{\mathsf{T}}_{\emptyset,j'}(\widehat{\boldsymbol{\beta}}^{(\bullet)}_{\mathsf{LASSO}}) \mathbb{W}_{\mathcal{S}_{\mathsf{full}}}(\widehat{\boldsymbol{\beta}}^{(\bullet)}_{\mathsf{LASSO}}) \left[\mathbb{W}^{\mathsf{T}}_{\mathcal{S}_{\mathsf{full}}}(\widehat{\boldsymbol{\beta}}^{(\bullet)}_{\mathsf{LASSO}}) \mathbb{W}_{\mathcal{S}_{\mathsf{full}}}(\widehat{\boldsymbol{\beta}}^{(\bullet)}_{\mathsf{LASSO}})\right]^{-1} \boldsymbol{\eta}_{\mathcal{S}_{\mathsf{full}}} \right\|_{\widetilde{\mathbb{T}}} \\
\leq& \frac{\epsilon}{2} + 1 - \epsilon = 1 - \frac{\epsilon}{2} < 1.
\end{aligned}
$$

Therefore, we have

$$
\mathsf{P}(\forall\ j \in \mathcal{S}_\mu^c,\ \|\boldsymbol{\eta}_{j,\emptyset}\|_\infty < 1;\ \forall j \in \mathcal{S}_\alpha^c,\ \|\boldsymbol{\eta}_{\emptyset,j}\|_{\widetilde{\mathbb{T}}} < \lambda_g) \to 1,
$$

and Theorem 3 thus follows. □

A.3.6 Technical Lemmas

In this section, we present the technical lemmas used in the proofs. Some of them are simple consequences of the existing results, and we provide brief introductions and outline their proofs.

Lemma A1. *Under Condition 2 and assume* $\log p = o(N/M)$*, there exists* $\lambda_{01} = \Theta\{(\log p/N)^{1/2}\}$ *and* $\lambda_{02} = \Theta\{[(M + \log p)/N]^{1/2}\}$ *such that, with probability approaching* 1*,*

$$
2\left\| N^{-1} \sum_{m=1}^{M} n_m \nabla \widehat{\mathcal{L}}_m(\boldsymbol{\beta}_0^{(\mathsf{m})}) \right\|_\infty \leq \lambda_{01}; \quad 2\|\nabla \widehat{\mathcal{L}}_\bullet(\boldsymbol{\beta}_0^{(\bullet)})\|_{2,\infty} \leq \lambda_{02}/M^{1/2}.
$$

Proof. Let $\mathbf{\Phi}_m := \left\{ f_1'(\mathbf{X}_i^{(\mathsf{m})\mathsf{T}} \boldsymbol{\beta}_0^{(\mathsf{m})}, Y_i^{(\mathsf{m})}) \right\}_{i=1}^{n_m}$ and $\mathbf{\Phi} = \left(\mathbf{\Phi}_1^{\mathsf{T}}, \mathbf{\Phi}_2^{\mathsf{T}}, \ldots, \mathbf{\Phi}_M^{\mathsf{T}} \right)^{\mathsf{T}}$. Note that

$$\mathsf{E}[n_m \nabla \widehat{\mathcal{L}}_m(\boldsymbol{\beta}_0^{(\mathsf{m})})] = \mathsf{E}[\mathbb{X}^{(\mathsf{m})\mathsf{T}} \mathbf{\Phi}_m] = \mathbf{0}.$$

Under Condition 2, each element of $\mathbf{X}_i^{(\mathsf{m})} f_1'(\mathbf{X}_i^{(\mathsf{m})\mathsf{T}} \boldsymbol{\beta}_0^{(\mathsf{m})}, Y_i^{(\mathsf{m})})$ is sub-Gaussian. Then by $\log p = o(N/M)$, there exists $\lambda_{01} = \Theta\{(\log p/N)^{1/2}\}$ that with probability approaching 1,

$$2 \left\| N^{-1} \sum_{m=1}^{M} n_m \nabla \widehat{\mathcal{L}}_m(\boldsymbol{\beta}_0^{(\mathsf{m})}) \right\|_\infty = \left\| 2N^{-1} \sum_{m=1}^{M} \mathbb{X}^{(\mathsf{m})\mathsf{T}} \mathbf{\Phi}_m \right\|_\infty \leq \lambda_{01}.$$

Referring to Theorem 1 of Hsu et al. (2012), under Condition 2, there exists $\lambda_{02} = \Theta\{[(\log p + M)/N]^{1/2}\}$, with probability approaching 1, $2\|\nabla \widehat{\mathcal{L}}_\bullet(\boldsymbol{\beta}_0^{(\bullet)})\|_{2,\infty} \leq \lambda_{02}/M^{1/2}$. □

