HIGH-DIMENSIONAL FACTOR REGRESSION FOR HETEROGENEOUS SUBPOPULATIONS

Peiyao Wang[1], Quefeng Li[1], Dinggang Shen[1,2] and Yufeng Liu[1]

[1] *University of North Carolina at Chapel Hill and* [2] *Korea University*

Supplementary Materials

Section S1 gives proofs of Theorems 1–4, Corollaries 1.1–4.1, and the supporting Lemmas. Section S2 provides a rule of thumb to choose between our proposed model and the group-specific model in practice. Section S3 presents additional simulation results. Section S4 contains additional results from the ADNI data analysis. Section S5 gives the results when we apply our method to a combined microarray dataset.

S1 Proofs

S1.1 Proof of Theorem 1

For (a), we have

$$\begin{aligned}\hat{\boldsymbol{\gamma}}_g - \mathbf{H}_g\boldsymbol{\gamma}_g^* &= \frac{1}{n_g}(\hat{\mathbf{F}}_g - \mathbf{F}_g\mathbf{H}_g')'\boldsymbol{Y}_g + \mathbf{H}_g(\frac{1}{n_g}\mathbf{F}_g'\mathbf{F}_g - \mathbf{I})\boldsymbol{\gamma}_g^* + \frac{1}{n_g}\mathbf{H}_g\mathbf{F}_g'\mathbf{U}_g\boldsymbol{\beta}^* \\ &+ \frac{1}{n_g}\mathbf{H}_g\mathbf{F}_g'\boldsymbol{\epsilon}_g \\ &= I + II + III + IV.\end{aligned}$$

Since $\mathbb{E}(y_{g,i}^2) < \infty$, we have $(1/n_g)\sum_{i=1}^{n_g} y_{g,i}^2 = O_P(1)$. Moreover, it follows from (4.1) that $\|\boldsymbol{\gamma}_g^*\| \leq C\mathbb{E}(y_{g,i}^2) < \infty$ and $\|\boldsymbol{\beta}^*\| \leq C\mathbb{E}(y_{g,i}^2) < \infty$ for some $C > 0$. These results together with Lemma 1 imply that

$$\|I\| \leq \sqrt{\frac{1}{n_g}\sum_{i=1}^{n_g}\|\hat{\mathbf{f}}_{g,i} - \mathbf{H}_g\mathbf{f}_{g,i}\|^2 \frac{1}{n_g}\sum_{i=1}^{n_g} y_{g,i}^2} = O_P(\frac{1}{\sqrt{n_g}} + \frac{1}{\sqrt{p}}).$$

Similarly, by Lemma 2, we have

$$\|II\| \leq \|\mathbf{H}_g\|\sqrt{\sum_{k=1}^{K_g}(\frac{1}{n_g}\sum_{i=1}^{n_g} f_{g,ik}\mathbf{f}'_{g,i}\boldsymbol{\gamma}_g^* - \gamma_{g,k}^*)^2} = O_P(\frac{1}{\sqrt{n_g}}\|\boldsymbol{\gamma}_g^*\|),$$

$$\|III\| \leq \frac{1}{n_g}\|\mathbf{H}_g\|\sqrt{\sum_{k=1}^{K_g}(\sum_{i=1}^{n_g} f_{g,ik}\mathbf{u}'_{g,i}\boldsymbol{\beta}^*)^2} = O_P(\frac{1}{\sqrt{n_g}}\|\boldsymbol{\beta}^*\|),$$

$$\|IV\| \leq \frac{1}{n_g}\|\mathbf{H}_g\|\sqrt{\sum_{k=1}^{k_g}(\sum_{i=1}^{n_g} f_{g,ik}\epsilon_{g,i})^2} = O_P(\frac{1}{\sqrt{n_g}}).$$

Hence we conclude that

$$\|\hat{\gamma}_g - \mathbf{H}_g\boldsymbol{\gamma}_g^*\| = O_P(\frac{1}{\sqrt{n_g}} + \frac{1}{\sqrt{p}}).$$

For (b), let $\boldsymbol{\beta}_\lambda^* = \boldsymbol{\Sigma}_{u,\lambda}^{-1}\boldsymbol{\Sigma}_u\boldsymbol{\beta}^*, \boldsymbol{\Sigma}_{u,\lambda} = \boldsymbol{\Sigma}_u + 2\lambda\mathbf{I}$, and $\hat{\boldsymbol{\Sigma}}_{u,\lambda} = \hat{\boldsymbol{\Sigma}}_u + 2\lambda\mathbf{I}$. Since $\hat{\boldsymbol{\beta}}_\lambda^{ridge} = \hat{\boldsymbol{\Sigma}}_{u,\lambda}^{-1}\hat{\mathbf{U}}'\boldsymbol{Y}/n$, we have

$$\|\hat{\boldsymbol{\beta}}_\lambda^{ridge} - \boldsymbol{\beta}\| \leq \|\hat{\boldsymbol{\beta}}_\lambda^{ridge} - \boldsymbol{\beta}_\lambda^*\| + \|\boldsymbol{\beta}_\lambda^* - \boldsymbol{\beta}^*\|.$$

Note that,

$$\begin{aligned}\hat{\boldsymbol{\beta}}_\lambda^{ridge} - \boldsymbol{\beta}_\lambda^* &= (\hat{\boldsymbol{\Sigma}}_{u,\lambda}^{-1} - \boldsymbol{\Sigma}_{u,\lambda}^{-1})(\hat{\mathbf{U}} - \mathbf{U})'\boldsymbol{Y}/n + (\hat{\boldsymbol{\Sigma}}_{u,\lambda}^{-1} - \boldsymbol{\Sigma}_{u,\lambda}^{-1})\mathbf{U}'\boldsymbol{Y}/n \\ &+ \boldsymbol{\Sigma}_{u,\lambda}^{-1}(\hat{\mathbf{U}} - \mathbf{U})'\boldsymbol{Y}/n + \boldsymbol{\Sigma}_{u,\lambda}^{-1}(\frac{1}{n}\mathbf{U}'\boldsymbol{Y} - \boldsymbol{\Sigma}_u\boldsymbol{\beta}^*) \\ &= I + II + III + IV.\end{aligned}$$

By Lemma 4 (a), we have

$$\begin{aligned}\|(\hat{\mathbf{U}}_g - \mathbf{U}_g)'\boldsymbol{Y}_g\| = \sqrt{\sum_{j=1}^p \Big\{\sum_{i=1}^{n_g}(\hat{u}_{g,ij} - u_{g,ij})y_{g,i}\Big\}^2} &\le \sqrt{p\max_{i,j}|\hat{u}_{g,ij} - u_{g,ij}|^2\sum_{i=1}^{n_g} y_{g,i}^2} \\ &= O_P(\sqrt{p\log n_g \log p} + n_g^{3/4}).\end{aligned}$$

Hence,

$$\|(\hat{\mathbf{U}} - \mathbf{U})'\boldsymbol{Y}\| \le \sum_{g=1}^G \|(\hat{\mathbf{U}}_g - \mathbf{U}_g)'\boldsymbol{Y}_g\| = O_P(\sqrt{p\log n_{\max}\log p} + n_{\max}^{3/4}). \tag{S1.1}$$

By Lemma 3, we have

$$\|\mathbf{U}_g'\mathbf{F}_g\boldsymbol{\gamma}_g^*\| = \sqrt{\sum_{j=1}^p \Big\{\sum_{i=1}^{n_g} u_{g,ij}\mathbf{f}_{g,i}'\boldsymbol{\gamma}_g^*\Big\}^2} = O_P(\sqrt{n_g p}\|\boldsymbol{\gamma}_g^*\|),$$

$$\begin{aligned}\|(\frac{1}{n_g}\mathbf{U}_g'\mathbf{U}_g - \boldsymbol{\Sigma}_u)\boldsymbol{\beta}^*\| &= \sqrt{\sum_{j=1}^p \left\{\frac{1}{n_g}\sum_{i=1}^{n_g} u_{g,ij}\mathbf{u}_{g,i}'\boldsymbol{\beta}^* - \sum_{\ell=1}^p \sigma_{u,j\ell}\beta_\ell^*\right\}^2} \\ &= O_P(\sqrt{\frac{p}{n_g}}\|\boldsymbol{\beta}^*\|),\end{aligned} \tag{S1.2}$$

$$\|\mathbf{U}_g'\boldsymbol{\epsilon}_g\| = \sqrt{\sum_j \Big\{\sum_{i=1}^{n_g} u_{g,ij}\epsilon_{g,i}\Big\}^2} = O_P(\sqrt{n_g p}),$$

$$\|\frac{1}{n_g}\mathbf{U}_g'\mathbf{U}_g\boldsymbol{\beta}^*\| \leq \|(\frac{1}{n_g}\mathbf{U}_g'\mathbf{U}_g - \boldsymbol{\Sigma}_u)\boldsymbol{\beta}^*\| + \|\boldsymbol{\Sigma}_u\boldsymbol{\beta}^*\| = O_P\left((\sqrt{\frac{p}{n_g}}+1)\|\boldsymbol{\beta}^*\|\right).$$

Noting that $\|\boldsymbol{\gamma}_g^*\| = O(1)$ and $\|\boldsymbol{\beta}^*\| = O(1)$, we have

$$\|\mathbf{U}'\boldsymbol{Y}\| \leq \sum_{g=1}^{G}\|\mathbf{U}_g'\mathbf{F}_g\boldsymbol{\gamma}_g^*\| + \sum_{g=1}^{G}\|\mathbf{U}_g'\mathbf{U}_g\boldsymbol{\beta}^*\| + \sum_{g=1}^{G}\|\mathbf{U}_g'\boldsymbol{\epsilon}_g\| = O_P(\sqrt{n_{\max}p} + n_{\max}). \quad \text{(S1.3)}$$

$$\begin{aligned}\|\frac{1}{n}\mathbf{U}'\boldsymbol{Y} - \boldsymbol{\Sigma}_u\boldsymbol{\beta}^*\| &\leq \frac{1}{n}\sum_{g=1}^{G}\|\mathbf{U}_g'\mathbf{F}_g\boldsymbol{\gamma}_g^*\| + \sum_{g=1}^{G}\frac{n_g}{n}\|(\frac{1}{n_g}\mathbf{U}_g'\mathbf{U}_g - \boldsymbol{\Sigma}_u)\boldsymbol{\beta}^*\| \\ &+ \frac{1}{n}\sum_{g=1}^{G}\|\mathbf{U}_g'\boldsymbol{\epsilon}_g\| \qquad \text{(S1.4)} \\ &= O_P(\frac{\sqrt{n_{\max}p}}{n}).\end{aligned}$$

It follows from Theorem 1 of Fan et al. (2013) that $\|\hat{\boldsymbol{\Sigma}}_u - \boldsymbol{\Sigma}_u\| = O_P(m_p\omega_n)$. This result together with (S1.1) and (S1.3) implies that

$$\|I\| \leq \|\hat{\boldsymbol{\Sigma}}_{u,\lambda}^{-1} - \boldsymbol{\Sigma}_{u,\lambda}^{-1}\|\|\frac{1}{n}(\hat{\mathbf{U}} - \mathbf{U})'\boldsymbol{Y}\| = O_P\left(m_p\omega_n(\frac{\sqrt{p\log n_{\max}\log p}}{n} + \frac{n_{\max}^{3/4}}{n})\right),$$

$$\|II\| \leq \|\hat{\boldsymbol{\Sigma}}_{u,\lambda}^{-1} - \boldsymbol{\Sigma}_{u,\lambda}^{-1}\|\|\frac{1}{n}\mathbf{U}'\boldsymbol{Y}\| = O_P(m_p\omega_n(\frac{\sqrt{n_{\max}p}}{n} + \frac{n_{\max}}{n})),$$

where we use the fact that

$$\|\hat{\boldsymbol{\Sigma}}_{u,\lambda}^{-1} - \boldsymbol{\Sigma}_{u,\lambda}^{-1}\| = O_P(\|\hat{\boldsymbol{\Sigma}}_{u,\lambda} - \boldsymbol{\Sigma}_{u,\lambda}\|) = O_P(\|\hat{\boldsymbol{\Sigma}}_u - \boldsymbol{\Sigma}_u\|) = O_P(m_p\omega_n),$$

Then, applying the Weyl's Theorem with the stated choice of λ gives

$$\|\boldsymbol{\Sigma}_{u,\lambda}^{-1}\| = \lambda_{\max}(\boldsymbol{\Sigma}_{u,\lambda}^{-1}) = 1/\lambda_{\min}(\boldsymbol{\Sigma}_{u,\lambda}) \leq \frac{1}{2\lambda + C_{\min}} = O(1).$$

This result together with (S1.1) and (S1.4) implies that

$$\|III\| \le \|\mathbf{\Sigma}_{u,\lambda}^{-1}\|\|\frac{1}{n}(\hat{\mathbf{U}} - \mathbf{U})'\boldsymbol{Y}\| = O_P(\frac{\sqrt{p\log n_{\max}\log p}}{n} + \frac{n_{\max}^{3/4}}{n}),$$

$$\|IV\| \le \|\mathbf{\Sigma}_{u,\lambda}^{-1}\|\|\frac{1}{n}\mathbf{U}'\boldsymbol{Y} - \mathbf{\Sigma}_u\boldsymbol{\beta}^*\| = O_P(\frac{\sqrt{n_{\max}p}}{n}).$$

Since $m_p\omega_n = o(1)$, we have

$$\|\hat{\boldsymbol{\beta}}_\lambda^{ridge} - \boldsymbol{\beta}_\lambda^*\| = O_P(\frac{n_{\max}^{3/4}}{n} + \frac{\sqrt{n_{\max}p}}{n} + m_p\omega_n\frac{n_{\max}}{n}). \tag{S1.5}$$

On the other hand, since $\boldsymbol{\beta}_\lambda^* - \boldsymbol{\beta}^* = -2\lambda\mathbf{\Sigma}_{u,\lambda}^{-1}\boldsymbol{\beta}^*$, we have

$$\|\boldsymbol{\beta}^* - \boldsymbol{\beta}_\lambda^*\| \le 2\lambda\|\mathbf{\Sigma}_{u,\lambda}^{-1}\|\|\boldsymbol{\beta}^*\| \le \frac{2\lambda}{2\lambda + C_{\min}}\|\boldsymbol{\beta}^*\| = O(\lambda\|\boldsymbol{\beta}^*\|). \tag{S1.6}$$

Then, (S1.5) and (S1.6) together with the stated choice of λ prove (b).

For (c), we rely on the general high-dimensional M-estimator theory (Negahban et al., 2012) to prove the result. As shown in Negahban et al. (2012), to obtain the convergence rate of $\|\hat{\boldsymbol{\beta}}_\lambda^{lasso} - \boldsymbol{\beta}^*\|$, the key is to bound $\|\nabla\ell(\boldsymbol{\beta}^*)\|_\infty$, where

$$\ell(\boldsymbol{\beta}) = \frac{1}{2}\boldsymbol{\beta}'\hat{\mathbf{\Sigma}}_u\boldsymbol{\beta} - \frac{1}{n}\boldsymbol{Y}'\hat{\mathbf{U}}\boldsymbol{\beta}.$$

We have

$$\begin{aligned}\nabla\ell(\boldsymbol{\beta}^*) &= \hat{\mathbf{\Sigma}}_u\boldsymbol{\beta}^* - \frac{1}{n}\hat{\mathbf{U}}'\boldsymbol{Y}\\ &= (\hat{\mathbf{\Sigma}}_u - \mathbf{\Sigma}_u)\boldsymbol{\beta}^* + (\mathbf{\Sigma}_u - \frac{1}{n}\mathbf{U}'\mathbf{U})\boldsymbol{\beta}^* - \frac{1}{n}\left((\hat{\mathbf{U}} - \mathbf{U})'\boldsymbol{Y} + \sum_{g=1}^{G}\mathbf{U}_g'\mathbf{F}_g\boldsymbol{\gamma}_g^* + \mathbf{U}'\boldsymbol{\epsilon}\right)\\ &= I + II - \frac{1}{n}(III + IV + V).\end{aligned}$$

From Fan et al. (2013), we have $\|\hat{\boldsymbol{\Sigma}}_u - \boldsymbol{\Sigma}_u\| = O_p(m_p\omega_n)$. By Lemmas 4 snd 5, we have

$$\|I\|_\infty \leq \|\hat{\boldsymbol{\Sigma}}_u - \boldsymbol{\Sigma}_u\|\|\boldsymbol{\beta}^*\| = O_P(m_p\omega_n\|\boldsymbol{\beta}^*\|),$$

$$\|II\|_\infty \leq \|(\boldsymbol{\Sigma}_u - \frac{1}{n}\mathbf{U}'\mathbf{U})\boldsymbol{\beta}^*\|_\infty = O_P(\sqrt{\frac{\log p}{n}}\|\boldsymbol{\beta}^*\|).$$

By Lemma 4 (c) and $\sum_{i=1}^{n_g}\|\mathbf{f}_{g,i}\|^2 = O_P(n_g)$, we have

$$\begin{aligned}\|(\hat{\mathbf{U}}_g - \mathbf{U}_g)'\mathbf{F}_g\boldsymbol{\gamma}_g^*\|_\infty &= \max_j \big|\sum_{i=1}^{n_g}(\hat{u}_{g,ij} - u_{g,ij})\mathbf{f}_{g,i}'\boldsymbol{\gamma}_g^*\big| \\ &\leq \sqrt{\max_j \sum_{i=1}^{n_g}(\hat{u}_{g,ij} - u_{g,ij})^2 \sum_{i=1}^{n_g}\|\mathbf{f}_{g,i}\|^2\|\boldsymbol{\gamma}_g^*\|^2} \\ &= O_P(\sqrt{n_g n}\omega_n\|\boldsymbol{\gamma}_g^*\|).\end{aligned} \tag{S1.7}$$

Similarly, we have $\|(\hat{\mathbf{U}}_g - \mathbf{U}_g)'\mathbf{U}_g\boldsymbol{\beta}^*\|_\infty = O_P(\sqrt{n_g n}\omega_n\|\boldsymbol{\beta}^*\|)$ and $\|(\hat{\mathbf{U}}_g - \mathbf{U}_g)'\boldsymbol{\epsilon}_g\|_\infty = O_P(\sqrt{n_g n}\omega_n)$. Hence,

$$\begin{aligned}\|III\|_\infty &\leq \sum_{g=1}^G \|(\hat{\mathbf{U}}_g - \mathbf{U}_g)'\mathbf{F}_g\boldsymbol{\gamma}_g^*\|_\infty + \sum_{g=1}^G \|(\hat{\mathbf{U}}_g - \mathbf{U}_g)'\mathbf{U}_g\boldsymbol{\beta}^*\|_\infty + \sum_{g=1}^G \|(\hat{\mathbf{U}}_g - \mathbf{U}_g)'\boldsymbol{\epsilon}_g\|_\infty \\ &= O_P\big(\sqrt{n n_{\max}}\omega_n(\sum_{g=1}^G \|\boldsymbol{\gamma}_g^*\| + \|\boldsymbol{\beta}^*\| + 1)\big),\end{aligned}$$

