Supplementary Materials of "Second-Order Sparse Sufficient Dimension Reduction with Applications to Quadratic Discriminant Analysis"

Abstract

The Supplementary Materials are organized as follows. In Section S.1, we present the additional results for sparse DR, paralleling with the results for sparse SAVE in the paper. The additional assumptions required by Theorem 4 is given in Section S.2. In Section S.3, we provide a toy example illustrating the phenomenon in Assumption (A5). In Section S.4, we provide the improved rate of estimation error of sparse SAVE. The thorough comparison with existing literature is given in Section S.5. In Section S.6, we illustrate the integrated formulation of our two-step procedure. The extension of our sparse SDR framework to the third-order method is described in Section S.7. Summaries of our algorithms and the tuning parameter selection details are provided in Section S.8. In Section S.9, we demonstrate the effect of nuclear-norm regularization with empirical findings. Additional numerical results, complementary to those presented in the main paper, are displayed in Section S.10. In Section S.11, we provide the analysis of spaceflight data. The technical proofs for the theoretical results in the paper, namely, Theorems 1–4 and Corollary S.1, are presented from Section S.12 to Section S.17. In Section S.18, we provide the proof for Theorem S.2. In each section, the proof of the auxiliary lemmas is given in subsections. Some auxiliary definitions and lemmas are provided in Section S.19.

S.1 Extension: sparse DR

The directional regression (DR; Li and Wang, 2007) method is another second-order SDR method, proposed to integrate the strengths of SIR and SAVE. In the paper, we thoroughly developed the estimation procedure (Section 3) and theoretical properties (Section 4) for the sparse SAVE method. We have also demonstrated the empirical performance of the proposed sparse DR through the numerical studies. Due to the space limit of the paper, we relegate the the complete procedure and theoretical properties of sparse DR in this section.

Recall that $(\mathbf{X}', Y')$ denotes an independent copy of $(\mathbf{X}, Y)$, and $\widetilde{Y}$ and $\widetilde{Y}'$ denote the discretized versions of Y and Y', taking values in $\{1, \ldots, H\}$. For $h, h' \in \{1, \ldots, H\}$ and $h > h'$, we define the new index $\bar{h} = (h-1)(h-2)/2 + h'$ such that (h, h') correspond to $\bar{h}$ one-by-one. For $1 \le \bar{h} \le \bar{H}$, where $\bar{H} = H(H-1)/2$, define $\mathbf{\Lambda}_{\bar{h}} = \mathrm{E}\{(\mathbf{X} - \mathbf{X}')(\mathbf{X} - \mathbf{X}')^\top \mid \widetilde{Y} = h, \widetilde{Y}' = h'\}$ and $\mathbf{\Theta}_{\bar{h}} = \mathbf{\Lambda}_{\bar{h}} - 2\mathbf{\Sigma}$. The DR subspace is

$$\mathcal{S}_{\mathrm{DR}} = \mathrm{span}(\mathbf{\Sigma}^{-1}\mathbf{\Theta}_1, \ldots, \mathbf{\Sigma}^{-1}\mathbf{\Theta}_{\bar{H}}). \tag{S.1}$$

Our development of sparse DR is parallel to that of sparse SAVE, including parsimonious reparameterization, variable selection, and subspace estimation.

Let $\mathbf{M}_{\bar{h}} = \mathbf{\Sigma}^{-1}\mathbf{\Theta}_{\bar{h}}$, for $\bar{h} = 1, \ldots, \bar{H}$. The DR subspace can be expressed as

$$\mathcal{S}_{\mathrm{DR}} = \mathrm{span}(\mathbf{M}_1, \ldots, \mathbf{M}_{\bar{H}}).$$

Given the group sparsity assumption for the central subspace $\mathcal{S}_{Y|\mathbf{X}} = \mathrm{span}(\boldsymbol{\beta})$ such that

$$\boldsymbol{\beta} = (\boldsymbol{\beta}_{\mathcal{A}}^\top, \mathbf{0}^\top)^\top,$$

and the equivalence $\mathcal{S}_{\mathrm{DR}} = \mathcal{S}_{Y|\mathbf{X}}$ under the linearity condition and the constant variance condition, each matrix $\mathbf{M}_{\bar{h}}$ exhibits row-wise sparsity as follows,

$$\mathbf{M}_{\bar{h}} = (\mathbf{M}_{\bar{h},\mathcal{A}*}^\top, \mathbf{0}^\top)^\top, \quad \bar{h} = 1, \ldots, \bar{H}. \tag{S.2}$$

Furthermore, the non-zero component $\mathbf{M}_{\bar{h},\mathcal{A}*}$ possesses an intrinsic low-rank structure in the sense that $\mathrm{span}(\mathbf{M}_{\bar{h},\mathcal{A}*}) = \mathrm{span}(\mathbf{M}_{\bar{h},\mathcal{A}\mathcal{A}})$. The combination of row-wise sparsity and low-rank structure facilitates the parsimonious reparameterization of the DR subspace as follows.

Theorem S.1. *Let $\mathcal{A}$ denote the active set. We have $\mathbf{M}_{\bar{h},\mathcal{A}\mathcal{A}} = (\mathbf{\Sigma}_{\mathcal{A}\mathcal{A}})^{-1}\mathbf{\Theta}_{\bar{h},\mathcal{A}\mathcal{A}}$ and $\mathbf{M}_{\bar{h},\mathcal{A}\mathcal{A}^c} = (\mathbf{\Sigma}_{\mathcal{A}\mathcal{A}})^{-1}\mathbf{\Theta}_{\bar{h},\mathcal{A}\mathcal{A}^c}$. Furthermore, we have*

$$\mathrm{span}(\mathbf{M}_{\bar{h}}) = \mathrm{span}\left\{\begin{pmatrix}\mathbf{M}_{\bar{h},\mathcal{A}\mathcal{A}}\\ \mathbf{0}\end{pmatrix}\right\}, \quad \mathcal{S}_{\mathrm{DR}} = \mathrm{span}\left\{\begin{pmatrix}\mathbf{M}_{1,\mathcal{A}\mathcal{A}}, \ldots, \mathbf{M}_{\bar{H},\mathcal{A}\mathcal{A}}\\ \mathbf{0}\end{pmatrix}\right\}.$$

The proof of Theorem S.1 follows similar arguments to the proof of Theorem 1 for the SAVE subspace and is thus omitted. As indicated by Theorem S.1, the reparameterization of the DR subspace leads to a reduction in the number of parameters from $p^2H(H-1)/2$ to $s^2H(H-1)/2$. Moreover, the parameterization of the DR subspace involves only the variables indexed by $\mathcal{A}$. Under the parsimonious reparameterization of the DR subspace, our sparse DR proposal also comprises of two steps, namely, variable selection in the first step and subspace estimation in the second step.

Variable selection. With some abuse of notation, we still use $\mathcal{M} \in \mathbb{R}^{\bar{H}\times p\times p}$ to denote the tensor parameter of sparse DR, constructed by stacking $\bar{H}$ blocks $\mathbf{M}_{\bar{h}}$ along the third direction. The optimization problem we solve in the first step follows a similar form to that

in sparse SAVE, as shown below,

$$\underset{\mathcal{G}\in\mathbb{R}^{\bar{H}\times p\times p}}{\operatorname{argmin}} \sum_{\bar{h}=1}^{\bar{H}} \left\{ \frac{1}{2}\operatorname{tr}\left(\mathcal{G}_{[\bar{h},:,:]}^{\top}\widehat{\boldsymbol{\Sigma}}\mathcal{G}_{[\bar{h},:,:]}\right) - \operatorname{tr}(\mathcal{G}_{[\bar{h},:,:]}^{\top}\widehat{\boldsymbol{\Theta}}_{\bar{h}}) \right\} + \lambda_1 \sum_{i=1}^{p}\sum_{j=1}^{p}\left(\sum_{\bar{h}=1}^{\bar{H}}\mathcal{G}_{[\bar{h},i,j]}^2\right)^{1/2}, \quad \text{(S.3)}$$

where $\widehat{\boldsymbol{\Theta}}_{\bar{h}} = \widehat{\boldsymbol{\Lambda}}_{\bar{h}} - 2\widehat{\boldsymbol{\Sigma}}$ is the sample estimate of $\boldsymbol{\Theta}_{\bar{h}}$ with $\widehat{\boldsymbol{\Lambda}}_{\bar{h}} = \widehat{\mathbf{K}}_h + \widehat{\mathbf{K}}_{h'} - \overline{\mathbf{X}}_h\overline{\mathbf{X}}_{h'}^{\top} - \overline{\mathbf{X}}_{h'}\overline{\mathbf{X}}_h^{\top}$, $\widehat{\mathbf{K}}_h = \sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i^{\top} I(Y_i = h)/n_h$, $\overline{\mathbf{X}}_h = \sum_{i=1}^n \mathbf{X}_i I(Y_i = h)/n_h$, and $n_h = \sum_{i=1}^H I(Y_i = h)$.

Let $\widehat{\mathcal{M}} \in \mathbb{R}^{(H(H-1)/2)\times p\times p}$ denote the solution of (S.3). The important variables are selected by identifying non-zero slices $\widehat{\mathcal{M}}_{[:,i,:]}$, $i = 1,\ldots,p$. We estimate the active set $\mathcal{A}$ via $\widehat{\mathcal{A}} = \{i \mid \|\widehat{\mathcal{M}}_{[:,i,:]}\|_{2,\infty} > 0\}$, where $\|\widehat{\mathcal{M}}_{[:,i,:]}\|_{2,\infty} = \max_j \|\widehat{\mathcal{M}}_{[:,i,j]}\|_2$, and the estimated sparsity level $\widehat{s} = |\widehat{\mathcal{A}}|$. The optimization problem (S.3) can also be solved by SMORE algorithm in Algorithm 1, where we only need to replace $\widehat{\boldsymbol{\delta}} = (\operatorname{vec}(\widehat{\boldsymbol{\Delta}}_1),\ldots,\operatorname{vec}(\widehat{\boldsymbol{\Delta}}_H)) \in \mathbb{R}^{p^2\times H}$ with $\widehat{\boldsymbol{\delta}} = (\operatorname{vec}(\widehat{\boldsymbol{\Theta}}_1),\ldots,\operatorname{vec}(\widehat{\boldsymbol{\Theta}}_{\bar{H}})) \in \mathbb{R}^{p^2\times\bar{H}}$.

Subspace estimation. In the second step, with the estimated active set $\widehat{\mathcal{A}}$, we shift the focus to the estimation of the lower-dimensional matrix $(\mathbf{M}_{1,\widehat{\mathcal{A}}\widehat{\mathcal{A}}},\ldots,\mathbf{M}_{\bar{H},\widehat{\mathcal{A}}\widehat{\mathcal{A}}})$. We solve the following nuclear-norm penalized optimization problem,

$$\underset{\mathbf{B}_1,\ldots,\mathbf{B}_{\bar{H}}\in\mathbb{R}^{\widehat{s}\times\widehat{s}}}{\operatorname{argmin}} \sum_{\bar{h}=1}^{\bar{H}} \left\{ \frac{1}{2}\operatorname{tr}(\mathbf{B}_{\bar{h}}^{\top}\widehat{\boldsymbol{\Sigma}}_{\widehat{\mathcal{A}}\widehat{\mathcal{A}}}\mathbf{B}_{\bar{h}}) - \operatorname{tr}(\mathbf{B}_{\bar{h}}^{\top}\widehat{\boldsymbol{\Theta}}_{\bar{h},\widehat{\mathcal{A}}\widehat{\mathcal{A}}}) \right\} + \lambda_2\|(\mathbf{B}_1,\ldots,\mathbf{B}_{\bar{H}})\|_*,$$

and denote the solution by $\widehat{\mathbf{B}} = (\widehat{\mathbf{B}}_1,\ldots,\widehat{\mathbf{B}}_{\bar{H}})$. Let $\widehat{\sigma}_j$ and $\widehat{\boldsymbol{\eta}}_j \in \mathbb{R}^{\widehat{s}}$ denote the j-th largest eigenvalue and the corresponding left singular vector of $\widehat{\mathbf{B}}$. The structural dimension d is estimated by

$$\widehat{d} = |\{j \mid \widehat{\sigma}_j \geq \delta\}| \wedge d_{\max}, \quad \text{(S.4)}$$

for some user-defined constants $\delta > 0$ and $d_{\max}$. Then, we expand the vector $\widehat{\boldsymbol{\eta}}_j$ to the p-dimensional vector $\widehat{\boldsymbol{\beta}}_j$, such that $\widehat{\boldsymbol{\beta}}_{j,\widehat{\mathcal{A}}} = \widehat{\boldsymbol{\eta}}_j$ and $\widehat{\boldsymbol{\beta}}_{j,\widehat{\mathcal{A}}^c} = \mathbf{0}$, $j = 1,\ldots,\widehat{d}$. The estimate of the basis $\boldsymbol{\beta}$ is given by $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_1,\ldots,\widehat{\boldsymbol{\beta}}_{\widehat{d}}) \in \mathbb{R}^{p\times\widehat{d}}$.