We remark here that the bound of $2\|\nabla \widehat{\mathcal{L}}_\bullet(\boldsymbol{\beta}_0^{(\bullet)})\|_{2,\infty}$ relies on maximum chi-squared tail of the sub-Gaussian noise, which is different from the commonly used maximum Gaussian tail inequality, in ultra-high dimensional regime. Detailed proof of this result is given by Hsu et al. (2012). Here we provide a simplified example to intuitively explain the results in Lemma A1. Let $\boldsymbol{\epsilon}^{(\mathsf{m})} = (\epsilon_1^{(\mathsf{m})}, \ldots, \epsilon_{n_m}^{(\mathsf{m})})^{\mathsf{T}}$ and $\nabla \widehat{\mathcal{L}}_\bullet(\boldsymbol{\beta}_0^{(\bullet)}) = (\boldsymbol{\epsilon}^{(1)\mathsf{T}}, \ldots, \boldsymbol{\epsilon}^{(\mathsf{m})\mathsf{T}})^{\mathsf{T}}/N^{1/2}$, where the $\epsilon_i^{(\mathsf{m})}$ are i.i.d N$(0,1)$. For $j \in [p]$, we let $z_j = \sum_{m=1}^{M} \{\epsilon_j^{(\mathsf{m})}\}^2$. Since $z_j \sim \chi_M^2$, which is sub-exponential with mean M, we have

$$\|\nabla \widehat{\mathcal{L}}_\bullet(\boldsymbol{\beta}_0^{(\bullet)})\|_{2,\infty}^2 = \frac{\max_{j\in[p]}(z_j - M) + M}{N} \leq \frac{c \log p + M}{N},$$

for some constant c. Therefore, we have $\|\nabla \widehat{\mathcal{L}}_\bullet(\boldsymbol{\beta}_0^{(\bullet)})\|_{2,\infty} = \Theta_{\mathsf{P}}\{[(\log p + M)/N]^{1/2}\}$.

A.4. OUTLINE OF THE THEORETICAL ANALYSIS WITH OTHER PENALTY FUNCTIONS

In this section, we outline the theoretical analyses for the risk bounds of SHIR with the following penalty functions $\rho(\cdot)$. (i) Group LASSO: $\rho(\boldsymbol{\beta}^{(\bullet)}) = \sum_{j=2}^{p} \|\boldsymbol{\beta}_j\|_2$; (ii) Hierarchical LASSO (Zhou and Zhu, 2010): $\rho(\boldsymbol{\beta}^{(\bullet)}) = \sum_{j=2}^{p} \|\boldsymbol{\beta}_j\|_1^{1/2}$ and (iii) Mixture sparse penalty: $\rho(\boldsymbol{\beta}^{(\bullet)}) = \|\boldsymbol{\mu}_{-1}\|_1 + \lambda_g \sum_{m=1}^{M} \|\boldsymbol{\alpha}_{-1}^{(\mathsf{m})}\|_1$.

A.4.1 Penalty functions (i) and (iii)

We outline the technical analyses for (i) and (iii) together since they are all convex and decomposable as defined by Negahban et al. (2012). Again, start from the basic inequality (S9):

$$\begin{aligned}
&N^{-1}\sum_{m=1}^{M} n_m(\widehat{\boldsymbol{\beta}}^{(\mathsf{m})}_{\mathsf{SHIR}}-\boldsymbol{\beta}^{(\mathsf{m})}_0)^{\mathsf{T}}\widehat{\mathbb{H}}_m(\widehat{\boldsymbol{\beta}}^{(\mathsf{m})}_{\mathsf{SHIR}}-\boldsymbol{\beta}^{(\mathsf{m})}_0)+\lambda\rho(\widehat{\boldsymbol{\beta}}^{(\bullet)}_{\mathsf{SHIR}})\\
&\le -2N^{-1}\sum_{m=1}^{M} n_m(\widehat{\boldsymbol{\beta}}^{(\mathsf{m})}_{\mathsf{SHIR}}-\boldsymbol{\beta}^{(\mathsf{m})}_0)^{\mathsf{T}}\nabla\widehat{\mathcal{L}}_m(\boldsymbol{\beta}^{(\mathsf{m})}_0)+2N^{-1}\sum_{m=1}^{M} n_m(\widehat{\boldsymbol{\beta}}^{(\mathsf{m})}_{\mathsf{SHIR}}-\boldsymbol{\beta}^{(\mathsf{m})}_0)^{\mathsf{T}}\boldsymbol{\eta}^{(\mathsf{m})}_{\mathsf{SHIR}}+\lambda\rho(\boldsymbol{\beta}^{(\bullet)}_0)\\
&=: \xi_1+\xi_2+\lambda\rho(\boldsymbol{\beta}^{(\bullet)}_0),
\end{aligned}$$

