Multivariate Varying-coefficient Models via Tensor Decomposition

Fengyu Zhang, Ya Zhou, Kejun He, and Raymond K. W. Wong

Renmin University of China, Chinese Academy of Medical Sciences and Peking Union Medical College, Renmin University of China, and Texas A&M University

Supplementary Material

S.1 Computation Details

S.1.1 Details on Updating $\boldsymbol{S}$

In this subsection, we provide more details on solving (4.1) of the main paper through the MM algorithm. To construct the majorized function for $P_{\boldsymbol{A}^{(t)}}(\boldsymbol{S})$, we first let $\boldsymbol{S}_k$ and $\boldsymbol{S}_{(1),k}$ be the k-th iterations of $\boldsymbol{S}$ and $\boldsymbol{S}_{(1)}$ in the MM algorithm respectively. Using the concave property of the square root function, we have

$$\|\boldsymbol{x}\|_2 = \sqrt{\|\boldsymbol{x}\|_2^2} \leq \sqrt{\|\boldsymbol{x}_0\|_2^2} + \frac{\|\boldsymbol{x}\|_2^2 - \|\boldsymbol{x}_0\|_2^2}{2\sqrt{\|\boldsymbol{x}_0\|_2^2}} := g(\boldsymbol{x}|\boldsymbol{x}_0).$$

It can be verified that $g(\boldsymbol{x}|\boldsymbol{x}_0)$ is a majorization function for $\|\boldsymbol{x}\|_2$. Since $P_{\boldsymbol{A}^{(t)}}(\boldsymbol{S})$ is a summation over ℓ_2 norms, a majorized function for $P_{\boldsymbol{A}^{(t)}}(\boldsymbol{S})$ at $\boldsymbol{S}_k$ is

$$\Phi_{\boldsymbol{A}^{(t)}}(\boldsymbol{S}|\boldsymbol{S}_k) = \lambda \sum_{j=1}^{p} \left\{ \sqrt{\|(\boldsymbol{a}_j^{(t)})^\intercal \boldsymbol{S}_{(1),k}\|_2^2} + \frac{\|(\boldsymbol{a}_j^{(t)})^\intercal \boldsymbol{S}_{(1)}\|_2^2 - \|(\boldsymbol{a}_j^{(t)})^\intercal \boldsymbol{S}_{(1),k}\|_2^2}{2\|(\boldsymbol{a}_j^{(t)})^\intercal \boldsymbol{S}_{(1),k}\|_2} \right\}. \tag{S.1}$$

Denoting $\lambda_{j,k}^{(t)} = 1/\{2\|(\boldsymbol{a}_j^{(t)})^\intercal \boldsymbol{S}_{(1),k}\|_2\}$, $j = 1, \ldots, p$, and $\boldsymbol{\Sigma}_k^{(t)} = \mathrm{diag}(\lambda_{1,k}^{(t)}, \ldots, \lambda_{p,k}^{(5)})$, the majorized function $\Phi_{\boldsymbol{A}^{(t)}}(\boldsymbol{S}|\boldsymbol{S}_k)$ in (S.1) can be rewritten as

$$\begin{aligned}\Phi_{\boldsymbol{A}^{(t)}}(\boldsymbol{S}|\boldsymbol{S}_k) = P_{\boldsymbol{A}^{(t)}}(\boldsymbol{S}_k) + \lambda \Big[& \mathrm{vec}(\boldsymbol{S}_{(1)})^\intercal \{\boldsymbol{I}_{R_2 R_3} \otimes (\boldsymbol{A}^{(t)})^\intercal \boldsymbol{\Sigma}_k^{(t)} \boldsymbol{A}^{(t)}\} \mathrm{vec}(\boldsymbol{S}_{(1)}) \\ & - \mathrm{vec}(\boldsymbol{S}_{(1),k})^\intercal \{\boldsymbol{I}_{R_2 R_3} \otimes (\boldsymbol{A}^{(t)})^\intercal \boldsymbol{\Sigma}_k^{(t)} \boldsymbol{A}^{(t)}\} \mathrm{vec}(\boldsymbol{S}_{(1),k}) \Big].\end{aligned}$$

Note that $\Phi_{\boldsymbol{A}^{(t)}}(\boldsymbol{S}|\boldsymbol{S}_k)$ is well-defined except for $\|(\boldsymbol{a}_j^{(t)})^\intercal \boldsymbol{S}_{(1),k}\|_2 = 0$. When $\|(\boldsymbol{a}_j^{(t)})^\intercal \boldsymbol{S}_{(1),k}\|_2 = 0$, $j = 1, 2, \ldots, p$, we make a further discussion in two cases: (i) $\|\boldsymbol{a}_j^{(t)}\|_2 = 0$; (ii) $\|\boldsymbol{a}_j^{(t)}\|_2 \neq 0$, but $\boldsymbol{a}_j^{(t)}$ belongs to the null space of $\boldsymbol{S}_{(1),k}^\intercal$.

For case (i), denote $\widetilde{\mathcal{J}} = \{j : \|\boldsymbol{a}_j^{(t)}\|_2 = 0, j = 1, \ldots, p\}$. If $j \in \widetilde{\mathcal{J}}$, the j-th component in $P_{\boldsymbol{A}^{(t)}}(\boldsymbol{S})$ equals zero, which allows us to set the j-th diagonal element in $\boldsymbol{\Sigma}_k^{(t)}$ to be zero. Define $\widetilde{\boldsymbol{\Sigma}}_k^{(t)} = \mathrm{diag}(\widetilde{\lambda}_{1,k}^{(t)}, \ldots, \widetilde{\lambda}_{p,k}^{(t)})$ and $\widetilde{\lambda}_{j,k}^{(t)} = \lambda_{j,k}^{(t)} \cdot \mathbf{I}\{j \in \widetilde{\mathcal{J}}\}$, $j = 1, \ldots, p$, where $\mathbf{I}\{\cdot\}$ is the indicator function. Denote the refined quadratic function

$$\widetilde{\Phi}_{\boldsymbol{A}^{(t)}}(\boldsymbol{S}|\boldsymbol{S}_k) = P_{\boldsymbol{A}^{(t)}}(\boldsymbol{S}_k) + \mathrm{vec}(\boldsymbol{\Delta}_{(1),k})^\intercal \{\boldsymbol{I}_{R_2R_3} \otimes (\boldsymbol{A}^{(t)})^\intercal \widetilde{\boldsymbol{\Sigma}}_k^{(t)} \boldsymbol{A}^{(t)}\} \mathrm{vec}(\boldsymbol{\Delta}_{(1),k}).$$

Similar to $\Phi_{\boldsymbol{A}^{(t)}}(\boldsymbol{S}|\boldsymbol{S}_k)$, $\widetilde{\Phi}_{\boldsymbol{A}^{(t)}}(\boldsymbol{S}|\boldsymbol{S}_k)$ is also a majorized function for $P_{\boldsymbol{A}^{(t)}}(\boldsymbol{S})$. Now we proceed to the minimizing step of the MM algorithm, and the corresponding minimization problem is

$$\min_{\boldsymbol{S}} \{H_{\boldsymbol{A}^{(t)}, \boldsymbol{B}^{(t)}, \boldsymbol{C}^{(t)}}(\boldsymbol{S}) + \widetilde{\Phi}_{\boldsymbol{A}^{(t)}}(\boldsymbol{S}|\boldsymbol{S}_k)\}.$$

Taking gradient of the objective function and setting it to zero gives

$$\mathrm{vec}(\boldsymbol{S}_{(1),k+1}) = \left[\sum_{i=1}^n (\boldsymbol{U}_{\boldsymbol{S},i}^{(t)})^\intercal \boldsymbol{U}_{\boldsymbol{S},i}^{(t)} + \boldsymbol{I}_{R_2R_3} \otimes \{(\boldsymbol{A}^{(t)})^\intercal \widetilde{\boldsymbol{\Sigma}}_k^{(t)} \boldsymbol{A}^{(t)}\}\right]^+ \left(\sum_{i=1}^n (\boldsymbol{U}_{\boldsymbol{S},i}^{(t)})^\intercal \boldsymbol{y}_i\right), \tag{S.2}$$

where $\boldsymbol{U}_{\boldsymbol{S},i}^{(t)} = \boldsymbol{C}^{(t)} \otimes \{\boldsymbol{b}^\intercal(t_i)\boldsymbol{B}^{(t)}\} \otimes (\boldsymbol{x}_i^\intercal \boldsymbol{A}^{(t)})$ and $(\cdot)^+$ denotes the Moore-Penrose inverse.

For case (ii), an improved version of local quadratic approximation has been suggested in Hunter and Li (2005). To handle the ill-posedness of (S.1) when $2\|(\boldsymbol{a}_j^{(t)})^\intercal \boldsymbol{S}_{(1),k}\|_2 = 0$, Hunter and Li (2005) proposed to replace $2\|(\boldsymbol{a}_j^{(t)})^\intercal \boldsymbol{S}_{(1),k}\|_2$ by $2(\varepsilon + \|(\boldsymbol{a}_j^{(t)})^\intercal \boldsymbol{S}_{(1),k}\|_2)$ for some $\varepsilon > 0$. After applying such replacement to all $(\boldsymbol{a}_j^{(t)})^\intercal \boldsymbol{S}_{(1),k}$ in (S.1), $j = 1, \ldots, p$, the adjusted version is no longer a majorizer of $P_{\boldsymbol{A}^{(t)}}(\boldsymbol{S})$ as required by the MM algorithm. Nonetheless, Hunter and Li (2005) showed that it majorizes a perturbed version of $P_{\boldsymbol{A}^{(t)}}(\boldsymbol{S})$. This would lead to an objective function that is similar to the one in case (i). In addition, the minimizer of this function should be close to the minimizer of the original function as long as ε is small enough and the original objective function is not extremely flat in the neighborhood of the minimizer. We thus define $\widetilde{\boldsymbol{\Sigma}}_{\varepsilon,k}^{(t)} = \mathrm{diag}\big(\widetilde{\lambda}_{1,k}^{(t)}(\varepsilon), \ldots, \widetilde{\lambda}_{p,k}^{(t)}(\varepsilon)\big)$ with $\widetilde{\lambda}_{j,k}^{(t)}(\varepsilon) = 1/\{2\varepsilon + 2\|(\boldsymbol{a}_j^{(t)})^\intercal \boldsymbol{S}_{(1),k}\|_2\}$. We get the perturbed version of the approximation function

$$\widetilde{\Phi}_{\boldsymbol{A}^{(t)}}^{\varepsilon}(\boldsymbol{S}|\boldsymbol{S}_k) = P_{\boldsymbol{A}^{(t)}}(\boldsymbol{S}_k) + \mathrm{vec}(\boldsymbol{\Delta}_{(1),k})^\intercal \{\boldsymbol{I}_{R_2R_3} \otimes (\boldsymbol{A}^{(t)})^\intercal \widetilde{\boldsymbol{\Sigma}}_{\varepsilon,k}^{(t)} \boldsymbol{A}^{(t)}\} \mathrm{vec}(\boldsymbol{\Delta}_{(1),k}).$$

Similar to case (i), we replace $\widetilde{\boldsymbol{\Sigma}}_k^{(t)}$ by $\widetilde{\boldsymbol{\Sigma}}_{\varepsilon,k}^{(t)}$ in (S.2) and obtain an updating rule for case (ii):

$$\mathrm{vec}(\boldsymbol{S}_{(1),k+1}) = \left[\sum_{i=1}^n (\boldsymbol{U}_{s,i}^{(t)})^\intercal \boldsymbol{U}_{\boldsymbol{S},i}^{(t)} + \boldsymbol{I}_{R_2R_3} \otimes \{(\boldsymbol{A}^{(t)})^\intercal \widetilde{\boldsymbol{\Sigma}}_{\varepsilon,k}^{(t)} \boldsymbol{A}^{(t)}\}\right]^+ \left(\sum_{i=1}^n (\boldsymbol{U}_{\boldsymbol{S},i}^{(t)})^\intercal \boldsymbol{y}_i\right). \tag{S.3}$$

The choice of ε can be

$$\varepsilon = \frac{\tau}{2n} \min \left\{ \|(\boldsymbol{a}_j^{(t)})^\intercal \boldsymbol{S}_{(1),0}\|_2 : (\boldsymbol{a}_j^{(t)})^\intercal \boldsymbol{S}_{(1),0} \neq 0,\, j = 1, \ldots, p \right\}$$

with $\tau = 10^{-4}$ as suggested in Hunter and Li (2005), where $\boldsymbol{S}_{(1),0}$ is the initial value of $\boldsymbol{S}_{(1)}$ for the MM algorithm to solve (4.1) of the main paper. We summarize the MM algorithm in Algorithm S.1.

Algorithm S.1: MM Algorithm for Updating $\boldsymbol{S}$.

Input : Data set $\{\boldsymbol{y}_i, \boldsymbol{X}_i, t_i\}_{i=1}^n$; Fixed $\boldsymbol{A}^{(t)}$, $\boldsymbol{B}^{(t)}$, and $\boldsymbol{C}^{(t)}$; Random initial $\boldsymbol{S}_{(1),0} \in \mathbb{R}^{R_1 \times R_2 R_3}$; $\delta > 0$.