Parallel to the theoretical development of sparse SAVE, we establish the corresponding properties for the sparse DR proposal. Specifically, we provide variable selection and subspace estimation consistencies for sparse DR under technical assumptions similar to those used in sparse SAVE.

Theoretical properties. The following assumption is a revised version of Assumption (A3) adapting to $\boldsymbol{\Theta}_{\bar{h}}$ in sparse DR.

(A3′) Assume that $\|(\boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}})^{-1}\|_\infty$, $\|\boldsymbol{\Sigma}_{\mathcal{A}^c\mathcal{A}}\|_\infty$, $\max_{\bar{h}}\|\boldsymbol{\Theta}_{\bar{h},\mathcal{A}*}\|_{\max}$, $\max_h\|\boldsymbol{\Theta}_{\bar{h},\mathcal{A}\mathcal{A}}\|_1$ are bounded from above.

With the adapted Assumption (A3′) and some other assumptions used in Theorem 2, we establish the variable selection consistency for sparse DR, stated in the following theorem.

Theorem S.2 (Variable selection consistency for sparse DR). *Suppose Assumptions (A1)(A2)(A3′)(A4) hold and that* $n \geq c_1 s^2 \log p$ *for some large enough constant* $c_1 > 0$*, there exist constants* $\kappa_1, \kappa_2, C, C' > 0$ *such that, by taking* $\kappa_1 \sqrt{s^2 \log p/n} \leq \lambda_1 \leq \kappa_2 \sqrt{s^2 \log p/n}$*, with probability at least* $1 - Cp^{-C'}$*, we have*

(i) $\widehat{\mathcal{A}} \subseteq \mathcal{A}$,

(ii) $\max_{i=1,\ldots,p} \|\widehat{\mathcal{M}}_{[:,i,:]} - \mathcal{M}_{[:,i,:]}\|_{2,\infty} \lesssim \sqrt{s^2 \log p/n}$,

(iii) $\widehat{\mathcal{A}} = \mathcal{A}$ *if we further assume that Assumption (A5) holds.*

Moreover, under the extra Assumption (A7), sparse DR also achieves the consistent subspace estimation.

Theorem S.3 (Subspace estimation consistency for sparse DR). *Under Assumptions (A1)(A2)(A3′) and (A4)–(A7), there exist constants* $\kappa_3, C, C' > 0$ *such that, by taking the thresholding value* $\delta = \sigma/2$ *and* $0 < \lambda_2 \leq \kappa_3 \sqrt{s \log p/n}$*, with probability at least* $1 - Cp^{-C'}$*, we have*

(i) $\widehat{d} = d$,

(ii) $D(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}) \lesssim \sqrt{s^2 \log p/(n\sigma^2)}$.

The proof of Theorem S.2 is provided in Section S.18. The proof of Theorem S.3 is similar to that of Theorem 3 and is omitted. By comparing Theorems S.2–S.3 and Theorems 2–3, we establish exactly the same consistency results and convergence rates for sparse DR and sparse SAVE under nearly the same assumptions. Therefore, Theorems 2, 3, S.2, and S.3 together provide a unified understanding of the theoretical properties of our second-order sparse SDR methods.

S.2 Additional assumptions

The following are additional assumptions required by Theorem 4.

(A8) There exists a constant $m > 0$ such that $m^{-1} \leq \varphi_{\min}(\boldsymbol{\Sigma}_k) \leq \varphi_{\max}(\boldsymbol{\Sigma}_k) \leq m$, $\|\boldsymbol{\mu}_k - \boldsymbol{\mu}_1\|_2 \leq m$ for $k = 1, \ldots, K$.

(A9) There exists some constant $T > 0$ such that $\|\boldsymbol{\Sigma}_k^{-1} - \boldsymbol{\Sigma}_1^{-1}\|_F \leq T$, for $k = 2, \ldots, K$.

(A10) Denote $Q_{k,k'}(\mathbf{X}) = Q_k(\mathbf{X}) - Q_{k'}(\mathbf{X})$, $1 \leq k < k' \leq K$, and let $f_{Q_{k,k'}}$ denote the density function of $Q_{k,k'}(\mathbf{X})$. Assume that for some constants $\delta, M > 0$, we have $\sup_{|x| \leq \delta} f_{Q_{k,k'}}(x) < M$.

S.3 A toy example illustrating Assumption (A5)

We provide a toy QDA example illustrating the phenomenon in Assumption (A5) that the signal for identifying an active variable may be weak inside $\mathbf{M}_{h,\mathcal{A}\mathcal{A}}$ but strong inside $\mathbf{M}_{h,\mathcal{A}\mathcal{A}^c}$, presented in the following lemma.

Lemma S.1. *Let* $Y \sim \mathrm{Unif}\{1,2\}$ *and* $\mathbf{X} = (X_1, X_2, X_3)^\top \in \mathbb{R}^3$, *and assume that* $(X_1, X_2)^\top \mid (Y = 1) \sim N(\mathbf{0}, \mathbf{\Sigma}_{1,\mathcal{A}\mathcal{A}})$ *and* $(X_1, X_2)^\top \mid (Y = 2) \sim N(\mathbf{0}, \mathbf{\Sigma}_{2,\mathcal{A}\mathcal{A}})$, *where* $\mathbf{\Sigma}_{1,\mathcal{A}\mathcal{A}} = \mathbf{I}_2 + \tau\mathbf{J}_2$, $\mathbf{\Sigma}_{2,\mathcal{A}\mathcal{A}} = \mathbf{I}_2 - \tau\mathbf{J}_2$, $\mathbf{J}_2 = \mathbf{1}_2\mathbf{1}_2^\top$, $\mathbf{1}_2 = (1,1)^\top$, *and* $\tau > 0$. *Define* $X_3 = \ell(X_1 + X_2) + \varepsilon$, *where* $\ell > 0$ *and* $\varepsilon \sim N(0,1)$ *is independent of* (X_1, X_2, Y). *Then:*

(1) $Y \perp\!\!\!\perp X_3 \mid (X_1, X_2)$, *and hence* $\mathcal{A} = \{1,2\}$ *and* $\mathcal{A}^c = \{3\}$;

(2) For each $i \in \mathcal{A}$, $\max_{j\in\mathcal{A}} \|\mathcal{M}_{[:,i,j]}\|_2 = \sqrt{2}\tau$ *and* $\max_{j\in\mathcal{A}^c} \|\mathcal{M}_{[:,i,j]}\|_2 = 2\sqrt{2}\ell\tau$.

(3) Let $r_n = c\sqrt{s^2 \log p/n}$. *If* $0 < \tau < r_n/\sqrt{2}$ *and* $\ell > r_n/(2\sqrt{2}\tau)$, *then*

$$\|\mathcal{M}_{[:,i,:]}\|_{2,\infty} = \max_{1\le j\le p} \|\mathcal{M}_{[:,i,j]}\|_2 > c\sqrt{s^2 \log p/n}, \quad \forall i \in \mathcal{A}$$

and

$$\max_{j\in\mathcal{A}} \|\mathcal{M}_{[:,i,j]}\|_2 < c\sqrt{s^2 \log p/n}, \quad \forall i \in \mathcal{A}.$$

The proof is presented in Section S.16. This example shows that though (X_1, X_2) contribute to the distribution of Y, it is nearly independent of Y as τ gets close to zero and the signal for identifying the active set $\mathcal{A} = \{1,2\}$ can be weak inside $\mathbf{M}_{h,\mathcal{A}\mathcal{A}}$. Though X_3 is not contributory in inferring Y, X_3 has a strong correlation with (X_1, X_2) when ℓ is large, which enhances the signal of $\mathcal{A}$ inside $\mathbf{M}_{h,\mathcal{A}\mathcal{A}^c}$. Therefore, this lemma provides a toy example illustrating the cases where an active variable may have weak signal in $\mathbf{M}_{h,\mathcal{A}\mathcal{A}}$ but strong signal in $\mathbf{M}_{h,\mathcal{A}\mathcal{A}^c}$. Hence Assumption (A5) is strictly weaker than a condition requiring the signal to come only from $\mathbf{M}_{h,\mathcal{A}\mathcal{A}}$, and this weakening is practically meaningful rather than merely technical.

S.4 Improved rate of estimation error

The convergence rate of the sparse SAVE estimator $\mathrm{span}(\widehat{\boldsymbol{\beta}})$ in Theorem 3 (ii) can be further improved under a weaker constraint on the eigen-gap σ, as described in the following variant of Assumption (A7).

(A7′) There exists a constant $\widetilde{c}$ that depends on parameters $m, \theta_0, \eta_1, H, \pi_h, \|X_j\|_{\psi_2}$, and $\|X_j I(\widetilde{Y} = h)\|_{\psi_2}$, for $j = 1, \ldots,$ and $p, h = 1, \ldots, H$, and does not depend on s, n and p, such that $\sigma > 2\widetilde{c}\sqrt{s^2 \log s/n}$.

Corollary S.1. *Suppose Assumptions (A1)–(A6) and (A7′) hold and that* $n \geq c_3 s^2 \log p$ *for some large enough constant* $c_3 > 0$*, there exist constants* $\kappa_1, \kappa_2, \kappa_4, C, C' > 0$ *such that, by taking* $\kappa_1\sqrt{s^2 \log p/n} \leq \lambda_1 \leq \kappa_2\sqrt{s^2 \log p/n}$*,* $0 < \lambda_2 \leq \kappa_4\sqrt{s \log s/n}$*, and the thresholding value* $\delta = \sigma/2$*, with probability at least* $1 - Cs^{-C'}$*, we have (i)* $\widehat{d} = d$*; (ii)* $D(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}) \lesssim \sqrt{s^2 \log s/(n\sigma^2)}$.

According to Corollary S.1, when s is diverging, a sharper rate $O\{s^2 \log s/(n\sigma^2)\}$ for subspace estimation is achieved. As n, p, s approach to infinity simultaneously with the rates $s^2 \log s = o(n\sigma^2)$ and $s^2 \log p = o(n)$, the estimated subspace $\text{span}(\widehat{\boldsymbol{\beta}})$ converges to the true parameter asymptotically. We remark that the sharper convergence rate should be interpreted as a refinement under diverging sparsity, whereas Theorem 3 remains a more meaningful non-asymptotic result in the fixed-s or moderate-s regimes.