where $\boldsymbol{\eta}^{(\mathsf{m})}_{\mathsf{SHIR}} := \int_0^1 \nabla^2\widehat{\mathcal{L}}_m\left(\boldsymbol{\beta}^{(\mathsf{m})}_0+t[\widehat{\boldsymbol{\beta}}^{(\mathsf{m})}_{\mathsf{LASSO}}-\boldsymbol{\beta}^{(\mathsf{m})}_0]\right)(\widehat{\boldsymbol{\beta}}^{(\mathsf{m})}_{\mathsf{LASSO}}-\boldsymbol{\beta}^{(\mathsf{m})}_0)dt$ and $\boldsymbol{\eta}^{(\bullet)}_{\mathsf{SHIR}} = (\boldsymbol{\eta}^{(1)\mathsf{T}}_{\mathsf{SHIR}},\ldots,\boldsymbol{\eta}^{(\mathsf{M})\mathsf{T}}_{\mathsf{SHIR}})^{\mathsf{T}}$. Following the paradigm for analyzing high dimensional regularized M-estimator (Bühlmann and Van De Geer, 2011; Negahban et al., 2012), one can bound ξ_1 by $|\xi_1| = O(M^{-1}\rho(\boldsymbol{\beta}^{(\bullet)})\rho^{\perp}\{\nabla\widehat{\mathcal{L}}_{\bullet}(\boldsymbol{\beta}^{(\bullet)}_0)\})$, where $\rho^{\perp}$ represents the conjugate norm of the convex and decomposable $\rho(\cdot)$. For (i), $\rho(\boldsymbol{\beta}^{(\bullet)}) = \sum_{j=2}^{p}\|\boldsymbol{\beta}_j\|_2$ and $M^{-1}\rho^{\perp}\{\nabla\widehat{\mathcal{L}}_{\bullet}(\boldsymbol{\beta}^{(\bullet)}_0)\} \simeq \|\nabla\widehat{\mathcal{L}}_{\bullet}(\boldsymbol{\beta}^{(\bullet)}_0)\|_{2,\infty}$. For (iii), we let $\lambda_g = \Theta(M^{-1/2})$ and have

$$M^{-1}\rho^{\perp}\{\nabla\widehat{\mathcal{L}}_{\bullet}(\boldsymbol{\beta}^{(\bullet)}_0)\} \simeq M^{-\frac{1}{2}}\|\nabla\widehat{\mathcal{L}}_{\bullet}(\boldsymbol{\beta}^{(\bullet)}_0)\|_{\infty}+M^{-1}\left\|\sum_{m=1}^{M}\nabla\widehat{\mathcal{L}}_m(\boldsymbol{\beta}^{(\mathsf{m})}_0)\right\|_{\infty}.$$

As a result, one can choose λ accordingly to control this term. For SHIR, we need to handle the additional error term ξ_2. Similar to $|\xi_1|$, we can bound ξ_2 by $|\xi_2| = O\{M^{-1}\rho(\boldsymbol{\beta}^{(\bullet)})\rho^{\perp}(\boldsymbol{\eta}^{(\bullet)}_{\mathsf{SHIR}})\}$. By (S8) and Condition 4, $\|\boldsymbol{\eta}^{(\bullet)}_{\mathsf{SHIR}}\|_{\infty} = O_p(1/n^{\mathsf{eff}}_m)$. Then we can further use $\|\boldsymbol{\eta}^{(\bullet)}_{\mathsf{SHIR}}\|_{\infty}$ to control $\rho^{\perp}(\boldsymbol{\eta}^{(\bullet)}_{\mathsf{SHIR}})$. For both (i) and (iii), we have $\rho^{\perp}(\boldsymbol{\eta}^{(\bullet)}_{\mathsf{SHIR}}) = O(\|\boldsymbol{\eta}^{(\bullet)}_{\mathsf{SHIR}}\|_{\infty})$. Consequently, to control the aggregation error, one can increase λ with $CM^{-1}\rho^{\perp}(\boldsymbol{\eta}^{(\bullet)}_{\mathsf{SHIR}}) = O_p(1/\{Mn^{\mathsf{eff}}_m\})$ for some large enough constant $C > 0$. Then the following procedures again fall into the paradigm of Negahban et al. (2012).

A.4.2 Penalty function (ii)

The technical details for analyzing hierarchical LASSO penalty $\rho(\boldsymbol{\beta}^{(\bullet)}) = \sum_{j=2}^{p}\|\boldsymbol{\beta}_j\|_1^{1/2}$, or the more general group bridge penalty (Huang et al., 2009), is different from (i) and (iii) because it is non-convex. Here, we follow Huang et al. (2009) and Zhou and Zhu (2010), and consider the regime where p grows in a polynomial rate of the sample size. Theorems 2 and 3 of Zhou and Zhu (2010) established that the convergence rate for the ℓ_2-error of hierarchical LASSO estimator is $(p/n)^{1/2}$. Consistent with them, we assume that $p^4/n = o(1)$ and the tuning parameter λ is taken to satisfy that $\lambda/n^{1/2} = O(1)$ and $n^{1/4}p/\lambda = o(1)$.