Output: $\widetilde{\boldsymbol{S}}_{(1)}^{(t+1)}$ (so is $\widetilde{\boldsymbol{S}}^{(t+1)}$).

for k *from* $0, 1, \ldots$ **do**

  if For any $j = 1, \ldots, p$, $\|(\boldsymbol{a}_j^{(t)})^\intercal \boldsymbol{S}_{(1),k}\|_2 = 0$ and $\|\boldsymbol{a}_j^{(t)}\|_2 \neq 0$ **then**

    Update $\mathrm{vec}(\boldsymbol{S}_{(1),k+1})$ by (S.3) and calculate

$$d_{\boldsymbol{S}}^{\varepsilon} = H_{\boldsymbol{A}^{(t)},\boldsymbol{B}^{(t)},\boldsymbol{C}^{(t)}}(\boldsymbol{S}_k) + \Phi_{\boldsymbol{A}^{(t)}}^{\varepsilon}(\boldsymbol{S}_k|\boldsymbol{S}_k) H_{\boldsymbol{A}^{(t)},\boldsymbol{B}^{(t)},\boldsymbol{C}^{(t)}}(\boldsymbol{S}_{k+1}) - \Phi_{\boldsymbol{A}^{(t)}}^{\varepsilon}(\boldsymbol{S}_{k+1}|\boldsymbol{S}_k).$$

  else

    Update $\mathrm{vec}(\boldsymbol{S}_{(1),k+1})$ by (S.2) and calculate

$$d_{\boldsymbol{S}} = H_{\boldsymbol{A}^{(t)},\boldsymbol{B}^{(t)},\boldsymbol{C}^{(t)}}(\boldsymbol{S}_k) + \Phi_{\boldsymbol{A}^{(t)}}(\boldsymbol{S}_k|\boldsymbol{S}_k) - H_{\boldsymbol{A}^{(t)},\boldsymbol{B}^{(t)},\boldsymbol{C}^{(t)}}(\boldsymbol{S}_{k+1}) - \Phi_{\boldsymbol{A}^{(t)}}(\boldsymbol{S}_{k+1}|\boldsymbol{S}_k).$$

  end

  If $d_{\boldsymbol{S}} \leq \delta$ or $d_{\boldsymbol{S}}^{\varepsilon} \leq \delta$, then output $\boldsymbol{S}_{(1),k+1}$ as $\widetilde{\boldsymbol{S}}_{(1)}^{(t+1)}$. If not, set $k \leftarrow k+1$ and continue the loop.

end

S.1.2 Details on Updating $\boldsymbol{A}$

In this subsection, we provide more details on solving (4.4) of the main paper through the Alternating Direction Method of Multipliers (ADMM, Gabay and Mercier, 1976) algorithm. Let $\boldsymbol{A}_{(k)}$, $\boldsymbol{\Gamma}_{(k)}$, and $\boldsymbol{\nu}_{(k)}$ be the k-th iteration of the ADMM algorithm. With a penalty parameter ρ, the ADMM solves the subproblem (4.4) by generating the

following iterates at the $(k+1)$-th step

$$\begin{aligned}
\boldsymbol{A}_{(k+1)} &= \arg\min_{\boldsymbol{A}} \left\{ H_{\widetilde{\boldsymbol{S}}^{(t+1)},\boldsymbol{B}^{(t)},\boldsymbol{C}^{(t)}}(\boldsymbol{A}) + \rho/2 \left\| \boldsymbol{A}\widetilde{\boldsymbol{S}}_{(1)}^{(t+1)} - \boldsymbol{\Gamma}_{(k)} + \boldsymbol{\nu}_{(k)}/\rho \right\|_2^2 \right\}, \\
\boldsymbol{\Gamma}_{(k+1)} &= \arg\min_{\boldsymbol{\Gamma}} \left\{ \lambda \sum_{j=1}^{p} g(\boldsymbol{\gamma}_j) + \rho/2 \left\| \boldsymbol{A}_{(k+1)}\widetilde{\boldsymbol{S}}_{(1)}^{(t+1)} - \boldsymbol{\Gamma} + \boldsymbol{\nu}_{(k)}/\rho \right\|_2^2 \right\}, \\
\boldsymbol{\nu}_{(k+1)} &= \boldsymbol{\nu}_{(k)} + \rho \left(\boldsymbol{A}_{(k+1)}\widetilde{\boldsymbol{S}}_{(1)}^{(t+1)} - \boldsymbol{\Gamma}_{(k+1)} \right).
\end{aligned} \tag{S.4}$$

For simplicity, denote $\bar{\boldsymbol{y}} = (\boldsymbol{y}_1^\intercal, \ldots, \boldsymbol{y}_n^\intercal)^\intercal \in \mathbb{R}^{nq}$ and $\boldsymbol{U}_{\boldsymbol{A}} = (\boldsymbol{U}_{\boldsymbol{A},1}^\intercal, \ldots, \boldsymbol{U}_{\boldsymbol{A},n}^\intercal)^\intercal \in \mathbb{R}^{nq \times pR_1}$, where $\boldsymbol{U}_{\boldsymbol{A},i} = \left\{ \left[\boldsymbol{C}^{(t)} \otimes \{\boldsymbol{b}^\intercal(t_i)\boldsymbol{B}^{(t)}\} \right] (\widetilde{\boldsymbol{S}}_{(1)}^{(t+1)})^\intercal \right\} \otimes \boldsymbol{x}_i^\intercal \in \mathbb{R}^{q \times pR_1}$, $i = 1, 2, \ldots, n$. $H_{\widetilde{\boldsymbol{S}}^{(t+1)},\boldsymbol{B}^{(t)},\boldsymbol{C}^{(t)}}(\boldsymbol{A})$ can be reformulated as

$$H_{\widetilde{\boldsymbol{S}}^{(t+1)},\boldsymbol{B}^{(t)},\boldsymbol{C}^{(t)}}(\boldsymbol{A}) = \frac{1}{2}\|\bar{\boldsymbol{y}} - \boldsymbol{U}_{\boldsymbol{A}} \operatorname{vec}(\boldsymbol{A})\|_2^2.$$

We first note that the first line of (S.4) has a closed-form solution

$$\begin{aligned}
\operatorname{vec}(\boldsymbol{A}) = &\left[\boldsymbol{U}_{\boldsymbol{A}}^\intercal \boldsymbol{U}_{\boldsymbol{A}} + \rho \left\{ \widetilde{\boldsymbol{S}}_{(1)}^{(t+1)} \otimes \boldsymbol{I}_p \right\} \left\{ (\widetilde{\boldsymbol{S}}_{(1)}^{(t+1)})^\intercal \otimes \boldsymbol{I}_p \right\} \right]^{-1} \\
&\times \left[\boldsymbol{U}_{\boldsymbol{A}}^\intercal \bar{\boldsymbol{y}} + \rho \left\{ \widetilde{\boldsymbol{S}}_{(1)}^{(t+1)} \otimes \boldsymbol{I}_p \right\} \left\{ \operatorname{vec}(\boldsymbol{\Gamma}_{(k)}) - \frac{1}{\rho}\operatorname{vec}(\boldsymbol{\nu}_{(k)}) \right\} \right].
\end{aligned} \tag{S.5}$$

As for the second line of (S.4), we can split it into p subproblems:

$$\boldsymbol{\gamma}_{j,(k+1)} = \arg\min_{\boldsymbol{\gamma} \in \mathbb{R}^{R_2 R_3}} \left\{ \lambda g(\boldsymbol{\gamma}) + \frac{\rho}{2} \left\| (\widetilde{\boldsymbol{S}}_{(1)}^{(t+1)})^\intercal \boldsymbol{a}_{j,(k+1)} - \boldsymbol{\gamma} + \frac{1}{\rho}\boldsymbol{\nu}_{j,(k)} \right\|_2^2 \right\}, \quad j = 1, 2, \ldots, p,$$

where $\boldsymbol{a}_{j,(k+1)}^\intercal$ and $\boldsymbol{\nu}_{j,(k)}^\intercal$ are the j-th rows of $\boldsymbol{A}_{(k+1)}$ and $\boldsymbol{\nu}_{(k)}$ respectively. We introduce a proximal operator Boyd et al. (2011) to simplify our problem. For a given function $h(\cdot)$, its proximal map is defined by

$$\operatorname{prox}_{h(\cdot)}(u) = \arg\min_{v} \left\{ h(v) + \frac{1}{2}\|u - v\|_2^2 \right\}.$$

The proximal map $\operatorname{prox}_{h(\cdot)}(u)$ exists and is unique for all u if $h(\cdot)$ is a closed and convex function. Using this notation, we observe that $\boldsymbol{\gamma}_{j,(k+1)} = \operatorname{prox}_{\lambda g(\cdot)/\rho}(\boldsymbol{u}_j)$ with $\boldsymbol{u}_j = (\widetilde{\boldsymbol{S}}_{(1)}^{(t+1)})^\intercal \boldsymbol{a}_{j,(k+1)} + \boldsymbol{\nu}_{j,(k)}/\rho$. Furthermore, we can obtain the analytic form of the proximal operator with respect to $g(\cdot)$,

$$\boldsymbol{\gamma}_{j,(k+1)} = \left(1 - \frac{\lambda}{\rho\|\boldsymbol{u}_j\|_2} \right)_+ \boldsymbol{u}_j, \tag{S.6}$$

where $(x)_+ = \max(0, x)$.

We use the standard stopping criterion for the ADMM algorithm based on primal and dual residuals as in Section 3.3 of Boyd et al. (2011). To be specific, we let $r_{(k+1)}$ and $s_{(k+1)}$ denote the primal and dual residuals, which

are defined as

$$\begin{cases} r_{(k+1)} = \|\boldsymbol{A}_{(k+1)}\widetilde{\boldsymbol{S}}_{(1)}^{(t+1)} - \boldsymbol{\Gamma}_{(k+1)}\|_2 \\ s_{(k+1)} = \|\rho(\boldsymbol{\Gamma}_{(k+1)} - \boldsymbol{\Gamma}_{(k)})(\widetilde{\boldsymbol{S}}_{(1)}^{(t+1)})^\intercal\|_2. \end{cases} \tag{S.7}$$

The ADMM algorithm is terminated when $r_{(k+1)} \leqslant \epsilon^{\text{pri}}$ and $s_{(k+1)} \leqslant \epsilon^{\text{dual}}$ are satisfied, where the primal and dual feasibility ϵ^{pri} and ϵ^{dual} are specified as suggested by Boyd et al. (2011). Theoretically, the primal and dual residuals will converge to zero for any fixed $\rho > 0$. However, different choices of ρ may lead to different convergence speeds. An efficient and stable way of varying penalty strategy is suggested by Zhu (2017). The updating rule for ρ is set to be

$$\rho = \begin{cases} \eta\rho, & \text{if } r_{(k)}/\epsilon^{\text{pri}} \geqslant \mu\, s_{(k)}/\epsilon^{\text{dual}} \\ \eta^{-1}\rho, & \text{if } s_{(k)}/\epsilon^{\text{dual}} \geqslant \mu\, r_{(k)}/\epsilon^{\text{pri}}, \end{cases} \tag{S.8}$$

where μ and η are set to be 10 and 2, respectively, as suggested by Boyd et al. (2011). The above discussions lead to Algorithm S.2 for updating $\boldsymbol{A}$.

S.1.3 Details on Updating $\boldsymbol{B}$

In this section, we provide detailed steps to update $\boldsymbol{B}$ using the manifold gradient algorithm. Recall that

$$\boldsymbol{B}^{(t+1)} = \underset{\boldsymbol{B}\in \text{St}(R_2,K)}{\arg\min}\, H_{\boldsymbol{S}^{(t+1)},\boldsymbol{A}^{(t+1)},\boldsymbol{C}^{(t)}}(\boldsymbol{B}),$$

where $\text{St}(R_2, K)$ denotes the Stiefel manifold

$$\text{St}(R_2, K) = \{\boldsymbol{B} \in \mathbb{R}^{K\times R_2} : \boldsymbol{B}^\intercal\boldsymbol{B} = \boldsymbol{I}_{R_2}\}.$$

Note that $H_{\boldsymbol{S}^{(t+1)},\boldsymbol{A}^{(t+1)},\boldsymbol{C}^{(t)}}(\boldsymbol{B})$ has the form as

$$H_{\boldsymbol{S}^{(t+1)},\boldsymbol{A}^{(t+1)},\boldsymbol{C}^{(t)}}(\boldsymbol{B}) = \sum_{i=1}^{n}\left\|\boldsymbol{y}_i - ([\{\boldsymbol{C}^{(t)} \otimes (\boldsymbol{x}_i^\intercal\boldsymbol{A}^{(t+1)})\}\{\boldsymbol{S}_{(2)}^{(t+1)}\}^\intercal] \otimes \boldsymbol{b}^\intercal(t_i))\text{vec}(\boldsymbol{B})\right\|_2^2.$$

Let $\boldsymbol{B}_{(k)}$ denote the k-th iteration of the manifold gradient algorithm. First observe that the Euclidean gradient of $H_{\boldsymbol{S}^{(t+1)},\boldsymbol{A}^{(t+1)},\boldsymbol{C}^{(t)}}(\boldsymbol{B})$ can be directly obtained as

$$\nabla H_{\boldsymbol{S}^{(t+1)},\boldsymbol{A}^{(t+1)},\boldsymbol{C}^{(t)}}(\boldsymbol{B}) = -2\text{Mat}_{K,R_2}\left[\sum_{i=1}^{n}\left\{(\boldsymbol{U}_{\boldsymbol{B},i}^{(t)})^\intercal\boldsymbol{y}_i - (\boldsymbol{U}_{\boldsymbol{B},i}^{(t)})^\intercal\boldsymbol{U}_{\boldsymbol{B},i}^{(t)}\,\text{vec}(\boldsymbol{B})\right\}\right],$$

Algorithm S.2: ADMM for Updating $\boldsymbol{A}$.

Input: Data set $\{\boldsymbol{y}_i, \boldsymbol{X}_i, t_i\}_{i=1}^n$; Fixed $\widetilde{\boldsymbol{S}}^{(t+1)}$, $\boldsymbol{B}^{(t)}$, and $\boldsymbol{C}^{(t)}$; Primal and dual feasibility ϵ^{pri} and ϵ^{dual}; Parameters of varying penalty strategy μ and η; Random initial $\boldsymbol{A} \in \mathbb{R}^{p \times R_1}$.

Output: $\widetilde{\boldsymbol{A}}^{(t+1)}$ and $\boldsymbol{\Gamma}^{(t+1)}$ (as for the altertive of $\widetilde{\boldsymbol{A}}^{(t+1)}\widetilde{\boldsymbol{S}}_{(1)}^{(t+1)}$).

for $k = 0, 1, \ldots$ **do**

1. Update $\boldsymbol{A}_{(k+1)}$ using (S.5).
2. Update $\boldsymbol{\Gamma}_{(k+1)}$ through by rows through (S.6).
3. Update $\boldsymbol{\nu}_{(k+1)}$ through the third line of (S.4).
4. Calculate the primal and dual residuals $r_{(k+1)}$ and $s_{(k+1)}$ as in (S.7).

if $r_{(k+1)} \leqslant \epsilon^{\text{pri}}$ and $s_{(k+1)} \leqslant \epsilon^{\text{dual}}$ **then**
Output $\boldsymbol{A}_{(k+1)}$ as $\widetilde{\boldsymbol{A}}^{(t+1)}$ and $\boldsymbol{\Gamma}_{(k+1)}$ as $\boldsymbol{\Gamma}^{(t+1)}$.

else
Update ρ using the varying penalty strategy (S.8) and back to step 1 with $k \leftarrow k + 1$.

end

end

where $\boldsymbol{U}_{\boldsymbol{B},i}^{(t)} = [\{\boldsymbol{C}^{(t)} \otimes (\boldsymbol{x}_i^\intercal \boldsymbol{A}^{(t+1)})\}(\boldsymbol{S}_{(2)}^{(t+1)})^\intercal] \otimes \boldsymbol{b}^\intercal(t_i)$ and $\text{Mat}_{K,R_2}(\cdot)$ transforms a vector with length $K \cdot R_2$ into a matrix with $K \times R_2$ dimension by its columns. We denote $\text{grad}\, H_{\boldsymbol{S}^{(t+1)},\boldsymbol{A}^{(t+1)},\boldsymbol{C}^{(t)}}(\boldsymbol{B})$ as the Riemannian gradient of $H_{\boldsymbol{S}^{(t+1)},\boldsymbol{A}^{(t+1)},\boldsymbol{C}^{(t)}}(\boldsymbol{B})$, and define a projection operator $\mathcal{P}_{\boldsymbol{B}}(\boldsymbol{\xi})$ be

$$\mathcal{P}_{\boldsymbol{B}}(\boldsymbol{\xi}) = (\boldsymbol{I} - \boldsymbol{B}\boldsymbol{B}^\intercal)\boldsymbol{\xi} + \frac{1}{2}\boldsymbol{B}(\boldsymbol{B}^\intercal\boldsymbol{\xi} - \boldsymbol{\xi}^\intercal\boldsymbol{B}),$$

where $\boldsymbol{\xi}$ is a generic matrix with dimension $K \times R_2$. According to Absil et al. (2009), the Riemannian gradient $\text{grad}\, H_{\boldsymbol{S}^{(t+1)},\boldsymbol{A}^{(t+1)},\boldsymbol{C}^{(t)}}(\boldsymbol{B})$ at $\boldsymbol{B}$ is simply the projection of $\nabla H_{\boldsymbol{S}^{(t+1)},\boldsymbol{A}^{(t+1)},\boldsymbol{C}^{(t)}}(\boldsymbol{B})$ onto the tangent space, that is,

$$\text{grad}\, H_{\boldsymbol{S}^{(t+1)},\boldsymbol{A}^{(t+1)},\boldsymbol{C}^{(t)}}(\boldsymbol{B}) = \mathcal{P}_{\boldsymbol{B}}\{\nabla H_{\boldsymbol{S}^{(t+1)},\boldsymbol{A}^{(t+1)},\boldsymbol{C}^{(t)}}(\boldsymbol{B})\}. \tag{S.9}$$