By Lemma 5, we have

$$\|IV\|_\infty \leq \sum_{g=1}^G \|\mathbf{U}_g'\mathbf{F}_g\boldsymbol{\gamma}_g^*\|_\infty = O_P(\sqrt{n_{\max}\log p}\sum_{g=1}^G \|\boldsymbol{\gamma}_g^*\|),$$

$$\|V\|_\infty \leq \sum_{g=1}^G \|\mathbf{U}_g'\boldsymbol{\epsilon}_g\|_\infty = O_P(\sqrt{n_{\max}\log p}).$$

Hence, we have

$$\|\nabla \ell(\boldsymbol{\beta}^*)\|_\infty = O_P(m_p\omega_n + \sqrt{\frac{n_{\max}}{n}}\omega_n).$$

Next, we prove that the RE condition holds for $\hat{\boldsymbol{\Sigma}}_u$ with probability tending to 1. Indeed, from Fan et al. (2013), we have $\|\hat{\boldsymbol{\Sigma}}_u - \boldsymbol{\Sigma}_u\|_{\max} = O_P(\omega_n)$. Hence for all $\boldsymbol{\beta} \in \mathbb{C}(S)$, we have

$$|\boldsymbol{\beta}'(\hat{\boldsymbol{\Sigma}}_u - \boldsymbol{\Sigma}_u)\boldsymbol{\beta}| \leq \|\boldsymbol{\beta}\|_1\|(\hat{\boldsymbol{\Sigma}}_u - \boldsymbol{\Sigma}_u)\boldsymbol{\beta}\|_\infty \leq \|\hat{\boldsymbol{\Sigma}}_u - \boldsymbol{\Sigma}_u\|_{\max}\|\boldsymbol{\beta}\|_1^2 = O_P(s\|\boldsymbol{\beta}\|^2\omega_n),$$

as $\|\boldsymbol{\beta}\|_1 = \|\boldsymbol{\beta}_{S^c}\|_1 + \|\boldsymbol{\beta}_S\|_1 \leq 4\|\boldsymbol{\beta}_S\|_1 \leq 4\sqrt{s}\|\boldsymbol{\beta}_S\| \leq 4\sqrt{s}\|\boldsymbol{\beta}\|$. Since $s\omega_n = o(1)$, it follows from (S1.1) that $|\boldsymbol{\beta}'(\hat{\boldsymbol{\Sigma}}_u - \boldsymbol{\Sigma}_u)\boldsymbol{\beta}| = o_P(\|\boldsymbol{\beta}\|^2)$, hence $\boldsymbol{\beta}'\hat{\boldsymbol{\Sigma}}_u\boldsymbol{\beta} \geq \boldsymbol{\beta}'\boldsymbol{\Sigma}_u\boldsymbol{\beta} - |\boldsymbol{\beta}'(\hat{\boldsymbol{\Sigma}}_u - \boldsymbol{\Sigma}_u)\boldsymbol{\beta}| = \boldsymbol{\beta}'\boldsymbol{\Sigma}_u\boldsymbol{\beta} + o_P(\|\boldsymbol{\beta}\|^2)$, which proves the desired result. Then, it follows from Corollary 2 of Negahban et al. (2012) that

$$\|\hat{\boldsymbol{\beta}}_\lambda^{lasso} - \boldsymbol{\beta}^*\| = O_P\left(\sqrt{s}(m_p\omega_n + \sqrt{\frac{n_{\max}}{n}}\omega_n)\right),$$

if we choose $\lambda = C\omega_n\big(m_p + \sqrt{n_{\max}/n}\big)$ for some large enough constant C.

S1.2 Proof of Corollary 1.1

As the proofs for Ridge and Lasso estimators are similar, we denote $\hat{\boldsymbol{Y}}_{g,\lambda} = \hat{\mathbf{F}}_g\hat{\boldsymbol{\gamma}}_g + \mathbf{U}_g\hat{\boldsymbol{\beta}}_\lambda$, where $\hat{\boldsymbol{\beta}}_\lambda$ can be either $\hat{\boldsymbol{\beta}}_\lambda^{lasso}$ or $\hat{\boldsymbol{\beta}}_\lambda^{ridge}$. Then, we have

$$\frac{1}{n_g}\big\{\hat{\boldsymbol{Y}}_{g,\lambda} - \mathbb{E}(\boldsymbol{Y}_g|\mathbf{F}_g, \mathbf{U}_g)\big\} = \frac{1}{n_g}(\hat{\mathbf{F}}_g\hat{\boldsymbol{\gamma}}_g - \mathbf{F}_g\boldsymbol{\gamma}_g^*) + \frac{1}{n_g}(\hat{\mathbf{U}}_g\hat{\boldsymbol{\beta}}_\lambda - \mathbf{U}_g\boldsymbol{\beta}^*).$$

For the first term, we have

$$\frac{1}{n_g}(\hat{\mathbf{F}}_g\hat{\boldsymbol{\gamma}}_g - \mathbf{F}_g\boldsymbol{\gamma}_g^*) = \frac{1}{n_g}\big\{(\hat{\mathbf{F}}_g - \mathbf{F}_g\mathbf{H}_g')(\hat{\boldsymbol{\gamma}}_g - \mathbf{H}_g\boldsymbol{\gamma}_g^*) + \mathbf{F}_g\mathbf{H}_g'(\hat{\boldsymbol{\gamma}}_g - \mathbf{H}_g\boldsymbol{\gamma}_g^*)$$
$$+ (\hat{\mathbf{F}}_g - \mathbf{F}_g\mathbf{H}_g')\mathbf{H}_g\boldsymbol{\gamma}_g^* + \mathbf{F}_g(\mathbf{H}_g'\mathbf{H}_g - \mathbf{I})\boldsymbol{\gamma}_g^*\big\}$$
$$= \frac{1}{n_g}(I + II + III + IV),$$

whose dominating terms are II and III. By Theorem 1, Lemma 1, we have

$$\|II\| \leq \|\mathbf{H}_g\|\sqrt{\sum_{i=1}^{n_g}\|\mathbf{f}_{g,i}\|^2\|\hat{\boldsymbol{\gamma}}_g - \mathbf{H}_g\boldsymbol{\gamma}_g^*\|^2} = O_P(1 + \sqrt{\frac{n_g}{p}}),$$

$$\|III\| \leq \|\mathbf{H}_g\|\sqrt{\sum_{i=1}^{n_g}\|\hat{\mathbf{f}}_{g,i} - \mathbf{H}_g\mathbf{f}_{g,i}\|^2\|\boldsymbol{\gamma}_g^*\|^2} = O_P(1 + \sqrt{\frac{n_g}{p}}).$$

Therefore,

$$\|\frac{1}{n_g}(\hat{\mathbf{F}}_g\hat{\boldsymbol{\gamma}}_g - \mathbf{F}_g\boldsymbol{\gamma}_g^*)\| = O_P(\frac{1}{n_g} + \frac{1}{\sqrt{n_g p}}). \tag{S1.1}$$

For the second term, we have

$$\frac{1}{n_g}(\hat{\mathbf{U}}_g\hat{\boldsymbol{\beta}}_\lambda - \mathbf{U}_g\boldsymbol{\beta}^*) = \frac{1}{n_g}(\hat{\mathbf{U}}_g - \mathbf{U}_g)(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^*) + \frac{1}{n_g}\mathbf{U}_g(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^*) + \frac{1}{n_g}(\hat{\mathbf{U}}_g - \mathbf{U}_g)\boldsymbol{\beta}^*$$
$$= I + II + III.$$

By Assumption 2, $\|\boldsymbol{\Sigma}_u\| \leq \|\boldsymbol{\Sigma}_u\|_1 = O(1)$, hence $(1/n_g)\sum_{i=1}^{n_g}\|\mathbf{u}_{g,i}\|^2 = O_P(1)$. Then, it follows from Lemma 4 (b) that

$$\|I\| = O_P\bigg((\frac{\sqrt{\log n_g \log p}}{n_g} + \frac{1}{n_g^{1/4}\sqrt{p}})\|\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^*\|\bigg),$$

$$\|II\| \leq \frac{1}{n_g}\sqrt{\sum_{i=1}^{n_g}\|\mathbf{u}_{g,i}\|^2\|\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^*\|^2} = O_P\big(\frac{1}{\sqrt{n_g}}\|\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^*\|\big),$$

$$\|III\| = O_P\left(\frac{\sqrt{\log n_g \log p}}{n_g} + \frac{1}{n_g^{1/4}\sqrt{p}}\right).$$

Since $n = o(p^2)$ and $\log p = o(n^{2/39})$, these results together with (S1.1) implies that

$$\|\frac{1}{n_g}\big\{\hat{\boldsymbol{Y}}_{g,\lambda} - \mathbb{E}(\boldsymbol{Y}_g|\mathbf{F}_g, \mathbf{U}_g)\big\}\| = O_P\left(\frac{1}{\sqrt{n_g}}\|\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^*\|\right) + O_P\left(\frac{\sqrt{\log n_g \log p}}{n_g} + \frac{1}{n_g^{1/4}\sqrt{p}}\right).$$

Plugging the convergence rates of the corresponding estimators established in Theorem 1 completes the proof.

S1.3 Proof of Theorem 2

For (a), the Ridge estimator of (4.10) has

$$\hat{\boldsymbol{\beta}}_{g,\lambda}^{ridge} = \frac{1}{n_g}(\hat{\boldsymbol{\Sigma}}_{\tilde{x},g} + 2\lambda\mathbf{I})^{-1}\tilde{\mathbf{X}}_g'\boldsymbol{Y}_g,$$

where $\hat{\boldsymbol{\Sigma}}_{\tilde{x},g} = (1/n_g)\tilde{\mathbf{X}}_g'\tilde{\mathbf{X}}_g$. Hence, we have

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{g,\lambda}^{ridge} - \boldsymbol{\beta}^* =& -2\lambda(\hat{\boldsymbol{\Sigma}}_{\tilde{x},g} + 2\lambda\mathbf{I})^{-1}\boldsymbol{\beta}^* + \frac{1}{n_g}(\hat{\boldsymbol{\Sigma}}_{\tilde{x},g} + 2\lambda\mathbf{I})^{-1}\tilde{\mathbf{X}}_g'\mathbf{F}_g\boldsymbol{\delta}_g \\ &+ \frac{d_p}{n_g}(\hat{\boldsymbol{\Sigma}}_{\tilde{x},g} + 2\lambda\mathbf{I})^{-1}\tilde{\mathbf{X}}_g'\mathbf{U}_g\boldsymbol{\beta}^* + \frac{1}{n_g}(\hat{\boldsymbol{\Sigma}}_{\tilde{x},g} + 2\lambda\mathbf{I})^{-1}\tilde{\mathbf{X}}_g'\boldsymbol{\epsilon}_g \\ =& \, I + II + III + IV.\end{aligned}$$

First, we prove that with probability tending to 1, $\lambda_{\min}(\hat{\mathbf{\Sigma}}_{\tilde{x},g}) > c_1/p$. Indeed, by Weyl's Theorem, we have

$$\lambda_{\min}(\hat{\mathbf{\Sigma}}_{\tilde{x},g}) \geq \frac{1}{p}\{\lambda_{\min}(\mathbf{\Lambda}_g'\mathbf{\Lambda}_g) + \lambda_{\min}(\mathbf{\Lambda}_g'(\frac{1}{n_g}\mathbf{F}_g'\mathbf{F}_g - \mathbf{I})\mathbf{\Lambda}_g) + \frac{1}{n_g}\lambda_{\min}(\mathbf{\Lambda}_g'\mathbf{F}_g'\mathbf{U}_g + \mathbf{U}_g'\mathbf{F}_g\mathbf{\Lambda}_g)$$
$$+ \lambda_{\min}(\frac{1}{n_g}\mathbf{U}_g'\mathbf{U}_g - \mathbf{\Sigma}_u) + \lambda_{\min}(\mathbf{\Sigma}_u)\}.$$

From Assumption 3 (a), we have $\|\mathbf{\Lambda}_g\|_1 = O(1)$. Therefore, from Lemma 5, we have

$$|\frac{1}{p}\lambda_{\min}(\mathbf{\Lambda}_g'(\frac{1}{n_g}\mathbf{F}_g'\mathbf{F}_g - \mathbf{I})\mathbf{\Lambda}_g)| \leq \frac{1}{p}\|\mathbf{\Lambda}_g'(\frac{1}{n_g}\mathbf{F}_g'\mathbf{F}_g - \mathbf{I})\mathbf{\Lambda}_g\| \leq \|\frac{1}{n_g}\mathbf{F}_g'\mathbf{F}_g - \mathbf{I}\|_{\max}\|\mathbf{\Lambda}_g\|_1^2 = O_P(\frac{1}{\sqrt{n_g}}),$$
$$|\frac{1}{n_g p}\lambda_{\min}(\mathbf{\Lambda}_g'\mathbf{F}_g'\mathbf{U}_g)| = |\frac{1}{n_g p}\lambda_{\min}(\mathbf{U}_g'\mathbf{F}_g\mathbf{\Lambda}_g)| \leq \frac{1}{n_g}\|\mathbf{U}_g'\mathbf{F}_g\|_{\max}\|\mathbf{\Lambda}_g\|_1 = O_P(\frac{1}{\sqrt{n_g}}),$$
$$|\frac{1}{p}\lambda_{\min}(\frac{1}{n_g}\mathbf{U}_g'\mathbf{U}_g - \mathbf{\Sigma}_u)| \leq \|\frac{1}{n_g}\mathbf{U}_g'\mathbf{U}_g - \mathbf{\Sigma}_u\|_{\max} = O_P(\sqrt{\frac{\log p}{n_g}}).$$

Thus, it follows from Assumption 2 that, with probability tending to 1,

$$\lambda_{\min}(\hat{\mathbf{\Sigma}}_{\tilde{x},g}) \dot{\geq} \frac{1}{p}\lambda_{\min}(\mathbf{\Lambda}_g'\mathbf{\Lambda}_g) + \frac{1}{p}\lambda_{\min}(\mathbf{\Sigma}_u) \geq \frac{1}{p}\lambda_{\min}(\mathbf{\Sigma}_u) > \frac{c_1}{p}.$$

This result implies that

$$\|(\hat{\mathbf{\Sigma}}_{\tilde{x},g} + 2\lambda\mathbf{I})^{-1}\| = \lambda_{\max}\big((\hat{\mathbf{\Sigma}}_{\tilde{x},g} + 2\lambda\mathbf{I})^{-1}\big) = 1/\lambda_{\min}(\hat{\mathbf{\Sigma}}_{\tilde{x},g} + 2\lambda\mathbf{I}) \leq 1/\big(\lambda_{\min}(\hat{\mathbf{\Sigma}}_{\tilde{x},g}) + 2\lambda\big)$$
$$\leq \frac{1}{c_1/p + 2\lambda}.$$

It follows from Lemma 2 that

$$\|\frac{1}{n_g}\mathbf{F}_g'\mathbf{F}_g\boldsymbol{\delta}_g\| \leq \|(\frac{1}{n_g}\mathbf{F}_g'\mathbf{F}_g - \mathbf{I})\boldsymbol{\delta}_g\| + \|\boldsymbol{\delta}_g\| = \sqrt{\sum_{k=1}^{K_g}\big(\frac{1}{n_g}\sum_{i=1}^{n_g} f_{g,ik}\mathbf{f}_{g,i}'\boldsymbol{\delta}_g - \delta_{g,k}\big)^2} + \|\boldsymbol{\delta}_g\|$$
$$= O_P(\|\boldsymbol{\delta}_g\|). \tag{S1.1}$$

By Lemma 3, we have

$$\|\frac{1}{n_g}\mathbf{U}_g'\mathbf{F}_g\boldsymbol{\delta}_g\| = \frac{1}{n_g}\sqrt{\sum_{j=1}^{p}\big(\sum_{i=1}^{n_g} u_{g,ij}\mathbf{f}_{g,i}'\boldsymbol{\delta}_g\big)^2} = O_P(\sqrt{\frac{p}{n_g}}\|\boldsymbol{\delta}_g\|), \tag{S1.2}$$

where $|\sum_{i=1}^{n_g} u_{g,ij}\mathbf{f}_{g,i}'\boldsymbol{\delta}_g| = O_P(\sqrt{n_g}\|\boldsymbol{\delta}_g\|)$. By Lemma 2, we have

$$\|\frac{1}{n_g}\mathbf{F}_g'\mathbf{U}_g\boldsymbol{\beta}^*\| = \frac{1}{n_g}\sqrt{\sum_{k=1}^{K_g}\big(\sum_{i=1}^{n_g} f_{g,ik}\mathbf{u}_{g,i}'\boldsymbol{\beta}^*\big)^2} = O_P(\frac{1}{\sqrt{n_g}}\|\boldsymbol{\beta}^*\|), \tag{S1.3}$$

where $|\sum_{i=1}^{n_g} f_{g,ik}\mathbf{u}_{g,i}'\boldsymbol{\beta}^*| = O_P(\sqrt{n_g}\|\boldsymbol{\beta}^*\|)$. By (S1.2) and Assumption 2, we have

$$\|\frac{1}{n_g}\mathbf{U}_g'\mathbf{U}_g\boldsymbol{\beta}^*\| \leq \|(\frac{1}{n_g}\mathbf{U}_g'\mathbf{U}_g - \boldsymbol{\Sigma}_u)\boldsymbol{\beta}^*\| + \|\boldsymbol{\Sigma}_u\|\|\boldsymbol{\beta}^*\| = O_P\bigg((\sqrt{\frac{p}{n_g}} + 1)\|\boldsymbol{\beta}^*\|\bigg). \tag{S1.4}$$

By Lemma 2, we have

$$\|\frac{1}{n_g}\mathbf{F}_g'\boldsymbol{\epsilon}_g\| = \frac{1}{n_g}\sqrt{\sum_{k=1}^{k_g}\big(\sum_{i=1}^{n_g} f_{g,ik}\epsilon_{g,i}\big)^2} = O_P(\frac{1}{\sqrt{n_g}}). \tag{S1.5}$$

By Lemma 3, we have

$$\|\frac{1}{n_g}\mathbf{U}_g'\boldsymbol{\epsilon}_g\| = \frac{1}{n_g}\sqrt{\sum_{j=1}^{p}\big(\sum_{i=1}^{n_g} u_{g,ij}\epsilon_{g,i}\big)^2} = O_P(\sqrt{\frac{p}{n_g}}). \tag{S1.6}$$