Roughly speaking, the proofs of Theorems 2 and 3 in Zhou and Zhu (2010) also compared their estimator and the true coefficients on the penalized loss function via the basic inequality (S9). Again, the additional challenge of analyzing SHIR is to handle $\xi_2 = 2N^{-1}\sum_{m=1}^{M} n_m(\widehat{\boldsymbol{\beta}}^{(\mathsf{m})}_{\mathsf{SHIR}} - \boldsymbol{\beta}^{(\mathsf{m})}_0)^{\mathsf{T}}\boldsymbol{\eta}^{(\mathsf{m})}_{\mathsf{SHIR}}$. Inspired by their way to deal with ξ_1, we propose to control ξ_2 by

$$|\xi_2| = O\{p^{1/2}\|\boldsymbol{\eta}^{(\bullet)}_{\mathsf{SHIR}}\|_\infty\|\widehat{\boldsymbol{\beta}}^{(\bullet)}_{\mathsf{SHIR}} - \boldsymbol{\beta}^{(\bullet)}_0\|_2\} = O_p(p^{1/2}/n^{\mathsf{eff}}_m)\cdot\|\widehat{\boldsymbol{\beta}}^{(\bullet)}_{\mathsf{SHIR}} - \boldsymbol{\beta}^{(\bullet)}_0\|_2,$$

which is equal to $o_p\{(p/n)^{1/2}\}\|\widehat{\boldsymbol{\beta}}^{(\bullet)}_{\mathsf{SHIR}} - \boldsymbol{\beta}^{(\bullet)}_0\|_2$ since it is assumed that $p^4/n = o(1)$. Then combining this with the proofs in Zhou and Zhu (2010), we obtain that the error term incurred by ξ_2 is asymptotically negligible, and consequently, SHIR has the same error rate as IPD.

A.5. ADDITIONAL TABLES AND FIGURES

In this section, we first present the pseudo-algorithm of our proposed method in Algorithm A1 and then summarize some additional simulation settings and results as supplements to the main text. In specific, under our simulation Setting (iii) and when the number of sites $M = 4$, we take $\varphi_\mu = 0.6$, $\varphi_\alpha = 0.45$, and let the coefficients $\boldsymbol{\mu}_0$ and $\boldsymbol{\alpha}_0^{(\bullet)}$ be the following:

$$\begin{aligned}
\boldsymbol{\mu}_0 &= \varphi_\mu[\mathbf{1}^\intercal_{9\times1}, -\mathbf{1}^\intercal_{9\times1}, \mathbf{0}^\intercal_{(p-18)\times1}]^\intercal;\\
\boldsymbol{\alpha}_0^{(1)} &= \varphi_\alpha[\mathbf{0}^\intercal_{6\times1}, \mathbf{0}^\intercal_{3\times1}, \quad \mathbf{1.8}^\intercal_{3\times1}, \quad \mathbf{0.7}^\intercal_{3\times1}, -\mathbf{1.3}^\intercal_{3\times1}, -\mathbf{0.85}^\intercal_{3\times1}, \mathbf{1.15}^\intercal_{3\times1}, \mathbf{0}^\intercal_{(p-24)\times1}]^\intercal;\\
\boldsymbol{\alpha}_0^{(2)} &= \varphi_\alpha[\mathbf{0}^\intercal_{6\times1}, \mathbf{0}^\intercal_{3\times1}, -\mathbf{1.8}^\intercal_{3\times1}, \quad \mathbf{1.3}^\intercal_{3\times1}, -\mathbf{0.7}^\intercal_{3\times1}, -\mathbf{1.15}^\intercal_{3\times1}, \mathbf{0.85}^\intercal_{3\times1}, \mathbf{0}^\intercal_{(p-24)\times1}]^\intercal;\\
\boldsymbol{\alpha}_0^{(3)} &= \varphi_\alpha[\mathbf{0}^\intercal_{6\times1}, -\mathbf{1.8}^\intercal_{3\times1}, \mathbf{0}^\intercal_{3\times1}, -\mathbf{0.85}^\intercal_{3\times1}, \mathbf{1.15}^\intercal_{3\times1}, \mathbf{0.7}^\intercal_{3\times1}, -\mathbf{1.3}^\intercal_{3\times1}, \mathbf{0}^\intercal_{(p-24)\times1}]^\intercal;\\
\boldsymbol{\alpha}_0^{(4)} &= \varphi_\alpha[\mathbf{0}^\intercal_{6\times1}, \quad \mathbf{1.8}^\intercal_{3\times1}, \mathbf{0}^\intercal_{3\times1}, -\mathbf{1.15}^\intercal_{3\times1}, \mathbf{0.85}^\intercal_{3\times1}, \mathbf{1.3}^\intercal_{3\times1}, -\mathbf{0.7}^\intercal_{3\times1}, \mathbf{0}^\intercal_{(p-24)\times1}]^\intercal.
\end{aligned}$$