S.5 Detailed comparison with existing literature

S.5.1 Comparison with Qian et al. (2019) and Yu et al. (2016)

Qian et al. (2019) developed a unified framework of sparse SDR methods based on the penalized minimum discrepancy approach. Specifically, their approach solves the following optimization problem:

$$\min_{\boldsymbol{\Gamma} \in \mathbb{R}^{p\times d}, \mathbf{V} \in \mathbb{R}^{d\times l}} \frac{1}{2}\text{tr}\{(\widehat{\mathbf{M}} - \widehat{\boldsymbol{\Sigma}}\boldsymbol{\Gamma}\mathbf{V})^\top \widehat{\boldsymbol{\Omega}}(\widehat{\mathbf{M}} - \widehat{\boldsymbol{\Sigma}}\boldsymbol{\Gamma}\mathbf{V})\} + \text{pen}(\boldsymbol{\Gamma}), \quad \text{subject to } \mathbf{V}\mathbf{V}^\top = \mathbf{I}_d,$$

where $\widehat{\mathbf{M}} \in \mathbb{R}^{p\times l}$ is a method-specific kernel matrix for some l such that $d \leq l \leq p$. Their approach relies on the introduction of a constrained parameter $\mathbf{V}$, which makes their procedure non-convex. Thus, the attainment of the global optimum is not guaranteed. Moreover, the parameter $\mathbf{V}$ is itself redundant since only the solution of $\boldsymbol{\Gamma}$ is the target, i.e., the estimator of basis matrix of $\mathcal{S}_{Y|\mathbf{X}}$. The redundancy in parameters increases the computation burden. Theoretically, though they established variable selection and subspace estimation consistency, their second-order SDR requires more assumptions than ours, such as the sub-Gaussianity of $\mathbf{X} \mid (\widetilde{Y} = h)$ and (approximate) sparsity on $\text{Cov}(\mathbf{X} \mid \widetilde{Y} = h)$.

Another related work is Yu et al. (2016), which proposed a model-free variable selection approach via constructing test statistics based on both first- and second-order SDR methods. For high-dimensional scenarios, their method enjoys the variable screening property: the selected set of variables contains the truly relevant variables with probability goes to one under the normality assumption of the predictors. However, our method only requires sub-Gaussianity for achieving the variable selection consistency.

S.5.2 Comparison with Jin and Luo (2025)

We elaborate on the differences between our work and Jin and Luo (2025) in terms of subspace estimation error rates and regularity assumptions. For ease of reference, we prefix all theorems, propositions, and assumptions from Jin and Luo (2025) with "JL-". In addition, we adopt the notation system used in our paper to facilitate a consistent comparison.

(I) Convergence rate. Jin and Luo (2025) showed in their Theorem JL-3 (ii) that $D(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}) = O_p\{(H^{1/2}s^{1/2} + s)(\log p/n)^{1/2}n^{\xi}\}$. For a fair comparison, we assume that H is bounded for both methods. Thus, the error bound in Jin and Luo (2025) becomes

$$D(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}) = O_p\{s(\log p/n)^{1/2}n^{\xi}\},$$

where the constant $\xi < \zeta/\min\{2, d\}$. According to Assumption (JL-A5), $(Hs+s^2)\log p/n = O(n^{-2\zeta})$, then $n^{-\xi} = O(\{\sqrt{s^2 \log p/n}\}^{1/\min\{2,d\}})$. In our Theorem 3, we establish the subspace estimation consistency rate that

$$D(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}) = O_p\{s(\log p/n)^{1/2}\sigma^{-1}\}.$$

Under our Assumption (A7), we further assume that $\sigma = O(\sqrt{s^2 \log p/n})$. Therefore, $\sigma = O(n^{-\xi})$ when $d = 1$ and $\sigma = o(n^{-\xi})$ when $d \geq 2$.

Consequently, when $d = 1$, our subspace estimation rate $D(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta})$ matches the rate in Jin and Luo (2025). When $d \geq 2$, our rate becomes slightly slower due to the weaker subspace signal σ. However, if a stronger signal condition is imposed such that $\sigma = O(n^{-\xi}) = O(\{\sqrt{s^2 \log p/n}\}^{1/\min\{2,d\}})$, then our convergence rate matches that of Jin and Luo (2025) for any value of d.

(II) Regularity assumptions.

(i) Distribution and moment assumption (A1 vs. JL-A4). Assumptions (A1) and (JL-A4) concern the distributional and moment properties of the predictors. Our Assumption (A1) imposes a slightly stronger vector-wise sub-Gaussianity of $\mathbf{X}$, whereas Assumption (JL-A4) assumes uniform sub-Gaussianity for individual elements X_i, $i = 1, \ldots, p$. Moreover, the finite moment assumption $\max_i \mathrm{E}(X_i^2) = O(1)$ in Assumption (JL-A4) is used to derive that $\|\boldsymbol{\Sigma}\|_{\max} = O(1)$, a condition weaker than our bounded eigenvalues assumption $\|\boldsymbol{\Sigma}\| = O(1)$ in Assumption (A1).

(ii) Consistency of moment matrix (A1 vs. JL-A7). Assumption (JL-A7) imposes a requirement on the error bound for the estimation of the moment matrix $(\boldsymbol{\Delta}_1, \ldots, \boldsymbol{\Delta}_H)$, which holds under the sub-Gaussianity of $X_i \mid (\widetilde{Y} = h)$ (cf. Proposition JL-1). In fact, the sub-Gaussianity of $X_i \mid (\widetilde{Y} = h)$ is needed to derive the error bound of $\|\widehat{\boldsymbol{\Sigma}}_h - \boldsymbol{\Sigma}_h\|_{\max}$, a key component of the error bound of $(\boldsymbol{\Delta}_1, \ldots, \boldsymbol{\Delta}_H)$ (cf. (B.8) in Jin and Luo (2025)).

In contrast, we derive the error bound of $\|\widehat{\mathbf{\Sigma}}_h - \mathbf{\Sigma}_h\|_{\max}$ without additionally assuming the sub-Gaussianity of $X_i \mid (\widetilde{Y} = h)$. Instead, we only require the bounded eigenvalues assumption for $\mathrm{Cov}\{\mathbf{X}I(\widetilde{Y} = h)\}$, as stated in Assumption (A1).

(iii) Boundedness assumption (A3 vs. JL-A6). Assumption (JL-A6) is required for deriving the estimation consistency of each individual column of the kernel matrix $(\mathbf{M}_1, \ldots, \mathbf{M}_H)$. Since each column $\mathbf{M}_{h,*j} = \mathbf{\Sigma}^{-1}\mathbf{\Delta}_{h,*j}$, by assuming the eigenvalues of $\mathbf{\Sigma}$ are bounded above in Assumption (A1), Assumption (JL-A6) is equivalent to the boundedness of $\max_{h,j} \|\mathbf{\Delta}_{h,*j}\|_2$. However, our Assumption (A3) only requires that $\max_h \|\mathbf{\Delta}_{h,\mathcal{A}*}\|_{\max}$ is bounded above, which is weaker than Assumption (JL-A6).

(iv) Variable-wise signal (A5 vs. JL-S1 & JL-S3). Assumptions (JL-S1) and (JL-S3) regulate the row-wise signals of the kernel matrix $(\mathbf{M}_1, \ldots, \mathbf{M}_H)$, which is crucial for establishing variable selection of their proposal (cf. Theorem JL-1(ii) & Theorem JL-3(iii)). In fact, Assumption (JL-S3) is a relaxed version of (JL-S1). In comparison, we impose Assumption (A5) to achieve variable selection in Step 1 (cf. Theorem 2). By rephrasing Assumption (JL-S3) in our notation, it states that

$$\|\mathcal{M}_{[:,i,:]}\|_{\max} \geq C\sqrt{(Hs + s^2)\log p/n}, \quad \forall i \in \mathcal{A},$$

for some constant $C > 0$, where $\|\mathcal{M}_{[:,i,:]}\|_{\max} = \max_{h,j} |\mathcal{M}_{[h,i,j]}|$, while our Assumption (A5) states that

$$\|\mathcal{M}_{[:,i,:]}\|_{2,\infty} \geq c\sqrt{s^2 \log p/n}, \quad \forall i \in \mathcal{A},$$

for some constant $c > 0$, where $\|\mathcal{M}_{[:,i,:]}\|_{2,\infty} = \max_j \|\mathcal{M}_{[:,i,j]}\|_2$. Since $\|\mathcal{M}_{[:,i,:]}\|_{2,\infty} \geq \|\mathcal{M}_{[:,i,:]}\|_{\max}$, our Assumption (A5) is weaker than Assumption (JL-S3).

(v) Subspace signal (A7 vs. JL-S2). Assumption (JL-S2) is imposed for column selection to ensure that the selected columns span the central subspace at the population level (cf. Theorem JL-2), which is a key result for establishing subspace estimation consistency in Theorem JL-3 (i) & (ii). Recall that in our paper, $\sigma > 0$ denotes the minimal non-zero singular values of the matrix $(\mathbf{M}_{1,\mathcal{A}\mathcal{A}}, \ldots, \mathbf{M}_{H,\mathcal{A}\mathcal{A}})$. Proposition JL-2 states that Assumption (JL-S2) is satisfied for SAVE- and DR-based methods if

$$\sigma = O(\{\sqrt{(Hs + s^2)\log p/n}\}^{1/(\min\{2,d\})}(Hs)^{1/2}).$$

In contrast, to guarantee the subspace estimation consistency in Step 2 of our proposal, we only require in Assumption (A7) that

$$\sigma = O(\sqrt{s^2 \log p/n}).$$

Thus, we impose a weaker assumption on σ than Jin and Luo (2025).

(vi) Symmetry condition (JL-A3) Assumption (JL-A3) is the symmetry condition, specifically required by the inverse third moment method. Since our paper only focus on

the SAVE and DR based methods, we do not impose this assumption.

(vii) Dimensionality assumption (JL-A5). By assuming the bounded H, Assumption (JL-A5) implies that $s^2 \log p = O(n^{1-2\zeta})$ for some constant $\zeta > 0$. In comparison, we assume that $s^2 \log p = o(n)$. Therefore, for the same values of s and n, our assumption permits a higher dimensionality for $\mathbf{X}$.

(viii) Irrepresentability condition (A2). Our Assumption (A2) is akin to the irrepresentability condition (Zhao and Yu, 2006), which is a standard assumption for establishing variable selection consistency. In contrast, Jin and Luo (2025) avoids this condition by employing the adaptive (group-) Lasso penalty in their penalized optimization problem.

(ix) Other regularity conditions (A3 & A6). To derive variable selection consistency in Theorem 2, we additionally require the boundedness of $\|(\mathbf{\Sigma}_{\mathcal{A}\mathcal{A}})^{-1}\|_\infty$ and $\|\mathbf{\Sigma}_{\mathcal{A}^c\mathcal{A}}\|_\infty$ in Assumption (A3). Furthermore, to guarantee subspace estimation consistency in Theorem 3, we impose Assumption (A6), which assumes the boundedness of $\max_h \|\mathbf{\Delta}_{h,\mathcal{A}\mathcal{A}}\|_1$. These assumptions are the cost we incur for handling the large $p \times (pH)$ matrix $(\mathbf{M}_1, \ldots, \mathbf{M}_H)$ in proving Theorem 2 and the matrix $(\mathbf{M}_{1,\mathcal{A}\mathcal{A}}, \ldots, \mathbf{M}_{H,\mathcal{A}\mathcal{A}})$ in proving Theorem 3.

In contrast, these regularity assumptions are not required in Jin and Luo (2025), since all variable selection results (cf. Theorem JL-1 (ii) & Theorem JL-3 (iii)) and subspace estimation error bounds (cf. Theorem JL-1 & Theorem JL-3 (i)(ii)) are established based on selected columns of the matrix $(\mathbf{M}_1, \ldots, \mathbf{M}_H)$, where the number of columns are bounded above.

(x) Assumption of the priors (A4). On page 4, line 1, Jin and Luo (2025) assumed $\Pr(\widetilde{Y} = h) = 1/H$, while we assume in Assumption (A4) that $\Pr(\widetilde{Y} = h)$ is bounded below to simplify the proof. If a diverging H is in demand, Assumption (A4) can be replaced with $\Pr(\widetilde{Y} = h) = O(1/H)$. By doing this, H will explicitly appear in all rates in the assumptions and error bounds.

S.5.3 Comparison with Zeng et al. (2024)

In the following, we prefix all theorems, propositions, and assumptions from Zeng et al. (2024) with "ZMZ-".