Since Assumption 1 implies that $\|\mathbf{\Lambda}_g\| = O_P(\sqrt{p})$, it follows from (S1.1)–(S1.6) that

$$\|\frac{1}{n_g}\tilde{\mathbf{X}}_g'\mathbf{F}_g\boldsymbol{\delta}_g\| \leq \frac{1}{\sqrt{p}}\|\mathbf{\Lambda}_g\|\|\frac{1}{n_g}\mathbf{F}_g'\mathbf{F}_g\boldsymbol{\delta}_g\| + \frac{1}{\sqrt{p}}\|\frac{1}{n_g}\mathbf{U}_g'\mathbf{F}_g\boldsymbol{\delta}_g\| = O_P(\|\boldsymbol{\delta}_g\|),$$

$$\|\frac{1}{n_g}\tilde{\mathbf{X}}_g'\mathbf{U}_g\boldsymbol{\beta}^*\| \leq \frac{1}{\sqrt{p}}\|\mathbf{\Lambda}_g\|\|\frac{1}{n_g}\mathbf{F}_g'\mathbf{U}_g\boldsymbol{\beta}^*\| + \frac{1}{\sqrt{p}}\|\frac{1}{n_g}\mathbf{U}_g'\mathbf{U}_g\boldsymbol{\beta}^*\| = O_P\Big((\frac{1}{\sqrt{n_g}} + \frac{1}{\sqrt{p}})\|\boldsymbol{\beta}^*\|\Big),$$

$$\|\frac{1}{n_g}\tilde{\mathbf{X}}_g'\boldsymbol{\epsilon}_g\| \leq \frac{1}{\sqrt{p}}\|\mathbf{\Lambda}_g\|\|\frac{1}{n_g}\mathbf{F}_g'\boldsymbol{\epsilon}_g\| + \frac{1}{\sqrt{p}}\|\frac{1}{n_g}\mathbf{U}_g'\boldsymbol{\epsilon}_g\| = O_P(\frac{1}{\sqrt{n_g}}).$$

Hence, we have

$$\|I\| \leq \frac{2\lambda}{c_1/p + 2\lambda}\|\boldsymbol{\beta}^*\|,$$

$$\|II\| \leq \|(\hat{\mathbf{\Sigma}}_{\tilde{x},g} + 2\lambda\mathbf{I})^{-1}\|\|\frac{1}{n_g}\tilde{\mathbf{X}}_g'\mathbf{F}_g\boldsymbol{\delta}_g\| = O_P(\frac{1}{c_1/p + 2\lambda}\|\boldsymbol{\delta}_g\|),$$

$$\|III\| \leq \|(\hat{\mathbf{\Sigma}}_{\tilde{x},g} + 2\lambda\mathbf{I})^{-1}\|\|\frac{d_p}{n_g}\tilde{\mathbf{X}}_g'\mathbf{U}_g\boldsymbol{\beta}^*\| = O_P\Big(\frac{d_p}{c_1/p + 2\lambda}(\frac{1}{\sqrt{n_g}} + \frac{1}{\sqrt{p}})\|\boldsymbol{\beta}^*\|\Big),$$

$$\|IV\| \leq \|(\hat{\mathbf{\Sigma}}_{\tilde{x},g} + 2\lambda\mathbf{I})^{-1}\|\|\frac{1}{n_g}\tilde{\mathbf{X}}_g'\boldsymbol{\epsilon}_g\| = O_P\Big(\frac{1}{\sqrt{n_g(c_1/p + 2\lambda)}}\Big).$$

Then, by choosing $\lambda = C/\sqrt{p}$ and noting that $\|\boldsymbol{\beta}^*\| = O(1)$, we have

$$\|\hat{\boldsymbol{\beta}}_{g,\lambda}^{ridge} - \boldsymbol{\beta}^*\| = O_P\Big(\sqrt{p}\|\boldsymbol{\delta}_g\| + d_p(1 + \sqrt{\frac{p}{n_g}}) + \sqrt{\frac{p}{n_g}}\Big).$$

For (b), letting $\ell_g(\boldsymbol{\beta}^*) = \|\boldsymbol{Y}_g - \tilde{\mathbf{X}}_g\boldsymbol{\beta}^*\|^2/(2n_g)$, we have

$$\nabla\ell_g(\boldsymbol{\beta}^*) = -\frac{1}{n_g}(\tilde{\mathbf{X}}_g'\mathbf{F}_g\boldsymbol{\delta}_g + d_p\tilde{\mathbf{X}}_g'\mathbf{U}_g\boldsymbol{\beta}^* + \tilde{\mathbf{X}}_g'\boldsymbol{\epsilon}_g) = -\frac{1}{n_g}(I + II + III).$$

It follows from Lemma 5 that

$$\|I\|_\infty \leq \frac{1}{\sqrt{p}}\|\mathbf{\Lambda}_g'\mathbf{F}_g'\mathbf{F}_g\boldsymbol{\delta}_g\|_\infty + \frac{1}{\sqrt{p}}\|\mathbf{U}_g'\mathbf{F}_g\boldsymbol{\delta}_g\|_\infty = O_P\Big((\frac{n_g}{\sqrt{p}} + \sqrt{\frac{n_g\log p}{p}})\|\boldsymbol{\delta}_g\|\Big).$$

Noting that $\|\mathbf{\Sigma}_u\| = O(1)$, we have

$$\|\mathbf{U}_g'\mathbf{U}_g\boldsymbol{\beta}^*\|_\infty \le \max_j |\sum_{i=1}^{n_g} u_{g,ij}\mathbf{u}_{g,i}'\boldsymbol{\beta}^* - n_g\sum_{\ell=1}^{p}\sigma_{u,j\ell}\beta_\ell^*| + \max_j |n_g\sum_{\ell=1}^{p}\sigma_{u,j\ell}\beta_\ell^*|$$
$$= O_P(\sqrt{n_g\log p}\|\boldsymbol{\beta}^*\|) + O_P(n_g\|\boldsymbol{\beta}^*\|) = O_P(n_g\|\boldsymbol{\beta}^*\|).$$

Therefore,

$$\|II\|_\infty \le \frac{d_p}{\sqrt{p}}\|\mathbf{\Lambda}_g'\mathbf{F}_g'\mathbf{U}_g\boldsymbol{\beta}^*\|_\infty + \frac{d_p}{\sqrt{p}}\|\mathbf{U}_g'\mathbf{U}_g\boldsymbol{\beta}^*\|_\infty = O_P\left(d_p(\frac{n_g}{\sqrt{p}} + \sqrt{\frac{n_g\log p}{p}})\|\boldsymbol{\beta}^*\|\right),$$

$$\|III\|_\infty \le \frac{1}{\sqrt{p}}\|\mathbf{\Lambda}_g'\mathbf{F}_g'\boldsymbol{\epsilon}_g\|_\infty + \frac{1}{\sqrt{p}}\|\mathbf{U}_g'\boldsymbol{\epsilon}_g\|_\infty = O_P(\sqrt{\frac{n_g\log p}{p}}).$$

Since $\|\boldsymbol{\beta}^*\| = O(1)$, $\|\nabla\ell_g(\boldsymbol{\beta}^*)\|_\infty = O_P\big((1/\sqrt{p} + \sqrt{\log p/n_g p})(d_p + \|\boldsymbol{\delta}_g\|) + \sqrt{\log p/n_g}\big)$.

Next, we verify that the RE condition holds for $\sqrt{p}\hat{\mathbf{\Sigma}}_{\tilde{x},g}$ with probability tending to 1. Indeed, we have

$$\begin{aligned}\boldsymbol{\beta}'\hat{\mathbf{\Sigma}}_{\tilde{x},g}\boldsymbol{\beta} &= \frac{1}{n_g}\boldsymbol{\beta}'\tilde{\mathbf{X}}_g'\tilde{\mathbf{X}}_g\boldsymbol{\beta}\\ &= \frac{1}{p}\boldsymbol{\beta}'\{\mathbf{\Lambda}_g'\mathbf{\Lambda}_g + \mathbf{\Lambda}_g'(\frac{1}{n_g}\mathbf{F}_g'\mathbf{F}_g - \mathbf{I})\mathbf{\Lambda}_g + \frac{1}{n_g}\mathbf{\Lambda}_g'\mathbf{F}_g'\mathbf{U}_g + \frac{1}{n_g}\mathbf{U}_g'\mathbf{F}_g\mathbf{\Lambda}_g + \frac{1}{n_g}\mathbf{U}_g'\mathbf{U}_g\}\boldsymbol{\beta}\\ &\ge \frac{1}{p}\boldsymbol{\beta}'\mathbf{\Lambda}_g'\mathbf{\Lambda}_g\boldsymbol{\beta} + \frac{1}{p}\boldsymbol{\beta}'\mathbf{\Lambda}_g'(\frac{1}{n_g}\mathbf{F}_g'\mathbf{F}_g - \mathbf{I})\mathbf{\Lambda}_g\boldsymbol{\beta} + \frac{1}{n_g p}\boldsymbol{\beta}'(\mathbf{\Lambda}_g'\mathbf{F}_g'\mathbf{U}_g + \mathbf{U}_g'\mathbf{F}_g\mathbf{\Lambda}_g)\boldsymbol{\beta}\\ &= \frac{1}{p}\boldsymbol{\beta}'\mathbf{\Lambda}_g'\mathbf{\Lambda}_g\boldsymbol{\beta} + I + II.\end{aligned} \tag{S1.7}$$

By Assumption 3(a), we have $\|\mathbf{\Lambda}_g\|_1 = O_P(1)$. Hence, it follows from Lemma 5 that

$$\|\mathbf{\Lambda}_g'(\frac{1}{n_g}\mathbf{F}_g'\mathbf{F}_g - \mathbf{I})\mathbf{\Lambda}_g\|_{\max} \le \|\frac{1}{n_g}\mathbf{F}_g'\mathbf{F}_g - \mathbf{I}\|_{\max}\|\mathbf{\Lambda}_g\|_1^2 = O_P(\frac{1}{\sqrt{n_g}}),$$

$$\|\frac{1}{n_g}\mathbf{\Lambda}_g'\mathbf{F}_g'\mathbf{U}_g\|_{\max} = \|\frac{1}{n_g}\mathbf{U}_g'\mathbf{F}_g\mathbf{\Lambda}_g\|_{\max} \le \frac{1}{n_g}\|\mathbf{F}_g'\mathbf{U}_g\|_{\max}\|\mathbf{\Lambda}_g\|_1 = O_P(\sqrt{\frac{\log p}{n_g}}).$$

For all $\boldsymbol{\beta} \in \mathbb{C}(S)$, it follows from an analogous argument as in the proof of Theorem 1 that $\|\boldsymbol{\beta}\|_1 \le 4\sqrt{s}\|\boldsymbol{\beta}\|$. Therefore,

$$|I| \le \frac{1}{p}\|\mathbf{\Lambda}_g'(\frac{1}{n_g}\mathbf{F}_g'\mathbf{F}_g - \mathbf{I})\mathbf{\Lambda}_g\|_{\max}\|\boldsymbol{\beta}\|_1^2 = O_P(\frac{16s\|\boldsymbol{\beta}\|^2}{\sqrt{n_g}p}),$$

$$|II| \le \frac{1}{n_g p}(\|\mathbf{\Lambda}_g'\mathbf{F}_g'\mathbf{U}_g\|_{\max} + \|\mathbf{U}_g'\mathbf{F}_g\mathbf{\Lambda}_g\|_{\max})\|\boldsymbol{\beta}\|_1^2 = O_P(\frac{16s\|\boldsymbol{\beta}\|^2}{p}\sqrt{\frac{\log p}{n_g}}).$$

Hence, we have $\sqrt{p}(|I| + |II|) = o_P(\|\boldsymbol{\beta}\|^2)$ since $s\sqrt{\log p/(n_g p)} = o(1)$. It follows from (S1.7) that $\sqrt{p}\boldsymbol{\beta}'\hat{\mathbf{\Sigma}}_{\tilde{x},g}\boldsymbol{\beta} \ge \boldsymbol{\beta}'\mathbf{\Lambda}_g'\mathbf{\Lambda}_g\boldsymbol{\beta}/\sqrt{p} + \sqrt{p}(|I| + |II|) = \boldsymbol{\beta}'\mathbf{\Lambda}_g'\mathbf{\Lambda}_g\boldsymbol{\beta}/\sqrt{p} + o_P(\|\boldsymbol{\beta}\|^2)$. Hence the RE condition holds for $\sqrt{p}\hat{\mathbf{\Sigma}}_{\tilde{x},g}$ with probability tending to 1 as it holds on $\mathbf{\Lambda}_g'\mathbf{\Lambda}_g/\sqrt{p}$. Then, it follows from Corollary 2 of Negahban et al. (2012) that

$$\|\hat{\boldsymbol{\beta}}_{g,\lambda}^{lasso} - \boldsymbol{\beta}^*\| = O_P\left(\sqrt{s}\big\{(1 + \sqrt{\frac{\log p}{n_g}})(d_p + \|\boldsymbol{\delta}_g\|) + \sqrt{\frac{\log p}{n_g}}\big\}\right),$$

if we choose $\lambda = C\{(1 + \sqrt{\log p/n_g})(d_p + \|\boldsymbol{\delta}_g\|) + \sqrt{\log p/n_g}\}/\sqrt{p}$.

S1.4 Proof of Corollary 2.1

Without loss of generality, denote $\hat{\boldsymbol{Y}}_{g,\lambda} = \tilde{\mathbf{X}}_g\hat{\boldsymbol{\beta}}_{g,\lambda}$, where $\hat{\boldsymbol{\beta}}_{g,\lambda}$ can be either $\hat{\boldsymbol{\beta}}_{g,\lambda}^{lasso}$ or $\hat{\boldsymbol{\beta}}_{g,\lambda}^{ridge}$. Then,

$$\frac{1}{n_g}\big\{\hat{\boldsymbol{Y}}_{g,\lambda} - \mathbb{E}(\boldsymbol{Y}_g|\mathbf{F}_g, \mathbf{U}_g)\big\} = \frac{1}{n_g}\big\{\tilde{\mathbf{X}}_g(\hat{\boldsymbol{\beta}}_{g,\lambda} - \boldsymbol{\beta}^*) - \mathbf{F}_g\boldsymbol{\delta}_g - d_p\mathbf{U}_g\boldsymbol{\beta}^*\big\}.$$

Since $\|\mathbb{E}(\tilde{\mathbf{x}}_{g,i}\tilde{\mathbf{x}}'_{g,i})\| < \infty$, it implies that $(1/n_g)\sum_{i=1}^{n_g}\|\tilde{\mathbf{x}}_{g,i}\|^2 = O_P(1)$. Similarly, $(1/n_g)\sum_{i=1}^{n_g}\|\mathbf{f}_{g,i}\|^2 = O_P(1)$ and $(1/n_g)\sum_{i=1}^{n_g}\|\mathbf{u}_{g,i}\|^2 = O_P(1)$. Hence,

$$\begin{aligned}\|\frac{1}{n_g}\big\{\hat{\boldsymbol{Y}}_{g,\lambda} - \mathbb{E}(\boldsymbol{Y}_g|\mathbf{F}_g, \mathbf{U}_g)\big\}\| &\leq \frac{1}{n_g}\sqrt{\sum_{i=1}^{n_g}\|\tilde{\mathbf{x}}_{g,i}\|^2}\|\hat{\boldsymbol{\beta}}_{g,\lambda} - \boldsymbol{\beta}^*\| + \frac{1}{n_g}\sqrt{\sum_{i=1}^{n_g}\|\mathbf{f}_{g,i}\|^2}\|\boldsymbol{\delta}_g\| \\ &\quad + \frac{1}{n_g}\sqrt{\sum_{i=1}^{n_g}\|\mathbf{u}_{g,i}\|^2}\|\boldsymbol{\beta}^*\| \\ &= O_P\bigg(\frac{1}{\sqrt{n_g}}(\|\hat{\boldsymbol{\beta}}_{g,\lambda} - \boldsymbol{\beta}^*\| + \|\boldsymbol{\delta}_g\| + \|\boldsymbol{\beta}^*\|)\bigg).\end{aligned}$$

Plugging the convergence rates of $\|\hat{\boldsymbol{\beta}}_{g,\lambda} - \boldsymbol{\beta}^*\|$ established in Theorem 2 completes the proof.

S1.5 Proof of Theorem 3

For (a), the Ridge estimator of (4.11) has

$$\hat{\boldsymbol{\beta}}^{ridge}_{\lambda,global} = \frac{1}{n}(\hat{\boldsymbol{\Sigma}}_{\tilde{x}} + 2\lambda\mathbf{I})^{-1}\tilde{\mathbf{X}}'\boldsymbol{Y},$$

where $\hat{\boldsymbol{\Sigma}}_{\tilde{x}} = \frac{1}{n}\tilde{\mathbf{X}}'\tilde{\mathbf{X}}$. Then,

$$\begin{aligned}\hat{\boldsymbol{\beta}}^{ridge}_{\lambda,global} - \boldsymbol{\beta}^* &= -2\lambda(\hat{\boldsymbol{\Sigma}}_{\tilde{x}} + 2\lambda\mathbf{I})^{-1}\boldsymbol{\beta}^* + \frac{1}{n}(\hat{\boldsymbol{\Sigma}}_{\tilde{x}} + 2\lambda\mathbf{I})^{-1}\sum_{g=1}^{G}\tilde{\mathbf{X}}'_g\mathbf{F}_g\boldsymbol{\delta}_g \\ &\quad + \frac{1}{n}(\hat{\boldsymbol{\Sigma}}_{\tilde{x}} + 2\lambda\mathbf{I})^{-1}\sum_{g=1}^{G}d_p\tilde{\mathbf{X}}'_g\mathbf{U}_g\boldsymbol{\beta}^* + \frac{1}{n}(\hat{\boldsymbol{\Sigma}}_{\tilde{x}} + 2\lambda\mathbf{I})^{-1}\sum_{g=1}^{G}\tilde{\mathbf{X}}'_g\boldsymbol{\epsilon}_g \\ &= I + II + III + IV.\end{aligned}$$

Since

$$\lambda_{\min}(\hat{\mathbf{\Sigma}}_{\tilde{x}}) = \sum_{g=1}^{G} \frac{n_g}{n} \lambda_{\min}(\hat{\mathbf{\Sigma}}_{\tilde{x},g}) > \frac{c_1}{p},$$

it follows from Weyl's Theorem that

$$\|(\hat{\mathbf{\Sigma}}_{\tilde{x}} + 2\lambda \mathbf{I})^{-1}\| \leq \frac{1}{c_1/p + 2\lambda},$$

with probability tending to 1. Using similar arguments as in the proof of Theorem 2, we have

$$\|I\| = \frac{2\lambda}{c_1/p + 2\lambda}\|\boldsymbol{\beta}^*\|,$$

$$\|II\| \leq \|(\hat{\mathbf{\Sigma}}_{\tilde{x}} + 2\lambda \mathbf{I})^{-1}\| \sum_{g=1}^{G} \|\frac{1}{n}\tilde{\mathbf{X}}_g' \mathbf{F}_g \boldsymbol{\delta}_g\| = O_P(\frac{1}{c_1/p + 2\lambda}\frac{n_{\max}}{n}\sum_{g=1}^{G}\|\boldsymbol{\delta}_g\|),$$

$$\|III\| \leq \|(\hat{\mathbf{\Sigma}}_{\tilde{x}} + 2\lambda \mathbf{I})^{-1}\| \sum_{g=1}^{G} \|\frac{d_p}{n}\tilde{\mathbf{X}}_g' \mathbf{U}_g \boldsymbol{\beta}^*\| = O_P\left(\frac{d_p}{c_1/p + 2\lambda}(\frac{\sqrt{n_{\max}}}{n} + \frac{n_{\max}}{n\sqrt{p}})\|\boldsymbol{\beta}^*\|\right),$$

$$\|IV\| \leq \|(\hat{\mathbf{\Sigma}}_{\tilde{x}} + 2\lambda \mathbf{I})^{-1}\| \sum_{g=1}^{G} \|\frac{1}{n}\tilde{\mathbf{X}}_g' \boldsymbol{\epsilon}_g\| = O_P(\frac{1}{c_1/p + 2\lambda}\frac{\sqrt{n_{\max}}}{n}).$$