After getting the Riemannian gradient, we perform a line-search on the negative Riemannian gradient and retract the update back onto the Stiefel manifold. One of the commonly used retractions of Stiefel manifold is (Absil et al., 2009)

$$R_{\boldsymbol{B}}(\boldsymbol{\xi}) = \mathrm{qf}(\boldsymbol{B} + \boldsymbol{\xi}),$$

where $\mathrm{qf}(\cdot)$ denotes the Q factor of the QR decomposition for a matrix. Therefore, we update

$$\boldsymbol{B}_{(k+1)} = \mathrm{qf}\{\boldsymbol{B}_{(k)} - \tau_{\boldsymbol{B}_{(k)}} \mathrm{grad}\, H_{\boldsymbol{S}^{(t+1)},\boldsymbol{A}^{(t+1)},\boldsymbol{C}^{(t)}}(\boldsymbol{B}_{(k)})\}, \tag{S.10}$$

where $\tau_{\boldsymbol{B}_{(k)}}$ is the Armijo step size. We select $\tau_{\boldsymbol{B}_{(k)}}$ such that, for some fixed $\epsilon_{\boldsymbol{B}} \in (0,1)$, the following inequality holds

$$\begin{aligned} H_{\boldsymbol{S}^{(t+1)},\boldsymbol{A}^{(t+1)},\boldsymbol{C}^{(t)}}(\boldsymbol{B}_{(k+1)}) \leq\ & H_{\boldsymbol{S}^{(t+1)},\boldsymbol{A}^{(t+1)},\boldsymbol{C}^{(t)}}(\boldsymbol{B}_{(k)}) \\ & - \epsilon_{\boldsymbol{B}} \tau_{\boldsymbol{B}_{(k)}} \|\mathrm{grad}\, H_{\boldsymbol{S}^{(t+1)},\boldsymbol{A}^{(t+1)},\boldsymbol{C}^{(t)}}(\boldsymbol{B}_{(k)})\|_F^2. \end{aligned} \tag{S.11}$$

The above discussion leads to Algorithm S.3 as follows.

Algorithm S.3: Stiefel Manifold Optimization for Updating $\boldsymbol{B}$.

Input : Data set $\{\boldsymbol{y}_i, \boldsymbol{X}_i, t_i\}_{i=1}^n$; Fixed $\boldsymbol{S}^{(t+1)}$, $\boldsymbol{A}^{(t+1)}$, and $\boldsymbol{C}^{(t)}$; Random initial $\boldsymbol{B}_{(0)} \in \mathbb{R}^{K \times R_2}$; $\epsilon_{\boldsymbol{B}} > 0$ and $\epsilon > 0$.

Output: $\boldsymbol{B}^{(t+1)}$

for k *from* $0, 1, \ldots$ **do**

1. Compute the Riemannian gradient at $\boldsymbol{B}_{(k)}$ as in (S.9).
2. Obtain $\boldsymbol{B}_{(k+1)}$ as in (S.10) with step size $\tau_{\boldsymbol{B}_{(k)}}$, where $\tau_{\boldsymbol{B}_{(k)}}$ is chosen according to (S.11).
3. Calculate the descent of the objective function

$$d_{\boldsymbol{B}} = H_{\boldsymbol{S}^{(t+1)},\boldsymbol{A}^{(t+1)},\boldsymbol{C}^{(t)}}(\boldsymbol{B}_{(k)}) - H_{\boldsymbol{S}^{(t+1)},\boldsymbol{A}^{(t+1)},\boldsymbol{C}^{(t)}}(\boldsymbol{B}_{(k+1)}).$$

If $d_{\boldsymbol{B}} \leq \epsilon$, then output $\boldsymbol{B}_{(k+1)}$ as $\boldsymbol{B}^{(t+1)}$. If not, return to step 1.

end

S.2 Proof of Theorem 1

To present the proof, we use C with or without subscripts to represent generic positive constants that may change values from line to line. Recall that $\boldsymbol{Z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)^\intercal \in \mathbb{R}^{n \times pK}$, where $\boldsymbol{z}_i = \boldsymbol{b}(t_i) \otimes \boldsymbol{x}_i$. For any $\mathcal{I} \subset \{1, \ldots, p\}$, define $\boldsymbol{Z}_\mathcal{I} \in \mathbb{R}^{|\mathcal{I}| \times pK}$ such that it has the same columns as $\boldsymbol{Z}$ for those predictors within the index $\mathcal{I}$ and zero columns otherwise. Denote $\mathcal{P}_\mathcal{I}$ as the projection matrix onto the column space of $\boldsymbol{Z}_\mathcal{I}$.

With the notation of $\boldsymbol{z}_i$, we can write

$$\{\boldsymbol{G} \bar{\times}_2 \boldsymbol{b}(t_i)\}^\intercal \boldsymbol{x}_i = (\boldsymbol{A} \otimes \boldsymbol{B}) \boldsymbol{S}_{(3)}^\intercal \boldsymbol{C}^\intercal \boldsymbol{z}_i, \quad i = 1, \ldots, n,$$

in (3.5) of the main paper by using the matricization operator of a tensor (Kolda and Bader, 2009). Let $\boldsymbol{D} = (\boldsymbol{A} \otimes \boldsymbol{B}) \boldsymbol{S}_{(3)}^\intercal \boldsymbol{C}^\intercal \in \mathbb{R}^{pK \times q}$. It can be seen that $\boldsymbol{D} = \boldsymbol{G}_{(3)}^\intercal$, where $\boldsymbol{G}_{(3)}$ is the mode-3 matricization of $\boldsymbol{G}$. Let $\boldsymbol{D}_j \in \mathbb{R}^{K \times q}$ be the collection of rows of $\boldsymbol{D}$ associated with the predictor j.

The optimization problem (3.6) of the main paper can then be rewritten as

$$\arg\min_{\boldsymbol{D}} \|\boldsymbol{Y} - \boldsymbol{Z}\boldsymbol{D}\|_F^2 + \lambda \sum_{j=1}^{p} \|\boldsymbol{D}_j\|_F. \tag{S.12}$$

Respectively denoting $\boldsymbol{D}_0 = \boldsymbol{G}_{0,(3)}^\intercal$ and $\widehat{\boldsymbol{D}}$ as the solution of (S.12) (which is equal to $\widehat{\boldsymbol{G}}_{(3)}^\intercal$), we have $\Delta_{\boldsymbol{G}}^2 = \|\boldsymbol{Z}(\boldsymbol{D} - \boldsymbol{D}_0)\|_F^2$ and $\Delta_{\widehat{\boldsymbol{G}}}^2 = \|\boldsymbol{Z}(\widehat{\boldsymbol{D}} - \boldsymbol{D}_0)\|_F^2$. Recall that $\mathcal{J}(\boldsymbol{G}) = \{j : \boldsymbol{D}_j \neq \boldsymbol{0}\}$ and $\mathcal{J}(\widehat{\boldsymbol{G}}) = \{j : \widehat{\boldsymbol{D}}_j \neq \boldsymbol{0}\}$. Denote $\widetilde{\mathcal{J}} = \mathcal{J}(\boldsymbol{G}) \cup \mathcal{J}(\widehat{\boldsymbol{G}})$ for simplicity.

By definition, we can obtain

$$\|\boldsymbol{Y} - \boldsymbol{Z}\widehat{\boldsymbol{D}}\|_F^2 + \lambda \sum_{j \in \mathcal{J}(\widehat{\boldsymbol{G}})} \|\widehat{\boldsymbol{D}}_j\|_F \leq \|\boldsymbol{Y} - \boldsymbol{Z}\boldsymbol{D}\|_F^2 + \lambda \sum_{j \in \mathcal{J}(\boldsymbol{G})} \|\boldsymbol{D}_j\|_F,$$

which implies

$$\begin{aligned} \Delta_{\widehat{\boldsymbol{G}}}^2 + \lambda \sum_{j \in \mathcal{J}(\boldsymbol{G})^c} \|\widehat{\boldsymbol{D}}_j - \boldsymbol{D}_j\|_F \leq \Delta_{\boldsymbol{G}}^2 &+ 2\langle \boldsymbol{E}, \boldsymbol{Z}_{\widetilde{\mathcal{J}}}(\widehat{\boldsymbol{D}} - \boldsymbol{D}) \rangle \\ &+ 2\langle \boldsymbol{R}, \boldsymbol{Z}_{\widetilde{\mathcal{J}}}(\widehat{\boldsymbol{D}} - \boldsymbol{D}) \rangle + \lambda \sum_{j \in \mathcal{J}(\boldsymbol{G})} \|\boldsymbol{D}_j - \widehat{\boldsymbol{D}}_j\|_F. \end{aligned} \tag{S.13}$$

Denote $d_1(\cdot)$ as the largest singular value of a matrix. Since $\operatorname{rank}(\boldsymbol{D}) \leq R_3$ and $\operatorname{rank}(\boldsymbol{S}_{(3)}) \leq R_1 R_2$, we have

$$\operatorname{rank}\{\boldsymbol{Z}_{\widetilde{\mathcal{J}}}(\widehat{\boldsymbol{D}} - \boldsymbol{D})\} \leq 2R = 2\min\{R_3, R_1 R_2\}.$$

It yields

$$\begin{aligned}
2\langle \boldsymbol{E}, \boldsymbol{Z}_{\tilde{\mathcal{J}}}(\widehat{\boldsymbol{D}}-\boldsymbol{D})\rangle &= 2\langle \mathcal{P}_{\tilde{\mathcal{J}}}\boldsymbol{E}, \boldsymbol{Z}_{\tilde{\mathcal{J}}}(\widehat{\boldsymbol{D}}-\boldsymbol{D})\rangle \\
&\le 2d_1(\mathcal{P}_{\tilde{\mathcal{J}}}\boldsymbol{E})\|\boldsymbol{Z}_{\tilde{\mathcal{J}}}(\widehat{\boldsymbol{D}}-\boldsymbol{D})\|_* \\
&\le 2d_1(\mathcal{P}_{\tilde{\mathcal{J}}}\boldsymbol{E})\sqrt{2R}\|\boldsymbol{Z}_{\tilde{\mathcal{J}}}(\widehat{\boldsymbol{D}}-\boldsymbol{D})\|_F \\
&\le 2d_1(\mathcal{P}_{\tilde{\mathcal{J}}}\boldsymbol{E})\sqrt{2R}(\Delta_{\widehat{\boldsymbol{G}}}+\Delta_{\boldsymbol{G}}) \\
&\le 16Rd_1^2(\mathcal{P}_{\tilde{\mathcal{J}}}\boldsymbol{E})+\frac{1}{4}(\Delta_{\widehat{\boldsymbol{G}}}^2+\Delta_{\boldsymbol{G}}^2),
\end{aligned} \tag{S.14}$$

where $\mathcal{P}_{\tilde{\mathcal{J}}}$ is the projection matrix of the space spanned by $\boldsymbol{Z}_{\tilde{\mathcal{J}}}$ and $\|\cdot\|_*$ denotes the nuclear norm. The second, third, fourth, and fifth lines of (S.14) are due to the Cauchy-Schwarz inequality, the relationship between the nuclear norm and Frobenius norm, the triangle inequality, and the inequality of arithmetic and geometric means, respectively. It follows from (S.13) and (S.14),

$$\begin{aligned}
\frac{3}{4}\Delta_{\widehat{\boldsymbol{G}}}^2+\lambda\sum_{j\in\mathcal{J}(\boldsymbol{G})^c}\|\widehat{\boldsymbol{D}}_j-\boldsymbol{D}_j\|_F \le{}& \frac{5}{4}\Delta_{\boldsymbol{G}}^2+16Rd_1^2(\mathcal{P}_{\tilde{\mathcal{J}}}\boldsymbol{E}) \\
&+2\langle \boldsymbol{R}, \boldsymbol{Z}_{\tilde{\mathcal{J}}}(\widehat{\boldsymbol{D}}-\boldsymbol{D})\rangle+\lambda\sum_{j\in\mathcal{J}(\boldsymbol{G})}\|\boldsymbol{D}_j-\widehat{\boldsymbol{D}}_j\|_F.
\end{aligned}$$

Now, there are two cases as follows.

Case 1. If

$$\lambda\sum_{j\in\mathcal{J}(\boldsymbol{G})}\|\boldsymbol{D}_j-\widehat{\boldsymbol{D}}_j\|_F \ge \frac{5}{4}\Delta_{\boldsymbol{G}}^2+16Rd_1^2(\mathcal{P}_{\tilde{\mathcal{J}}}\boldsymbol{E})+2\langle \boldsymbol{R}, \boldsymbol{Z}_{\tilde{\mathcal{J}}}(\widehat{\boldsymbol{D}}-\boldsymbol{D})\rangle, \tag{S.15}$$

it straightforwardly implies

$$\sum_{j\in\mathcal{J}(\boldsymbol{G})^c}\|\widehat{\boldsymbol{D}}_j-\boldsymbol{D}_j\|_F \le 2\sum_{j\in\mathcal{J}(\boldsymbol{G})}\|\boldsymbol{D}_j-\widehat{\boldsymbol{D}}_j\|_F.$$

By Condition $\mathcal{M}(\mathcal{J}(\boldsymbol{G}),\delta_{\mathcal{J}}(\boldsymbol{G}))$, we have

$$\begin{aligned}
\lambda\sum_{j\in\mathcal{J}(\boldsymbol{G})}\|\boldsymbol{D}_j-\widehat{\boldsymbol{D}}_j\|_F &\le \frac{4}{n\delta_{\mathcal{J}(\boldsymbol{G})}}\lambda^2|\mathcal{J}(\boldsymbol{G})|+\frac{n\delta_{\mathcal{J}(\boldsymbol{G})}}{16}\sum_{j\in\mathcal{J}(\boldsymbol{G})}\|\boldsymbol{D}_j-\widehat{\boldsymbol{D}}_j\|_F^2 \\
&\le \frac{4}{n\delta_{\mathcal{J}(\boldsymbol{G})}}\lambda^2|\mathcal{J}(\boldsymbol{G})|+\frac{1}{16}\|\boldsymbol{Z}\widehat{\boldsymbol{D}}-\boldsymbol{Z}\boldsymbol{D}\|_F^2 \\
&\le \frac{4}{n\delta_{\mathcal{J}(\boldsymbol{G})}}\lambda^2|\mathcal{J}(\boldsymbol{G})|+\frac{1}{8}(\Delta_{\widehat{\boldsymbol{G}}}^2+\Delta_{\boldsymbol{G}}^2).
\end{aligned} \tag{S.16}$$