For (iii) with $M = 8$, we set $\boldsymbol{\mu}_0$, $\boldsymbol{\alpha}_0^{(1)}$, $\boldsymbol{\alpha}_0^{(2)}$, $\boldsymbol{\alpha}_0^{(3)}$ and $\boldsymbol{\alpha}_0^{(4)}$ to be the same as above, and additionally set the rest of the coefficients as below:

$$\begin{aligned}
\boldsymbol{\alpha}_0^{(5)} &= \varphi_\alpha[\mathbf{0}^\intercal_{6\times1}, \quad \mathbf{0}^\intercal_{3\times1}, -\mathbf{1.5}^\intercal_{3\times1}, \mathbf{0.5}^\intercal_{3\times1}, -\mathbf{1.1}^\intercal_{3\times1}, \mathbf{0.8}^\intercal_{3\times1}, -\mathbf{1}^\intercal_{3\times1}, \mathbf{0}^\intercal_{(p-24)\times1}]^\intercal;\\
\boldsymbol{\alpha}_0^{(6)} &= \varphi_\alpha[\mathbf{0}^\intercal_{6\times1}, \quad \mathbf{0}^\intercal_{3\times1}, \mathbf{1.5}^\intercal_{3\times1}, \quad \mathbf{1.2}^\intercal_{3\times1}, -\mathbf{0.6}^\intercal_{3\times1}, \mathbf{0.9}^\intercal_{3\times1}, -\mathbf{0.7}^\intercal_{3\times1}, \mathbf{0}^\intercal_{(p-24)\times1}]^\intercal;\\
\boldsymbol{\alpha}_0^{(7)} &= \varphi_\alpha[\mathbf{0}^\intercal_{6\times1}, -\mathbf{1.5}^\intercal_{3\times1}, \mathbf{0}^\intercal_{3\times1}, -\mathbf{0.8}^\intercal_{3\times1}, \quad \mathbf{1}^\intercal_{3\times1}, -\mathbf{0.5}^\intercal_{3\times1}, \mathbf{1.1}^\intercal_{3\times1}, \mathbf{0}^\intercal_{(p-24)\times1}]^\intercal;\\
\boldsymbol{\alpha}_0^{(8)} &= \varphi_\alpha[\mathbf{0}^\intercal_{6\times1}, \quad \mathbf{1.5}^\intercal_{3\times1}, \mathbf{0}^\intercal_{3\times1}, -\mathbf{0.9}^\intercal_{3\times1}, \quad \mathbf{0.7}^\intercal_{3\times1}, -\mathbf{1.2}^\intercal_{3\times1}, \mathbf{0.6}^\intercal_{3\times1}, \mathbf{0}^\intercal_{(p-24)\times1}]^\intercal.
\end{aligned}$$

For Setting (iv), we let the directions of $\boldsymbol{\mu}_0$ and $\boldsymbol{\alpha}_0^{(\bullet)}$ be exactly the same as those of Setting (iii), and choose weaker signal strengths: $\varphi_\mu = 0.35$ and $\varphi_\alpha = 0.25$. Note that under Settings (iii) and (iv), the heterogeneous effects $\boldsymbol{\alpha}_0^{(\bullet)}$ show more heterogeneity than those in (i) and (ii), and the distributed model coefficients $\boldsymbol{\beta}_0^{(1)}$,...,$\boldsymbol{\beta}_0^{(M)}$ are pairwise different.

Finally, we present the true positive rate (TPR) and false discovery rate (FDR) on detecting $\boldsymbol{\beta}^{(\bullet)}$ under the simulation Settings (i)–(iv) in Figures A1 and A2, respectively. Similarly as observed in the paper, SMA performs poorly under nearly all the settings with either low TPR or high FDR, especially when $p = 800, 1500$. Both IPDpool and SHIR have good support recovery performance with all TPRs above 0.80 and FDRs below 0.13 under the strong signal setting, and all TPRs above 0.74 and FDRs below 0.05 under the weak signal setting. The $\mathsf{IPDpool}$ and SHIR attain similar TPRs and FDRs with absolute differences less than 0.02 across all settings. In comparison, $\text{Debias}_{\mathsf{L\&B}}$ has worse performance than $\mathsf{IPDpool}$ and SHIR. For example, under Setting (i), the TPR of $\text{Debias}_{\mathsf{L\&B}}$ is consistently lower than that of SHIR by about 0.13 while the FDR of $\text{Debias}_{\mathsf{L\&B}}$ is generally higher than that of SHIR, except for the case when $p = 100$ where $\text{Debias}_{\mathsf{L\&B}}$ attains very low FDR due to over shrinkage. Under the weak and sparse signal Setting (ii) with $M = 4$, $\text{Debias}_{\mathsf{L\&B}}$ is substantially less powerful than SHIR in recovering true signals (lower TPR by around 0.52), while its average FDR is comparable to that of SHIR. When $M = 8$, $\text{Debias}_{\mathsf{L\&B}}$ attains TPR comparable to that of SHIR but generally has substantially higher FDR.