Hence, by choosing $\lambda = C/\sqrt{p}$ we have

$$\|\hat{\boldsymbol{\beta}}_{\lambda,global}^{ridge} - \boldsymbol{\beta}^*\| = O_P\left(\frac{n_{\max}\sqrt{p}}{n}\sum_{g=1}^{G}\|\boldsymbol{\delta}_g\| + d_p(\frac{n_{\max}}{n} + \frac{\sqrt{n_{\max}p}}{n}) + \frac{\sqrt{n_{\max}p}}{n}\right).$$

For (b), letting $\ell(\boldsymbol{\beta}^*) = (2n)^{-1}\|\boldsymbol{Y} - \tilde{\mathbf{X}}\boldsymbol{\beta}^*\|^2$, we have

$$\nabla \ell(\boldsymbol{\beta}^*) = -\frac{1}{n}(\sum_{g=1}^{G}\tilde{\mathbf{X}}_g'\mathbf{F}_g\boldsymbol{\delta}_g + \sum_{g=1}^{G} d_p\tilde{\mathbf{X}}_g'\mathbf{U}_g\boldsymbol{\beta}^* + \sum_{g=1}^{G}\tilde{\mathbf{X}}_g'\boldsymbol{\epsilon}_g) = -\frac{1}{n}(I + II + III).$$

Similarly as in the proof of Theorem 2,

$$\|I\|_\infty \le \frac{1}{\sqrt{p}}\sum_{g=1}^G \|\mathbf{\Lambda}_g'\mathbf{F}_g'\mathbf{F}_g\boldsymbol{\delta}_g\|_\infty + \frac{1}{\sqrt{p}}\sum_{g=1}^G \|\mathbf{U}_g'\mathbf{F}_g\boldsymbol{\delta}_g\|_\infty = O_P\Big((\frac{n_{\max}}{\sqrt{p}} + \sqrt{\frac{n_{\max}\log p}{p}}) \sum_{g=1}^G \|\boldsymbol{\delta}_g\| \Big),$$

$$\|II\|_\infty \le \frac{d_p}{\sqrt{p}}\sum_{g=1}^G \|\mathbf{\Lambda}_g'\mathbf{F}_g'\mathbf{U}_g\boldsymbol{\beta}^*\|_\infty + \frac{d_p}{\sqrt{p}}\sum_{g=1}^G \|\mathbf{U}_g'\mathbf{U}_g\boldsymbol{\beta}^*\|_\infty = O_P\Big(d_p(\frac{n_{\max}}{\sqrt{p}} + \sqrt{\frac{n_{\max}\log p}{p}}) \|\boldsymbol{\beta}^*\| \Big),$$

$$\|III\|_\infty \le \frac{1}{\sqrt{p}}\sum_{g=1}^G \|\mathbf{\Lambda}_g'\mathbf{F}_g'\boldsymbol{\epsilon}_g\|_\infty + \frac{1}{\sqrt{p}}\sum_{g=1}^G \|\mathbf{U}_g'\boldsymbol{\epsilon}_g\|_\infty = O_P(\sqrt{\frac{n_{\max}\log p}{p}}).$$

Therefore,

$$\|\nabla \ell(\boldsymbol{\beta}^*)\|_\infty = O_P\Big((\frac{n_{\max}}{n\sqrt{p}} + \frac{1}{n}\sqrt{\frac{n_{\max}\log p}{p}})\big(d_p + \sum_{g=1}^G \|\boldsymbol{\delta}_g\|\big) + \frac{1}{n}\sqrt{\frac{n_{\max}\log p}{p}} \Big).$$

Next, we show that the RE condition holds for hold for $\sqrt{p}\hat{\mathbf{\Sigma}}_{\tilde{x}}$ with probability tending to 1. Indeed, it follows from the proof of Theorem 2 that the RE condition holds for $\sqrt{p}\hat{\mathbf{\Sigma}}_{\tilde{x},g}$ with probability tending to 1 for any $g \in [G]$. Since $\hat{\mathbf{\Sigma}}_{\tilde{x}} = \sum_{g=1}^G (n_g/n)\hat{\mathbf{\Sigma}}_{\tilde{x},g}$, the same RE condition also holds for $\hat{\mathbf{\Sigma}}_{\tilde{x}}$. Then, with the stated choice of λ, it follows from Corollary 2 of Negahban et al. (2012) that

$$\|\hat{\boldsymbol{\beta}}^{lasso}_{\lambda,global} - \boldsymbol{\beta}^*\| = O_P\Big(\sqrt{s}\big\{ (\frac{n_{\max}}{n} + \frac{\sqrt{n_{\max}\log p}}{n})\big(d_p + \sum_{g=1}^G \|\boldsymbol{\delta}_g\|\big) + \frac{\sqrt{n_{\max}\log p}}{n} \big\} \Big).$$

S1.6 Proof of Corollary 3.1

The proof follows similar arguments as in the proof of Corollary 2.1, by noting that for $\hat{\boldsymbol{Y}}_{g,\lambda} = \tilde{\mathbf{X}}_g \hat{\boldsymbol{\beta}}_{\lambda,global}$,

$$\|\frac{1}{n_g}\{\hat{\boldsymbol{Y}}_{g,\lambda} - \mathbb{E}(\boldsymbol{Y}_g|\mathbf{F}_g, \mathbf{U}_g)\}\| \leq \frac{1}{n_g}\{\|\tilde{\mathbf{X}}_g(\hat{\boldsymbol{\beta}}_{\lambda,global} - \boldsymbol{\beta}^*)\| + \|\mathbf{F}_g\boldsymbol{\delta}_g\| + \|d_p\mathbf{U}_g\boldsymbol{\beta}^*\|\}$$
$$= O_P\bigg(\frac{1}{\sqrt{n_g}}(\|\hat{\boldsymbol{\beta}}_{\lambda,global} - \boldsymbol{\beta}^*\| + \|\boldsymbol{\delta}_g\| + \|\boldsymbol{\beta}^*\|)\bigg).$$

S1.7 Proof of Theorem 4

For (a), by the same arguments as in proving Theorem 1(a), we conclude that

$$\|\hat{\boldsymbol{\gamma}}_g - \frac{1}{\sqrt{p}}\mathbf{H}_g\boldsymbol{\Lambda}_g\boldsymbol{\beta}_g^*\| = O_P(\frac{1}{\sqrt{n_g}} + \frac{1}{\sqrt{p}}).$$

For (b), letting $\boldsymbol{\beta}^*_{g,\lambda} = \boldsymbol{\Sigma}_{u,\lambda}^{-1}\boldsymbol{\Sigma}_u\boldsymbol{\beta}_g^*$ and $\hat{\boldsymbol{\beta}}_\lambda^{ridge} = \hat{\boldsymbol{\Sigma}}_{u,\lambda}^{-1}\hat{\mathbf{U}}'\boldsymbol{Y}/n$, for any $g \in [G]$ we have

$$\hat{\boldsymbol{\beta}}_\lambda^{ridge} - \frac{1}{\sqrt{p}}\boldsymbol{\beta}^*_{g,\lambda} = \frac{1}{n}(\hat{\boldsymbol{\Sigma}}_{u,\lambda}^{-1} - \boldsymbol{\Sigma}_{u,\lambda}^{-1})(\hat{\mathbf{U}} - \mathbf{U})'\boldsymbol{Y} + \frac{1}{n}(\hat{\boldsymbol{\Sigma}}_{u,\lambda}^{-1} - \boldsymbol{\Sigma}_{u,\lambda}^{-1})\mathbf{U}'\boldsymbol{Y}$$
$$+ \frac{1}{n}\boldsymbol{\Sigma}_{u,\lambda}^{-1}(\hat{\mathbf{U}} - \mathbf{U})'\boldsymbol{Y} + \boldsymbol{\Sigma}_{u,\lambda}^{-1}(\frac{1}{n}\mathbf{U}'\boldsymbol{Y} - \frac{1}{\sqrt{p}}\boldsymbol{\Sigma}_u\boldsymbol{\beta}_g^*)$$
$$= I + II + III + IV.$$

Similarly as in the proof of Theorem 1, $\|(\hat{\mathbf{U}} - \mathbf{U})'\boldsymbol{Y}\| = O_P(\sqrt{p\log n_{\max}\log p} + n_{\max}^{3/4})$. From Assumption 3 (a), we have $\|\boldsymbol{\lambda}_{g,j}\| < \infty$, for all $j \in [p]$. Since $\|\boldsymbol{\Lambda}_g\boldsymbol{\beta}_g^*/\sqrt{p}\| \leq$

$(1/\sqrt{p})\sqrt{\sum_{j=1}^p \|\boldsymbol{\lambda}_{g,j}\|^2}\|\boldsymbol{\beta}^*\| = O(1)$, from (S1.2), (S1.4) and (S1.6), we have

$$\begin{aligned}\|\mathbf{U}'\boldsymbol{Y}\| &\leq \sum_{g=1}^G \|\mathbf{U}_g'\mathbf{F}_g\boldsymbol{\Lambda}_g\boldsymbol{\beta}_g^*/\sqrt{p}\| + \sum_{g=1}^G \frac{1}{\sqrt{p}}\|\mathbf{U}_g'\mathbf{U}_g\boldsymbol{\beta}_g^*\| + \sum_{g=1}^G \|\mathbf{U}_g'\boldsymbol{\epsilon}_g\| \\ &= O_P(\sqrt{n_{\max}p}) + O_P(\sqrt{n_{\max}} + \frac{n_{\max}}{\sqrt{p}}) + O_P(\sqrt{n_{\max}p}) \\ &= O_P\big(\sqrt{n_{\max}p} + \frac{n_{\max}}{\sqrt{p}}\big).\end{aligned}$$

Moreover, from (S1.2) and Assumption 2, we have

$$\begin{aligned}\|\frac{1}{n}\mathbf{U}'\boldsymbol{Y} - \frac{1}{\sqrt{p}}\boldsymbol{\Sigma}_u\boldsymbol{\beta}_g^*\| \leq{}& \frac{1}{n}\sum_{g'=1}^G \|\mathbf{U}_{g'}'\mathbf{F}_{g'}\boldsymbol{\Lambda}_{g'}\boldsymbol{\beta}_{g'}^*/\sqrt{p}\| + \sum_{g'=1}^G \frac{n_{g'}}{n\sqrt{p}}\|(\frac{1}{n_{g'}}\mathbf{U}_{g'}'\mathbf{U}_{g'} - \boldsymbol{\Sigma}_u)\boldsymbol{\beta}_{g'}^*\| \\ &+ \sum_{g'=1}^G \frac{n_{g'}}{n\sqrt{p}}\|\boldsymbol{\Sigma}_u\|\|\boldsymbol{\beta}_{g'}^* - \boldsymbol{\beta}_g^*\| + \frac{1}{n}\sum_{g'=1}^G \|\mathbf{U}_{g'}'\boldsymbol{\epsilon}_{g'}\| \\ ={}& O_P(\frac{\sqrt{n_{\max}p}}{n}) + O_P(\frac{\sqrt{n_{\max}}}{n}) + \sum_{g'=1}^G O_P(\frac{n_{g'}}{n\sqrt{p}}\|\boldsymbol{\beta}_{g'}^* - \boldsymbol{\beta}_g^*\|) + O_P(\frac{\sqrt{n_{\max}p}}{n}) \\ ={}& O_P(\frac{\sqrt{n_{\max}p}}{n}) + \sum_{g'=1}^G O_P(\frac{n_{g'}}{n\sqrt{p}}\|\boldsymbol{\beta}_{g'}^* - \boldsymbol{\beta}_g^*\|).\end{aligned}$$

Since $\|\hat{\boldsymbol{\Sigma}}_{u,\lambda}^{-1} - \boldsymbol{\Sigma}_{u,\lambda}^{-1}\| = O_P(m_p\omega_n)$ and $\|\boldsymbol{\Sigma}_{u,\lambda}^{-1}\| = O(1)$, we have

$$\|I\| \leq \|\hat{\boldsymbol{\Sigma}}_{u,\lambda}^{-1} - \boldsymbol{\Sigma}_{u,\lambda}^{-1}\|\|\frac{1}{n}(\hat{\mathbf{U}} - \mathbf{U})'\boldsymbol{Y}\| = O_P\left(m_p\omega_n(\frac{\sqrt{p\log n_{\max}\log p}}{n} + \frac{n_{\max}^{3/4}}{n})\right),$$

$$\|II\| \leq \|\hat{\boldsymbol{\Sigma}}_{u,\lambda}^{-1} - \boldsymbol{\Sigma}_{u,\lambda}^{-1}\|\|\frac{1}{n}\mathbf{U}'\boldsymbol{Y}\| = O_P\left(m_p\omega_n\big(\frac{\sqrt{n_{\max}p}}{n} + \frac{n_{\max}}{n\sqrt{p}}\big)\right),$$

$$\|III\| \leq \|\boldsymbol{\Sigma}_{u,\lambda}^{-1}\|\|\frac{1}{n}(\hat{\mathbf{U}} - \mathbf{U})'\boldsymbol{Y}\| = O_P\left(\frac{\sqrt{p\log n_{\max}\log p}}{n} + \frac{n_{\max}^{3/4}}{n}\right),$$

$$\|IV\| \leq \|\boldsymbol{\Sigma}_{u,\lambda}^{-1}\|\|\|\frac{1}{n}\mathbf{U}'\boldsymbol{Y} - \frac{1}{\sqrt{p}}\boldsymbol{\Sigma}_u\boldsymbol{\beta}_g^*\| = O_P\left(\frac{\sqrt{n_{\max}p}}{n}\right) + \sum_{g'=1}^G O_P\left(\frac{n_{g'}}{n\sqrt{p}}\|\boldsymbol{\beta}_{g'}^* - \boldsymbol{\beta}_g^*\|\right).$$

By the assumption that $m_p\omega_n = o(1)$, and $n = o(p^2)$, we have

$$\|\hat{\boldsymbol{\beta}}_\lambda^{ridge} - \frac{1}{\sqrt{p}}\boldsymbol{\beta}_{g,\lambda}^*\| = O_P\left(\frac{\sqrt{n_{\max}p}}{n} + \frac{n_{\max}^{3/4}}{n}\right) + \sum_{g'=1}^G O_P\left(\frac{n_{g'}}{n\sqrt{p}}\|\boldsymbol{\beta}_{g'}^* - \boldsymbol{\beta}_g^*\|\right).$$

Since

$$\|\boldsymbol{\beta}_g^* - \boldsymbol{\beta}_{g,\lambda}^*\| \le 2\lambda\|\boldsymbol{\Sigma}_{u,\lambda}^{-1}\|\|\boldsymbol{\beta}_g^*\| = O(\lambda\|\boldsymbol{\beta}_g^*\|),$$

and

$$\|\hat{\boldsymbol{\beta}}_\lambda^{ridge} - \frac{1}{\sqrt{p}}\boldsymbol{\beta}_g^*\| \le \|\hat{\boldsymbol{\beta}}_\lambda^{ridge} - \frac{1}{\sqrt{p}}\boldsymbol{\beta}_{g,\lambda}^*\| + \frac{1}{\sqrt{p}}\|\boldsymbol{\beta}_{g,\lambda}^* - \boldsymbol{\beta}_g^*\|,$$

it follows from the stated choice of λ that

$$\|\hat{\boldsymbol{\beta}}_\lambda^{ridge} - \frac{1}{\sqrt{p}}\boldsymbol{\beta}_g^*\| = O_P\left(\frac{\sqrt{n_{\max}p}}{n} + \frac{n_{\max}^{3/4}}{n}\right) + \sum_{g'=1}^G O_P\left(\frac{n_{g'}}{n\sqrt{p}}\|\boldsymbol{\beta}_{g'}^* - \boldsymbol{\beta}_g^*\|\right).$$

For (c), we first bound $\|\nabla\ell(\boldsymbol{\beta}_g^*)\|_\infty$. We have

$$\begin{aligned}
\nabla\ell(\boldsymbol{\beta}_g^*) &= \frac{1}{\sqrt{p}}\hat{\boldsymbol{\Sigma}}_u\boldsymbol{\beta}_g^* - \frac{1}{n}\hat{\mathbf{U}}'\boldsymbol{Y} \\
&= \frac{1}{\sqrt{p}}\left\{(\hat{\boldsymbol{\Sigma}}_u - \boldsymbol{\Sigma}_u)\boldsymbol{\beta}_g^* + \sum_{g'=1}^G \frac{n_{g'}}{n}(\boldsymbol{\Sigma}_u - \frac{1}{n_{g'}}\mathbf{U}_{g'}'\mathbf{U}_{g'})\boldsymbol{\beta}_{g'}^* + \sum_{g'=1}^G \frac{n_{g'}}{n}\boldsymbol{\Sigma}_u(\boldsymbol{\beta}_g^* - \boldsymbol{\beta}_{g'}^*)\right\} \\
&- \frac{1}{n}\left\{(\hat{\mathbf{U}} - \mathbf{U})'\boldsymbol{Y} + \mathbf{U}'\boldsymbol{\epsilon} + \sum_{g'=1}^G \mathbf{U}_{g'}'\mathbf{F}_{g'}\boldsymbol{\Lambda}_{g'}\boldsymbol{\beta}_{g'}^*/\sqrt{p}\right\} \\
&= \frac{1}{\sqrt{p}}(I + II + III) - \frac{1}{n}(IV + V + VI).
\end{aligned}$$

Similarly as in the proof of Theorem 1,

$$\|I\|_\infty \le \|\hat{\boldsymbol{\Sigma}}_u - \boldsymbol{\Sigma}_u\|\|\boldsymbol{\beta}_g^*\| = O_P(m_p\omega_n\|\boldsymbol{\beta}_g^*\|),$$

$$\|II\|_\infty \leq \sum_{g'=1}^{G} \frac{n_{g'}}{n} \|(\boldsymbol{\Sigma}_u - \frac{1}{n_{g'}}\mathbf{U}'_{g'}\mathbf{U}_{g'})\boldsymbol{\beta}^*_{g'}\|_\infty = O_P(\frac{\sqrt{n_{\max}\log p}}{n}\sum_{g'=1}^{G}\|\boldsymbol{\beta}^*_{g'}\|),$$

$$\|III\|_\infty \leq \sum_{g'=1}^{G} \frac{n_{g'}}{n}\|\boldsymbol{\Sigma}_u\|\|\boldsymbol{\beta}^*_g - \boldsymbol{\beta}^*_{g'}\| = O_P(\frac{n_{\max}}{n}\sum_{g'=1}^{G}\|\boldsymbol{\beta}^*_g - \boldsymbol{\beta}^*_{g'}\|).$$

By an analogous argument as in proving (S1.7), $\|(\hat{\mathbf{U}}_g - \mathbf{U}_g)'\mathbf{F}_g\boldsymbol{\Lambda}_g\boldsymbol{\beta}^*_g/\sqrt{p}\|_\infty = O_P(\sqrt{n_g n}\omega_n\|\boldsymbol{\beta}^*_g\|)$, $\|(\hat{\mathbf{U}}_g - \mathbf{U}_g)'\mathbf{U}_g\boldsymbol{\beta}^*_g\|_\infty = O_P(\sqrt{n_g n}\omega_n\|\boldsymbol{\beta}^*_g\|)$, and $\|(\hat{\mathbf{U}}_g - \mathbf{U}_g)'\boldsymbol{\epsilon}_g\|_\infty = O_P(\sqrt{n_g n}\omega_n)$.