(i) Distribution assumption (A1 vs. ZMZ-A3). The sub-Gaussianity of $\mathbf{X}$ and the bounded eigenvalue assumption of $\mathbf{\Sigma}$ in Assumption (A1) are also required by Zeng et al. (2024) (cf. Assumption (ZMZ-A3)). Nevertheless, the bounded eigenvalues assumption of $\mathrm{Cov}(\mathbf{X}I(Y = h))$ in Assumption (A1) is uniquely required by sparse SAVE/DR, since it is used for establishing the error bound of $\|\widehat{\mathbf{\Sigma}}_h - \mathbf{\Sigma}_h\|_{\max}$, which is not needed in sparse SIR method.

(ii) Irrepresentability condition (A2 vs. ZMZ-A10). The irrepresentability Assumption (A2) is also required by Zeng et al. (2024) (cf. Assumption (ZMZ-A10)), while Assumption (A2) is stronger than that in Zeng et al. (2024) for simplifying the proof. Specifically, we assume that $\max_{j\in A^c} \|\mathbf{\Sigma}_{j\mathcal{A}}(\mathbf{\Sigma}_{\mathcal{A}\mathcal{A}})^{-1}\|_1 = 1-\alpha$ for some constant α, while Assumption (ZMZ-A10) assumes that such that $\max_{j\in A^c} \|\mathbf{\Sigma}_{j\mathcal{A}}(\mathbf{\Sigma}_{\mathcal{A}\mathcal{A}})^{-1}\mathbf{R}\|_2 \leq 1-\alpha$ for some constant $0 < \alpha \leq 1$, where $\mathbf{R}$ is a sub-gradient of the $L_{2,1}$-norm penalty. Since $\|\mathbf{\Sigma}_{j\mathcal{A}}(\mathbf{\Sigma}_{\mathcal{A}\mathcal{A}})^{-1}\mathbf{R}\|_2 \leq \|\mathbf{\Sigma}_{j\mathcal{A}}(\mathbf{\Sigma}_{\mathcal{A}\mathcal{A}})^{-1}\|_1$, our Assumption (A2) is stronger.

(iii) Variable-wise signal (A5 vs. ZMZ-A2). Assumption (A5) is similar to Assumption (ZMZ-A2), both of which regulate the variable-wise signal but differ in the parameter matrix and the signal rate, due to different SDR framework the two works are based on.

(iv) Subspace signal (A7 vs. ZMZ-A1). Assumption (A7) is similar to Assumption (ZMZ-A1). These two assumptions concern the subspace signals for establishing the subspace estimation consistency. They only differ in the parameterization of signal and signal rate.

(vi) Boundedness assumption (A3 & A6 vs. ZMZ-A9). The boundedness of the quantities $\|(\mathbf{\Sigma}_{\mathcal{A}\mathcal{A}})^{-1}\|_\infty$ and $\|\mathbf{\Sigma}_{\mathcal{A}^c\mathcal{A}}\|_\infty$ in Assumption (A3) are similar to Assumption (ZMZ-A9), which states that $\|\mathbf{\Sigma}_{\mathcal{A}^c\mathcal{A}}(\mathbf{\Sigma}_{\mathcal{A}\mathcal{A}})^{-1}\|_\infty < \infty$. Since $\|\mathbf{\Sigma}_{\mathcal{A}^c\mathcal{A}}(\mathbf{\Sigma}_{\mathcal{A}\mathcal{A}})^{-1}\|_\infty \leq \|\mathbf{\Sigma}_{\mathcal{A}^c\mathcal{A}}\|_\infty \|(\mathbf{\Sigma}_{\mathcal{A}\mathcal{A}})^{-1}\|_\infty$, our Assumption (A3) is stronger. However, both assumptions restrict that the association between $\mathbf{X}_{\mathcal{A}^c}$ and $\mathbf{X}_{\mathcal{A}}$ is not arbitrarily large and are used for establishing variable selection consistency.

The boundedness of $\max_h \|\mathbf{\Delta}_{h,\mathcal{A}*}\|_{\max}$ in Assumption (A3) and $\max_h \|\mathbf{\Delta}_{h,\mathcal{A}\mathcal{A}}\|_1$ in Assumption (A6) are uniquely required by sparse SAVE/DR since the parameters $\mathbf{\Delta}_h$ only appears in the SAVE and DR methods.

S.6 Integrated formulation

To provide a unified picture of our proposed two-step procedure, we integrate it into the formulation as follows,

$$\min_{\mathcal{G}\in\mathbb{R}^{H\times p\times p}} \sum_{h=1}^{H} \left\{ \frac{1}{2}\mathrm{tr}\left(\mathcal{G}_{[h,:,:]}^\top \widehat{\mathbf{\Sigma}}\mathcal{G}_{[h,:,:]}\right) - \mathrm{tr}(\mathcal{G}_{[h,:,:]}^\top \widehat{\mathbf{\Delta}}_h) \right\} + \mathrm{pen}(\mathcal{G}).$$

Obviously, in Step 1, the penalty function $\mathrm{pen}(\mathcal{G})$ is taken as the $\mathrm{pen}(\mathcal{G}) = \mathrm{pen}_{\lambda_1}(\mathcal{G}) = \lambda_1 \sum_{i=1}^p \sum_{j=1}^p (\sum_{h=1}^H \mathcal{G}_{[h,i,j]}^2)^{1/2}$. Given the estimated active set $\widehat{\mathcal{A}}$ from Step 1, the penalty function $\mathrm{pen}(\mathcal{G})$ in Step 2 is $\mathrm{pen}(\mathcal{G}) = \sum_{h=1}^H I_{\widehat{\mathcal{A}}\widehat{\mathcal{A}}}(\mathcal{G}_{[h,:,:]}) + \lambda_2 \|(\mathcal{G}_{[1,:,:]}, \ldots, \mathcal{G}_{[H,:,:]})\|_*$, where $I_{\widehat{\mathcal{A}}\widehat{\mathcal{A}}} : \mathbb{R}^{p\times p} \to \{0, \infty\}$ is an indicator function, defined as $I_{\widehat{\mathcal{A}}\widehat{\mathcal{A}}}(\mathcal{G}_{[h,:,:]}) = 0$ if $\mathcal{G}_{[h,i,j]} = 0$ for all $(i,j) \notin \widehat{\mathcal{A}} \times \widehat{\mathcal{A}}$, and $I_{\widehat{\mathcal{A}}\widehat{\mathcal{A}}}(\mathcal{G}_{[h,i,j]}) = \infty$ if $\mathcal{G}_{[h,i,j]} \neq 0$ for some $(i,j) \notin \widehat{\mathcal{A}} \times \widehat{\mathcal{A}}$. Therefore,

the formulation in each step is essentially a quadratic objective function regularized by a task-specific convex penalty function.

S.7 Extensions to third-order methods

Our sparse SDR framework can be also extended to the third-order method. For $h = 1, \ldots, H$, define the tensor $\mathcal{T}_h = \mathrm{E}\{(\mathbf{X} - \boldsymbol{\mu}_h) \circ (\mathbf{X} - \boldsymbol{\mu}_h) \circ (\mathbf{X} - \boldsymbol{\mu}_h) \mid \widetilde{Y} = h\}$, $h = 1, \ldots, H$, and define the h-th tensor block $\mathcal{D}_h \in \mathbb{R}^{p \times p \times p}$ as $\mathcal{D}_h = \mathcal{T}_h \times_1 \boldsymbol{\Sigma}^{-1}$ such that

$$\mathcal{D}_h = \boldsymbol{\Sigma}^{-1} \mathrm{E}\left[(\mathbf{X} - \boldsymbol{\mu}_h) \circ (\mathbf{X} - \boldsymbol{\mu}_h) \circ (\mathbf{X} - \boldsymbol{\mu}_h) \mid \widetilde{Y} = h\right],$$

where $\boldsymbol{\mu}_h = \mathrm{E}(\mathbf{X} \mid \widetilde{Y} = h) \in \mathbb{R}^p$, $\times_1$ denotes the 1-mode product of a tensor and a matrix, and " $\circ$ " denotes the outer-product. We further define

$$\mathbf{D}_h = (\mathcal{D}_{h,[:,:,1]}, \ldots, \mathcal{D}_{h,[:,:,p]}) \in \mathbb{R}^{p \times p^2},$$

which concatenates all slices of $\mathcal{D}_h$ at the third mode, where each slice of $\mathcal{D}_h$ is $\mathcal{D}_{h,[:,:,k]} = \boldsymbol{\Sigma}^{-1}\mathrm{E}[(\mathbf{X}-\boldsymbol{\mu}_h)(\mathbf{X}-\boldsymbol{\mu}_h)^\top(X_k-\mu_{h,k}) \mid \widetilde{Y} = h] \in \mathbb{R}^{p \times p}$, $k = 1, \ldots, p$. Then the third-moment subspace is defined as

$$\mathcal{S}_{\mathrm{TM}} = \mathrm{span}(\mathbf{D}_1, \ldots, \mathbf{D}_H).$$

Under the linearity, constant variance and symmetry conditions, it has been shown that $\mathcal{S}_{\mathrm{TM}} = \mathcal{S}_{Y|\mathbf{X}}$ (Yin and Cook, 2003). Similar to subspaces $\mathcal{S}_{\mathrm{SAVE}}$ and $\mathcal{S}_{\mathrm{DR}}$, the sparsity structure of $\mathcal{S}_{\mathrm{TM}}$ is equivalent to the row sparsity of the matrix $(\mathbf{D}_1, \ldots, \mathbf{D}_H)$.

We reformulate the basis $(\mathbf{D}_1, \ldots, \mathbf{D}_H)$ as a high-order tensor $\mathcal{M} \in \mathbb{R}^{H \times p \times p \times p}$ by stacking H blocks $\mathcal{D}_h$, $h = 1, \ldots, H$, along the first mode, such that $\mathcal{M}_{[h,:,:,:]} = \mathcal{D}_h$. Therefore, the second mode of $\mathcal{M}$ corresponds to the elements of predictors. Specifically, the i-th slice $\mathcal{M}_{[:,i,:,:]} \in \mathbb{R}^{H \times p \times p}$ stacks all vectors $\mathcal{D}_{h,[i,:,k]}$, for $h = 1, \ldots, H$, and $k = 1, \ldots, p$, together, corresponding to the i-th row of the matrix $(\mathbf{D}_1, \ldots, \mathbf{D}_H)$. Consequently, the important variables can be identified by seeking non-zero slices $\mathcal{M}_{[:,i,:,:]}$. We provide a toy example with $H = 2$, $p = 5$, and $\mathcal{A} = \{1, 2, 3\}$ to illustrate the correspondence between the important variables and the sparsity structure of tensor parameter $\mathcal{M}$ in Figure S.1. The colored slices in $\mathcal{D}_1$ and $\mathcal{D}_2$ indicate the positions of important variables.