Algorithm A1 SHIR Method.

```
Input: Observed individual data $\{\mathbf{X}^{\mathsf{(m)}}, \mathbf{Y}^{\mathsf{(m)}}\}$ at the $m^{\mathsf{th}}$ local site for $m \in [M]$.
  • For $m \in [M]$, at the $m$-th local site:
      1. Fit $\widehat{\boldsymbol{\beta}}^{\mathsf{(m)}}_{\mathsf{LASSO}} = \text{argmin}_{\boldsymbol{\beta}^{\mathsf{(m)}}} \widehat{\mathcal{L}}_m(\boldsymbol{\beta}^{\mathsf{(m)}}) + \lambda_m \|\boldsymbol{\beta}^{\mathsf{(m)}}_{-1}\|_1$;
      2. Calculate $\widehat{\mathbb{H}}_m = \nabla^2 \widehat{\mathcal{L}}_m(\widehat{\boldsymbol{\beta}}^{\mathsf{(m)}}_{\mathsf{LASSO}})$ and $\widehat{\mathbf{g}}_m = \widehat{\mathbb{H}}_m \widehat{\boldsymbol{\beta}}^{\mathsf{(m)}}_{\mathsf{LASSO}} - \nabla \widehat{\mathcal{L}}_m(\widehat{\boldsymbol{\beta}}^{\mathsf{(m)}}_{\mathsf{LASSO}})$. Send the summary
         statistics $\widehat{\mathcal{D}}_m = \{n_m, \widehat{\mathbb{H}}_m, \widehat{\mathbf{g}}_m\}$ to the central node.
  • At the central node, obtain $\widehat{\boldsymbol{\beta}}^{(\bullet)}_{\mathsf{SHIR}}$ by minimizing:
        $\widehat{Q}_{\mathsf{SHIR}}(\boldsymbol{\beta}^{(\bullet)}) = N^{-1} \sum_{m=1}^{M} n_m \left\{ \boldsymbol{\beta}^{\mathsf{(m)T}} \widehat{\mathbb{H}}_m \boldsymbol{\beta}^{\mathsf{(m)}} - 2\boldsymbol{\beta}^{\mathsf{(m)T}} \widehat{\mathbf{g}}_m \right\} + \lambda \rho(\boldsymbol{\beta}^{(\bullet)}).$
Output: The SHIR estimator $\widehat{\boldsymbol{\beta}}^{(\bullet)}_{\mathsf{SHIR}}$.
```

REFERENCES

Battey, H., Fan, J., Liu, H., Lu, J., Zhu, Z., et al. (2018). Distributed testing and estimation under sparse high dimensional models. *The Annals of Statistics*, 46(3):1352–1382.

Bühlmann, P. and Van De Geer, S. (2011). *Statistics for high-dimensional data: methods, theory and applications.* Springer Science & Business Media.

Feng, C., Wang, H., Chen, T., Tu, X. M., et al. (2014). On exact forms of taylors theorem for vector-valued functions. *Biometrika*, 101(4):1003–1003.

Hsu, D., Kakade, S., and Zhang, T. (2012). A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability*, 17.

Huang, J., Ma, S., Xie, H., and Zhang, C.-H. (2009). A group bridge approach for variable selection. *Biometrika*, 96(2):339–355.

Lee, J. D., Liu, Q., Sun, Y., and Taylor, J. E. (2017). Communication-efficient sparse regression. *Journal of Machine Learning Research*, 18(5):1–30.

Mendelson, S., Pajor, A., and Tomczak-Jaegermann, N. (2007). Reconstruction and subgaussian operators in asymptotic geometric analysis. *Geometric and Functional Analysis*, 17(4):1248–1282.

Mendelson, S., Pajor, A., and Tomczak-Jaegermann, N. (2008). Uniform uncertainty principle for bernoulli and subgaussian ensembles. *Constructive Approximation*, 28(3):277–289.

Nardi, Y., Rinaldo, A., et al. (2008). On the asymptotic properties of the group lasso estimator for linear models. *Electronic Journal of Statistics*, 2:605–633.

Negahban, S. N., Ravikumar, P., Wainwright, M. J., Yu, B., et al. (2012). A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557.