Hence,

$$\begin{aligned}\|IV\|_\infty &= \|(\hat{\mathbf{U}} - \mathbf{U})'\boldsymbol{Y}\|_\infty \\ &\leq \sum_{g'=1}^{G}\|(\hat{\mathbf{U}}_{g'} - \mathbf{U}_{g'})'\mathbf{F}_{g'}\boldsymbol{\Lambda}_{g'}\boldsymbol{\beta}^*_{g'}/\sqrt{p}\|_\infty + \sum_{g'=1}^{G}\|\frac{1}{\sqrt{p}}(\hat{\mathbf{U}}_{g'} - \mathbf{U}_{g'})'\mathbf{U}_{g'}\boldsymbol{\beta}^*_{g'}\|_\infty \\ &+ \sum_{g'=1}^{G}\|(\hat{\mathbf{U}}_{g'} - \mathbf{U}_{g'})'\boldsymbol{\epsilon}_{g'}\|_\infty \\ &= O_P\big(\sqrt{n n_{\max}}\omega_n(\sum_{g'=1}^{G}\|\boldsymbol{\beta}^*_{g'}\| + 1)\big).\end{aligned}$$

Then, it follows from Lemma 5 that

$$\|V\|_\infty \leq \sum_{g'=1}^{G}\|\mathbf{U}'_{g'}\boldsymbol{\epsilon}_{g'}\|_\infty = O_P(\sqrt{n_{\max}\log p}),$$

$$\|VI\|_\infty \leq \sum_{g'=1}^{G}\|\mathbf{U}'_{g'}\mathbf{F}_{g'}\boldsymbol{\Lambda}_{g'}\boldsymbol{\beta}^*_{g'}/\sqrt{p}\|_\infty = O_P\big(\sqrt{n_{\max}\log p}\sum_{g'=1}^{G}\|\boldsymbol{\beta}^*_{g'}\|\big).$$

Since $\|\boldsymbol{\beta}^*_g\| = O(1)$ and $m_p\omega_n = o(1)$, we have

$$\|\nabla\ell(\boldsymbol{\beta}^*_g)\|_\infty = O_P\left(\sqrt{\frac{n_{\max}}{n}}\omega_n + \frac{n_{\max}}{n\sqrt{p}}\sum_{g'=1}^{G}\|\boldsymbol{\beta}^*_g - \boldsymbol{\beta}^*_{g'}\|\right).$$

Then, given the RE condition on $\hat{\boldsymbol{\Sigma}}_u$, which was proved in Theorem 1, it follows from

Corollary 2 of Negahban et al. (2012) that

$$\|\hat{\boldsymbol{\beta}}_{\lambda}^{lasso} - \frac{1}{\sqrt{p}}\boldsymbol{\beta}_g^*\| = O_P\bigg(\sqrt{s}\big(\sqrt{\frac{n_{\max}}{n}}\omega_n + \frac{n_{\max}}{n\sqrt{p}}\sum_{g'=1}^{G}\|\boldsymbol{\beta}_g^* - \boldsymbol{\beta}_{g'}^*\|\big)\bigg),$$

if we choose $\lambda = C\{\omega_n\sqrt{n_{\max}/n} + n_{\max}/(n\sqrt{p})\sum_{g'=1}^{G}\|\boldsymbol{\beta}_g^* - \boldsymbol{\beta}_{g'}^*\|\}$.

S1.8 Proof of Corollary 4.1

Similarly as in the proof of Corollary 1.1, we have

$$\frac{1}{n_g}\big\{\hat{\boldsymbol{Y}}_{g,\lambda} - \mathbb{E}(\boldsymbol{Y}_g|\tilde{\mathbf{X}}_g)\big\} = \frac{1}{n_g}(\hat{\mathbf{F}}_g\hat{\boldsymbol{\gamma}}_g - \frac{1}{\sqrt{p}}\mathbf{F}_g\boldsymbol{\Lambda}_g\boldsymbol{\beta}_g^* + \hat{\mathbf{U}}_g\hat{\boldsymbol{\beta}}_{\lambda} - \frac{1}{\sqrt{p}}\mathbf{U}_g\boldsymbol{\beta}_g^*).$$

First,

$$\|\frac{1}{n_g}(\hat{\mathbf{F}}_g\hat{\boldsymbol{\gamma}}_g - \frac{1}{\sqrt{p}}\mathbf{F}_g\boldsymbol{\Lambda}_g\boldsymbol{\beta}_g^*)\| = O_P(\frac{1}{n_g} + \frac{1}{\sqrt{n_g p}}). \tag{S1.1}$$

On the other hand,

$$\begin{aligned}\frac{1}{n_g}(\hat{\mathbf{U}}_g\hat{\boldsymbol{\beta}}_{\lambda} - \frac{1}{\sqrt{p}}\mathbf{U}_g\boldsymbol{\beta}_g^*) &= \frac{1}{n_g}(\hat{\mathbf{U}}_g - \mathbf{U}_g)(\hat{\boldsymbol{\beta}}_{\lambda} - \frac{1}{\sqrt{p}}\boldsymbol{\beta}_g^*) + \frac{1}{n_g}\mathbf{U}_g(\hat{\boldsymbol{\beta}}_{\lambda} - \frac{1}{\sqrt{p}}\boldsymbol{\beta}_g^*)\\ &\quad + \frac{1}{n_g\sqrt{p}}(\hat{\mathbf{U}}_g - \mathbf{U}_g)\boldsymbol{\beta}_g^*\\ &= I + II + III.\end{aligned}$$

By Lemma 4 (b),

$$\|I\| = O_P\bigg(\big(\frac{\sqrt{\log n_g \log p}}{n_g} + \frac{1}{n_g^{1/4}\sqrt{p}}\big)\|\hat{\boldsymbol{\beta}}_{\lambda} - \frac{1}{\sqrt{p}}\boldsymbol{\beta}_g^*\|\bigg),$$

$$\|II\| = O_P(\frac{1}{\sqrt{n_g}}\|\hat{\boldsymbol{\beta}}_{\lambda} - \frac{1}{\sqrt{p}}\boldsymbol{\beta}_g^*\|),$$

$$\|III\| = \frac{1}{n_g\sqrt{p}}\|(\hat{\mathbf{U}}_g - \mathbf{U}_g)\boldsymbol{\beta}_g^*\| = O_P\left(\frac{1}{n_g}\sqrt{\frac{\log n_g \log p}{p}} + \frac{1}{n_g^{1/4}p}\right).$$

Therefore, for any $g \in [G]$,

$$\|\frac{1}{n_g}(\hat{\mathbf{U}}_g\hat{\boldsymbol{\beta}}_\lambda - \mathbf{U}_g\boldsymbol{\beta}_g^*)\| = O_P\left(\frac{1}{\sqrt{n_g}}\|\hat{\boldsymbol{\beta}}_\lambda - \frac{1}{\sqrt{p}}\boldsymbol{\beta}_g^*\|\right) + O_P\left(\frac{1}{n_g}\sqrt{\frac{\log n_g \log p}{p}} + \frac{1}{n_g^{1/4}p}\right). \tag{S1.2}$$

Then, (S1.1) and (S1.2) together imply that

$$\|\frac{1}{n_g}\{\hat{\boldsymbol{Y}}_{g,\lambda} - \mathbb{E}(\boldsymbol{Y}_g|\tilde{\mathbf{X}}_g)\}\| = O_P(\frac{1}{n_g}) + O_P(\frac{1}{\sqrt{n_g p}}) + O_P\left(\frac{1}{\sqrt{n_g}}\|\hat{\boldsymbol{\beta}}_\lambda - \frac{1}{\sqrt{p}}\boldsymbol{\beta}_g^*\|\right).$$

S1.9 Supporting Lemmas and their proofs

Lemma 1. *Under Assumptions 1–3, for any $g \in [G]$, we have*

(a) $(1/n_g)\sum_{i=1}^{n_g}\|\hat{\mathbf{f}}_{g,i} - \mathbf{H}_g\mathbf{f}_{g,i}\|^2 = O_P(1/n_g + 1/p)$.

(b) $\|\mathbf{H}_g'\mathbf{H}_g - \mathbf{I}\| \leq K_g\|\mathbf{H}_g'\mathbf{H}_g - \mathbf{I}\|_{\max} = O_P(1/n_g + 1/p)$.

(c) $\|\mathbf{H}_g'\mathbf{H}_g - \mathbf{I}\|_F = O_P(1/\sqrt{n_g} + 1/\sqrt{p})$

(d) $\|\mathbf{H}_g\| = O_P(1)$.

Proof. These results directly follow from Lemmas 10, 11 (b) in Fan et al. (2013) and Lemma C.1 (iv) in Fan et al. (2018). □

Lemma 2. *Under Assumption 1, for any $k \in [K_g]$, $j \in [p]$, and $g \in [G]$, we have*

(a) $|(1/n_g)\sum_{i=1}^{n_g} f_{g,ik}\mathbf{f}_{g,i}'\boldsymbol{\gamma}_g^* - \gamma_{g,k}^*| = O_P(\|\boldsymbol{\gamma}_g^*\|/\sqrt{n_g})$.

(b) $|\sum_{i=1}^{n_g} f_{g,ik}\mathbf{u}_{g,i}'\boldsymbol{\beta}^*| = O_P(\sqrt{n_g}\|\boldsymbol{\beta}^*\|)$.

(c) $|\sum_{i=1}^{n_g} f_{g,ik}\epsilon_{g,i}| = O_P(\sqrt{n_g})$.

(d) $|(1/n_g)\sum_{i=1}^{n_g} f_{g,ik}\mathbf{f}'_{g,i}\boldsymbol{\delta}_g - \delta_{gk}| = O_P(\|\boldsymbol{\delta}_g\|/\sqrt{n_g})$.

Proof. (a) Since both $\{f_{g,ik}\}_{i\le n_g}$ and $\{\mathbf{f}'_{g,i}\boldsymbol{\gamma}^*_g\}_{i\le n_g}$ are i.i.d. sub-Gaussian random variables, $\{f_{g,ik}\mathbf{f}'_{g,i}\boldsymbol{\gamma}^*_g\}_{i\le n_g}$ are i.i.d. sub-exponential (Vershynin, 2018), with $\mathbb{E}(f_{g,ik}\mathbf{f}'_{g,i}\boldsymbol{\gamma}^*_g) = \gamma^*_{g,k}$. Hence for a sufficiently small positive number s, there exists a sufficiently large constant $C > 0$, such that

$$\mathbb{P}(\big|\frac{1}{n_g}\sum_{i=1}^{n_g} f_{g,ik}\mathbf{f}'_{g,i}\boldsymbol{\gamma}^*_g - \gamma^*_{g,k}\big| > s) \le 2e^{-\frac{Cn_g s^2}{\|\boldsymbol{\gamma}^*_g\|^2}},$$

which concludes the result.

(b) Since both $\{f_{g,ik}\}_{i\le n_g}$ and $\{\mathbf{u}'_{g,i}\boldsymbol{\beta}^*\}_{i\le n_g}$ are i.i.d. sub-Gaussian random variables, hence $\{f_{g,ik}\mathbf{u}'_{g,i}\boldsymbol{\beta}^*\}_{i\le n_g}$ are i.i.d. sub-exponential with mean zero. Since we have $\|\boldsymbol{\Sigma}_u\| = O(1)$, hence for a sufficiently small positive number s, there exists a constant $C > 0$, such that

$$\mathbb{P}(\big|\sum_{i=1}^{n_g} f_{g,ik}\mathbf{u}'_{g,i}\boldsymbol{\beta}^*\big| > s) \le 2e^{-\frac{Cs^2}{n_g\|\boldsymbol{\beta}^*\|^2}},$$

which concludes the result.

The proofs of (c) and (d) follow similar arguments as in (a) and (b). □

Lemma 3. *Under Assumption 1, for any $j \in [p]$ and $g \in [G]$, we have*

(a) $|\sum_{i=1}^{n_g} u_{g,ij}\mathbf{f}'_{g,i}\boldsymbol{\gamma}^*_g| = O_P(\sqrt{n_g}\|\boldsymbol{\gamma}^*_g\|)$.

(b) $|(1/n_g)\sum_{i=1}^{n_g} u_{g,ij}\mathbf{u}'_{g,i}\boldsymbol{\beta}^* - \sum_{\ell=1}^{p}\sigma_{u,j\ell}\beta^*_\ell| = O_P(\|\boldsymbol{\beta}^*\|/\sqrt{n_g})$.

(c) $|\sum_{i=1}^{n_g} u_{g,ij}\mathbf{f}'_{g,i}\boldsymbol{\delta}_g| = O_P(\sqrt{n_g}\|\boldsymbol{\delta}_g\|)$.

(d) $|\sum_{i=1}^{n_g} u_{g,ij}\epsilon_{g,i}| = O_P(\sqrt{n_g})$.

Proof. (a) Since both $\{u_{g,ij}\}_{i\le n_g}$ and $\{\mathbf{f}'_{g,i}\boldsymbol{\gamma}^*_g\}_{i\le n_g}$ are i.i.d. sub-Gaussian random variables, $\{u_{g,ij}\mathbf{f}'_{g,i}\boldsymbol{\gamma}^*_g\}_{i\le n_g}$ are i.i.d. sub-exponential random variables with mean zero. Since we have $\|\boldsymbol{\Sigma}_u\| = O(1)$, it implies that for a sufficiently small positive number s, there exists a constant $C > 0$, such that

$$\mathbb{P}(\big|\sum_{i=1}^{n_g} u_{g,ij}\mathbf{f}'_{g,i}\boldsymbol{\gamma}^*_g\big| > s) \le 2e^{-\frac{Cs^2}{n_g\|\boldsymbol{\gamma}^*_g\|^2}},$$

which concludes the result.

(b) Since both $\{u_{g,ij}\}_{i\le n_g}$ and $\{\mathbf{u}'_{g,i}\boldsymbol{\beta}^*\}_{i\le n_g}$ are i.i.d. sub-Gaussian random variables, $\{u_{g,ij}\mathbf{u}'_{g,i}\boldsymbol{\beta}^*\}_{i\le n_g}$ are i.i.d. sub-exponential random variables with $\mathbb{E}(u_{g,ij}\mathbf{u}'_{g,i}\boldsymbol{\beta}^*) = \sum_{\ell=1}^{p}\sigma_{u,j\ell}\beta^*_\ell$. Hence for a sufficiently small positive s, there exists a constant $C > 0$, such that

$$\mathbb{P}(\big|\frac{1}{n_g}\sum_{i=1}^{n_g} u_{g,ij}\mathbf{u}'_{g,i}\boldsymbol{\beta}^* - \sum_{\ell=1}^{p}\sigma_{u,j\ell}\beta^*_\ell\big| > s) \le 2e^{-\frac{Cn_g s^2}{\|\boldsymbol{\beta}^*\|^2}},$$

which concludes the result.

The proofs of (c) and (d) follow similar arguments as in (a) and (b). □

Lemma 4. *Under Assumption 3, we have*

(a) $\max_{i,j}|\hat{u}_{g,ij} - u_{g,ij}| = \max_{i,j}|(\hat{\mathbf{F}}_g\hat{\boldsymbol{\Lambda}}_g - \mathbf{F}_g\boldsymbol{\Lambda}_g)_{ij}| = O_P(\sqrt{\log n_g \log p/n_g} + n_g^{1/4}/\sqrt{p})$.

(b) $\sup\{\|(\hat{\mathbf{U}}_g - \mathbf{U}_g)\boldsymbol{\alpha}\| : \boldsymbol{\alpha}\in\mathbb{R}^p, \|\boldsymbol{\alpha}\| = 1\} = O_P(\sqrt{\log n_g \log p} + n_g^{3/4}/\sqrt{p})$.

(c) $\max_{j\le p}(1/n)\sum_{g=1}^{G}\sum_{i=1}^{n_g}(\hat{u}_{g,ij} - u_{g,ij})^2 = O_P(\omega_n^2)$.

Proof. Statements (a) and (c) directly follows from Corollary 1 and Lemma 12 of Fan et al. (2013).

For (b), let $\boldsymbol{\alpha} \in \mathbb{R}^p$ be a vector such that $\|\boldsymbol{\alpha}\| = 1$. Then, we have for all $i \in [n]$,

$$|((\hat{\mathbf{U}}_g - \mathbf{U}_g)\boldsymbol{\alpha})_i| = |\sum_{j=1}^{p}(\hat{u}_{g,ij} - u_{g,ij})\alpha_j| \leq \sqrt{\sum_{j=1}^{p}(\hat{u}_{g,ij} - u_{g,ij})^2\alpha_j^2} \leq \sqrt{\max_{i,j}(\hat{u}_{g,ij} - u_{g,ij})^2 \sum_{j=1}^{p}\alpha_j^2}$$
$$= O_P\Big(\sqrt{\frac{\log n_g \log p}{n_g}} + \frac{n_g^{1/4}}{\sqrt{p}}\Big).$$

Hence, we have

$$\|(\hat{\mathbf{U}}_g - \mathbf{U}_g)\boldsymbol{\alpha}\| = O_P\bigg(\sqrt{n_g}(\sqrt{\frac{\log n_g \log p}{n_g}} + \frac{n_g^{1/4}}{\sqrt{p}})\bigg) = O_P\bigg(\sqrt{\log n_g \log p} + \frac{n_g^{3/4}}{\sqrt{p}}\bigg).$$

□

Lemma 5. *Under Assumptions 1 and 2, for any $g \in [G]$, we have,*

(a) $\|\boldsymbol{\Lambda}_g'\mathbf{F}_g'\mathbf{F}_g\boldsymbol{\delta}_g\|_\infty = O_P(n_g\|\boldsymbol{\delta}_g\|)$.

(b) $\|\boldsymbol{\Lambda}_g'\mathbf{F}_g'\mathbf{U}_g\boldsymbol{\beta}^*\|_\infty = O_P(\sqrt{n_g \log p}\|\boldsymbol{\beta}^*\|)$.

(c) $\|\boldsymbol{\Lambda}_g'\mathbf{F}_g'\boldsymbol{\epsilon}_g\|_\infty = O_P(\sqrt{n_g \log p})$.

(d) $\|\mathbf{U}_g'\mathbf{F}_g\boldsymbol{\delta}_g\|_\infty = O_P(\sqrt{n_g \log p}\|\boldsymbol{\delta}_g\|)$.