Since $\mathcal{M}_{[h,:,:,k]} = \boldsymbol{\Sigma}^{-1}\mathcal{T}_{h,[:,:,k]}$ for each h and k, the tensor parameter $\mathcal{M}$ can be recovered from the following quadratic optimization problem,

$$\mathcal{M} = \underset{\mathcal{G} \in \mathbb{R}^{H \times p \times p \times p}}{\mathrm{argmin}} \sum_{h=1}^{H} \sum_{k=1}^{p} \left\{ \frac{1}{2} \mathrm{tr}(\mathcal{G}_{[h,:,:,k]}^\top \boldsymbol{\Sigma} \mathcal{G}_{[h,:,:,k]}) - \mathrm{tr}(\mathcal{G}_{[h,:,:,k]}^\top \mathcal{T}_{h,[:,:,k]}) \right\}.$$

To encourage the sparsity structure in $\mathcal{M}$, we propose to adopt the group-Lasso type

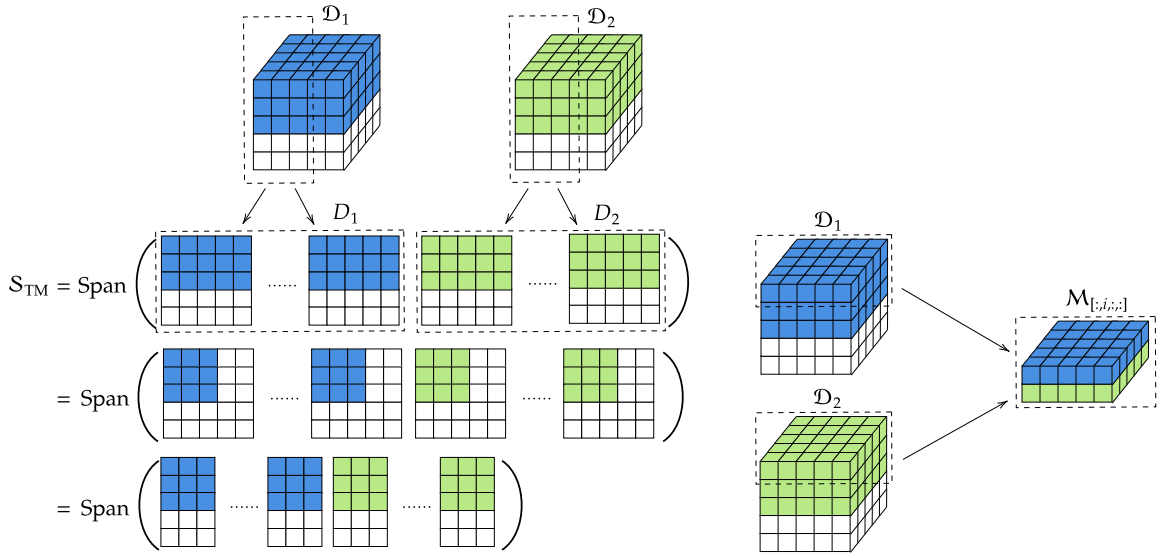

Figure S.1: (Left) The illustration of the parsimonious reparameterization of the third-moment subspace, with $H = 2$, $p = 5$, and $\mathcal{A} = \{1, 2, 3\}$; (Right) The illustration of the tensor $\mathcal{M} \in \mathbb{R}^{H\times p\times p\times p}$, and its sub-tensor $\mathcal{M}_{[:,i,:,:]}$, corresponding to predictors.

penalty on each slice $\mathcal{M}_{[:,i,:,:]}$, which corresponds to the i-th predictor X_i, i.e.,

$$\text{pen}(\mathcal{M}) = \lambda_1 \sum_{i=1}^{p} \text{pen}(\mathcal{M}_{[:,i,:,:]}) = \lambda_1 \sum_{i=1}^{p} \left\{ \sum_{j=1}^{p}\sum_{k=1}^{p} \left(\sum_{h=1}^{H} \mathcal{M}^2_{[h,i,j,k]} \right)^{1/2} \right\}.$$

This penalization is similar to the group-Lasso type penalty in our sparse SAVE/DR method. Then, we solve the following penalized convex optimization problem in the first step,

$$\widehat{\mathcal{M}} = \underset{\mathcal{G}\in\mathbb{R}^{H\times p\times p\times p}}{\text{argmin}} \sum_{h=1}^{H}\sum_{k=1}^{p} \left\{ \frac{1}{2}\text{tr}(\mathcal{G}^\top_{[h,:,:,k]}\boldsymbol{\Sigma}\mathcal{G}_{[h,:,:,k]}) - \text{tr}(\mathcal{G}^\top_{[h,:,:,k]}\mathcal{T}_{h,[:,:,k]}) \right\}$$
$$+ \lambda_1 \sum_{i=1}^{p} \left\{ \sum_{j=1}^{p}\sum_{k=1}^{p} \left(\sum_{h=1}^{H} \mathcal{G}^2_{[h,i,j,k]} \right)^{1/2} \right\}.$$

The active set $\mathcal{A}$ is defined by $\widehat{\mathcal{A}} = \{i \mid \|\widehat{\mathcal{M}}_{[:,i,:,:]}\|_{2,\infty} > 0\}$, where $\|\widehat{\mathcal{M}}_{[:,i,:,:]}\|_{2,\infty} = \max_{j,k} \|\widehat{\mathcal{M}}_{[:,i,j,k]}\|_2$.

Similar to Theorem 1, once we are given the oracle active set $\mathcal{A}$, the third-moment subspace can be reformulated as

$$\mathcal{S}_{\text{TM}} = \text{span}\left\{ \begin{pmatrix} \mathbf{D}_1^{(\mathcal{A})}, \ldots, \mathbf{D}_H^{(\mathcal{A})} \\ \mathbf{0} \end{pmatrix} \right\},$$

where $\mathbf{D}_h^{(\mathcal{A})} = (\mathcal{D}_{h,[\mathcal{A},\mathcal{A},1]}, \ldots, \mathcal{D}_{h,[\mathcal{A},\mathcal{A},p]})$, for $h = 1, \ldots, H$. It can be shown that $\mathcal{D}_{h,[\mathcal{A},\mathcal{A},k]} = \boldsymbol{\Sigma}^{-1}_{\mathcal{A}\mathcal{A}}\mathcal{T}_{h,[\mathcal{A},\mathcal{A},k]}$. The non-zero part $(\mathbf{D}_1^{(\mathcal{A})}, \ldots, \mathbf{D}_H^{(\mathcal{A})})$ can be further recovered by the following

optimization problem,

$$(\mathbf{D}_1^{(\mathcal{A})}, \ldots, \mathbf{D}_H^{(\mathcal{A})}) = \underset{\mathbf{B}_{11},\ldots,\mathbf{B}_{Hp}\in\mathbb{R}^{s\times s}}{\operatorname{argmin}} \sum_{h=1}^{H}\sum_{k=1}^{p}\left\{\frac{1}{2}\operatorname{tr}(\mathbf{B}_{hk}^\top \boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}}\mathbf{B}_{hk}) - \operatorname{tr}(\mathbf{B}_{hk}^\top \mathcal{T}_{h,[\mathcal{A},\mathcal{A},k]})\right\}.$$

In the second step, once we obtain the estimated active set $\widehat{\mathcal{A}}$ from the first step, we estimate $(\mathbf{D}_1^{(\widehat{\mathcal{A}})}, \ldots, \mathbf{D}_H^{(\widehat{\mathcal{A}})})$ by solving the following nuclear-norm penalized problem,

$$\underset{\mathbf{B}_{11},\ldots,\mathbf{B}_{Hp}\in\mathbb{R}^{\widehat{s}\times\widehat{s}}}{\operatorname{argmin}} \sum_{h=1}^{H}\sum_{k=1}^{p}\left\{\frac{1}{2}\operatorname{tr}(\mathbf{B}_{hk}^\top \boldsymbol{\Sigma}_{\widehat{\mathcal{A}}\widehat{\mathcal{A}}}\mathbf{B}_{hk}) - \operatorname{tr}(\mathbf{B}_{hk}^\top \mathcal{T}_{h,[\widehat{\mathcal{A}},\widehat{\mathcal{A}},k]})\right\} + \lambda_2\|(\mathbf{B}_{11},\ldots,\mathbf{B}_{1p},\ldots,\mathbf{B}_{H1},\ldots,\mathbf{B}_{Hp})\|_*.$$

Let $\widehat{\mathbf{B}} = (\widehat{\mathbf{B}}_{11},\ldots,\widehat{\mathbf{B}}_{Hp})$ denote the solution, where $\widehat{\mathbf{B}}_{hk} \in \mathbb{R}^{\widehat{s}\times\widehat{s}}$, $h = 1,\ldots,H$, and $k = 1,\ldots,p$. Let $\widehat{\boldsymbol{\eta}}_j \in \mathbb{R}^{\widehat{s}}$ denote the j-th leading left singular vector of $\widehat{\mathbf{B}}$. We then expand each $\widehat{\boldsymbol{\eta}}_j$ to the p-dimensional vector $\widehat{\boldsymbol{\beta}}_j$, such that $\widehat{\boldsymbol{\beta}}_{j,\widehat{\mathcal{A}}} = \widehat{\boldsymbol{\eta}}_j$ and $\widehat{\boldsymbol{\beta}}_{j,\widehat{\mathcal{A}}^c} = \mathbf{0}$, $j = 1,\ldots,\widehat{d}$, where $\widehat{d}$ is the estimated structural dimension. The estimated basis matrix of $\mathcal{S}_{\mathrm{TM}}$ is defined as $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_1,\ldots,\widehat{\boldsymbol{\beta}}_{\widehat{d}}) \in \mathbb{R}^{p\times\widehat{d}}$.

The variable selection consistency of the third-order sparse SDR can be similarly derived utilizing the proof techniques similar in our work, subject to appropriate assumptions on the parameter space and the tail distribution of $\mathbf{X}$. Specifically, the definitions of $\boldsymbol{\Delta}_h$ and $\mathcal{M}$ in Assumptions (A3) and (A5) should be updated to those in the third-order SDR. Moreover, under the same high-dimensional setting that $s^2\log p = o(n)$ as in our paper, it is essential to verify that with high probability, we have

$$\max_{h=1,\ldots,H}\max_{j=1,\ldots,p}\|\widehat{\boldsymbol{\Delta}}_{h,j} - \boldsymbol{\Delta}_{h,j}\|_{\max} \lesssim \sqrt{\log p/n}. \tag{S.5}$$

Since $\widehat{\boldsymbol{\Delta}}_{h,j}$ involves the product of three random variables, which has heavier tails than the product terms we bound in sparse SAVE/DR, the stronger distribution assumption on $\mathbf{X}$ is needed for establishing (S.5). For example, (S.5) can be derived under the assumption that $\mathbf{X}$ is a bounded random vector.

S.8 Algorithms and implementation details

S.8.1 Full description

We solve two penalized optimization problems in sparse SAVE: the group-Lasso-type convex optimization (5) and the nuclear-norm penalized convex optimization (6). This section provides the complete description of algorithms for solving these two problems, along with the implementation details of the algorithms. For two matrices $\mathbf{A} = (a_{ij}) \in \mathbb{R}^{p\times q}$ and $\mathbf{B} \in \mathbb{R}^{s\times t}$, the Kronecker-product of $\mathbf{A}$ and $\mathbf{B}$, denoted by $\mathbf{A}\otimes\mathbf{B} \in \mathbb{R}^{(ps)\times(qt)}$, is the block matrix composed of $p\times q$ blocks, with the (i,j)-th block being $a_{ij}\mathbf{B}$, for $i = 1,\ldots,p$, $j = 1,\ldots,q$.

We first rewrite (5) in the vectorized form and develop a blockwise coordinate descent algorithm similar to an existing algorithm (Pan and Mai, 2020) for sparse estimation problems in multiple differential networks and quadratic discriminant analysis. Let $\widehat{\boldsymbol{\delta}}_h = \text{vec}(\widehat{\boldsymbol{\Delta}}_h)$, $h = 1, \ldots, H$, where $\text{vec}(\widehat{\boldsymbol{\Delta}}_h)$ denotes the vectorization of $\widehat{\boldsymbol{\Delta}}_h$ by concatenating its columns, and denote $\widehat{\boldsymbol{\delta}} = (\widehat{\boldsymbol{\delta}}_1, \ldots, \widehat{\boldsymbol{\delta}}_H) \in \mathbb{R}^{p^2 \times H}$. Also, we vectorize $\mathcal{G}_{[h,:,:]}$ into $\mathbf{g}_h = \text{vec}(\mathcal{G}_{[h,:,:]}) \in \mathbb{R}^{p^2}$ and denote $\mathbf{g} = (\mathbf{g}_1, \ldots, \mathbf{g}_H) \in \mathbb{R}^{p^2 \times H}$. The vectorized form of (5) is

$$\underset{\mathbf{g} \in \mathbb{R}^{p^2 \times H}}{\text{argmin}} \frac{1}{2} \text{tr}\{\mathbf{g}^\top (\mathbf{I}_p \otimes \widehat{\boldsymbol{\Sigma}}) \mathbf{g}\} - \text{tr}(\mathbf{g}^\top \widehat{\boldsymbol{\delta}}) + \lambda_1 \sum_{j=1}^{p^2} \|\mathbf{g}_{j*}\|_2. \tag{S.6}$$

We introduce the new index k which has a one-to-one corresponding to $(i, j) \in \{1, \ldots, p\} \times \{1, \ldots, p\}$ such that $k = i + (j-1)p$. The following lemma provides the update of each row $\mathbf{g}_{k*}$ while fixing all other rows of $\mathbf{g}$.