Using (S.15), (S.16), and the Cauchy-Schwarz inequality, we further obtain

$$\frac{3}{4}\Delta_{\widehat{\boldsymbol{G}}}^2 \le 2\lambda\sum_{j\in\mathcal{J}(\boldsymbol{G})}\|\boldsymbol{D}_j-\widehat{\boldsymbol{D}}_j\|_F \le \frac{8}{n\delta_{\mathcal{J}(\boldsymbol{G})}}\lambda^2|\mathcal{J}(\boldsymbol{G})|+\frac{1}{4}(\Delta_{\widehat{\boldsymbol{G}}}^2+\Delta_{\boldsymbol{G}}^2).$$

Thus,

$$\Delta_{\widehat{\boldsymbol{G}}}^2 \leq 2\Delta_{\boldsymbol{G}}^2 + \frac{16\lambda^2|\mathcal{J}(\boldsymbol{G})|}{n\delta_{\mathcal{J}(\boldsymbol{G})}}. \tag{S.17}$$

For notational simplicity, we let

$$\gamma = 24n\lambda_{\max}(\boldsymbol{\Sigma})K\sigma^2\{1+\log(p)\}. \tag{S.18}$$

Taking $\lambda^2 = 768\gamma R_3 R$ in (S.17) will then lead to

$$\Delta_{\widehat{\boldsymbol{G}}}^2 \leq C_1\Delta_{\boldsymbol{G}}^2 + C_2\|\boldsymbol{R}\|_F^2 + \frac{C_3R_3R\lambda_{\max}(\boldsymbol{\Sigma})K\sigma^2\{1+\log(p)\}|\mathcal{J}(\boldsymbol{G})|}{\delta_{\mathcal{J}(\boldsymbol{G})}}. \tag{S.19}$$

Case 2. If

$$\lambda\sum_{j\in\mathcal{J}(\boldsymbol{G})}\|\boldsymbol{D}_j - \widehat{\boldsymbol{D}}_j\|_F < \frac{5}{4}\Delta_{\boldsymbol{G}}^2 + 16Rd_1^2(\mathcal{P}_{\widetilde{\mathcal{J}}}\boldsymbol{E}) + 2\langle\boldsymbol{R}, \boldsymbol{Z}_{\widetilde{\mathcal{J}}}(\widehat{\boldsymbol{D}} - \boldsymbol{D})\rangle,$$

using the Cauchy-Schwarz inequality, we have

$$\begin{aligned} 3\Delta_{\widehat{\boldsymbol{G}}}^2 &\leq 10\Delta_{\boldsymbol{G}}^2 + 128Rd_1^2(\mathcal{P}_{\widetilde{\mathcal{J}}}\boldsymbol{E}) + 16\langle\boldsymbol{R}, \boldsymbol{Z}_{\widetilde{\mathcal{J}}}(\widehat{\boldsymbol{D}} - \boldsymbol{D})\rangle \\ &\leq 10\Delta_{\boldsymbol{G}}^2 + 128Rd_1^2(\mathcal{P}_{\widetilde{\mathcal{J}}}\boldsymbol{E}) + 16\|\boldsymbol{R}\|_F(\Delta_{\widehat{\boldsymbol{G}}} + \Delta_{\boldsymbol{G}}). \end{aligned} \tag{S.20}$$

The quadratic form of inequality (S.20) implies that

$$\Delta_{\widehat{\boldsymbol{G}}} \leq \frac{16\|\boldsymbol{R}\|_F + \sqrt{\zeta_2}}{6}, \tag{S.21}$$

where

$$\zeta_2 = 256\|\boldsymbol{R}\|_F^2 + 12\{10\Delta_{\boldsymbol{G}}^2 + 128Rd_1^2(\mathcal{P}_{\widetilde{\mathcal{J}}}\boldsymbol{E}) + 16\|\boldsymbol{R}\|_F\Delta_{\boldsymbol{G}}\}.$$

Plugging (S.21) into (S.20) shows that

$$9\Delta_{\widehat{\boldsymbol{G}}}^2 \leq 304\|\boldsymbol{R}\|_F^2 + 108\Delta_{\boldsymbol{G}}^2 + 768Rd_1^2(\mathcal{P}_{\widetilde{\mathcal{J}}}\boldsymbol{E}). \tag{S.22}$$

According to Lemma 8 of Bunea et al. (2012), we have

$$\mathbb{P}(d_1^2(\mathcal{P}_{\widetilde{\mathcal{J}}}\boldsymbol{E}) - 6K\sigma^2\{|\widetilde{\mathcal{J}}| + |\widetilde{\mathcal{J}}|\log(ep/|\widetilde{\mathcal{J}}|)\} - 6\sigma^2 q > \tilde{t}) \leq \frac{16\sigma^2\exp(-q/2)}{\tilde{t}}$$

for $\tilde{t} > 0$. Taking $\tilde{t} = 6K\sigma^2\log(p)$, we then have

$$\begin{aligned} d_1^2(\mathcal{P}_{\widetilde{\mathcal{J}}}\boldsymbol{E}) &\leq \tilde{t} + 6K\sigma^2\{|\widetilde{\mathcal{J}}| + |\widetilde{\mathcal{J}}|\log(ep/|\widetilde{\mathcal{J}}|)\} + 6\sigma^2 q \\ &\leq 12K\sigma^2\{1+\log(p)\}\{|\mathcal{J}(\boldsymbol{G})| + |\mathcal{J}(\widehat{\boldsymbol{G}})|\} + 6\sigma^2 q \end{aligned} \tag{S.23}$$

with probability at least

$$1 - \frac{8\exp(-q/2)}{3K\log(p)}. \tag{S.24}$$

Next, observe that when $\widehat{\boldsymbol{C}}$ of the solution of (3.6) of the main paper is given, the other components $(\widehat{\boldsymbol{S}}, \widehat{\boldsymbol{A}}, \widehat{\boldsymbol{B}})$ can be obtained through solving

$$\underset{\boldsymbol{S},\boldsymbol{A},\boldsymbol{B}}{\arg\min} \|\boldsymbol{Y}\widehat{\boldsymbol{C}} - \boldsymbol{Z}(\boldsymbol{A}\otimes\boldsymbol{B})\boldsymbol{S}_{(3)}^{\intercal}\|_F^2 + \lambda\sum_{j=1}^{p}\|\boldsymbol{a}_j^{\intercal}\boldsymbol{S}_{(1)}\|_F,$$

where $\boldsymbol{a}_j^{\intercal}$ is the j-th row of $\boldsymbol{A}$. By a QR decomposition, we can represent $\boldsymbol{A}\boldsymbol{S}_{(1)} = \boldsymbol{U}\boldsymbol{V}^{\intercal}$ such that $\boldsymbol{U} \in \mathbb{R}^{p\times R_1}$, $\boldsymbol{V} \in \mathbb{R}^{R_2R_3\times R_1}$ and $\boldsymbol{V}^{\intercal}\boldsymbol{V} = \boldsymbol{I}_{R_1}$. Thus, the solution of (3.6) of the main paper, which is in the form of $(\boldsymbol{S}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})$, can be reparametrized into the form of $(\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{B}, \boldsymbol{C})$. In other words, we can find a $(\widehat{\boldsymbol{V}}, \widehat{\boldsymbol{U}})$ such that $\widehat{\boldsymbol{G}}_{(1)} = \widehat{\boldsymbol{A}}\widehat{\boldsymbol{S}}_{(1)}(\widehat{\boldsymbol{C}}\otimes\widehat{\boldsymbol{B}})^{\intercal} = \widehat{\boldsymbol{U}}\widehat{\boldsymbol{V}}^{\intercal}(\widehat{\boldsymbol{C}}\otimes\widehat{\boldsymbol{B}})^{\intercal}$ Let

$$\widetilde{\boldsymbol{Z}}_i = \big[\boldsymbol{x}_i^{\intercal}\otimes\{\boldsymbol{I}_{R_3}\otimes\boldsymbol{b}^{\intercal}(t_i)\widehat{\boldsymbol{B}}\}\big](\boldsymbol{I}_p\otimes\widehat{\boldsymbol{V}}) \in \mathbb{R}^{R_3\times pR_1}, \quad i = 1,\ldots,n. \tag{S.25}$$

Therefore, $\widehat{\boldsymbol{U}}$ can be obtained from solving

$$\widehat{\boldsymbol{U}} = \underset{\boldsymbol{U}}{\arg\min}\,\|\boldsymbol{Y}\widehat{\boldsymbol{C}} - (\widetilde{\boldsymbol{Z}}_1\boldsymbol{u},\ldots,\widetilde{\boldsymbol{Z}}_n\boldsymbol{u})^{\intercal}\|_F^2 + \lambda\sum_{j=1}^{p}\|\boldsymbol{u}_j\|_2, \tag{S.26}$$

where $\boldsymbol{u} = \mathrm{vec}(\boldsymbol{U}^{\intercal}) \in \mathbb{R}^{pR_1}$ and $\boldsymbol{u}_j$ is the j-th row of $\boldsymbol{U}$. Since $\boldsymbol{Y}\widehat{\boldsymbol{C}} \in \mathbb{R}^{n\times R_3}$, we write the i-th row of $\boldsymbol{Y}\widehat{\boldsymbol{C}}$ as $\boldsymbol{\gamma}_i^{\intercal}$, $i = 1,\ldots,n$, that is,

$$\boldsymbol{Y}\widehat{\boldsymbol{C}} = (\boldsymbol{\gamma}_1,\ldots,\boldsymbol{\gamma}_n)^{\intercal}.$$

Using this notation, (S.26) can be written as

$$\widehat{\boldsymbol{U}} = \underset{\boldsymbol{U}}{\arg\min}\sum_{i=1}^{n}\|\boldsymbol{\gamma}_i - \widetilde{\boldsymbol{Z}}_i\boldsymbol{u}\|_2^2 + \lambda\sum_{j=1}^{p}\|\boldsymbol{u}_j\|_2.$$

Analogously, let $\widehat{\boldsymbol{u}} = \mathrm{vec}(\widehat{\boldsymbol{U}})$ and $\widehat{\boldsymbol{u}}_j$ be the j-th row of $\widehat{\boldsymbol{U}}$. To solve $\widehat{\boldsymbol{u}}_j$ from the above displayed objective function, we have

$$\sum_{i=1}^{n}\widetilde{\boldsymbol{Z}}_{ij}^{\intercal}(\boldsymbol{\gamma}_i - \widetilde{\boldsymbol{Z}}_i\widehat{\boldsymbol{u}}) = \lambda\frac{\widehat{\boldsymbol{u}}_j}{\|\widehat{\boldsymbol{u}}_j\|_2} \in \mathbb{R}^{R_1},$$

where $\widetilde{\boldsymbol{Z}}_{ij} \in \mathbb{R}^{R_3\times R_1}$ is the submatrix of $\widetilde{\boldsymbol{Z}}_i$ defined in (S.25), associated the j-th predictor, $j \in \mathcal{J}(\widehat{\boldsymbol{G}})$. For simplicity, denote

$$\widetilde{\boldsymbol{Z}}_\sharp = (\widetilde{\boldsymbol{Z}}_1^{\intercal},\ldots,\widetilde{\boldsymbol{Z}}_n^{\intercal})^{\intercal} = (\widetilde{\boldsymbol{Z}}_{\sharp,1},\ldots,\widetilde{\boldsymbol{Z}}_{\sharp,p}) \in \mathbb{R}^{nR_3\times pR_1},$$

where $\widetilde{\boldsymbol{Z}}_{\sharp,j} = (\widetilde{\boldsymbol{Z}}_{1j}^\intercal, \ldots, \widetilde{\boldsymbol{Z}}_{nj}^\intercal)^\intercal \in \mathbb{R}^{nR_3 \times R_1}$. We then have

$$\Big\| \widetilde{\boldsymbol{Z}}_{\sharp,j}^\intercal \big[(\boldsymbol{\gamma}_1^\intercal, \ldots, \boldsymbol{\gamma}_n^\intercal)^\intercal - \{ (\widetilde{\boldsymbol{Z}}_1 \widehat{\boldsymbol{u}})^\intercal, \ldots, (\widetilde{\boldsymbol{Z}}_n \widehat{\boldsymbol{u}})^\intercal \}^\intercal \big] \Big\|_2^2 = \lambda^2, \quad j \in \mathcal{J}(\widehat{\boldsymbol{G}}).$$

Thus,

$$\begin{aligned}
\lambda^2 |\mathcal{J}(\widehat{\boldsymbol{G}})| &= \sum_{j \in \mathcal{J}(\widehat{\boldsymbol{G}})} \Big\| \widetilde{\boldsymbol{Z}}_{\sharp,j}^\intercal \big[(\boldsymbol{\gamma}_1^\intercal, \ldots, \boldsymbol{\gamma}_n^\intercal)^\intercal - \{ (\widetilde{\boldsymbol{Z}}_1 \widehat{\boldsymbol{u}})^\intercal, \ldots, (\widetilde{\boldsymbol{Z}}_n \widehat{\boldsymbol{u}})^\intercal \}^\intercal \big] \Big\|_2^2 \\
&\le 2\lambda_{\max}(\widetilde{\boldsymbol{Z}}_\sharp \widetilde{\boldsymbol{Z}}_\sharp^\intercal) \big\{ \|\boldsymbol{Z}(\boldsymbol{A}_0 \otimes \boldsymbol{B}_0) \boldsymbol{S}_{0,(3)}^\intercal \boldsymbol{C}_0^\intercal \widehat{\boldsymbol{C}} - \boldsymbol{Z}(\widehat{\boldsymbol{A}} \otimes \widehat{\boldsymbol{B}}) \widehat{\boldsymbol{S}}_{(3)}^\intercal \widehat{\boldsymbol{C}}^\intercal \widehat{\boldsymbol{C}} \|_F^2 + \|\mathcal{P}_{\widetilde{\mathcal{J}}} \boldsymbol{E} \widehat{\boldsymbol{C}}\|_F^2 + \|\boldsymbol{R}\widehat{\boldsymbol{C}}\|_F^2 \big\} \\
&\le 2\lambda_{\max}(\widetilde{\boldsymbol{Z}}_\sharp \widetilde{\boldsymbol{Z}}_\sharp^\intercal) \big\{ R_3 \Delta_{\widehat{\boldsymbol{G}}}^2 + R_3 d_1^2(\mathcal{P}_{\widetilde{\mathcal{J}}} \boldsymbol{E}) + R_3 \|\boldsymbol{R}\|_F^2 \big\}. \qquad \text{(S.27)}
\end{aligned}$$

Note that the eigenvalues of

$$\widetilde{\boldsymbol{Z}}_\sharp^\intercal \widetilde{\boldsymbol{Z}}_\sharp = \sum_{i=1}^n \widetilde{\boldsymbol{Z}}_i^\intercal \widetilde{\boldsymbol{Z}}_i = \sum_{i=1}^n (\boldsymbol{I}_p \otimes \widehat{\boldsymbol{V}}^\intercal) \big\{ \boldsymbol{x}^\intercal \otimes (\boldsymbol{I}_{R_3} \otimes \boldsymbol{b}^\intercal \widehat{\boldsymbol{B}}) \big\}^\intercal \big\{ \boldsymbol{x}^\intercal \otimes (\boldsymbol{I}_{R_3} \otimes \boldsymbol{b}^\intercal \widehat{\boldsymbol{B}}) \big\} (\boldsymbol{I}_p \otimes \widehat{\boldsymbol{V}})$$

are the same as those of

$$\sum_{i=1}^n \big\{ \boldsymbol{x}^\intercal \otimes (\boldsymbol{I}_{R_3} \otimes \boldsymbol{b}^\intercal \widehat{\boldsymbol{B}}) \big\}^\intercal \big\{ \boldsymbol{x}^\intercal \otimes (\boldsymbol{I}_{R_3} \otimes \boldsymbol{b}^\intercal \widehat{\boldsymbol{B}}) \big\}. \qquad \text{(S.28)}$$