(e) $\|((1/n_g)\mathbf{U}_g'\mathbf{U}_g - \boldsymbol{\Sigma}_u)\boldsymbol{\beta}^*\|_\infty = O_P(\sqrt{\log p/n_g}\|\boldsymbol{\beta}^*\|)$.

(f) $\|\mathbf{U}_g'\mathbf{F}_g\boldsymbol{\gamma}_g^*\|_\infty = O_P(\sqrt{n_g \log p}\|\boldsymbol{\gamma}_g^*\|)$.

(g) $\|\mathbf{U}_g'\boldsymbol{\epsilon}_g\|_\infty = O_P(\sqrt{n_g \log p})$.

(h) $\|(1/n_g)\mathbf{F}_g'\mathbf{F}_g - \mathbf{I}\|_{\max} = O_P(1/\sqrt{n_g})$.

(i) $\|(1/n_g)\mathbf{U}_g'\mathbf{U}_g - \boldsymbol{\Sigma}_u\|_{\max} = O_P(\sqrt{\log p/n_g})$.

Proof. (a) Note that the j-th element of $\boldsymbol{\Lambda}_g'\mathbf{F}_g'\mathbf{F}_g\boldsymbol{\delta}_g$ satisfies $|\{\boldsymbol{\Lambda}_g'\mathbf{F}_g'\mathbf{F}_g\boldsymbol{\delta}_g\}_j| = |\sum_{i=1}^{n_g}\mathbf{f}_{g,i}'\boldsymbol{\lambda}_{g,j}\mathbf{f}_{g,i}'\boldsymbol{\delta}_g|$.

By Assumptions 2 and 3, $\mathbf{f}'_{g,i}\boldsymbol{\lambda}_{g,j}$ is sub-Gaussian with a finite since $\text{var}(\mathbf{f}'_{g,i}\boldsymbol{\lambda}_{g,j}) = \boldsymbol{\lambda}'_{g,j}\boldsymbol{\lambda}_{g,j} \le K_g M^2 < \infty$ as it is assumed in Assumption 3 that $\|\boldsymbol{\lambda}_{g,j}\|_\infty < M$. Similarly, $\mathbf{f}'_{g,i}\boldsymbol{\delta}_g$ is sub-Gaussian. Then, $\mathbf{f}'_{g,i}\boldsymbol{\lambda}_{g,j}\mathbf{f}'_{g,i}\boldsymbol{\delta}_g$ is sub-exponential with $\mathbb{E}(\mathbf{f}'_{g,i}\boldsymbol{\lambda}_{g,j}\mathbf{f}'_{g,i}\boldsymbol{\delta}_g) = \boldsymbol{\lambda}'_{g,j}\boldsymbol{\delta}_g$. Hence, we have

$$\mathbb{P}(\max_j \big|\frac{1}{n_g}\sum_{i=1}^{n_g}\mathbf{f}'_{g,i}\boldsymbol{\lambda}_{g,j}\mathbf{f}'_{g,i}\boldsymbol{\delta}_g - \boldsymbol{\lambda}'_{g,j}\boldsymbol{\delta}_g\big| > s) \le 2pe^{-\frac{Cn_g s^2}{K_g M^2\|\boldsymbol{\delta}_g\|^2}},$$

which implies that $\max_j \big|\frac{1}{n_g}\sum_{i=1}^{n_g}\mathbf{f}'_{g,i}\boldsymbol{\lambda}_{g,j}\mathbf{f}'_{g,i}\boldsymbol{\delta}_g - \boldsymbol{\lambda}'_{g,j}\boldsymbol{\delta}_g\big| = O_P(\sqrt{\log p/n_g}\|\boldsymbol{\delta}_g\|)$. This result together with $\|\boldsymbol{\lambda}_{g,j}\| < K_g M$ implies that

$$\|\boldsymbol{\Lambda}'_g\mathbf{F}'_g\mathbf{F}_g\boldsymbol{\delta}_g\|_\infty = \max_j \big|\sum_{i=1}^{n_g}\mathbf{f}'_{g,i}\boldsymbol{\lambda}_{g,j}\mathbf{f}'_{g,i}\boldsymbol{\delta}_g\big| \le \max_j \big|\sum_{i=1}^{n_g}\mathbf{f}'_{g,i}\boldsymbol{\lambda}_{g,j}\mathbf{f}'_{g,i}\boldsymbol{\delta}_g - n_g\boldsymbol{\lambda}'_{g,j}\boldsymbol{\delta}_g\big| + \max_j \big|n_g\boldsymbol{\lambda}'_{g,j}\boldsymbol{\delta}_g\big|$$

$$= O_P(\sqrt{n_g\log p}\|\boldsymbol{\delta}_g\|) + O_P(n_g\|\boldsymbol{\delta}_g\|) = O_P(n_g\|\boldsymbol{\delta}_g\|).$$

The proofs of (b) and (c) follow similar arguments as in (a).

(d) For any $g \in [G]$ and $j \in [p]$, since $\{u_{g,ij}\}_{i\le n_g}$ and $\{\mathbf{f}'_{g,i}\boldsymbol{\delta}_g\}_{i\le n_g}$ are i.i.d. sub-Gaussian random variables, $\{u_{g,ij}\mathbf{f}'_{g,i}\boldsymbol{\delta}_g\}_{i\le n_g}$ are i.i.d. sub-exponential with mean zero. Since we have $\|\boldsymbol{\Sigma}_u\| = O(1)$, for a sufficiently small positive number s, there exists a constant $C > 0$, such that

$$\mathbb{P}(\max_j \big|\sum_{i=1}^{n_g}u_{g,ij}\mathbf{f}'_{g,i}\boldsymbol{\delta}_g\big| > s) \le \sum_{j=1}^{p}\mathbb{P}(\big|\sum_{i=1}^{n_g}u_{g,ij}\mathbf{f}'_{g,i}\boldsymbol{\delta}_g\big| > s) \le 2pe^{-\frac{Cs^2}{n_g\|\boldsymbol{\delta}_g\|^2}},$$

which concludes the result.

(e) For any $g \in [G]$ and $j \in [p]$, since $\{u_{g,ij}\}_{i\le n_g}$ and $\{\mathbf{u}'_{g,i}\boldsymbol{\beta}^*\}_{i\le n_g}$ are i.i.d. sub-Gaussian random variables, $\{u_{g,ij}\mathbf{u}'_{g,i}\boldsymbol{\beta}^*\}_{i\le n_g}$ are i.i.d. sub-exponential with $\mathbb{E}(u_{g,ij}\mathbf{u}'_{g,i}\boldsymbol{\beta}^*) =$

$\sum_{\ell=1}^{p} \sigma_{u,j\ell}\beta_\ell^*$. Hence, for a sufficiently small positive number s, there exists a constant $C > 0$, such that

$$\mathbb{P}\Big(\max_j \Big|\frac{1}{n_g}\sum_{i=1}^{n_g} u_{g,ij}\mathbf{u}_{g,i}'\boldsymbol{\beta}^* - \sum_{\ell=1}^{p}\sigma_{u,j\ell}\beta_\ell^*\Big| > s\Big) \le \sum_{j=1}^{p}\mathbb{P}\Big(\Big|\frac{1}{n_g}\sum_{i=1}^{n_g} u_{g,ij}\mathbf{u}_{g,i}'\boldsymbol{\beta}^* - \sum_{\ell=1}^{p}\sigma_{u,j\ell}\beta_\ell^*\Big| > s\Big)$$
$$\le 2pe^{-\frac{Cn_g s^2}{\|\boldsymbol{\beta}^*\|^2}}.$$

Hence, we have $\max_j |(1/n_g)\sum_{i=1}^{n_g} u_{g,ij}\mathbf{u}_{g,i}'\boldsymbol{\beta}^* - \sum_{\ell=1}^{p}\sigma_{u,j\ell}\beta_\ell^*| = O_P(\sqrt{\log p/n_g}\|\boldsymbol{\beta}^*\|)$.

The proofs of (f) – (i) follow similar arguments as in (d). □

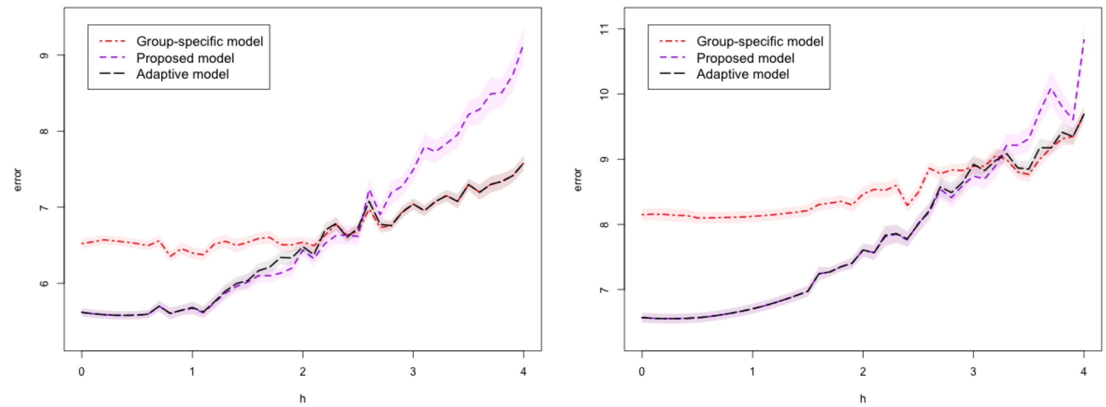
S2 A Rule of Thumb for Model Selection

In practice, it is desirable to have some guidance on choosing between the proposed factor regression model and the group-specific model. To this end, we propose a rule of thumb to choose between them. We define

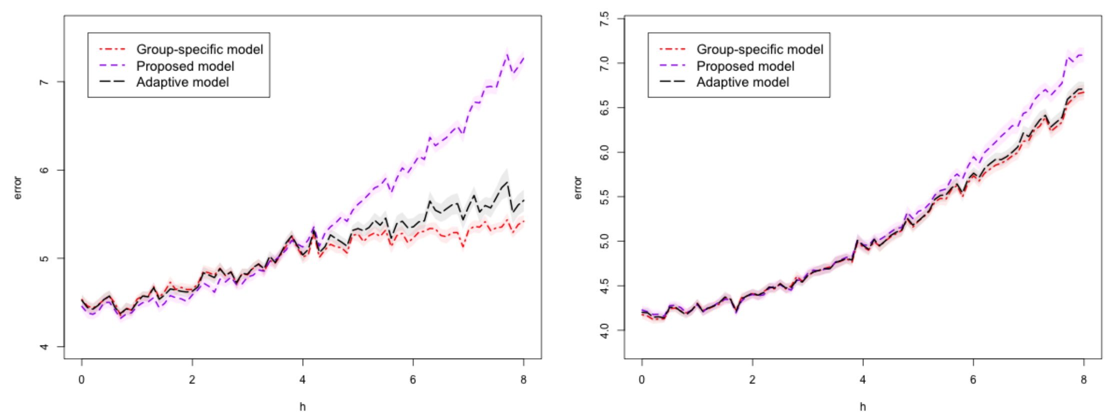
$$R = \frac{\sum_{g=1}^{G}\hat{\boldsymbol{\gamma}}_g'\hat{\mathbf{F}}_g'\hat{\mathbf{F}}_g\hat{\boldsymbol{\gamma}}_g}{\sum_{g=1}^{G}\hat{\boldsymbol{\gamma}}_g'\hat{\mathbf{F}}_g'\hat{\mathbf{F}}_g\hat{\boldsymbol{\gamma}}_g + \hat{\boldsymbol{\beta}}'\hat{\mathbf{U}}'\hat{\mathbf{U}}\hat{\boldsymbol{\beta}}}, \tag{S2.1}$$

where $\hat{\mathbf{F}}_g$, $\hat{\boldsymbol{\gamma}}_g$ and $\hat{\mathbf{U}}$ are as described in Section 3.1 of the main manuscript and $\hat{\boldsymbol{\beta}}$ is the solution to (3.4). We recommend to choose our model over the group-specific model when $R < 0.95$. More explicitly, we first run our method and check if $R < 0.95$. If so, we use our method. Otherwise, we switch to the group-specific model.

To assess the effectiveness of this rule of thumb, we rerun simulations in settings 1 and 2 by using our model only, the group-specific model only, and an adaptive model



(a) Setting 1



(b) Setting 2

Figure S1: The MSE curves given by our model, the group-specific model, the adaptive model using the rule of thumb The left panel represents results for a sparse $\boldsymbol{\beta}^*$ or $\boldsymbol{\beta}_g^*$ and the right panel represents results for a dense $\boldsymbol{\beta}^*$ or $\boldsymbol{\beta}_g^*$.

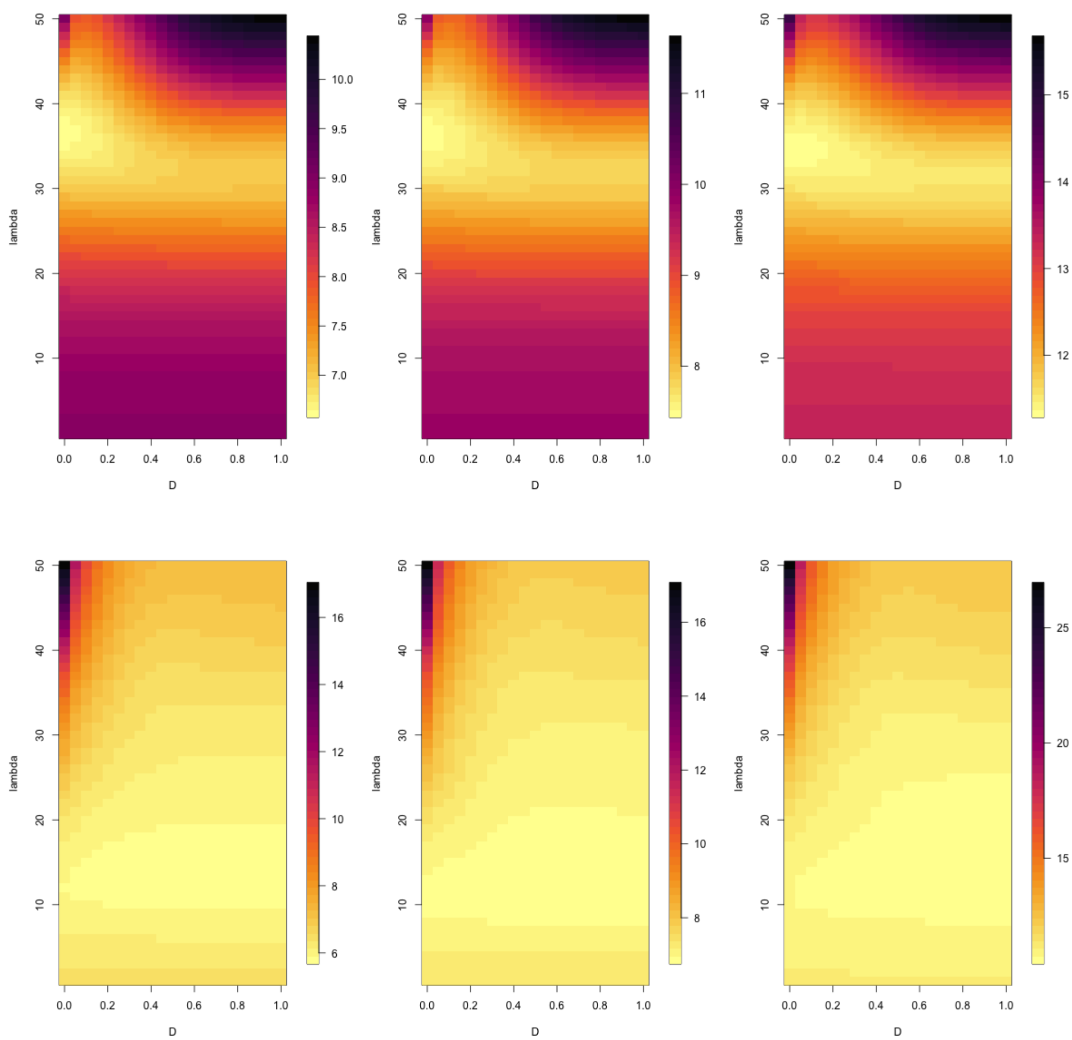
that follows the rule of thumb. Figure S1 gives the MSE curves of the three methods. Figure S1(a) shows that, the adaptive model can almost always catch the better performer between our and the group-specific models. This means that the rule of thumb can correctly pick the better performer in this setting. Figure S1(b) demonstrates a similar finding, even though its left panel shows that, when h is big, the adaptive model is a bit worse than the group-specific model but still much better than the proposed model. This indicates that, in Setting 2 when the underlying true model is the group-specific model, the rule of thumb can correctly choose to use group-specific model most of the time.

S3 Additional Simulation Studies

S3.1 Sensitivity Analysis on the Choice of D

To investigate the role of D in our proposed method, we apply our method to Setting 1 in Section 5.1 using different values of D and λ. In Figure S2, we present heatmaps of out-of-sample MSEs from both sparse and non-sparse cases with $h = 0, 2, 4$. In each heatmap, each grid point represents the MSE from the corresponding combination of D and λ. We find that our model's performance is much less affected by D than by λ, as colors change more along the vertical axis than along the horizontal axis. With an appropriate choice of λ, the minimal out-of-sample MSE is often attained when $D = 0$. This shows that when $p < n$, as in Setting 1 in Section 5.1, it is safe to choose $D = 0$.

Next, under the same setting, we further compare the prediction performance using $D = 0$ versus tuning D by ten-fold cross validation. In particular, in one case we fix $D = 0$ and tune λ by ten-fold cross validation and in another case we tune both D and λ by ten-fold cross validation. We summarize the out-of-sample MSEs by these two methods in Table S1. We find that the prediction performance by fixing $D = 0$ is comparable to that by tuning D. Thus, to save computational time, we recommend choosing a fixed D instead of tuning it.



(a) $h = 0$ (b) $h = 2$ (c) $h = 4$

Figure S2: Heatmaps of testing prediction MSEs from both the sparse and non-sparse case with $h = 0, 2, 4$ from setting 1 where data are generated from the proposed model. The top and bottom panels of plots are from dense and sparse $\boldsymbol{\beta}^*$ case respectively.

Table S1: Out-of-sample MSEs by different choices of D under Setting 1.

setting		D by CV	$D=0$
dense	$h=0$	6.700 (0.088)	6.695 (0.085)
	$h=2$	7.544 (0.103)	7.520 (0.107)
	$h=4$	11.395 (0.294)	11.417 (0.290)
sparse	$h=0$	5.563 (0.073)	5.557 (0.076)
	$h=2$	6.559 (0.112)	6.530 (0.116)
	$h=4$	10.348 (0.316)	10.310 (0.312)

S3.2 Sensitivity Analysis on Spiked Eigenvalues

In Section 5.1 and 5.2, we choose the values of $\lambda_{g,1},\ldots,\lambda_{g,K_g}$ to simulate the top K_g eigenvalues of $\mathbf{X}_g$. In this way, $\mathbf{X}_g$ generated from our simulations can satisfy the pervasiveness condition in Assumption 1, which is to ensure the latent factors can be well estimated by the PCA method in Section 3.1. More specifically, our $\lambda_{g,1},\ldots,\lambda_{g,K_g}$ are chosen as follows: $(\lambda_{1,1},\lambda_{1,2},\lambda_{1,3})=(1,\Gamma(2),\Gamma(3))*7$, $(\lambda_{2,1},\lambda_{2,2},\lambda_{2,3})=(1,\Gamma(2.25),\Gamma(3.25))*10$, $(\lambda_{2,1},\lambda_{2,2},\lambda_{2,3})=(1,\Gamma(2.5),\Gamma(3.5))*13$, where $\Gamma(\cdot)$ represents the gamma function.