Lemma S.2. *For index pair $(i', j') \in \{1, \ldots, p\} \times \{1, \ldots, p\}$, denote $k' = i' + (j'-1)p$. For $k \in \{1, \ldots, p^2\}$, given $\mathbf{g}_{k'*}$ for all $k' \neq k$, the solution to* (S.6) *is the same as*

$$\underset{\mathbf{g}_{k*} \in \mathbb{R}^H}{\text{argmin}} \frac{1}{2} \|\mathbf{g}_{k*} - \widetilde{\mathbf{g}}_{k*}\|_2^2 + \frac{\lambda_1}{\widehat{\Sigma}_{ii}} \|\mathbf{g}_{k*}\|_2, \quad \textit{where} \quad \widetilde{\mathbf{g}}_{k*} = (\widehat{\boldsymbol{\delta}}_{k*} - \sum_{k' \neq k} \widehat{\Sigma}_{i'i} \mathbf{g}_{k'*}) / \widehat{\Sigma}_{ii}. \tag{S.7}$$

The solution to (S.7) *is $\widehat{\mathbf{g}}_{k*} = \widetilde{\mathbf{g}}_{k*} \{1 - \lambda_1 / (\widehat{\Sigma}_{ii} \|\widetilde{\mathbf{g}}_{k*}\|_2)\}_+$, where the operator $(x)_+ = x$ if $x > 0$ and $(x)_+ = 0$ otherwise.*

By Lemma S.2, $\mathbf{g}_{k*}$ is updated by solving a group-Lasso problem. Since problem (S.6) is convex, according to Pan and Mai (2020), the algorithm converges to the global solution while the computation cost is relatively low (scaling well with p).

We solve (6) by the accelerated proximal gradient descent algorithm. Denote $\widehat{\boldsymbol{\Gamma}} = (\widehat{\boldsymbol{\Delta}}_{1,\widehat{\mathcal{A}}\widehat{\mathcal{A}}}, \ldots, \widehat{\boldsymbol{\Delta}}_{H,\widehat{\mathcal{A}}\widehat{\mathcal{A}}}) \in \mathbb{R}^{\widehat{s} \times (H\widehat{s})}$ and $\mathbf{B} = (\mathbf{B}_1, \ldots, \mathbf{B}_H) \in \mathbb{R}^{\widehat{s} \times (H\widehat{s})}$, then (6) is rewritten as

$$\underset{\mathbf{B} \in \mathbb{R}^{\widehat{s} \times (H\widehat{s})}}{\text{argmin}} \frac{1}{2} \text{tr}(\mathbf{B}^\top \widehat{\boldsymbol{\Sigma}}_{\widehat{\mathcal{A}}\widehat{\mathcal{A}}} \mathbf{B}) - \text{tr}(\mathbf{B}^\top \widehat{\boldsymbol{\Gamma}}) + \lambda_2 \|\mathbf{B}\|_*.$$

In the initialization step, we set $\mathbf{B}^{(0)} = \mathbf{0}$ and define the auxiliary matrix $\mathbf{C}^{(0)} = \mathbf{0}$. At the t-th iteration, for $t = 0, 1, \ldots$, we update $\mathbf{B}^{(t+1)}$ as $\mathbf{B}^{(t+1)} = S_{s_t \lambda_2}\{\mathbf{C}^{(t)} - s_t(\widehat{\boldsymbol{\Sigma}}_{\widehat{\mathcal{A}}\widehat{\mathcal{A}}} \mathbf{C}^{(t)} - \widehat{\boldsymbol{\Gamma}})\}$. Here, $s_t > 0$ is the stepsize, the operator $S_{s_t \lambda_2}(\mathbf{A}) = \sum_{j=1}^p (\sigma_j - s_t \lambda_2)_+ \mathbf{u}_j \mathbf{v}_j^\top$ for some matrix $\mathbf{A} \in \mathbb{R}^{p \times p}$, where $\sum_{j=1}^p \sigma_j \mathbf{u}_j \mathbf{v}_j^\top$ denotes the singular value decomposition of $\mathbf{A}$. Then, we take the acceleration step and update the auxiliary matrix $\mathbf{C}^{(t+1)}$ by $\mathbf{C}^{(t+1)} = \mathbf{B}^{(t+1)} + \{t/(t+3)\}(\mathbf{B}^{(t+1)} - \mathbf{B}^{(t)})$. The iteration stops when $\|\mathbf{C}^{(t+1)} - \mathbf{C}^{(t)}\|_{\max} \leq \delta$ is met, where $\delta > 0$ is some user-defined small value. We set $\delta = 10^{-3}$ in our numerical studies. We take a fixed stepsize $s = s_t = 1/L$ in this algorithm, where $L = \|\widehat{\boldsymbol{\Sigma}}_{\widehat{\mathcal{A}}\widehat{\mathcal{A}}}\|$, which guarantees a linear rate of algorithmic convergence (Nesterov, 2013). The blockwise coordinate descent algorithm and accelerated proximal gradient descent algorithm are summarized in Algorithm 1 and Algorithm 2, respectively.

Algorithm 1 Blockwise coordinate descent

1. Initialization: $\mathbf{g}^{(0)} = \mathbf{0}$.
2. For $t = 1, 2, \ldots,$ repeat the following until the convergence of $\mathbf{g}^{(t)}$: For each $k = 1, \ldots, p^2$,
   (i) Update $\widetilde{\mathbf{g}}_{k*}^{(t-1)}$: $\widetilde{\mathbf{g}}_{k*}^{(t-1)} = (\widehat{\Sigma}_{ii})^{-1}\{\widehat{\boldsymbol{\delta}}_{k*} - (\sum_{k'<k} \widehat{\Sigma}_{i'i}\mathbf{g}_{k'*}^{(t)} + \sum_{k'>k} \widehat{\Sigma}_{i'i}\mathbf{g}_{k'*}^{(t-1)})\}$.
   (ii) Update $\mathbf{g}_{k*}^{(t)}$: $\mathbf{g}_{k*}^{(t)} = \widetilde{\mathbf{g}}_{k*}^{(t-1)} \left(1 - \frac{\lambda_1}{\widehat{\Sigma}_{ii}\|\widetilde{\mathbf{g}}_{k*}^{(t-1)}\|_2}\right)_+$.
3. Output $\widehat{\mathcal{M}}$: At termination, let $\mathbf{g}^{(t)}$ denote the converged solution. We stack each column of $\mathbf{g}^{(t)}$ back to the p-by-p matrix $\widehat{\mathbf{M}}_h$, and stack H blocks $\widehat{\mathbf{M}}_h$ along the first mode, resulting $\widehat{\mathcal{M}}$.

Algorithm 2 Accelerated proximal gradient descent

1. Initialization: $\mathbf{B}^{(0)} = \mathbf{C}^{(0)} = \mathbf{0}$, $\delta > 0$, $s = s_t = 1/L$, where $L = \|\widehat{\boldsymbol{\Sigma}}_{\widehat{\mathcal{A}}\widehat{\mathcal{A}}}\|$.
2. For $t = 0, 1, \ldots,$ repeat the following steps until the criteria $\|\boldsymbol{\Theta}^{(t+1)} - \boldsymbol{\Theta}^{(t)}\|_{\max} \leq \delta$ is met:
   (i) $\mathbf{B}^{(t+1)} = S_{s_t\lambda_2}\left\{\mathbf{C}^{(t)} - s_t(\widehat{\boldsymbol{\Sigma}}_{\widehat{\mathcal{A}}\widehat{\mathcal{A}}}\mathbf{C}^{(t)} - \widehat{\boldsymbol{\Gamma}})\right\}$.
   (ii) $\mathbf{C}^{(t+1)} = \mathbf{B}^{(t+1)} + \frac{t}{t+3}(\mathbf{B}^{(t+1)} - \mathbf{B}^{(t)})$.
3. Output $\widehat{\mathbf{B}} = \mathbf{B}^{(t)}$ at convergence.

S.8.2 Tuning parameter selection

Our algorithms involve two tuning parameters: λ_1 for encouraging sparsity and facilitating variable selection in the first step, and λ_2 employed in the second step to encourage a low-rank structure. While the optimal tuning parameters can be determined through a grid search, we propose a sequential selection procedure to reduce the computation time. We first fix λ_2 at a predetermined value $\widetilde{\lambda}_2 > 0$ and identify the optimal value for λ_1, denoted as λ_1^*. Then, with λ_1 fixed at λ_1^*, we tune λ_2 to determine the optimal value, denoted as λ_2^*. Consequently, λ_1^* and λ_2^* serve as the optimal tuning parameters. The criterion for tuning parameter selection in our methodology is based on distance covariance (Székely et al., 2007), which has been shown effective in parameter estimation in SDR problems (Sheng and Yin, 2016; Zeng et al., 2024). Let $\mathbf{Y}^{\text{train}} \in \mathbb{R}^{n_1}$ and $\mathbf{X}^{\text{train}} \in \mathbb{R}^{n_1 \times p}$ be the training set, while $\mathbf{Y}^{\text{test}} \in \mathbb{R}^{n_2}$ and $\mathbf{X}^{\text{test}} \in \mathbb{R}^{n_2 \times p}$ denote the testing set. For a given pair $\lambda =$

(λ_1, λ_2), let $\widehat{\boldsymbol{\beta}}_\lambda$ be the estimated basis of $\mathcal{S}_{\mathrm{SAVE}}$ computed from the training set, and $\widehat{\boldsymbol{\Sigma}}^{\mathrm{test}}$ be the sample covariance matrix obtained from the testing set. Construct the matrix $\boldsymbol{\eta}_\lambda = \widehat{\boldsymbol{\beta}}_\lambda(\widehat{\boldsymbol{\beta}}_\lambda^\top \widehat{\boldsymbol{\Sigma}}^{\mathrm{test}} \widehat{\boldsymbol{\beta}}_\lambda)^{-1/2}$ such that $\mathrm{span}(\boldsymbol{\eta}_\lambda) = \mathrm{span}(\widehat{\boldsymbol{\beta}}_\lambda)$ and $\boldsymbol{\eta}_\lambda^\top \boldsymbol{\Sigma}^{\mathrm{test}} \boldsymbol{\eta}_\lambda = \mathbf{I}$. The tuning parameter λ is chosen to maximize the distance covariance between $\mathbf{Y}^{\mathrm{test}}$ and $\mathbf{X}^{\mathrm{test}}\boldsymbol{\eta}_\lambda$.

We also compare our sequential tuning strategy with the grid-search tuning strategy. We record the overall execution time for two strategies, which includes the time required for the tuning parameter selection process and the final model fitting. We also compare the two strategies from the perspective of subspace estimation and variable selection accuracy. The comparison results under Models (M1)-(M5) with $n = 300$ and $p = 700$ are presented in Table S.17. The results imply that the sequential tuning strategy yields almost identical results to the grid-search tuning strategy in terms of both subspace estimation and variable selection accuracy, but is computationally more efficient. We remark that though the sequential tuning strategy is faster than the grid-search tuning strategy, but the improvement in computation time is not orders of magnitude. The reason is that our algorithm is a two-step procedure. The computationally expensive first step for variable selection is implemented the same number of times in both strategies. Only the second step for subspace estimation, which is essentially a low-dimensional problem and is much cheaper, is implemented much more times in the grid-search tuning strategy.