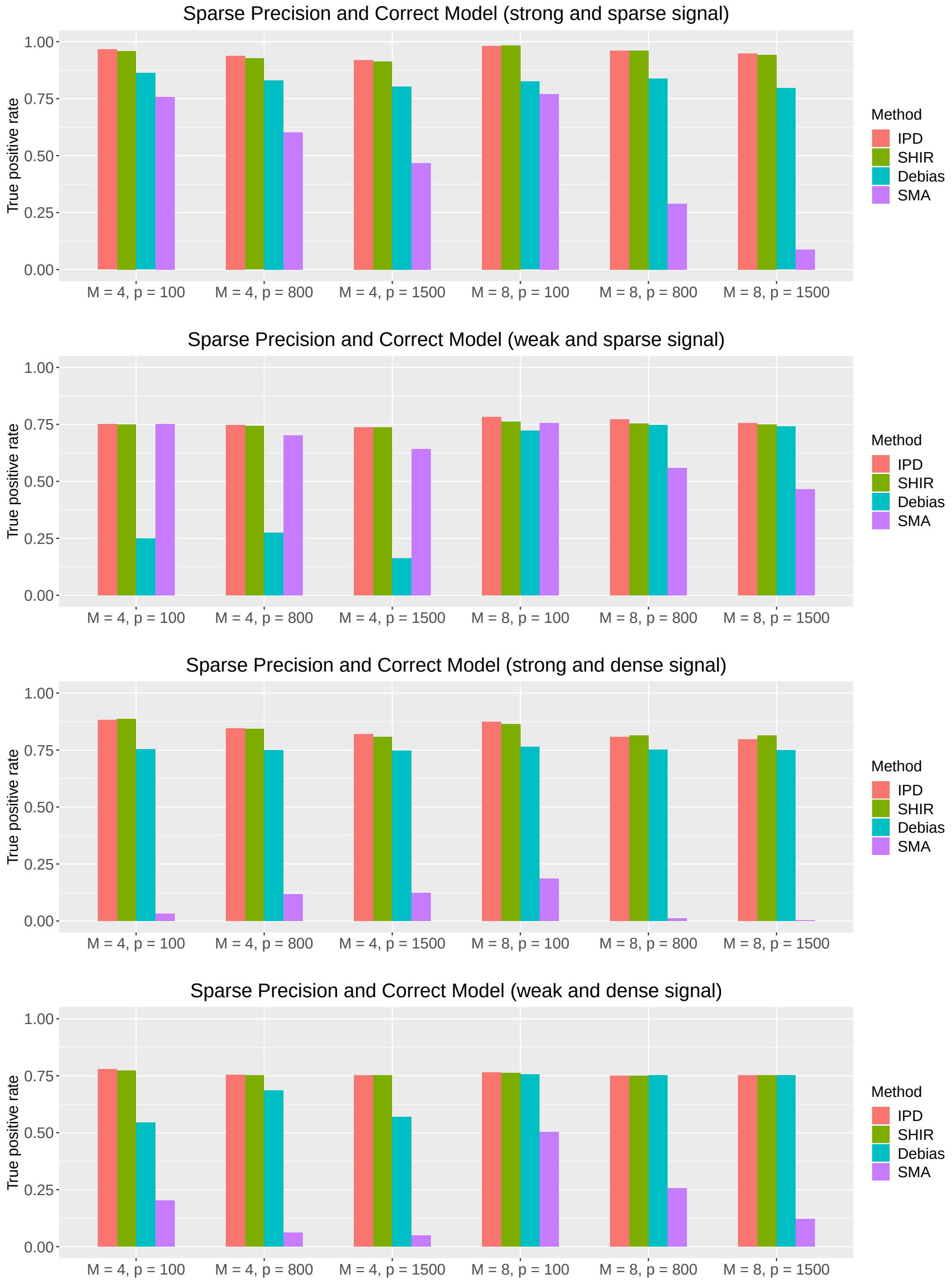
Rivasplata, O. (2012). Subgaussian random variables: An expository note. *Internet publication, PDF.*

Rudelson, M. and Zhou, S. (2012). Reconstruction from anisotropic random measurements. In *Conference on Learning Theory*, pages 10–1.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine learning research*, 7(Nov):2541–2563.

Zhou, N. and Zhu, J. (2010). Group variable selection via a hierarchical lasso and its oracle property. *arXiv preprint arXiv:1006.2871.*

Figure A1: The average true positive rate (TPR) on the original coefficients $\boldsymbol{\beta}^{(\bullet)}$ of IPDpool (IPD), SHIR, $\text{Debias}_{\text{L\&B}}$ (Debias) and SMA, different $M \in \{4, 8\}$, $p \in \{100, 800, 1500\}$ and data generation mechanisms (i)–(iv) introduced in Section 5.