As suggested by one of our reviewers, in this section, we conduct a sensitivity study on choice of $\lambda_{g,1},\ldots,\lambda_{g,K_g}$. In particular, we conduct two additional sets of simulations with new choice of $\lambda_{g,1},\ldots,\lambda_{g,K_g}$ on whole numbers:

(i) $(\lambda_{1,1},\lambda_{1,2},\lambda_{1,3})=(10,5,2)$, $(\lambda_{2,1},\lambda_{2,2},\lambda_{2,3})=(7,2,1)$, $(\lambda_{2,1},\lambda_{2,2},\lambda_{2,3})=(12,3,1)$;

(ii) $(\lambda_{1,1},\lambda_{1,2},\lambda_{1,3})=(5,3,1)$, $(\lambda_{2,1},\lambda_{2,2},\lambda_{2,3})=(6,3,2)$, $(\lambda_{2,1},\lambda_{2,2},\lambda_{2,3})=(7,4,2)$.

For each set of $\lambda_{g,1},\ldots,\lambda_{g,K_g}$, we conduct the same simulations for both Settings 1 and 2 as in Sections 5.1 and 5.2. We visualize the MSE curves for (i) and (ii) in Figures S3 and S4 respectively. In particular, each figure contains both the sparse and dense cases from both Settings 1 and 2. We can see very similar patterns of the plots from (i) and (ii) compared to the MSE curves given by the original choice of $\lambda_{g,1}, \ldots, \lambda_{g,K_g}$. Hence, we conclude that our simulation results are not sensitive to the choice of eigen values $\lambda_{g,1}, \ldots, \lambda_{g,K_g}$.

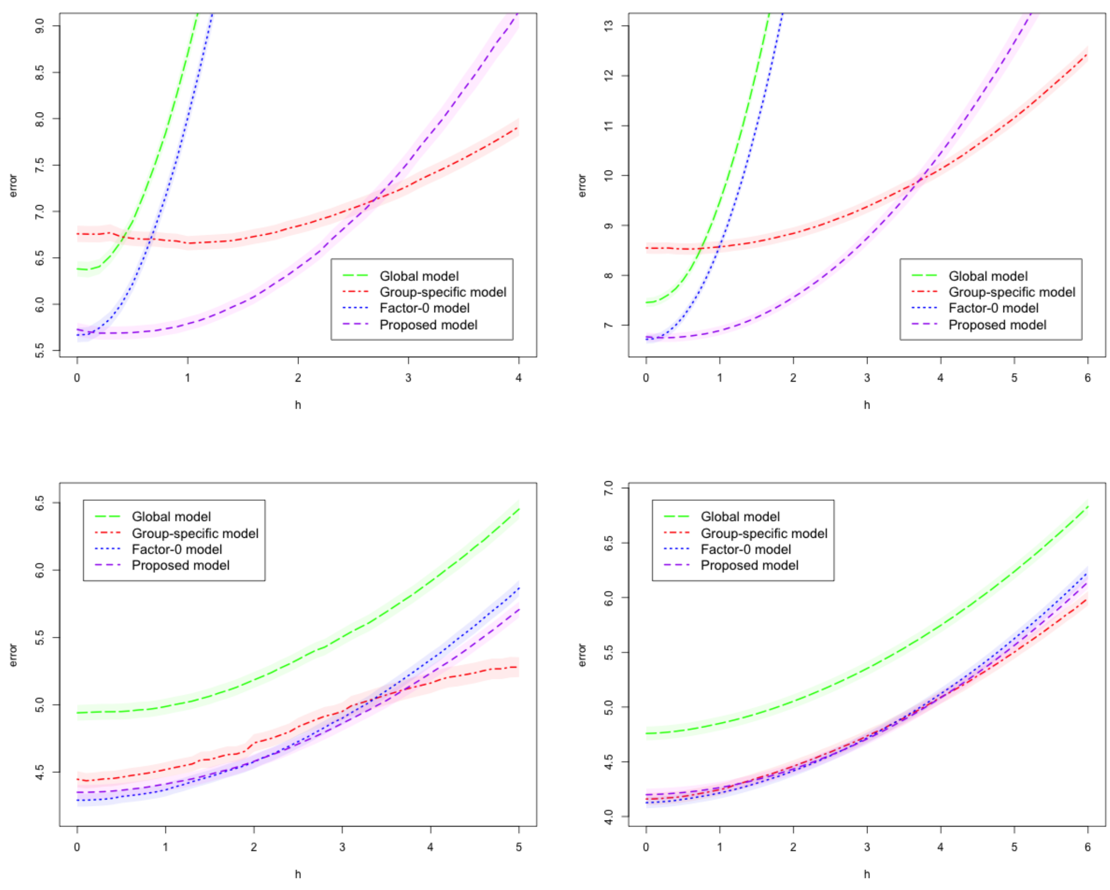


Figure S3: The MSE curves given by the four models for settings 1 and 2 with choice of $\lambda_{g,1}, \ldots, \lambda_{g,K_g}$ on (i). The left panel represents results for a sparse $\boldsymbol{\beta}^*$ or $\boldsymbol{\beta}_g^*$ and the right panel represents results for a dense $\boldsymbol{\beta}^*$ or $\boldsymbol{\beta}_g^*$. The top and bottom panels represent the MSE curves from setting 1 and setting 2 respectively.

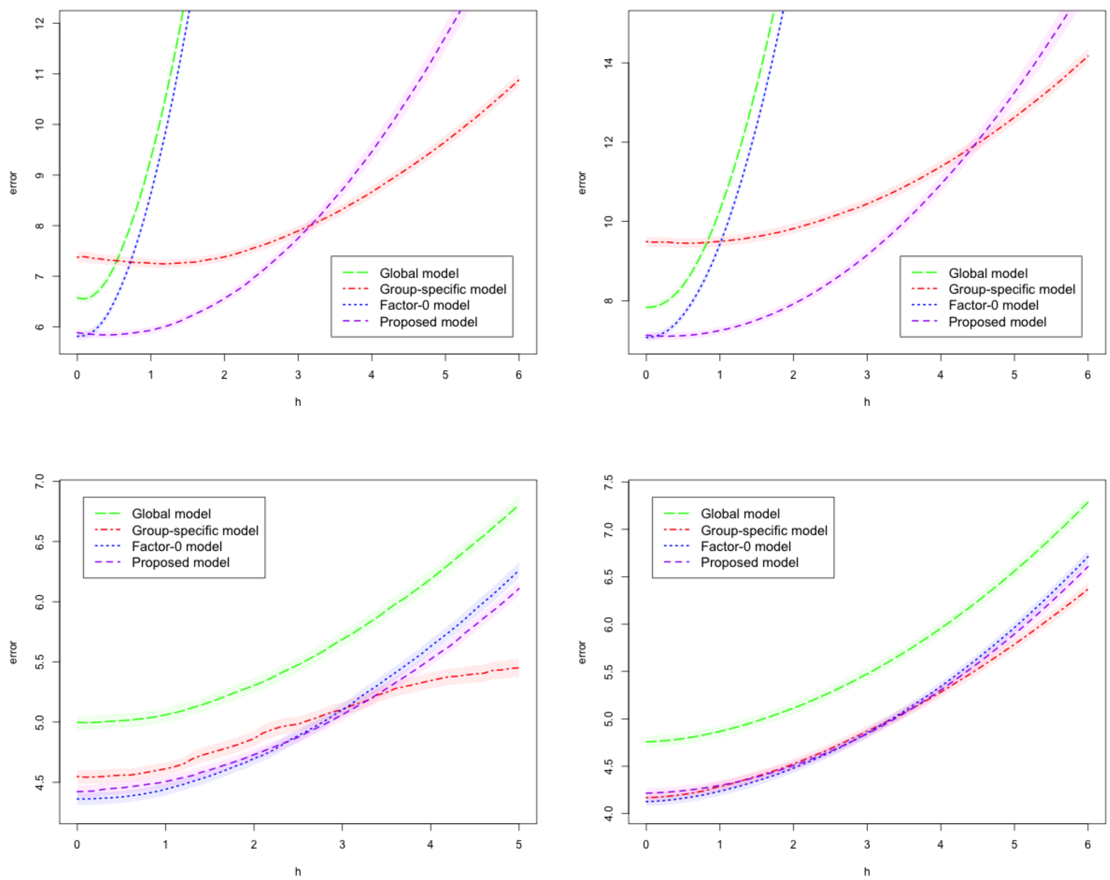


Figure S4: The MSE curves given by the four models for settings 1 and 2 with choice of $\lambda_{g,1}, \ldots, \lambda_{g,K_g}$ on (ii). The left panel represents results for a sparse $\boldsymbol{\beta}^*$ or $\boldsymbol{\beta}_g^*$ and the right panel represents results for a dense $\boldsymbol{\beta}^*$ or $\boldsymbol{\beta}_g^*$. The top and bottom panels represent the MSE curves from setting 1 and setting 2 respectively.

S3.3 Other settings of $\mathbf{f}_{g,i}$ and $\mathbf{u}_{g,i}$

As suggested by one of the reviewers, in this section, we consider the same settings as in Sections 5.1 and 5.2, except that we generate $\{\mathbf{f}_{g,i}\}_{i\leq n_g}$ and $\{\mathbf{u}_{g,i}\}_{i\leq n_g}$ as i.i.d. samples from a t-distribution with 10 degrees of freedom. Figures S5 and S6 show the MSE curves of four methods in Settings 1 and 2 respectively. We have similar conclusion as in the cases where data are generated from a multi-normal distribution. Under setting 1, it is seen from Figure S5 that our method still outperforms the other methods when h is small. When h becomes larger, the between-group heterogeneity increases. In that case, our model's performance gets worse than the group-specific model, which is also observed when $\mathbf{f}_{g,i}$ and $\mathbf{u}_{g,i}$ follow multi-normal distribution. Under setting 2, it is seen from Figure S5 that our method still has similar performance as the group-specific model, when the group-specific model is the truth.

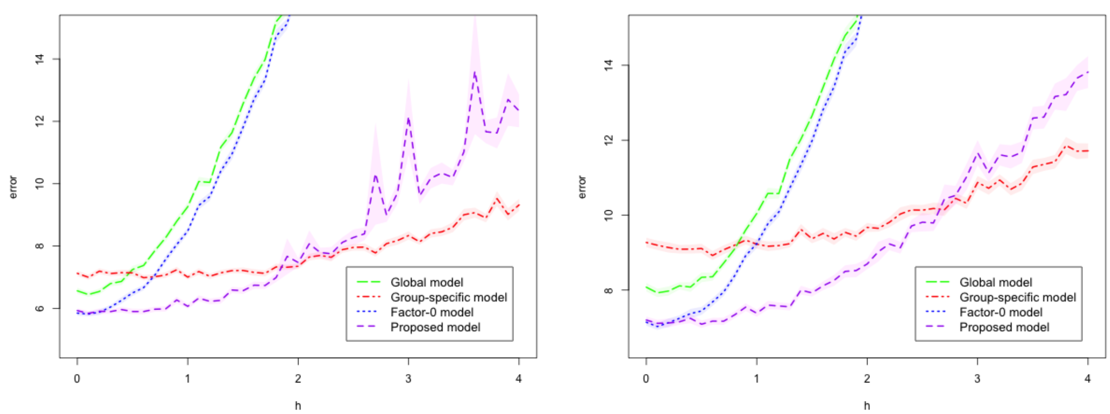


Figure S5: The MSE curves given by the four models under setting 1 when $\mathbf{f}_{g,i}$ and $\mathbf{u}_{g,i}$ follow the t-distribution with 10 degrees of freedom. The left panel represents results for a sparse $\boldsymbol{\beta}^*$ and the right panel represents results for a dense $\boldsymbol{\beta}^*$.

S3.4 Other choices of K_g

As suggested by one of the reviewers, in this section, we let $K_g = g$ for $g = 1, 2, 3$ so that it varies across groups. We consider two data generating schemes similar to

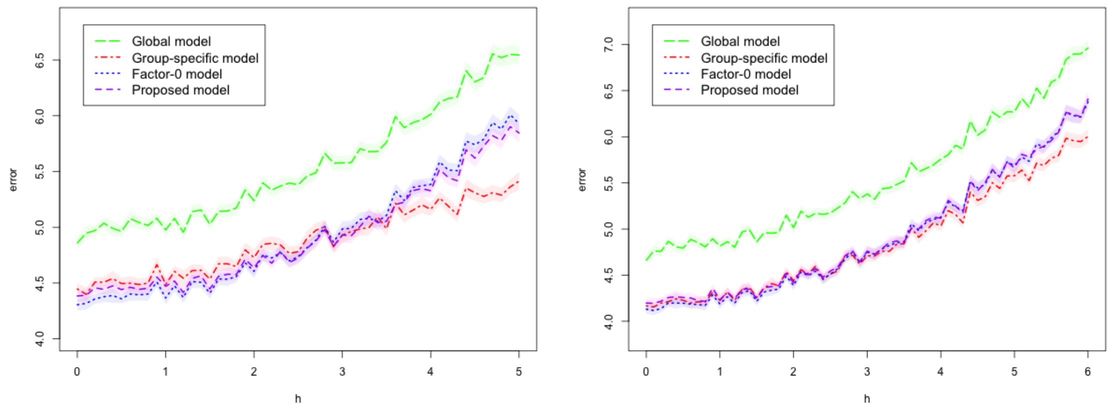


Figure S6: The MSE curves given by the four models for under setting 2 when $\mathbf{f}_{g,i}$ and $\mathbf{u}_{g,i}$ follow the t-distribution with 10 degrees of freedom. The left panel represents results for a sparse $\boldsymbol{\beta}^*$ and the right panel represents results for a dense $\boldsymbol{\beta}^*$.

Settings 1 and 2 described in Sections 5.1 and 5.2. In both cases, we keep choosing $G = 3, p = 200, n_g = 100$, $\mu_g^* = g$ for $g = 1, 2, 3$ and generate $\boldsymbol{\epsilon}$ as i.i.d. samples from $\mathcal{N}(0, 4)$.

For Setting 1, we use the same method to generate $\boldsymbol{\Lambda}_g$ as in Section 5.1, and let $\lambda_{1,1} = 10$, $(\lambda_{2,1}, \lambda_{2,2}) = (9, 4)$, $(\lambda_{3,1}, \lambda_{3,2}, \lambda_{3,3}) = (8, 3, 2)$. Moreover, we let $\boldsymbol{\Sigma}_u$ be the diagonal matrix with all diagonal elements equal to $2/15$. We set $\gamma_1^* = h$, $\boldsymbol{\gamma}_2^* = (h, h)'$ and $\boldsymbol{\gamma}_3^* = (h, h, h)'$. For a sparse $\boldsymbol{\beta}^*$, we let $\boldsymbol{\beta}^* = (\mathbf{0.7}_{10}, \mathbf{0}_{90}, -\mathbf{0.7}_{10}, \mathbf{0}_{90})'$. For a dense $\boldsymbol{\beta}^*$, we set $\boldsymbol{\beta}^* = (\mathbf{0.3}_{80}, \mathbf{0}_{20}, -\mathbf{0.3}_{80}, \mathbf{0}_{20})'$.

For Setting 2, we use the same method to generate $\boldsymbol{\Lambda}_g$ and $\boldsymbol{\Sigma}_u$ as in Section 5.2, and use the same $\lambda_{g,1}, \ldots, \lambda_{g,K_g}$ as in Setting 1 in above. For a sparse $\boldsymbol{\beta}_g^*$, we set $\boldsymbol{\beta}_1^* = (10h, 0, 0, \mathbf{0.25}_5, \mathbf{0}_{187}, -\mathbf{0.25}_5)'$, $\boldsymbol{\beta}_2^* = (-10h, 10h, 0, \mathbf{0.25}_5, \mathbf{0}_{187}, -\mathbf{0.25}_5)'$ and

$\boldsymbol{\beta}_3^* = (10h, -10h, 10h, \mathbf{0.25}_5, \mathbf{0}_{187}, -\mathbf{0.25}_5)'$. For a dense $\boldsymbol{\beta}_g^*$, we set $\boldsymbol{\beta}_1^* = (10h, 0, 0, \mathbf{0.25}_{80}, \mathbf{0}_{37}, -\mathbf{0.25}_{80})'$, $\boldsymbol{\beta}_2^* = (-10h, 10h, 0, \mathbf{0.25}_{80}, \mathbf{0}_{37}, -\mathbf{0.25}_{80})'$ and $\boldsymbol{\beta}_3^* = (10h, -10h, 10h, \mathbf{0.25}_{80}, \mathbf{0}_{37}, -\mathbf{0.25}_{80})'$.

Figures S7 and S8 show MSE curves of the four methods in Settings 1 and 2 respectively. We observe similar finding as in Sections 5.1 and 5.2. For Setting 1, the global model and the Factor-0 model cannot capture between-group heterogeneity, hence are significantly worse than the proposed model and the group-specific model. When the between-group heterogeneity is moderate, our proposed model is able to provide the best prediction by capturing the globally-shared and group-specific signals. The group-specific model gradually improves as h increases. For Setting 2, the differences among the three models other than the global model are small when h is small. As h increases, the predictions from the Factor-0 model and our model get worse, though the former deteriorates much faster. This is due to model mis-specification and the fact that between-group heterogeneity increases as h gets larger. Such a finding agrees with Corollary 4.

S3.5 Estimation errors

In this section, we provide parameter estimation errors from our, group-specific and global models in Tables S2 – S4 for both Settings 1 and 2. Tables S2 and S3 are for Setting 1, where data are generated from our proposed model. Table S4 is for Setting 2, where data are generated from the group-specific model. Table S2 summarizes

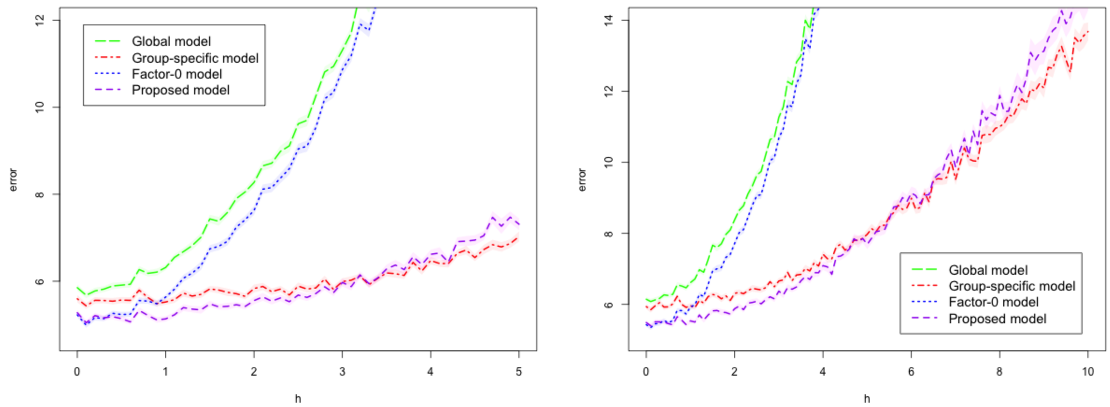


Figure S7: The MSE curves given by the four models under Setting 1. The left panel represents results for a sparse $\boldsymbol{\beta}^*$ and the right panel represents results for a dense $\boldsymbol{\beta}^*$.