Similarly, it can be shown that (S.28) has the same eigenvalues as of

$$\sum_{i=1}^n \big\{ \boldsymbol{x} \otimes \widehat{\boldsymbol{B}}^\intercal \boldsymbol{b} \big\} \big\{ \boldsymbol{x}^\intercal \otimes \boldsymbol{b}^\intercal \widehat{\boldsymbol{B}} \big\}.$$

By (S.27) and Lemma S.1 , we obtain

$$\lambda^2 |\mathcal{J}(\widehat{\boldsymbol{G}})| \le 2n\lambda_{\max}(\boldsymbol{\Sigma}) \big\{ R_3 \Delta_{\widehat{\boldsymbol{G}}}^2 + R_3 d_1^2(\mathcal{P}_{\widetilde{\mathcal{J}}} \boldsymbol{E}) + R_3 \|\boldsymbol{R}\|_F^2 \big\}. \qquad \text{(S.29)}$$

For simplicity, let

$$\alpha = 6\sigma^2 q + 16\sigma^2 \exp(-q/2) + 12K\sigma^2 |\mathcal{J}(\boldsymbol{G})| \{1 + \log(p)\}. \qquad \text{(S.30)}$$

Recall that $\gamma = 24n\lambda_{\max}(\boldsymbol{\Sigma}) K\sigma^2 \{1 + \log(p)\}$. It follows from (S.23) and (S.29) that

$$d_1^2(\mathcal{P}_{\widetilde{\mathcal{J}}} \boldsymbol{E}) \le \alpha + \frac{\gamma R_3}{\lambda^2} \big(\Delta_{\widehat{\boldsymbol{G}}}^2 + d_1^2(\mathcal{P}_{\widetilde{\mathcal{J}}} \boldsymbol{E}) + \|\boldsymbol{R}\|_F^2 \big),$$

with probability at least (S.24), which yields

$$\Big(1 - \frac{\gamma R_3}{\lambda^2} \Big) d_1^2(\mathcal{P}_{\widetilde{\mathcal{J}}} \boldsymbol{E}) \le \alpha + \frac{\gamma R_3}{\lambda^2} \big(\Delta_{\widehat{\boldsymbol{G}}}^2 + \|\boldsymbol{R}\|_F^2 \big),$$

that is,

$$d_1^2(\mathcal{P}_{\widetilde{\mathcal{J}}}\boldsymbol{E}) \leq \frac{\alpha}{1-\gamma R_3/\lambda^2} + \frac{\gamma R_3/\lambda^2}{1-\gamma R_3/\lambda^2}\big(\Delta^2_{\widehat{\boldsymbol{G}}} + \|\boldsymbol{R}\|_F^2\big), \tag{S.31}$$

Plugging (S.31) into (S.22) yields

$$\begin{aligned} &\frac{9-9\gamma R_3/\lambda^2 - 768\gamma R_3 R/\lambda^2}{1-\gamma R_3/\lambda^2}(\Delta^2_{\widehat{\boldsymbol{G}}}) \\ &\quad \leq 108(\Delta^2_{\boldsymbol{G}}) + \frac{768R\alpha}{1-\gamma R_3/\lambda^2} + \left(\frac{768\gamma R_3 R/\lambda^2}{1-\gamma R_3/\lambda^2} + 304\right)\|\boldsymbol{R}\|_F^2. \end{aligned} \tag{S.32}$$

Taking $\lambda^2 = 768\gamma R_3 R$ in (S.32) yields

$$\Delta^2_{\widehat{\boldsymbol{G}}} \leq C_1\Delta^2_{\boldsymbol{G}} + C_2 qR\sigma^2 + C_3 RK\sigma^2|\mathcal{J}(\boldsymbol{G})|\{1+\log(p)\} + C_4\|\boldsymbol{R}\|_F^2, \tag{S.33}$$

with probability at least (S.24). By definitions, it is obviously to see $\lambda_{\max}(\boldsymbol{\Sigma}) \geq \delta_{\mathcal{J}(\boldsymbol{G})}$. With (5.3) of the main paper, (S.19), and (S.33), it shows that

$$\Delta^2_{\widehat{\boldsymbol{G}}} \leq C_1\Delta^2_{\boldsymbol{G}} + C_2 qR\sigma^2 + C_3\frac{R_3 RK|\mathcal{J}(\boldsymbol{G})|\lambda_{\max}(\boldsymbol{\Sigma})\sigma^2\log(p)}{\delta_{\mathcal{J}(\boldsymbol{G})}} + C_4\frac{nsq}{K^{2\tau}},$$

with probability at least (S.24), which finishes the proof.

S.3 Proof of Corollaries

S.3.1 Proof of Corollary 1

By definitions, (5.3), (5.5), and (5.7) of the main paper, we have

$$\begin{aligned} \Delta^2_{\widehat{\boldsymbol{F}}} &= \sum_{i=1}^n \|\widehat{\boldsymbol{F}}(t_i)^\intercal\boldsymbol{x}_i - \boldsymbol{F}_0(t_i)^\intercal\boldsymbol{x}_i\|^2 \\ &= \sum_{i=1}^n \|\widehat{\boldsymbol{F}}(t_i)^\intercal\boldsymbol{x}_i - \{\boldsymbol{G}_0 \,\bar{\times}_2\, \boldsymbol{b}(t_i)\}^\intercal\boldsymbol{x}_i + \{\boldsymbol{G}_0 \,\bar{\times}_2\, \boldsymbol{b}(t_i)\}^\intercal\boldsymbol{x}_i - \boldsymbol{F}_0(t_i)^\intercal\boldsymbol{x}_i\|^2 \\ &\leq 2\sum_{i=1}^n \big\{\|\widehat{\boldsymbol{F}}(t_i)^\intercal\boldsymbol{x}_i - \{\boldsymbol{G}_0 \,\bar{\times}_2\, \boldsymbol{b}(t_i)\}^\intercal\boldsymbol{x}_i\|^2 + \|\{\boldsymbol{G}_0 \,\bar{\times}_2\, \boldsymbol{b}(t_i)\}^\intercal\boldsymbol{x}_i - \boldsymbol{F}_0(t_i)^\intercal\boldsymbol{x}_i\|^2\big\} \\ &\leq 2S_5\Delta^2_{\boldsymbol{G}} + 2S_6 qR\sigma^2 + 2S_7\frac{R_3 RK|\mathcal{J}(\boldsymbol{G})|\lambda_{\max}(\boldsymbol{\Sigma})\sigma^2\log(p)}{\delta_{\mathcal{J}(\boldsymbol{G})}} + (2S_8 + 2S_3)\frac{nsq}{K^{2\tau}}, \end{aligned} \tag{S.34}$$

with probability at least (S.24).

S.3.2 Proof of Corollary 2

Theorem 1 of the main paper implies that when $\boldsymbol{\Sigma}$ satisfies Condition $\mathcal{M}(\mathcal{J}(\boldsymbol{G}_0), \delta_{\mathcal{J}(\boldsymbol{G}_0)})$,

$$\Delta^2_{\widehat{\boldsymbol{G}}} \leq S_6 qR\sigma^2 + S_7 \frac{R_3 RKs\lambda_{\max}(\boldsymbol{\Sigma})\sigma^2 \log(p)}{\delta_{\mathcal{J}(\boldsymbol{G}_0)}} + S_8 \frac{nsq}{K^{2\tau}}, \tag{S.35}$$

with probability at least (S.24). Similarly, applying the arguments as in the proof of Corollary 1, it can be shown the convergence rate of $\Delta^2_{\widehat{\boldsymbol{F}}}$ has the upper bound

$$O_p\left(qR\sigma^2 + \frac{R_3 RKs\lambda_{\max}(\boldsymbol{\Sigma})\sigma^2 \log(p)}{\delta_{\mathcal{J}(\boldsymbol{G}_0)}} + \frac{nsq}{K^{2\tau}}\right),$$

by specifying $\boldsymbol{G} = \boldsymbol{G}_0$ in (S.34) under the assumption that $\boldsymbol{\Sigma}$ satisfies Condition $\mathcal{M}(\mathcal{J}(\boldsymbol{G}_0), \delta_{\mathcal{J}(\boldsymbol{G}_0)})$. In the above bound, let

$$K \asymp \left\{\frac{n\delta_{\mathcal{J}(\boldsymbol{G}_0)} q}{R_3 R\lambda_{\max}(\boldsymbol{\Sigma})\log(p)}\right\}^{1/(2\tau+1)}, \tag{S.36}$$

we then have

$$\Delta^2_{\widehat{\boldsymbol{F}}} = O_p\left(qR + \left\{\frac{R_3 R\lambda_{\max}(\boldsymbol{\Sigma})\log(p)}{\delta_{\mathcal{J}(\boldsymbol{G}_0)}}\right\}^{2\tau/(2\tau+1)} s(nq)^{1/(2\tau+1)}\right),$$

which finishes the proof of (5.9) of the main paper.

Next, we will show the relevant predictors will be selected consistently. Recall that $\boldsymbol{D}_0 = \boldsymbol{G}^{\intercal}_{0,(3)}$ and $\widehat{\boldsymbol{D}} = \widehat{\boldsymbol{G}}^{\intercal}_{(3)}$. By (5.1) of the main paper and the orthonormality of $\boldsymbol{b}(t)$, we observe that for $j \in \mathcal{J}(\boldsymbol{G}_0)$,

$$\begin{aligned}\sum_{l=1}^{q} \|f_{0,jl}\|_2^2 &\leq \sum_{l=1}^{q}\left(\left\|\sum_{k=1}^{K} G_{0,jkl} b_k\right\|_2 + \left\|\sum_{k=1}^{K} G_{0,jkl} b_k - f_{0,jl}\right\|_2\right)^2 \\ &\leq 2\|\boldsymbol{D}_{0,j}\|_F^2 + \frac{Cq}{K^{2\tau}},\end{aligned}$$

where $\boldsymbol{D}_{0,j} \in \mathbb{R}^{K\times q}$ is the submatrix of $\boldsymbol{D}_0$ associated with the predictor j. Now, with K chosen to be (S.36), and the assumption that $\sum_{l=1}^{q}\|f_{0,jl}\|_2^2 \geq S_{11}$, it follows that when (5.10) of the main paper is satisfied, we have

$$\|\boldsymbol{D}_{0,j}\|_F > C, \tag{S.37}$$

for some positive constant C. On the other hand, the triangular inequality implies

$$\|\boldsymbol{D}_{0,j}\|_F \leq \|\boldsymbol{D}_{0,j} - \widehat{\boldsymbol{D}}_j\|_F + \|\widehat{\boldsymbol{D}}_j\|_F. \tag{S.38}$$

Thus, if we can show as n tends to infinity

$$\mathbb{P}\big\{\|\boldsymbol{D}_{0,j} - \widehat{\boldsymbol{D}}_j\|_F < C/2,\, j \in \mathcal{J}(\boldsymbol{G}_0)\big\} \to 1, \tag{S.39}$$

then following (S.37)–(S.39), we obtain

$$\mathbb{P}\big\{\|\widehat{\boldsymbol{D}}_j\|_F > C/2,\, j \in \mathcal{J}(\boldsymbol{G}_0)\big\} \to 1.$$

Using the orthonormality of $\boldsymbol{b}(t)$, the above result will complete the proof of this corollary.

It remains to prove (S.39). Recall that

$$\begin{aligned}\frac{3}{4}\Delta^2_{\widehat{\boldsymbol{G}}} + \lambda \sum_{j\in\mathcal{J}(\boldsymbol{G}_0)^c} \|\widehat{\boldsymbol{D}}_j - \boldsymbol{D}_{0,j}\|_F \le 16Rd_1^2(\mathcal{P}_{\widetilde{\mathcal{J}}}\boldsymbol{E}) \\ + 2\langle \boldsymbol{R}, \boldsymbol{Z}_{\widetilde{\mathcal{J}}}(\widehat{\boldsymbol{D}} - \boldsymbol{D}_0)\rangle + \lambda \sum_{j\in\mathcal{J}(\boldsymbol{G}_0)} \|\boldsymbol{D}_{0,j} - \widehat{\boldsymbol{D}}_j\|_F,\end{aligned}$$

where $\widetilde{\mathcal{J}} = \mathcal{J}(\boldsymbol{G}_0) \cup \mathcal{J}(\widehat{\boldsymbol{G}})$, and there are two cases as follows.

Case 1. If

$$\lambda \sum_{j\in\mathcal{J}(\boldsymbol{G}_0)} \|\boldsymbol{D}_{0,j} - \widehat{\boldsymbol{D}}_j\|_F \ge 16Rd_1^2(\mathcal{P}_{\widetilde{\mathcal{J}}}\boldsymbol{E}) + 2\langle \boldsymbol{R}, \boldsymbol{Z}_{\widetilde{\mathcal{J}}}(\widehat{\boldsymbol{D}} - \boldsymbol{D}_0)\rangle,$$

it straightforwardly implies

$$\sum_{j\in\mathcal{J}(\boldsymbol{G}_0)^c} \|\widehat{\boldsymbol{D}}_j - \boldsymbol{D}_{0,j}\|_F \le 2 \sum_{j\in\mathcal{J}(\boldsymbol{G}_0)} \|\boldsymbol{D}_{0,j} - \widehat{\boldsymbol{D}}_j\|_F.$$

By Condition $\mathcal{M}(\mathcal{J}(\boldsymbol{G}_0), \delta_{\mathcal{J}}(\boldsymbol{G}_0))$, we have

$$\begin{aligned}\delta_{\mathcal{J}(\boldsymbol{G}_0)} \sum_{j\in\mathcal{J}(\boldsymbol{G}_0)} \|\boldsymbol{D}_{0,j} - \widehat{\boldsymbol{D}}_j\|_F^2 &\le \mathrm{tr}\{(\boldsymbol{D}_{0,j} - \widehat{\boldsymbol{D}}_j)^\intercal \boldsymbol{\Sigma} (\boldsymbol{D}_{0,j} - \widehat{\boldsymbol{D}}_j)\} \\ &\le \frac{\Delta^2_{\widehat{\boldsymbol{G}}}}{n}.\end{aligned} \tag{S.40}$$

Using (S.19), (S.36), and (5.3) of the main paper, we have

$$\begin{aligned}\Delta^2_{\widehat{\boldsymbol{G}}} &\le C_1 \frac{nsq}{K^{2\tau}} + \frac{C_2 R_3 RKs\lambda_{\max}(\boldsymbol{\Sigma})\sigma^2\{1+\log(p)\}}{\delta_{\mathcal{J}(\boldsymbol{G}_0)}} \\ &\le C_3 \bigg\{ \frac{R_3 R\lambda_{\max}(\boldsymbol{\Sigma})\log(p)}{\delta_{\mathcal{J}(\boldsymbol{G}_0)}} \bigg\}^{2\tau/(2\tau+1)} s(nq)^{1/(2\tau+1)}.\end{aligned} \tag{S.41}$$

With (5.10), we prove (S.39) by plugging (S.41) in (S.40).