Moreover, in our sequential tuning strategy, a predetermined value $\widetilde{\lambda}_2 > 0$ is needed for identifying the optimal value for λ_1. We examine the performance of our proposed methods under a wide range of choices of $\widetilde{\lambda}_2$. Specifically, we consider $\widetilde{\lambda}_2$ over $\{0.01, 0.1, 0.5, 1, 5\}$ under all forward Models (M1)–(M5) and discriminant Models (Q1)–(Q4), where we fix $n = 300$ and $p = 700$. The subspace estimation errors are reported in Table S.19. It is implied by Table S.19 that when $\widetilde{\lambda}_2 = 0.5, 1$, sparse SAVE/DR give the best performance in almost all models. Even when $\widetilde{\lambda}_2 = 0.1$, the estimation error does not greatly deviate from the best results. We additionally record the optimal value λ_2^* selected by our algorithm given different initialization values $\widetilde{\lambda}_2$ in Table S.20. The results suggest that λ_2^* is located in $[0.1, 1.5]$ for both sparse SAVE and sparse DR in most of settings, once the initialization value $\widetilde{\lambda}_2$ is properly selected. Therefore, in practice, we suggest to try with $\widetilde{\lambda}_2 \in [0.1, 1.5]$ for our proposal.

S.9 Effect of nuclear-norm regularization

Recall that we incorporate the nuclear-norm penalty in the optimization problem (6) to encourage the low-rankness of $\widehat{\mathbf{B}}$. Therefore, we begin with comparing the dimension determination accuracy between our proposal and oracle methods. For the sake of completeness, we also include SEAS-SIR and Lasso-SIR in the comparison. For oracle-SAVE, after we compute $\widehat{\mathbf{B}} = (\widehat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A}})^{-1}(\widehat{\boldsymbol{\Delta}}_{1,\mathcal{A}\mathcal{A}}, \ldots, \widehat{\boldsymbol{\Delta}}_{H,\mathcal{A}\mathcal{A}})$, the estimation of d is similar to the procedure

in our proposal. It is given by $\widehat{d} = |\{j \mid \widehat{\sigma}_j \geq \delta\}| \wedge d_{\max}$, where $\widehat{\sigma}_j$ is the j-th largest singular value of $\widehat{\mathbf{B}}$, and δ and $d_{\max}$ hold the same meanings as those in our selection procedure. The selection of d by oracle-DR is conducted in a similar fashion.

The dimension determination results are reported in Tables S.5 and S.6, where we vary p and n, respectively. It can be seen that our proposal provides the estimated structural dimension closest to the true value in almost all settings. TC-SAVE overly selects the structural dimension in Models (Q1) and (Q3). Sparse SIR methods completely break down in all models except for Models (M2) and (Q3), limited by its first-order nature. The oracle methods, across all settings, show the lack of ability to identify the correct low-rank structure. Therefore, our proposal indeed benefits from the nuclear-norm regularization in identifying the true structural dimension.

We also examine the subspace estimation error and the singular values of the estimator $\widehat{\mathbf{B}}$ for our proposal and oracle methods. For illustrative purposes, we run one replicate for sparse SAVE and oracle-SAVE under Models (M1)–(M5) with sample sizes $n = 200$, and 2000. The results are displayed in Figures S.3 and S.4. We assume that $\mathcal{A}$ and d are given, eliminating the bias from the variable selection and dimension determination and directly assess the specific effect of the nuclear-norm penalty. In the left penals, we display the singular values of the estimator $\widehat{\mathbf{B}}$ for oracle-SAVE and those of the best $\widehat{\mathbf{B}}$ on the solution path for sparse SAVE. In the right panels, we present the ratio of the estimation errors for sparse SAVE and oracle-SAVE over a sequence of λ_2.

The left panels of Figures S.3 and S.4 reveal that all singular values of $\widehat{\mathbf{B}}$ from oracle-SAVE are non-zero, even when n is as large as 2000. With the nuclear-norm regularization, singular values of sparse SAVE are greatly shrunken, resulting in sparse singular values with the first two being non-zero. This observation underscores the role of the nuclear-norm penalty in identifying the low-rank structure and correctly determining the dimension. Additionally, the nuclear-norm penalty also enhances the efficiency of the estimation. The right panels of Figures S.3 and S.4 clearly demonstrate that sparse SAVE, with a properly chosen λ_2, produces a more accurate subspace estimator compared to oracle-SAVE. As n increases, the curve is stretched and the advantage of sparse SAVE is weakened.

By introducing the nuclear-norm penalty, our proposed method utilizes the low-rank structure, somehow reducing the effect of H, which determines the number of parameters. In comparison, the performance of oracle methods should be highly related to the choice of H. We investigate the subspace estimation performance of oracle methods and our proposal with various H's in forward models (M1)–(M5), reported in Table S.8. We assume that $\mathcal{A}$ and d are given. It is observed that the oracle methods are sensitive to the selection of H. In contrast, our approach has almost the same performance in subspace estimation with various H's. Though the oracle methods achieve the best performance with $H = 3$ in most settings, they are still worse than our proposal.

S.10 Additional simulation results

S.10.1 Non-sub-Gaussian settings

In this section, we examine the performance of our sparse SAVE/DR methods under non-sub-Gaussian distributions, with particular attention to heavy-tailed predictors. Based on our observations, the proposed methods are sensitive to heavy-tailed predictors, but when the heavy-tailedness is mild, they can still provide favorable estimators.

In Tables S.10 and S.11, we report the subspace estimation and variable selection results for all methods under Models (M3)–(M5), where $\mathbf{X}$ follows either a multivariate t distribution or a multivariate normal distribution. Since the proposed methods deteriorate and may fail to produce reliable estimators when the multivariate t distribution has very heavy tails, we consider the degrees of freedom $\nu \in \{5, 6, 7, 8, \infty\}$ (the multivariate t reduces to multivariate normal when $\nu = \infty$). As shown in Tables S.10 and S.11, when $\nu = 5$, accurate variable selection is challenging for the proposed methods, which in turn leads to poor subspace estimation performance. As ν increases, that is, as the tails of the t distribution become lighter, the proposed methods yield more accurate estimators. In addition, Model (M3) is particularly challenging for the proposed methods, especially when ν is small. This is mainly due to inaccurate variable selection, as indicated by the results in Table S.11.

S.10.2 Higher structural dimension

We consider five Models (Q5-i)–(Q5-iv), where the structural dimension is increased from 3 to 6. Specifically, we first generate the complete basis matrix $(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_6) \in \mathbb{R}^{p\times 6}$, where the first $s = 10$ elements of its each column is from the uniform distribution supported on $[0.5, 1]$ and the rest of the elements is zero. We then normalize the complete basis matrix such that $\|\boldsymbol{\beta}_i\|_2 = 1$ and $\boldsymbol{\beta}_i^\top \boldsymbol{\beta}_j = 0$ for $i \neq j$. In each model, we construct the basis matrix $\boldsymbol{\beta}$ by taking the first d columns of the complete basis matrix such that $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_d) \in \mathbb{R}^{p\times d}$. Then, the generation of Y and $\mathbf{X}$ are similar to that in Model (Q2). Recall that $\mathbf{X} = \boldsymbol{\beta}\mathbf{Z} + \boldsymbol{\beta}_0\mathbf{Z}_0$ such that $\mathbf{Z} = \boldsymbol{\beta}^\top\mathbf{X}$ and $\mathbf{Z}_0 = \boldsymbol{\beta}_0^\top\mathbf{X}$. The distribution of $\mathbf{Z}$ in each model is as follows, where $\mathbf{0}_d \in \mathbb{R}^d$ denotes the all-zero vector:

(Q5-i) $(d = 3)$ $\mathbf{Z} \mid (Y = 1) \sim N(\mathbf{0}_3^\top, 0.1\mathbf{I}_3)$, $\mathbf{Z} \mid (Y = 2) \sim N((0, 0, 0.5)^\top, 0.5\mathbf{I}_3)$, and $\mathbf{Z} \mid (Y = 3) \sim N((1, 1, 0.5)^\top, 3\mathbf{I}_3)$;

(Q5-ii) $(d = 4)$ $\mathbf{Z} \mid (Y = 1) \sim N(\mathbf{0}_4^\top, 0.1\mathbf{I}_4)$, $\mathbf{Z} \mid (Y = 2) \sim N((0, 0, 0.5, 0)^\top, 0.5\mathbf{I}_4)$, and $\mathbf{Z} \mid (Y = 3) \sim N((1, 1, 0.5, 0)^\top, 3\mathbf{I}_4)$;

(Q5-iii) $(d = 5)$ $\mathbf{Z} \mid (Y = 1) \sim N(\mathbf{0}_5^\top, 0.1\mathbf{I}_5)$, $\mathbf{Z} \mid (Y = 2) \sim N((0, 0, 0.5, 0, 0)^\top, 0.5\mathbf{I}_5)$, and $\mathbf{Z} \mid (Y = 3) \sim N((1, 1, 0.5, 0, 0)^\top, 3\mathbf{I}_5)$;

(Q5-iv) $(d = 6)$ $\mathbf{Z} \mid (Y = 1) \sim N(\mathbf{0}_6^\top, 0.1\mathbf{I}_6)$, $\mathbf{Z} \mid (Y = 2) \sim N((0, 0, 0.5, 0, 0, 0)^\top, 0.5\mathbf{I}_6)$, and $\mathbf{Z} \mid (Y = 3) \sim N((1, 1, 0.5, 0, 0, 0)^\top, 3\mathbf{I}_6)$.

In Tables S.12–S.15, we report the subspace estimation errors, classification errors, variable selection results, and execution time, respectively. It can be seen from Table S.12 that our sparse methods still dominate other methods, except for the oracle methods. Based on the results in Table S.13, our methods achieve the most accurate predictive performance, and are the closest to the Bayes error. Table S.14 suggests that sparse SAVE/DR achieve better variable selection results over other methods. The results in Table S.15 show that sparse SAVE/DR enjoy fast computational speed and their execution times are comparable to SQDA.

S.10.3 Non-sparse models

We examine the performance of all competitors under the non-sparse version of Models (M1)–(M5) in this section. Let $\boldsymbol{\beta}_{\mathcal{A}} \in \mathbb{R}^{s\times d}$ denote the non-zero signal matrix of the basis $\boldsymbol{\beta} \in \mathbb{R}^{p\times d}$. We construct a non-sparse $\boldsymbol{\beta}$ by adding an additional non-zero signal $\boldsymbol{\eta} \in \mathbb{R}^{s'\times d}$ to $\boldsymbol{\beta}$ such that $\boldsymbol{\beta} = (\boldsymbol{\beta}_{\mathcal{A}}^\top, \boldsymbol{\eta}^\top, \mathbf{0}^\top)^\top$, where s' is set as $s' = 0.2n$. In our studies, we fix $p = 700$ and $n = 300$, then the overall sparsity level is $s + s' = 66$. To construct the additional signal matrix $\boldsymbol{\eta}$, we first generate the auxiliary orthonormal matrix $\widetilde{\boldsymbol{\eta}}$ in a similar way to $\boldsymbol{\beta}_{\mathcal{A}}$, and then take $\boldsymbol{\eta} = \rho\widetilde{\boldsymbol{\eta}}$, where ρ is the signal factor. We vary the signal factor ρ over $\{0.1, 0.15, 0.2, 0.3, 0.5, 0.7\}$ to investigate the effect of additional signal strength. Note that for the oracle methods, we still input $\mathcal{A} = \{1, \ldots, s\}$ as the active set.

The subspace estimation errors under the original sparse Models (M1)–(M5) and the their non-sparse versions are presented in Table S.16, and the corresponding error plots are displayed in Figure S.7. The results show that as the additional signal gets stronger, the performances of all methods get worse. Moreover, our sparse SAVE/DR methods remain accurate in subspace estimation even when models get denser. However, if we further increase the signal factor of the additional component $\boldsymbol{\eta}$, such as $\rho = 1$, all methods lose efficiency and fail to provide an accurate estimator of subspace.

S.10.4 Sensitivity of dimension selection to δ

To examine the sensitivity of dimension selection to this practical threshold, we conducted an additional sensitivity analysis by varying

$$\delta \in \{10^{-4},\, 5\times 10^{-4},\, 10^{-3},\, 5\times 10^{-3},\, 10^{-2}\}$$

under Models (M1)–(M5), with $(n, p) = (300, 700)$. The results for sparse SAVE and sparse DR are reported in Table S.18. The new results show that the estimated structural dimension is highly stable over this range of δ. For example, under Model (M3), the average

selected dimensions of sparse SAVE and sparse DR remain 2.08 and 2.01, respectively, for all five choices of δ. Under Model (M4), both sparse SAVE and sparse DR consistently select dimension 1.00 across all thresholds. Under Model (M5), sparse DR consistently selects dimension 2.00, while sparse SAVE only fluctuates slightly between 1.98 and 2.00. Similar stability is observed under Models (M1) and (M2). These results suggest that the dimension selection performance is not sensitive to the particular choice of δ within a reasonable range.