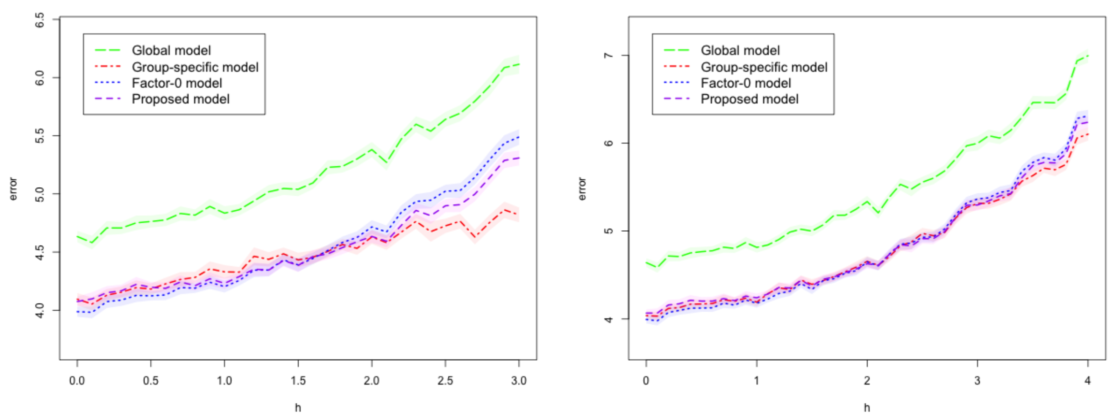


Figure S8: The MSE curves given by the four models under Setting 2. The left panel represents results for a sparse $\boldsymbol{\beta}^*$ and the right panel represents results for a dense $\boldsymbol{\beta}^*$.

estimation errors of $\hat{\boldsymbol{\gamma}}_g$ in terms of $\|\hat{\boldsymbol{\gamma}}_g - \mathbf{H}_g \boldsymbol{\gamma}_g^*\|$, where $\mathbf{H}_g$ is the rotation matrix defined in Theorem 1. Tables S3 and S4 summarize estimation errors of $\boldsymbol{\beta}^*$ (Setting 1) and $\boldsymbol{\beta}_2^*$ (Setting 2) respectively, where we denote the global model estimator as $\hat{\boldsymbol{\beta}}_{\lambda,global}$, the group-specific model estimator as $\hat{\boldsymbol{\beta}}_{g,\lambda}$ for $g = 1, 2, 3$; and our estimator as $\hat{\boldsymbol{\beta}}_{\lambda}$.

For each simulation setting, we pick three h values to represent low, medium and high between-group heterogeneity.

For Setting 1, as can be seen from Table S2, the estimation errors from our proposed estimator $\hat{\boldsymbol{\gamma}}_g$ are reasonably small considering the corresponding scales of true parameters $\boldsymbol{\gamma}_g^*$. In Table S3, our estimator $\hat{\boldsymbol{\beta}}_\lambda$ always achieves the best performance compared to estimators from the group-specific and global models, for all levels of between-group heterogeneity.

For Setting 2, it is expected that our proposed estimator $\hat{\boldsymbol{\beta}}_\lambda$ does not always perform the best, because we generate the underlying data from a mis-specified model. Unsurprisingly, when there is no between-group heterogeneity, i.e. $h = 0$, the estimator from the global model performs the best. Moreover, there is only a marginal gap from our proposed estimator to the best one. When there is high between-group heterogeneity, the estimator from the group-specific model achieves the smallest estimation errors. As indicated by Theorem 4, the estimation error from our proposed model is influenced by an extra term introduced from between-group heterogeneity, which explains the performance gap between our proposed estimator and group-specific model's estimator.

setting		$\|\hat{\gamma}_1 - \mathbf{H}_1\gamma_1^*\|^2$	$\|\hat{\gamma}_2 - \mathbf{H}_2\gamma_2^*\|^2$	$\|\hat{\gamma}_3 - \mathbf{H}_3\gamma_3^*\|^2$
dense	$h = 0$	0.088 (0.008)	0.082 (0.008)	0.092 (0.010)
	$h = 2$	0.411 (0.049)	0.380 (0.045)	0.308 (0.036)
	$h = 4$	1.151 (0.162)	1.478 (0.161)	1.307 (0.178)
sparse	$h = 0$	0.079 (0.008)	0.065 (0.008)	0.056 (0.006)
	$h = 2$	0.386 (0.046)	0.370 (0.044)	0.309 (0.038)
	$h = 4$	1.163 (0.165)	1.434 (0.155)	1.260 (0.169)

Table S2: The estimation errors of $\boldsymbol{\gamma}_g^*$ from our model in Setting 1.

setting		$\|\hat{\boldsymbol{\beta}}_{\lambda,global} - \boldsymbol{\beta}^*\|^2$	$\|\hat{\boldsymbol{\beta}}_{1,\lambda} - \boldsymbol{\beta}^*\|^2$	$\|\hat{\boldsymbol{\beta}}_{2,\lambda} - \boldsymbol{\beta}^*\|^2$	$\|\hat{\boldsymbol{\beta}}_{3,\lambda} - \boldsymbol{\beta}^*\|^2$	$\|\hat{\boldsymbol{\beta}}_{\lambda} - \boldsymbol{\beta}^*\|^2$
dense	$h = 0$	0.434 (0.006)	0.670 (0.011)	0.674 (0.010)	0.673 (0.013)	0.400 (0.005)
	$h = 2$	0.832 (0.013)	0.671 (0.011)	0.715 (0.010)	0.677 (0.010)	0.410 (0.007)
	$h = 4$	2.221 (0.052)	0.832 (0.013)	0.964 (0.016)	0.724 (0.007)	0.438 (0.005)
sparse	$h = 0$	0.252 (0.007)	0.424 (0.015)	0.394 (0.005)	0.405 (0.012)	0.247 (0.005)
	$h = 2$	0.807 (0.018)	0.373 (0.003)	0.413 (0.005)	0.396 (0.002)	0.237 (0.006)
	$h = 4$	2.649 (0.085)	0.507 (0.003)	0.624 (0.004)	0.449 (0.001)	0.270 (0.007)

Table S3: The estimation errors of $\boldsymbol{\beta}^*$ from the three models in Setting 1.

S4 Additional ADNI Data Analysis Results

We further represent brain connections using precision matrices estimated from Gaussian graphical models (Cai et al., 2011). Let $\boldsymbol{\Omega}_{x,g} = \boldsymbol{\Sigma}_{x,g}^{-1}$, $\boldsymbol{\Omega}_{\boldsymbol{\Lambda}'\boldsymbol{\Lambda},g} = (\boldsymbol{\Lambda}_g'\boldsymbol{\Lambda}_g)^{-1}$ and $\boldsymbol{\Omega}_u = \boldsymbol{\Sigma}_u^{-1}$. In Figure S9, we demonstrate the heatmaps of adjacency matrices corresponding to estimated precision matrices for the NC and AD groups. We choose two tuning parameters ($\nu = 0.2$ and $\nu = 0.3$) in the Gaussian graphical model to give graphs at different sparsity levels. The top, middle and bottom panels correspond to $\hat{\boldsymbol{\Omega}}_{x,g}$, $\hat{\boldsymbol{\Omega}}_{\boldsymbol{\Lambda}'\boldsymbol{\Lambda},g}$ and $\hat{\boldsymbol{\Omega}}_{u,g}$ respectively, where $\hat{\boldsymbol{\Omega}}_{u,g}$ is the Gaussian graphical estimators of $\boldsymbol{\Omega}_u$ by only using data from group g. We can see that the heatmaps of $\hat{\boldsymbol{\Omega}}_{u,g}$ in NC and AD groups are much more similar than those of $\hat{\boldsymbol{\Omega}}_{x,g}$. In both plots, it is interesting

setting		$\|\hat{\boldsymbol{\beta}}_{\lambda,global} - \boldsymbol{\beta}_2^*\|^2$	$\|\hat{\boldsymbol{\beta}}_{2,\lambda} - \boldsymbol{\beta}_2^*\|^2$	$\|\hat{\boldsymbol{\beta}}_{\lambda} - \boldsymbol{\beta}_2^*\|^2$
dense	$h = 0$	0.004 (0.000)	0.014 (0.005)	0.006 (0.001)
	$h = 3$	0.072 (0.000)	0.082 (0.008)	0.077 (0.001)
	$h = 6$	0.271 (0.001)	0.244 (0.002)	0.276 (0.002)
sparse	$h = 0$	0.025 (0.001)	0.047 (0.007)	0.028 (0.001)
	$h = 2.5$	0.077 (0.001)	0.091 (0.005)	0.082 (0.003)
	$h = 5$	0.215 (0.004)	0.181 (0.011)	0.223 (0.005)

Table S4: The estimation errors of $\boldsymbol{\beta}_2^*$ from the three models in Setting 2.

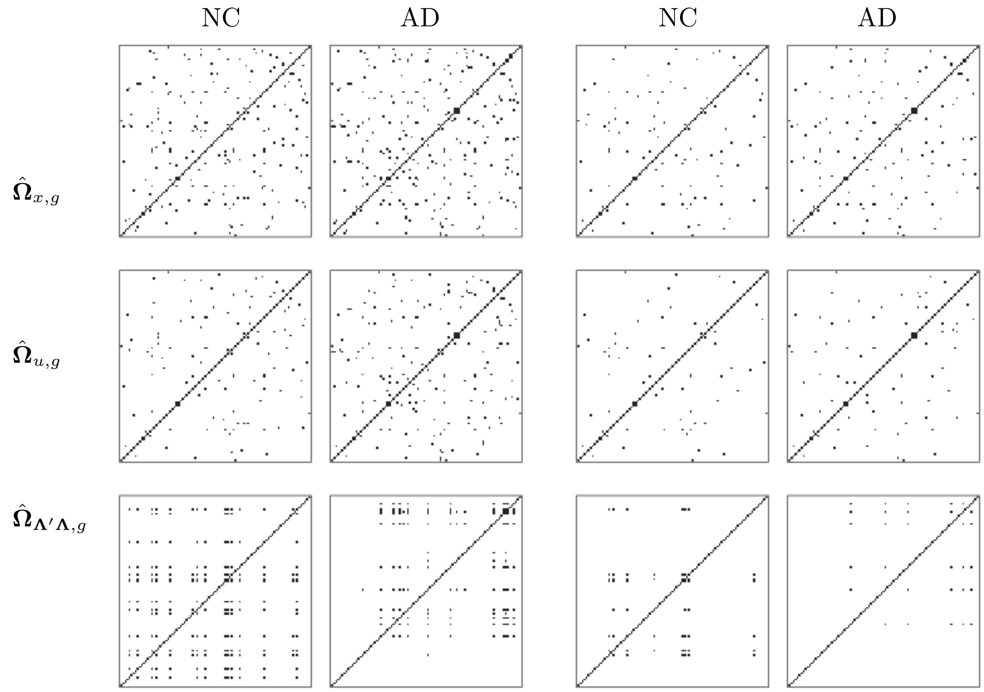
to note that the heatmaps of $\hat{\boldsymbol{\Omega}}_{\boldsymbol{\Lambda}'\boldsymbol{\Lambda},g}$ from the NC group are much denser than those from the AD group. This shows that the AD patients may have a significant loss of brain connections. To further investigate this difference, we explore the ROI connections that are selected in the NC group but not in the AD group on $\hat{\boldsymbol{\Omega}}_{\boldsymbol{\Lambda}'\boldsymbol{\Lambda},g}$. We find that the frontal, parietal and occipital lobes suffer significant loss of connections in the AD group, which is consistent with the previous findings in the Alzheimer's disease literature (Johnson et al., 2012; Zhang et al., 2015). We list the detected regions in Table S5.

S5 Application to Microarray Data Analysis

To further illustrate the effectiveness of our model, we apply it to the analysis of three microarray data and compare it with global and group-specific models. The microarray data come from three related cardiovascular disease studies on finding the key genes that mediate atherosclerotic and inflammatory process. The three microarray data are publicly available on Gene Expression Omnibus via the accession names as GSE12288,

Table S5: ROIs with loss of connections on $\hat{\mathbf{\Omega}}_{\mathbf{\Lambda}'\mathbf{\Lambda},g}$ from AD group compared to NC group.

	ROI1	ROI2
1	frontal lobe WM right	angular gyrus right
2	frontal lobe WM right	frontal lobe WM left
3	angular gyrus right	frontal lobe WM left
4	angular gyrus right	superior parietal lobule left
5	frontal lobe WM left	superior parietal lobule left
6	frontal lobe WM right	occipital lobe WM left
7	angular gyrus right	occipital lobe WM left
8	frontal lobe WM left	occipital lobe WM left
9	frontal lobe WM right	postcentral gyrus left
10	angular gyrus right	postcentral gyrus left
11	frontal lobe WM left	postcentral gyrus left
12	superior parietal lobule left	postcentral gyrus left
13	occipital lobe WM left	postcentral gyrus left
14	frontal lobe WM right	precentral gyrus left
15	angular gyrus right	precentral gyrus left
16	frontal lobe WM left	precentral gyrus left
17	superior parietal lobule left	precentral gyrus left
18	occipital lobe WM left	precentral gyrus left
19	postcentral gyrus left	precentral gyrus left
20	frontal lobe WM right	angular gyrus left
21	angular gyrus right	angular gyrus left
22	frontal lobe WM left	angular gyrus left
23	occipital lobe WM left	angular gyrus left
24	postcentral gyrus left	angular gyrus left
25	precentral gyrus left	angular gyrus left



(a) $\nu = 0.2$ (b) $\nu = 0.3$

Figure S9: Heatmaps of the adjacency matrices corresponding to $\hat{\mathbf{\Omega}}_{x,g}$, $\hat{\mathbf{\Omega}}_{u,g}$ and $\hat{\mathbf{\Omega}}_{\mathbf{\Lambda}'\mathbf{\Lambda},g}$ for NC and AD groups with the tuning parameter (a) $\nu = 0.2$, (b) $\nu = 0.3$. Each black dot in the plot indicates that the corresponding two covariates are partially correlated.

GSE16561 and GSE20129. There are 222, 63 and 119 samples in these three datasets. It was identified by Li et al. (2014) that the gene "IFNA4" plays a key role in the process. In order to understand how this gene connects to the other 464 genes in the 20 immunity-related signaling pathways, we propose to regress the expression of IFNA on the expression of the rest genes using four models. For the global model, we concatenate three datasets and run a single regression on the combined data. The group-specific model is applied separately to each of the three datasets. We also run the proposed and Factor-0 models on the three datasets together.

We randomly partition all samples into 75% for training and 25% for testing. We repeat the random split for 50 times. The overall and groupwise Mean Squared Errors (MSEs) with corresponding standard errors are reported in Tables S6 and S7 respectively. For each model, we use three penalty functions, namely the ℓ_2-penalty (Ridge), the ℓ_1-penalty (Lasso), and the Elastic Net (EN) penalty with the bridging parameter equals to 0.5.

As shown in Table S6, our proposed model outperforms the other three competitors by a fair margin in terms of the overall MSE. The global model performs poorly as it ignores the heterogeneity among the three studies and the group-specific model fails as the sample sizes in each study are not large. On the contrary, our method can find a good balance between the two and achieves better prediction. The Factor-0 model adjusts for group-specific means. Compared with our model, its prediction is still worse, which suggests that there are some additional latent factors need to be adjusted as done by our method. These factors could potentially be due to the batch effects. In terms of the groupwise MSE, our model performs the best in two of the three studies, even though the group-specific model has the best performance in GSE20129.

Table S6: Overall MSEs for the four models.

Penalty	Global	Group-specific	Factor-0	Proposed
Ridge	1.509 (0.029)	1.504 (0.031)	1.524 (0.030)	**1.461** (0.029)
EN	1.735 (0.031)	1.692 (0.039)	1.643 (0.032)	**1.582** (0.032)
Lasso	1.727 (0.031)	1.708 (0.042)	1.659 (0.032)	**1.619** (0.033)

Table S7: Groupwise MSEs for the four models.

	Global	Group-specific	Factor-0	Proposed
		Penalty = Ridge		
GSE12288	2.016 (0.050)	2.059 (0.049)	2.082 (0.052)	**2.000** (0.050)
GSE16561	1.334 (0.092)	1.501 (0.098)	1.308 (0.082)	**1.223** (0.080)
GSE20129	0.638 (0.025)	**0.452** (0.020)	0.578 (0.023)	0.562 (0.021)
		Penalty = Elastic Net		
GSE12288	2.259 (0.055)	2.361 (0.063)	2.234 (0.057)	**2.179** (0.058)
GSE16561	1.640 (0.102)	1.598 (0.139)	1.438 (0.089)	**1.277** (0.099)
GSE20129	0.790 (0.030)	**0.472** (0.023)	0.628 (0.027)	0.609 (0.025)
		Penalty = Lasso		
GSE12288	**2.233** (0.058)	2.386 (0.069)	2.254 (0.058)	2.244 (0.060)
GSE16561	1.714 (0.112)	1.581 (0.134)	1.464 (0.091)	**1.267** (0.089)
GSE20129	0.773 (0.031)	**0.487** (0.024)	0.632 (0.027)	0.615 (0.025)

References

Cai, T., W. Liu, and X. Luo (2011). A constrained $\ell1$ minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association 106*(494), 594–607.

Fan, J., Y. Liao, and M. Mincheva (2013). Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 75*(4), 603–680.

Fan, J., H. Liu, W. Wang, Z. Zhu, et al. (2018). Heterogeneity adjustment with applications to graphical model inference. *Electronic Journal of Statistics 12*(2), 3908–3952.

Johnson, K. A., N. C. Fox, R. A. Sperling, and W. E. Klunk (2012). Brain imaging in alzheimer disease. *Cold Spring Harbor perspectives in medicine 2*(4), a006213.

Li, Q., S. Wang, C.-C. Huang, M. Yu, and J. Shao (2014). Meta-analysis based variable selection for gene expression data. *Biometrics 70*(4), 872–880.

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

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, Volume 47. Cambridge university press.

Zhang, Y., Z. Dong, P. Phillips, S. Wang, G. Ji, J. Yang, and T.-F. Yuan (2015). Detection of subjects and brain regions related to alzheimer's disease using 3d mri scans based on eigenbrain and machine learning. *Frontiers in computational neuroscience 9*, 66.