Case 2. If

$$\lambda \sum_{j\in\mathcal{J}(\boldsymbol{G}_0)} \|\boldsymbol{D}_{0,j} - \widehat{\boldsymbol{D}}_j\|_F < 16Rd_1^2(\mathcal{P}_{\widetilde{\mathcal{J}}}\boldsymbol{E}) + 2\langle \boldsymbol{R}, \boldsymbol{Z}_{\widetilde{\mathcal{J}}}(\widehat{\boldsymbol{D}} - \boldsymbol{D}_0)\rangle, \tag{S.42}$$

we will use the following arguments to prove (S.39). By (S.31), we have with probability at least (S.24),

$$\begin{aligned}Rd_1^2(\mathcal{P}_{\widetilde{\mathcal{J}}}\boldsymbol{E}) &\le \frac{\alpha R}{1 - 1/(768R)} + \frac{1/768}{1 - 1/(768R)}\big(\Delta^2_{\widehat{\boldsymbol{G}}} + \|\boldsymbol{R}\|_F^2\big) \\ &\le C_1 \alpha R + C_2\big(\Delta^2_{\widehat{\boldsymbol{G}}} + \|\boldsymbol{R}\|_F^2\big),\end{aligned} \tag{S.43}$$

where α is defined as in (S.30). It follows from (S.42), (5.3) of the main paper, (S.35), and (S.43) that

$$\lambda \sum_{j\in\mathcal{J}(\boldsymbol{G}_0)} \|\boldsymbol{D}_{0,j} - \widehat{\boldsymbol{D}}_j\|_F < 16Rd_1^2(\mathcal{P}_{\widetilde{\mathcal{J}}}\boldsymbol{E}) + 2\langle \boldsymbol{R}, \boldsymbol{Z}_{\widetilde{\mathcal{J}}}(\widehat{\boldsymbol{D}} - \boldsymbol{D}_0)\rangle$$
$$\leq C_1 Rq + C_2\frac{nsq}{K^{2\tau}} + C_3\frac{R_3RKs\lambda_{\max}(\boldsymbol{\Sigma})\log(p)}{\delta_{\mathcal{J}(\boldsymbol{G}_0)}},$$

with probability at least (S.24). Using (5.6) of the main paper and (S.36), it implies

$$\begin{aligned}\lambda^2 &= S_4R_3Rn\lambda_{\max}(\boldsymbol{\Sigma})K\sigma^2\{1+\log(p)\}\\ &\geq S_4R_3Rn\lambda_{\max}(\boldsymbol{\Sigma})K\sigma^2\log(p)\\ &\asymp C\{R_3R\lambda_{\max}(\boldsymbol{\Sigma})\log(p)\}^{2\tau/(2\tau+1)}n^{2(\tau+1)/(2\tau+1)}\{q\delta_{\mathcal{J}(\boldsymbol{G}_0)}\}^{1/(2\tau+1)}.\end{aligned}$$

We then have

$$\begin{aligned}&\sum_{j\in\mathcal{J}(\boldsymbol{G}_0)}\|\boldsymbol{D}_{0,j} - \widehat{\boldsymbol{D}}_j\|_F\\ &\leq C_1 q^{(4\tau+1)/(4\tau+2)}R^{(\tau+1)/(2\tau+1)}\{R_3\lambda_{\max}(\boldsymbol{\Sigma})\log(p)\}^{-\tau/(2\tau+1)}n^{-(\tau+1)/(2\tau+1)}\delta_{\mathcal{J}(\boldsymbol{G}_0)}^{-1/(4\tau+2)}\\ &\quad + C_2 s\{RR_3\lambda_{\max}(\boldsymbol{\Sigma})\log(p)\}^{\tau/(2\tau+1)}n^{-\tau/(2\tau+1)}q^{1/(4\tau+2)}\delta_{\mathcal{J}(\boldsymbol{G}_0)}^{-(4\tau+1)/(4\tau+2)},\end{aligned} \tag{S.44}$$

with probability at least (S.24). Using (5.10) of the main paper, (S.36), and (S.44), we have (S.39), which completes the proof.

S.4 Random Design

This section is organized as follows. We first provide two auxiliary lemmas that are used as preliminary results (Lemmas S.1 and S.2). Lemma S.3 show that the Condition $\mathcal{M}(\mathcal{J}, \delta_{\mathcal{J}})$ can be satisfied with high probability (tending to one) when $\boldsymbol{x}$ and t are random, and the sample size n is large enough compared with $|\mathcal{J}|^2q^2K^2 + |\mathcal{J}|^2qK\log p$. A special application of this result is for $\mathcal{J}(\boldsymbol{G}_0)$ with $|\mathcal{J}(\boldsymbol{G}_0)| = s$. The term $\lambda_{\max}(\boldsymbol{\Sigma})/\delta_{\mathcal{J}}$ appeared in Theorem 1 of the main paper is determined in Lemmas S.4 and S.5 for the cases when p and q are diverging with the sample size n and fixed constants, respectively.

For notational simplicity, we use C with or without subscripts to represent generic constants that may change values from line to line.

Lemma S.1. *Let $\boldsymbol{z}_{i,\mathcal{I}} = \boldsymbol{x}_{i,\mathcal{I}} \otimes \boldsymbol{b}(t_i)$, where $\mathcal{I} \subset \{1, \ldots, p\}$, $i = 1, \ldots, n$. If the following eigenvalue bounds*

$$L_{\mathcal{I}} \leq \lambda_{\min}\bigg(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{z}_{i,\mathcal{I}}\boldsymbol{z}_{i,\mathcal{I}}^{\intercal}\bigg) \leq \lambda_{\max}\bigg(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{z}_{i,\mathcal{I}}\boldsymbol{z}_{i,\mathcal{I}}^{\intercal}\bigg) \leq U_{\mathcal{I}}$$

hold, then

$$L_{\mathcal{I}} \leq \lambda_{\min}\bigg(\frac{1}{n}\sum_{i=1}^{n} \bar{\boldsymbol{Z}}_{i,\mathcal{I}}\bar{\boldsymbol{Z}}_{i,\mathcal{I}}^{\intercal}\bigg) \leq \lambda_{\max}\bigg(\frac{1}{n}\sum_{i=1}^{n} \bar{\boldsymbol{Z}}_{i,\mathcal{I}}\bar{\boldsymbol{Z}}_{i,\mathcal{I}}^{\intercal}\bigg) \leq U_{\mathcal{I}},$$

where $\bar{\boldsymbol{Z}}_{i,\mathcal{I}} = \boldsymbol{x}_{i,\mathcal{I}} \otimes \boldsymbol{B}^{\intercal}\boldsymbol{b}(t_i)$, for any $\boldsymbol{B}$ satisfying $\boldsymbol{B}^{\intercal}\boldsymbol{B} = \boldsymbol{I}_{R_2}$.

Proof. By definition, for all $w_{jk} \in \mathbb{R}$, we have

$$\bigg\{\sum_{j\in\mathcal{I}}\sum_{k=1}^{K}\frac{1}{n}\sum_{i=1}^{n} x_{ij}b_k(t_i)w_{jk}\bigg\}^2 \leq U_{\mathcal{I}}\sum_{j\in\mathcal{I}}\sum_{k=1}^{K} w_{jk}^2.$$

Denote $\boldsymbol{B} = (\boldsymbol{B}_1, \ldots, \boldsymbol{B}_K)^{\intercal}$. We then have

$$\begin{aligned}
\bigg\{\sum_{j\in\mathcal{I}}\sum_{r_2=1}^{R_2}\frac{1}{n}\sum_{i=1}^{n} x_{ij}\sum_{k=1}^{K} B_{kr_2}b_k(t_i)\bar{w}_{jr_2}\bigg\}^2 &= \bigg\{\sum_{j\in\mathcal{I}}\sum_{k=1}^{K}\frac{1}{n}\sum_{i=1}^{n} x_{ij}b_k(t_i)\sum_{r_2=1}^{R_2} B_{kr_2}\bar{w}_{jr_2}\bigg\}^2 \\
&\leq U_{\mathcal{I}}\sum_{j\in\mathcal{I}}\sum_{k=1}^{K}(\bar{\boldsymbol{w}}_j^{\intercal}\boldsymbol{B}_k\boldsymbol{B}_k^{\intercal}\bar{\boldsymbol{w}}_j) \\
&= U_{\mathcal{I}}\sum_{j\in\mathcal{I}}(\bar{\boldsymbol{w}}_j^{\intercal}\boldsymbol{B}^{\intercal}\boldsymbol{B}\bar{\boldsymbol{w}}_j) \\
&= U_{\mathcal{I}}\sum_{j\in\mathcal{I}}\sum_{r_2=1}^{R_2}\bar{w}_{jr_2}^2,
\end{aligned}$$

where $\bar{w}_{jr_2} \in \mathbb{R}$ and $\bar{\boldsymbol{w}}_j = (\bar{w}_{j1}, \ldots, \bar{w}_{jR_2})^{\intercal}$. Therefore we finish the proof of the upper bound. The proof of lower bound can be obtained similarly. □

Lemma S.2. *Suppose the conditional density of t given $\boldsymbol{x}$, denoted as $f_{t|\boldsymbol{x}}(t)$, is bounded away from 0 and infinity by a constant, that is, there exists $C > 0$*

$$1/C < f_{t|\boldsymbol{x}}(t) < C, \quad t \in [0, 1].$$

If $\|\boldsymbol{x}\|_{\psi_2} \leq \kappa$, then

$$\|\boldsymbol{x} \otimes \boldsymbol{b}(t)\|_{\psi_2} \leq C\kappa\sqrt{K} \tag{S.45}$$

and

$$\lambda_{\max}\big[\mathbb{E}\{\boldsymbol{x} \otimes \boldsymbol{b}(t)\}\{\boldsymbol{x}^{\intercal} \otimes \boldsymbol{b}^{\intercal}(t)\}\big] \leq C\kappa^2. \tag{S.46}$$

Further, if $\lambda_{\min}\{\mathbb{E}(\boldsymbol{x}\boldsymbol{x}^\intercal)\} \geq C$, *then*

$$\lambda_{\min}\big[\mathbb{E}\{\boldsymbol{x}\otimes\boldsymbol{b}(t)\}\{\boldsymbol{x}^\intercal\otimes\boldsymbol{b}^\intercal(t)\}\big] \geq C. \tag{S.47}$$

Proof. Denote $\{\widetilde{b}_k(t)\}_{k=1}^K$ as the ordinary B-spline basis with the same knots and order as $\{b_k(t)\}$. By equivalence, there exists a nonnegative matrix $\boldsymbol{\Psi}\in\mathbb{R}^{K\times K}$ such that

$$\boldsymbol{b}(t) = \boldsymbol{\Psi}\widetilde{\boldsymbol{b}}(t), \tag{S.48}$$

where $\widetilde{\boldsymbol{b}}(t) = (\widetilde{b}_1(t),\ldots,\widetilde{b}_K(t))^\intercal$. If follows that

$$\boldsymbol{I}_K = \boldsymbol{\Psi}\bigg\{\int\widetilde{\boldsymbol{b}}(t)\widetilde{\boldsymbol{b}}^\intercal(t)dt\bigg\}\boldsymbol{\Psi}^\intercal$$

and

$$\boldsymbol{\Psi}^{-1}(\boldsymbol{\Psi}^\intercal)^{-1} = \int\widetilde{\boldsymbol{b}}(t)\widetilde{\boldsymbol{b}}^\intercal(t)dt.$$

By the property of B-spline (see, e.g., De Boor, 1973, 1976) and Assumption 3, for $1\leq q\leq+\infty$ and $\boldsymbol{v}\in\mathbb{R}^K$, we have

$$C_\zeta\|\boldsymbol{v}\|_q \leq h_n^{-1/q}\bigg\|\sum_{k=1}^K v_k\widetilde{b}_k\bigg\|_q \leq C\|\boldsymbol{v}\|_q, \tag{S.49}$$

where C_ζ and C are two positive constants and C_ζ depends on the order of B-spline ζ. Thus,

$$C_1\frac{1}{K}\leq\lambda_{\min}\big\{\boldsymbol{\Psi}^{-1}(\boldsymbol{\Psi}^\intercal)^{-1}\big\}\leq\lambda_{\max}\big\{\boldsymbol{\Psi}^{-1}(\boldsymbol{\Psi}^\intercal)^{-1}\big\}\leq C_2\frac{1}{K},$$

which yields

$$C_2K\leq\lambda_{\min}\big\{\boldsymbol{\Psi}\boldsymbol{\Psi}^\intercal\big\}\leq\lambda_{\max}\big\{\boldsymbol{\Psi}\boldsymbol{\Psi}^\intercal\big\}\leq C_1K. \tag{S.50}$$

Recall

$$\|\boldsymbol{x}\otimes\boldsymbol{b}(t)\|_{\psi_2} = \sup_{\|\boldsymbol{w}\|_2\leq1}\|\langle\boldsymbol{x}\otimes\boldsymbol{b}(t),\boldsymbol{w}\rangle\|_{\psi_2}.$$