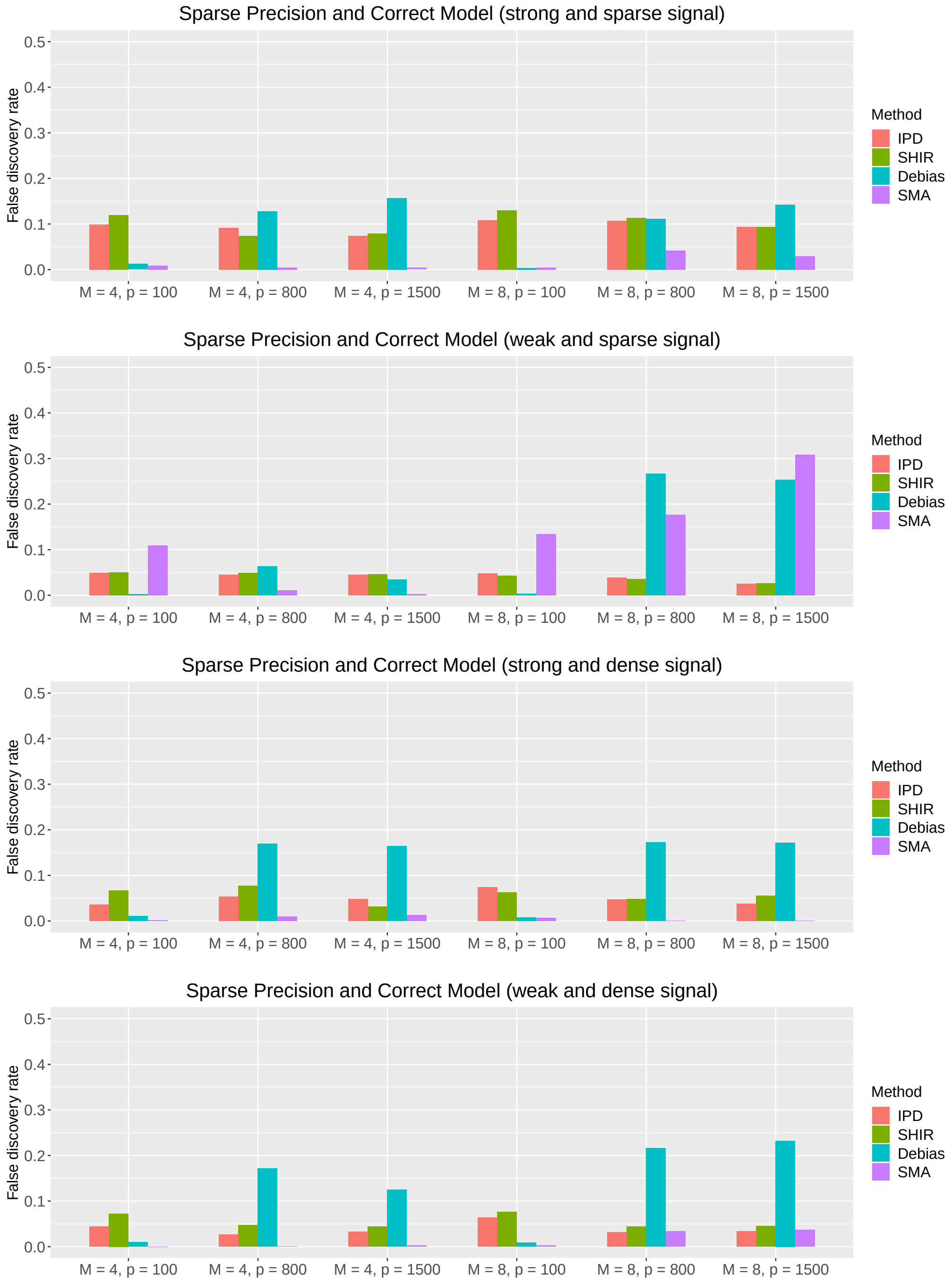
Sparse Precision and Correct Model (strong and sparse signal)

Sparse Precision and Correct Model (weak and sparse signal)

Sparse Precision and Correct Model (strong and dense signal)

Sparse Precision and Correct Model (weak and dense signal)

Figure A2: The average false discovery rate (FDR) on the original coefficients $\boldsymbol{\beta}^{(\bullet)}$ of IPDpool (IPD), SHIR, Debias$_{\mathrm{L\&B}}$ (Debias) and SMA, different $M \in \{4, 8\}$, $p \in \{100, 800, 1500\}$ and data generation mechanisms (i)–(iv) introduced in Section 5.


Sparse Precision and Correct Model (strong and sparse signal)
Sparse Precision and Correct Model (weak and sparse signal)
Sparse Precision and Correct Model (strong and dense signal)
Sparse Precision and Correct Model (weak and dense signal)
False discovery rate
M = 4, p = 100
M = 4, p = 800
M = 4, p = 1500
M = 8, p = 100
M = 8, p = 800
M = 8, p = 1500
Method
IPD
SHIR
Debias
SMA