For all $\boldsymbol{w}\in\mathbb{R}^{pK}$ with $\|\boldsymbol{w}\|_2\leq1$, we then have

$$\begin{aligned}\|\langle\boldsymbol{x}\otimes\boldsymbol{b}(t),\boldsymbol{w}\rangle\|_{L_q}^q &= \mathbb{E}_{\boldsymbol{x}}\int\bigg|\sum_{j,k}x_jb_k(t)w_{jk}\bigg|^q f_{t|\boldsymbol{x}}(t)dt\\ &= \mathbb{E}_{\boldsymbol{x}}\int\bigg|\sum_{j,k}x_j\Big(\sum_{\widetilde{k}}\Psi_{k,\widetilde{k}}\widetilde{b}_k(t)\Big)w_{jk}\bigg|^q f_{t|\boldsymbol{x}}(t)dt\end{aligned}$$

$$
\begin{aligned}
&\leq C_1 \mathbb{E}_{\boldsymbol{x}} \int \Big| \sum_{\widetilde{k}} \Big(\sum_{k,j} \Psi_{k,\widetilde{k}} x_j w_{jk} \Big) \widetilde{b}_{\widetilde{k}}(t) \Big|^q dt \\
&\leq C^q \frac{1}{K} \mathbb{E}_{\boldsymbol{x}} \left\| \left\{ \Big(\sum_{k,j} \Psi_{k,1} x_j w_{jk} \Big), \ldots, \Big(\sum_{k,j} \Psi_{k,K} x_j w_{jk} \Big) \right\} \right\|_q^q \\
&\leq C^q \frac{1}{K} \sum_{\widetilde{k}=1}^{K} \mathbb{E}_{\boldsymbol{x}} \Big| \sum_{k,j} \Psi_{k,\widetilde{k}} x_j w_{jk} \Big|^q \\
&\leq C^q \frac{1}{K} \sum_{\widetilde{k}=1}^{K} \mathbb{E}_{\boldsymbol{x}} \Big| \sum_{j} x_j \Big(\sum_{k} \Psi_{k,\widetilde{k}} w_{jk} \Big) \Big|^q \\
&\leq C^q \frac{1}{K} \sum_{k=1}^{K} \| \langle \boldsymbol{x}, \{ \boldsymbol{\Psi}_{:,k}^{\intercal}(\boldsymbol{w}_1, \ldots, \boldsymbol{w}_p) \}^{\intercal} \rangle, \|_{L_q}^q \\
&\leq C^q \frac{1}{K} \sum_{k=1}^{K} (\sqrt{q} \|\boldsymbol{x}\|_{\psi_2} \|\boldsymbol{\Psi}_{:,k}^{\intercal}(\boldsymbol{w}_1, \ldots, \boldsymbol{w}_p)\|_2)^q \\
&\leq C^q \frac{1}{K} q^{q/2} \kappa^q \sum_{k=1}^{K} \Big(\|\boldsymbol{\Psi}_{:,k}^{\intercal}(\boldsymbol{w}_1, \ldots, \boldsymbol{w}_p)\|_2^2 \Big)^{q/2} \\
&\leq C^q K^{q/2-1} q^{q/2} \kappa^q, \qquad \text{(S.51)}
\end{aligned}
$$

where $\boldsymbol{w}_j = (w_{j1}, \ldots, w_{jK})^{\intercal}$, w_{jk} is the $\{(j-1)K + k\}$-th element of $\boldsymbol{w}$ and $\boldsymbol{\Psi}_{:,k} = (\Psi_{1,k}, \ldots, \Psi_{K,k})^{\intercal}$. Therefore, we obtain

$$
\| \langle \boldsymbol{x} \otimes \boldsymbol{b}(t), \boldsymbol{w} \rangle \|_{L_q} \leq C \kappa \sqrt{q} K^{1/2-1/q} \leq C \kappa \sqrt{qK},
$$

which yields the result of (S.45). Taking $q = 2$ in (S.51) will lead to the result of (S.46). Similarly,

$$
\begin{aligned}
\| \langle \boldsymbol{x} \otimes \boldsymbol{b}(t), \boldsymbol{w} \rangle \|_{L_2}^2 &= \mathbb{E}_{\boldsymbol{x}} \int \left| \sum_{\widetilde{k}} \Big(\sum_{k,j} x_j \Psi_{k,\widetilde{k}} w_{jk} \Big) \widetilde{b}_{\widetilde{k}}(t) \right|^2 f_{t|\boldsymbol{x}}(t) dt \\
&\geq C \frac{1}{K} \mathbb{E}_{\boldsymbol{x}} \left\| \left\{ \Big(\sum_{k,j} \Psi_{k,1} x_j w_{jk} \Big), \ldots, \Big(\sum_{k,j} \Psi_{k,K} x_j w_{jk} \Big) \right\} \right\|_2^2 \\
&\geq C \frac{1}{K} \sum_{\widetilde{k}=1}^{K} \mathbb{E}_{\boldsymbol{x}} \Big| \sum_{j} x_j \Big(\sum_{k} \Psi_{k,\widetilde{k}} w_{jk} \Big) \Big|^2 \\
&\geq C \frac{1}{K} \sum_{\widetilde{k}=1}^{K} \| \boldsymbol{\Psi}_{:,k}^{\intercal}(\boldsymbol{w}_1, \ldots, \boldsymbol{w}_p) \|_2^2 \\
&\geq C \|\boldsymbol{w}\|_2^2,
\end{aligned}
$$

which proves (S.47). □

Before presenting the next lemmas, we first introduce γ_α-functionals defined in (Talagrand, 2005; Banerjee et al., 2015). In Lemmas S.3–S.5, γ_2 will be used in the proofs.

Definition S.1 (γ_α-functionals)**.** Consider a metric space (T, d) and for a finite set $\mathcal{A} \subset T$, let $|\mathcal{A}|$ denote its cardinality. An admissible sequence is an increasing sequence of subsets $\{\mathcal{A}_n, n \geq 0\}$ of T, such that $|\mathcal{A}_0| = 1$ and for $n \geq 1$, $|\mathcal{A}_n| = 2^{2^n}$. Given $\alpha > 0$, we define the γ_α-functional as

$$\gamma_\alpha(T, d) = \inf \sup_{t \in T} \sum_{n=0}^{\infty} \text{Diam}\{A_n(t)\},$$

where $A_n(t)$ is the unique element of $\mathcal{A}_n$ that contains t, $\text{Diam}\{A_n(t)\}$ is the diameter of A_n according to d, and the infimum is over all admissible sequences of T.

The next lemma shows that Condition $\mathcal{M}(\mathcal{J}, \delta_\mathcal{J})$ is satisfied with high probability under random models.

Lemma S.3. *If the conditional density of t given $\boldsymbol{x}$ is bounded by a constant, $\|\boldsymbol{x}\|_{\psi_2} \leq \kappa$ and $\lambda_{\min}\{\mathbb{E}(\boldsymbol{x}\boldsymbol{x}^\intercal)\} \geq C_1$, then Condition $\mathcal{M}(\mathcal{J}, \delta_\mathcal{J})$ holds for some $\delta_\mathcal{J} > 0$ with high probability when $n \geq C_3(|\mathcal{J}|^2 q^2 K^2 + |\mathcal{J}|^2 qK \log p)$, where C_3 depends on C_1 and $\delta_\mathcal{J}$.*

Proof. Recall $\boldsymbol{Z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)^\intercal$, where $\boldsymbol{z}_i = \boldsymbol{x}_i \otimes \boldsymbol{b}(t_i)$, which are i.i.d. copies of $\boldsymbol{z} = \boldsymbol{x} \otimes \boldsymbol{b}(t)$. Let $\boldsymbol{\Sigma}_z = \mathbb{E}(\boldsymbol{z}\boldsymbol{z}^\intercal)$. Consider the class of functions:

$$\mathscr{F} = \left[f_M\{\boldsymbol{x} \otimes \boldsymbol{b}(t)\} = \frac{1}{\sqrt{\text{tr}(\boldsymbol{M}^\intercal \boldsymbol{\Sigma}_z \boldsymbol{M})}} \sqrt{\text{tr}[\boldsymbol{M}^\intercal \{\boldsymbol{x} \otimes \boldsymbol{b}(t)\}\{\boldsymbol{x} \otimes \boldsymbol{b}(t)\}^\intercal \boldsymbol{M}]} : \text{vec}(\boldsymbol{M}) \in \mathcal{A} \right],$$

where

$$\mathcal{A} = \left\{ \text{vec}(\boldsymbol{M}) : \|\boldsymbol{M}\|_F^2 \leq 1 \text{ and } \ 2 \sum_{j \in \mathcal{J}} \|\boldsymbol{M}_j\|_F \geq \sum_{j \in \mathcal{J}^c} \|\boldsymbol{M}_j\|_F \right\}.$$

For $\boldsymbol{M} \in \mathcal{A}$, we have

$$\sum_{j=1}^{p} \|\boldsymbol{M}_j\|_F \leq 3 \sum_{j \in \mathcal{J}} \|\boldsymbol{M}_j\|_F \leq 3|\mathcal{J}| \|\boldsymbol{M}\|_F = 3|\mathcal{J}|.$$

Define

$$\bar{\mathcal{A}} = \left\{ \text{vec}(\boldsymbol{M}) : \sum_{j=1}^{p} \|\boldsymbol{M}_j\|_F \leq 3|\mathcal{J}| \text{ and } \ 2 \sum_{j \in \mathcal{J}} \|\boldsymbol{M}_j\|_F \geq \sum_{j \in J^c} \|\boldsymbol{M}_j\|_F \right\}.$$

We then have

$$\mathcal{A} \subset \bar{\mathcal{A}}. \tag{S.52}$$

Note that

$$\|f_M\|_{L_2}^2 = \frac{1}{\text{tr}(\boldsymbol{M}^\intercal \boldsymbol{\Sigma}_z \boldsymbol{M})} \mathbb{E}(\text{tr}[\boldsymbol{M}^\intercal \{\boldsymbol{x} \otimes \boldsymbol{b}(t)\}\{\boldsymbol{x} \otimes \boldsymbol{b}(t)\}^\intercal \boldsymbol{M}]) = 1.$$

We then have $\mathscr{F} \subset \mathscr{S}_{L_2} := \{f : \|f\|_{L_2} = 1\}$. By definitions, we have

$$\|f_M\|_{\psi_2} = \big\|\|\widetilde{\boldsymbol{z}}^\intercal \widetilde{\boldsymbol{M}}\|_2\big\|_{\psi_2} \le \sum_{l=1}^{q} \|\widetilde{\boldsymbol{z}}^\intercal \widetilde{\boldsymbol{M}}_l\|_{\psi_2} \le \|\widetilde{\boldsymbol{z}}\|_{\psi_2} \sum_{l=1}^{q} \|\widetilde{\boldsymbol{M}}_l\|_2 \le \sqrt{q}\|\widetilde{\boldsymbol{z}}\|_{\psi_2},$$

where $\widetilde{\boldsymbol{z}}^\intercal = \boldsymbol{z}^\intercal \boldsymbol{\Sigma}^{-1/2}$, $\widetilde{\boldsymbol{M}} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{M}$, and $\widetilde{\boldsymbol{M}}_l$ is the l-th column of $\widetilde{\boldsymbol{M}}$. Further, noting that

$$\langle \widetilde{\boldsymbol{z}}, \boldsymbol{w} \rangle = \langle \boldsymbol{z}^\intercal \boldsymbol{\Sigma}_z^{-1/2}, \boldsymbol{w}^\intercal \rangle = \langle \boldsymbol{z}, \boldsymbol{\Sigma}_z^{-1/2}\boldsymbol{w} \rangle,$$

and

$$\|\boldsymbol{\Sigma}_z^{-1/2}\boldsymbol{w}\|_2 \le C\|\boldsymbol{w}\|_2,$$

it follows from Lemma S.2 that

$$\|\widetilde{\boldsymbol{z}}\|_{\psi_2} \le C\sqrt{q}\|\boldsymbol{z}\|_{\psi_2} \le C\sqrt{qK}\kappa.$$

By Theorem 2.1.1 of Talagrand (2005), we have

$$\gamma_2(\mathscr{F} \cap \mathscr{S}_{L_2}, \|\cdot\|_{\psi_2}) \le C\sqrt{qK}\kappa\, \gamma_2(\mathscr{F} \cap \mathscr{S}_{L_2}, \|\cdot\|_{L_2}) \le C\sqrt{qK}\kappa w(\mathcal{A}),$$

where γ_2 is the γ_2-functional defined as in Definition S.1.

Using Theorem 10 of Banerjee et al. (2015), we chose

$$\theta = C\sqrt{qK}\kappa^2 \frac{w(\mathcal{A})}{\sqrt{n}} \ge C\kappa \frac{\gamma_2(\mathscr{F} \cap \mathscr{S}_{L_2}, \|\cdot\|_{\psi_2})}{\sqrt{n}}$$

to satisfy the lower bound in equation (125) of Banerjee et al. (2015). As a result,

$$\sup_{\mathcal{A}} \left| \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\mathrm{tr}(\boldsymbol{M}^\intercal \boldsymbol{\Sigma}_z \boldsymbol{M})} \mathrm{tr}[\boldsymbol{M}^\intercal \{\boldsymbol{x}_i \otimes \boldsymbol{b}(t_i)\}\{\boldsymbol{x}_i \otimes \boldsymbol{b}(t_i)\}^\intercal \boldsymbol{M}] - 1 \right| \le C\sqrt{qK}\kappa^2 \frac{w(\mathcal{A})}{\sqrt{n}},$$

with probability at least

$$1 - \exp\{-CqKw^2(\mathcal{A})\}.$$

Applying equation (53) of Banerjee et al. (2015) and (S.52), we have

$$w(\mathcal{A}) \le C|\mathcal{J}|\sqrt{Kq + \log p}.$$

The definition of $\mathcal{A}$ and Lemma S.2 together show that, for $\mathrm{vec}(\boldsymbol{M}) \in \mathcal{A}$,

$$\mathrm{tr}(\boldsymbol{M}^\intercal \boldsymbol{\Sigma}_z \boldsymbol{M}) \ge C\|\boldsymbol{M}\|_F^2 \ge C \sum_{j \in |\mathcal{J}|} \|\boldsymbol{M}_j\|_F^2.$$

When $n \geq C_3(|\mathcal{J}|^2 q^2 K^2 + |\mathcal{J}|^2 qK \log p)$, we have

$$\frac{1}{n}\sum_{i=1}^{n} \mathrm{tr}[\boldsymbol{M}^\intercal \{\boldsymbol{x}_i \otimes \boldsymbol{b}(t_i)\}\{\boldsymbol{x}_i \otimes \boldsymbol{b}(t_i)\}^\intercal \boldsymbol{M}] \geq C_4 \mathrm{tr}(\boldsymbol{M}^\intercal \boldsymbol{\Sigma}_z \boldsymbol{M}) \geq C_2 \sum_{j \in |\mathcal{J}|} \|\boldsymbol{M}_j\|_F^2$$

with high probability. Therefore, when $n \geq C_3(|\mathcal{J}|^2 q^2 K^2 + |\mathcal{J}|^2 qK \log p)$, Condition $\mathcal{M}(\mathcal{J}, \delta_{\mathcal{J}})$ is satisfied for some $\delta_{\mathcal{J}} = C_2 > 0$ with high probability. □

When q and p are allowed to be growing with the sample size n, the following lemma shows that $\lambda_{\max}(\boldsymbol{\Sigma})$ is bounded by $O(K)$ with probability tending to one. Therefore, together with Lemma S.3, it implies that the term $\lambda_{\max}(\boldsymbol{\Sigma})/\delta_{\mathcal{J}}$ in Theorem 1 of the main paper is upper bounded by $O(K)$ with high probability.

Lemma S.4. *Let* $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^\intercal$, $\boldsymbol{Z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)^\intercal$, *where* $\boldsymbol{z}_i = \boldsymbol{x}_i \otimes \boldsymbol{b}(t_i)$, $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^\intercal$. *If*

$$\lambda_{\max}\left(\frac{\boldsymbol{X}^\intercal \boldsymbol{X}}{n}\right) \leq U_x < \infty, \tag{S.53}$$

then

$$\lambda_{\max}\left(\frac{\boldsymbol{Z}^\intercal \boldsymbol{Z}}{n}\right) \leq C_1 K U_x. \tag{S.54}$$

Proof. For any $\boldsymbol{v} = (v_1, \ldots, v_p)^\intercal \in \mathbb{R}^p$, we have

$$\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{p} x_{ij} v_j\right)^2 \leq U_x \sum_{j=1}^{p} v_j^2 \tag{S.55}$$

due to the definition of $\lambda_{\max}(\cdot)$. It can be shown that for all $\widetilde{\boldsymbol{v}} = (\widetilde{v}_1, \ldots, \widetilde{v}_p)^\intercal \in \mathbb{R}^p$, we have

$$\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{p} |x_{ij}\widetilde{v}_j|\right)^2 \leq U_x \sum_{j=1}^{p} \widetilde{v}_j^2.$$

To see this, if $x_{ij} \leq 0$, we take $v_j = -|\widetilde{v}_j|$, and if $x_{ij} > 0$, we take $v_j = |\widetilde{v}_j|$. The aforementioned inequality is thus obtained by (S.55). Next, consider $\boldsymbol{w} = (\boldsymbol{w}_1^\intercal, \ldots, \boldsymbol{w}_p^\intercal)^\intercal \in \mathbb{R}^{pK}$, where $\boldsymbol{w}_j = (w_{j1}, \ldots, w_{jK})^\intercal \in \mathbb{R}^K$. We now borrow some notations used in the proof of Lemma S.2. Recall that $\{\widetilde{b}_k(t)\}_{k=1}^K$ is the B-spline basis with the same knots and order as $\{b_k(t)\}$, and there exists $\boldsymbol{\Psi} = (\Psi_{k,\widetilde{k}})_{K\times K}$ such that (S.48) holds. By definitions and (S.50), we have

$$\begin{aligned}
\frac{\boldsymbol{w}^\intercal \boldsymbol{Z}^\intercal \boldsymbol{Z} \boldsymbol{w}}{n} &= \frac{1}{n}\sum_{i=1}^{n}\left\{\sum_{j=1}^{p} x_{ij}\sum_{k=1}^{K} b_k(t) w_{jk}\right\}^2 \\
&\leq \frac{1}{n}\sum_{i=1}^{n}\left\{\sum_{j=1}^{p}\left|\sum_{k=1}^{K} b_k(t) w_{jk}\right| |x_{ij}|\right\}^2 \\
&\leq U_x \sum_{j=1}^{p}\left(\sum_{k=1}^{K} b_k(t) w_{jk}\right)^2 \\
&= C U_x K \|\boldsymbol{w}\|_2^2,
\end{aligned} \tag{S.56}$$

which finishes the proof. □

When q and p are fixed (that is, not growing with the sample size n), Lemma S.5 shows that with probability tending to one, $\lambda_{\max}(\boldsymbol{\Sigma})/\delta_{\mathcal{J}}$ is upper bounded by a constant.

Lemma S.5. *If the conditional density of t given $\boldsymbol{x}$ is bounded by a constant, $\|\boldsymbol{x}\|_{\psi_2} \leq \kappa$ and $\lambda_{\min}\{\mathbb{E}(\boldsymbol{x}\boldsymbol{x}^\intercal)\} \geq C_1$, then Condition $\mathcal{M}(\mathcal{J}, \delta_{\mathcal{J}})$ holds for some constant $\delta_{\mathcal{J}} > 0$ and $\lambda_{\max}(\boldsymbol{\Sigma})$ is upper bounded by a constant, with high probability when $n \geq CK^2$.*

Proof. The following proof is similar to that of Lemma S.3. Recall $\boldsymbol{Z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)^\intercal$, where $\boldsymbol{z}_i = \boldsymbol{x}_i \otimes \boldsymbol{b}(t_i)$, which are i.i.d. copies of $\boldsymbol{z} = \boldsymbol{x} \otimes \boldsymbol{b}(t)$. Let $\boldsymbol{\Sigma}_z = \mathbb{E}(\boldsymbol{z}\boldsymbol{z}^\intercal)$. Consider the class of functions:

$$\mathscr{F}_w = \left[f_w\{\boldsymbol{x} \otimes \boldsymbol{b}(t)\} = \frac{1}{\sqrt{\boldsymbol{w}^\intercal \boldsymbol{\Sigma}_z \boldsymbol{w}}} \sqrt{\boldsymbol{w}^\intercal \{\boldsymbol{x} \otimes \boldsymbol{b}(t)\}\{\boldsymbol{x} \otimes \boldsymbol{b}(t)\}^\intercal \boldsymbol{w}} : \boldsymbol{w} \in \mathcal{A}_w \right],$$

where

$$\mathcal{A}_w = \big\{ \boldsymbol{w} : \|\boldsymbol{w}\|_F^2 \leq 1 \big\}.$$

It is trivial to see $\|f_w\|_{L_2}^2 = 1$, which yields $\mathscr{F}_w \subset \mathscr{S}_{L_2} := \{f : \|f\|_{L_2} = 1\}$. By definitions and Lemma S.2, we have

$$\|f_w\|_{\psi_2} = \|\|\boldsymbol{z}^\intercal \boldsymbol{w}\|_2\|_{\psi_2} \leq \| \leq \|\boldsymbol{z}\|_{\psi_2}\|\boldsymbol{w}\|_2 \leq C\sqrt{K}\kappa.$$

By Theorem 2.1.1 of Talagrand (2005), we have

$$\gamma_2(\mathscr{F}_w \cap \mathscr{S}_{L_2}, \|\cdot\|_{\psi_2}) \leq C\sqrt{K}\kappa\, \gamma_2(\mathscr{F}_w \cap \mathscr{S}_{L_2}, \|\cdot\|_{L_2}) \leq C\sqrt{K}\kappa w(\mathcal{A}_w),$$

where γ_2 is the γ_2-functional defined as in Definition S.1.

Using Theorem 10 of Banerjee et al. (2015), we chose

$$\theta = C\sqrt{K}\kappa^2 \frac{w(\mathcal{A}_w)}{\sqrt{n}} \geq C\kappa \frac{\gamma_2(\mathscr{F}_w \cap \mathscr{S}_{L_2}, \|\cdot\|_{\psi_2})}{\sqrt{n}}$$

to satisfy the lower bound in equation (125) of Banerjee et al. (2015). As a result,

$$\sup_{\mathcal{A}} \left| \frac{1}{n} \sum_{i=1}^n \frac{1}{\boldsymbol{w}^\intercal \boldsymbol{\Sigma}_z \boldsymbol{w}} [\boldsymbol{w}^\intercal \{\boldsymbol{x}_i \otimes \boldsymbol{b}(t_i)\}\{\boldsymbol{x}_i \otimes \boldsymbol{b}(t_i)\}^\intercal \boldsymbol{w}] - 1 \right| \leq C\sqrt{K}\kappa^2 \frac{w(\mathcal{A}_w)}{\sqrt{n}}, \tag{S.57}$$

with probability at least

$$1 - \exp\{-CKw^2(\mathcal{A}_w)\}.$$

By the covering number argument, we have

$$N(\epsilon, \mathcal{A}_w, l_2) \leq (C_1/\epsilon)^{C_2 K}.$$

It follows from the Dudley's integral entropy bound (see, e.g., Theorem 3.1 of Koltchinskii, 2011) that the Gaussian width satisfies

$$w(\mathcal{A}_w) \leq C\sqrt{K}. \tag{S.58}$$

If $n \geq C\kappa^4 K^2$ for some constant C, by (S.46), (S.47), (S.57), and (S.58), we have

$$C_1 \leq \lambda_{\min}\left(\frac{\boldsymbol{Z}^\intercal \boldsymbol{Z}}{n}\right) \leq \lambda_{\max}\left(\frac{\boldsymbol{Z}^\intercal \boldsymbol{Z}}{n}\right) \leq C_2\kappa^2$$

with high probability, which finishes the proof. □

S.5 Additional Numerical Results

S.5.1 Additional Experiments of Synthetic Data

We extend our simulation setting to larger numbers of response variables to show the trend of the performance of the proposed method when q increases. To match the measure of estimation error used in our theoretical investigation (see (5.4) and (5.5) in Section 5 of the main paper), we calculate the integrated squared error (ISE) in this additional study, that is,

$$\text{ISE} = \sum_{j=1}^{p}\sum_{l=1}^{q}\int_0^1 \{\widehat{f}_{jl}(t) - f_{jl}(t)\}^2 \, \mathrm{d}t,$$

where $\widehat{f}_{jl}(t)$'s are the estimated coefficient functions of various methods. More precisely, we calculate the ISE for the scenarios of $q = 15$, 30, and 50 under our simulation setting of $n = 200$ or 400, $p = 51$, and SNR $= 20$. For each scenario, 50 replicates are generated and the proposed all-mode reduction model is trained, where the tuning parameters are determined as in Section 4.6 of the main paper. The results are shown in Figure S.1 as a line chart, with a bar denoted the standard errors of each scenario. Figure S.1 shows that the ISE has an rising trend as the number of responses increases, which is consistent with our main theorem.

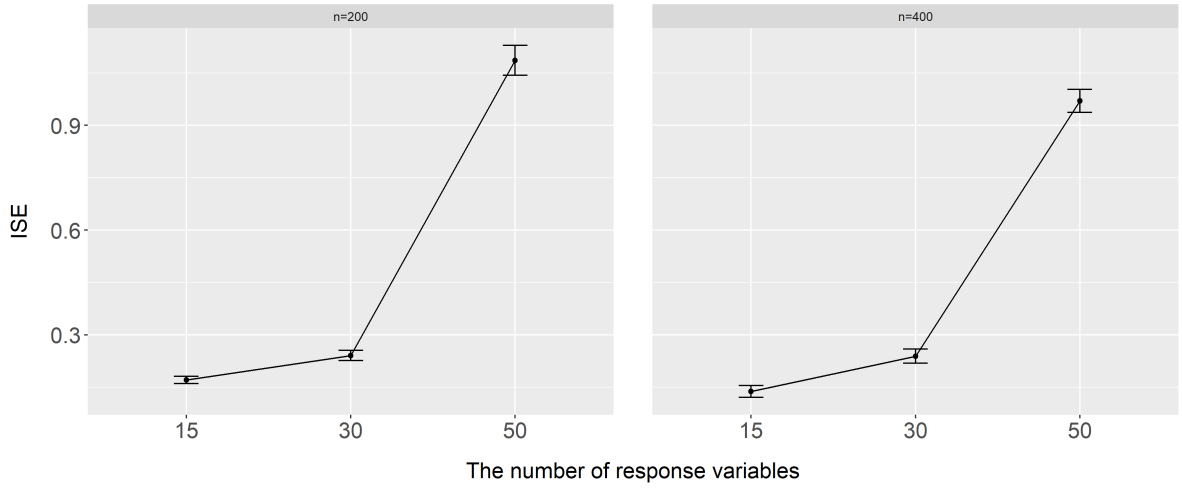

Figure S.1: The ISE for different numbers of the response variables.

S.5.2 Additional Results of Real Data

In this subsection, we plot the estimated coefficient functions of one relevant SNP, rs9321440, identified by the proposed all-mode reduction method. Figure S.2 depicts the fitted coefficient functions and their average based on 50 replicates of random splitting. From the estimation of varying coefficient functions, we observe that the SNP rs9321440 may have different varying effects on the phenotypes of *height*, *bi-deltoid girth*, *right arm girth-upper third*, *hip girth*, and *thigh girth* given distinct body weights. To be specific, for both *height* and *bi-deltoid girth*, the estimated coefficients show negative patterns, and the effects decrease first but then increase with respect to the increase of body weights. This similarity in patterns could be explained by the high correlations between these two phenotypes (Chalmers et al., 2021), where both are mainly due to the skeleton of a human. For *right arm girth-upper third*, *hip girth*, and *thigh girth*, these measurements are characterized by the body fat and highly correlated (Freedman and Rimm, 1989). Their corresponding estimated coefficients have similar patterns and show positive effects. As for the phenotype of *waist girth*, the effect of this SNP may not vary with body weights significantly.

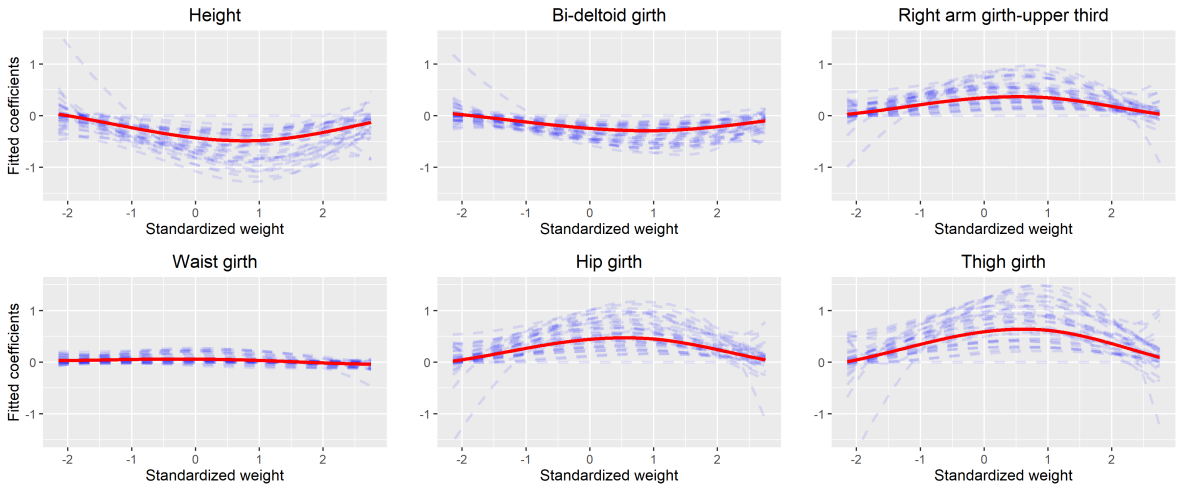

Figure S.2: Fitted coefficient functions of the biologically confirmed SNP rs9321440. In each panel, the blue dashed lines and the red solid line are the fitted functions and their average based on 50 replicates of random splitting, respectively.

References

Absil, P.-A., R. Mahony, and R. Sepulchre (2009). *Optimization algorithms on matrix manifolds*. Princeton, New Jersey: Princeton University Press.

Banerjee, A., S. Chen, F. Fazayeli, and V. Sivakumar (2015). Estimation with norm regularization. *arXiv preprint arXiv:1505.02294*.

Boyd, S., N. Parikh, E. Chu, B. Peleato, and J. Eckstein (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning 3*(1), 1–122.

Bunea, F., Y. She, and M. H. Wegkamp (2012). Joint variable and rank selection for parsimonious estimation of high-dimensional matrices. *The Annals of Statistics 40*(5), 2359–2388.

Chalmers, P. N., S. R. Lindsay, W. Smith, J. Kawakami, R. Hill, R. Z. Tashjian, and J. D. Keener (2021). Infraspinatus and deltoid length and patient height: implications for lateralization and distalization in reverse total shoulder arthroplasty. *Journal of Shoulder and Elbow Surgery 30*(4), 712–719.

De Boor, C. (1973). The quasi-interpolant as a tool in elementary polynomial spline theory. *Approximation Theory*, 269–276.

De Boor, C. (1976). Splines as linear combinations of b-splines. a survey. Technical report, Wisconsin Univ Madison Mathematics Research Center.

Freedman, D. S. and A. A. Rimm (1989). The relation of body fat distribution, as assessed by six girth measurements, to diabetes mellitus in women. *American Journal of Public Health 79*(6), 715–720.

Gabay, D. and B. Mercier (1976). A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Computers & Mathematics with Applications 2*(1), 17–40.

Hunter, D. R. and R. Li (2005). Variable selection using mm algorithms. *The Annals of Statistics 33*(4), 1617–1642.

Kolda, T. G. and B. W. Bader (2009). Tensor decompositions and applications. *SIAM Review 51*(3), 455–500.

Koltchinskii, V. (2011). *Oracle inequalities in empirical risk minimization and sparse recovery problems*. New York: Springer Science & Business Media.

Talagrand, M. (2005). *The generic chaining*. Berlin: Springer.

Zhu, Y. (2017). An augmented admm algorithm with application to the generalized lasso problem. *Journal of Computational and Graphical Statistics 26*(1), 195–204.