We therefore use $\delta = 10^{-3}$ as the default numerical threshold in our implementation, since it lies in this reasonable range.

S.10.5 Additional results

To save the space, we present additional simulation results in this part.

- **(Additional models)** Besides the models presented in the paper, we have considered additional forward and discriminant models, listed in the following.

  (M4) $Y = (\boldsymbol{\beta}^\top \mathbf{X})^2 + \varepsilon$;

  (M5) $Y = 2(\boldsymbol{\beta}_1^\top \mathbf{X})^2 + 5\boldsymbol{\beta}_2^\top \mathbf{X} + \varepsilon$;

  (Q3) $Y \sim \text{Unif}\{1, 2\}$, $\mathbf{Z} \mid (Y = 1) \sim N(0, 0.1)$ and $\mathbf{Z} \mid (Y = 2) \sim N(2, 3)$;

  (Q4) $Y \sim \text{Unif}\{1, 2, 3\}$, $\mathbf{Z} \mid (Y = 1) \sim N((0, 0)^\top, 0.1\mathbf{I}_2)$, $\mathbf{Z} \mid (Y = 2) \sim N((0, 0.5)^\top, 0.5\mathbf{I}_2)$, and $\mathbf{Z} \mid (Y = 3) \sim N((1, 1)^\top, 3\mathbf{I}_2)$;

  In Model (M4), $d = 1$ and one symmetric component is involved. In Model (M5), $d = 2$ and both symmetric and asymmetric components are considered. For discriminant models, $(K, d) = (2, 1)$ in Model (Q3) and $(K, d) = (3, 2)$ in Model (Q4). The complete results for subspace estimation error, classification error, and estimated structural dimension are shown in Tables S.1, S.3, and S.5.

- **(Complete results)** The complete results for subspace estimation error and classification error with $p = 700$ and $n = 300, 600$ are shown in Tables S.2 and S.4. The variable selection results under Models (M1)–(M5) and (Q1)–(Q4) are presented in Table S.7.

- **(Results with different number of slices)** The subspace estimation errors for oracle methods and our proposal with $H = 3, 5, 8$ under Models (M1)–(M5) are reported in Table S.8.

- **(Misspecification of active set)** The subspace estimation error of sparse SAVE with a misspecified active set from the first step is reported in Figure S.5, where d is given and $n = 300$. The results for sparse DR are similar and are omitted.

We consider different active sets $\{1, \ldots, a\}$ with $a = 3, \ldots, 10$. When $a < 6$, some important variables are missed. When $a > 6$, all important variables are included, but some false positives are present.

- **(Execution time)** The execution times of our proposal and other competitors, including TC-SAVE, sparse SIR methods (SEAS-SIR, Lasso-SIR), and sparse QDA methods (DA-QDA, SQDA) are recorded in Table S.9, where $n = 300$ and $p = 700, 1500$. Since the HOPG method relies on a fast variable screening step to reduce the dimension to $n/\log(n)$, essentially dealing with low-dimensional data ($p < n$), we exclude it from the comparison of computational efficiency. Meanwhile, we only consider discriminant models (Q1)–(Q4) such that sparse QDA methods can be included into comparison. For each method, we record the time for implementing the algorithm once, while the time for tuning parameter selection is not included. Moreover, for a fair comparison, the execution time of sparse SIR method is multiplied by p since the parameters to be estimated in SIR methods are on the order of $O(p)$, while those for second-order SDR methods are on the order of $O(p^2)$.

- **(Word cloud)** The word cloud of the top 30 words most frequently selected by sparse SAVE on the news report data is displayed in Figure S.6. The size of each word represents its relative frequency.

S.11 Spaceflight biological data analysis

On September 15, 2021, the SpaceX Inspiration4 mission launched from the Kennedy Space Center, sending four civilian astronauts into low Earth orbit. During the three-day mission, they faced several challenges inherent to spaceflight, such as radiation exposure, prolonged microgravity, and hostile space conditions. To understand the biological impacts of these conditions, various biological samples, including blood, saliva, skin swabs, skin biopsies, and capsule swabs, were collected from the crew at three distinct phases: pre-flight, in-flight, and post-flight. These samples underwent multiple assays to generate integrative multi-omics data, which revealed molecular changes across different biological levels.

One key focus of this research is to investigate the m6A RNA modification at different stages of the mission. The m6A RNA modification has been documented to play a crucial role in regulating RNA splicing, translation, stability, translocation, and high-level RNA structure (Yang et al., 2018). Recent studies have demonstrated that m6A modifications are essential for physiological and pathological processes in the hematopoietic, central nervous, and reproductive systems. Furthermore, regulators of m6A RNA methylation have been implicated in several human diseases, including nonalcoholic fatty liver disease, azoospermia, heart failure, and various cancers (Jiang et al., 2021). Therefore,

it is of vital importance to understand how spaceflight affects m6A methylation. In this study, our primary objectives are to identify differentially methylated m6A sites and to characterize the underlying patterns of m6A modification before and after low Earth orbit spaceflight using the discriminant analysis model.

During the Inspiration4 mission, whole blood samples were collected via venipuncture at three pre-flight timepoints (L-92, L-44, L-3) and four post-flight timepoints (R+1, R+45, R+82, R+194). Total RNA was extracted from these samples, and mRNA sequencing was performed using Oxford Nanopore Technologies' direct RNA-sequencing kit on the PromethION platform. These data were used to generate differential gene expression profiles and to identify m6A modification sites. The resulting m6A modification data comprises $n = 28$ observations, representing the four crew members at three pre-flight and four post-flight timepoints. Each observation records the probability that a given transcript contains an m6A modification at a particular site. The m6A modification data is publicly available via the Open Science Data Repository (OSDR) (https://osdr.nasa.gov/bio/), with the identifier OSD-569.

During data preparation, we identify many missing values that correspond to transcript positions with insufficient coverage to make a prediction (i.e., fewer than 20 reads). To address this, we keep only transcript positions where at least half of the records were available, resulting in a dataset of 6,540 genes. The remaining missing values are imputed using the sample mean. Following this, we retain the top 300 most relevant variables for further analysis based on distance correlation (Li et al., 2012).

The dataset is divided into two classes: pre-flight and post-flight data. We split the data into training, validation, and testing sets in an 6:2:2 ratio using stratified sampling. Each method is fitted to the training set, with tuning parameters selected based on the validation set. To ensure a fair comparison, all methods are constrained to the same level of sparsity, with a maximum of 10 selected variables. We first determine the structural dimension for the sparse SAVE, sparse DR, and TC-SAVE methods. Based on 100 data replicates, the average estimated dimensions are 2.4, 2.1, and 2.5, respectively, with standard errors are all less than 0.1. Consequently, we fix $d = 2$ for the second-order SDR methods in the subsequent analysis.

Then, we record the variables selected by sparse SAVE, sparse DR, HOPG, TC-SAVE, SEAS-SIR, and LassoSIR. The average sparsity levels are 7.0, 6.7, 10.0, 10.0, 7.7, and 8.3, with standard errors all below 0.4. To further assess the performance of variable selection, we measure the stability of variable selection using the Jaccard similarity coefficient (Jaccard, 1901), which are 0.26, 0.08, 0.14, 0.05, 0.10, and 0.07, respectively. Notably, sparse SAVE demonstrates a much higher stability value than the other methods, indicating its greater consistency in selecting variables.

Moreover, the literature confirms that m6A modifications in some genes frequently se-

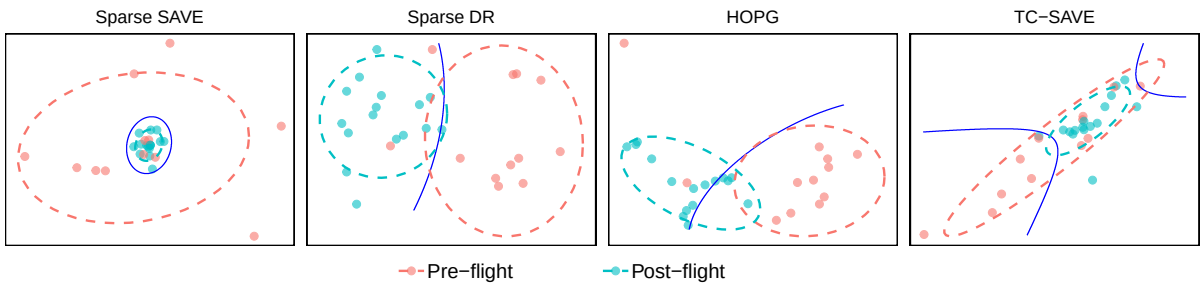

Figure S.2: The scatterplot of the first two reduced predictors in spaceflight data. The dashed circle represents the elliptical contour covering 85% of data, and the blue line means the decision boundary under the QDA model. Means of the classification errors of these methods followed by QDA classifier are 26.3%, 35.0%, 35.0%, and 37.0%, and the standard errors are less than 2.7%.

lected by our method are associated with the development of various human diseases. For example, the top five genes most frequently selected by sparse SAVE include *FGL2*, *PCGF5*, *UBE2R2*, *TXNIP*, and *RPL8*. Xiao et al. (2022) demonstrated that a deficiency in *FGL2* prevents fulminant hepatitis following MHV-3 (Murine hepatitis virus strain 3) infection. In addition, Yang et al. (2024) revealed that *PCGF5* promotes neural induction by regulating gene expression. Ceccarelli et al. (2011) found that silencing the ubiquitin-conjugating enzyme *UBE2R2* significantly inhibits cell proliferation, potentially contributing to disorders such as glutathionuria and subglottic angioma. Furthermore, Choi and Park (2023) identified the multifunctional role of *TXNIP* in the pathogenesis of diseases such as diabetes, chronic kidney disease, and Alzheimer's disease. Lebaron et al. (2022) reported that variants of *RPL8* encode functionally deficient proteins, which impact ribosome production and are linked to Diamond-Blackfan anemia.

Next, we analyze the discriminant patterns of m6A modification before and after low Earth orbit spaceflight. We make scatterplots of the first two reduced data from the sparse SAVE, sparse DR, HOPG, and TC-SAVE methods. To save the space, the density plots of the one-dimensional reduced data from SEAS-SIR and Lasso-SIR are omitted. The visualizations are presented in Figure S.2. All plots indicate a change in the distribution of m6A modification data after the spaceflight. In particular, in the sparse DR plot, two directions reveal the location and variation differences between the pre- and post-flight stages, separately. Notably, the post-flight m6A modification probabilities exhibit smaller variation. Regarding other methods, we have the following remarks. The first-order SDR methods, i.e., SEAS-SIR and Lasso-SIR, only capture the location difference between the two stages, failing to account for variance differences due to their inherent first-order limitations. As a result, these methods provide a limited view of the underlying patterns. The HOPG method fails to clearly capture variance differences between the two stages. The SAVE-based methods, specifically sparse SAVE and TC-SAVE, focus more on

variation rather than location. In contrast, sparse DR combines the strengths of both SIR and SAVE, successfully capturing both location and variance distinctions, which has also been observed in classical DR method.

S.12 Proof of Theorem 1

Since $\mathbf{M}_h = \mathbf{\Sigma}^{-1}\mathbf{\Delta}_h$, then $\mathbf{\Sigma}\mathbf{M}_h = \mathbf{\Delta}_h$. By the sparsity structure such that $\mathbf{M}_{h,\mathcal{A}^c*} = \mathbf{0}$, we have $\mathbf{M}_h = (\mathbf{M}_{h,\mathcal{A}*}^\top, \mathbf{0}^\top)^\top$ and

$$\mathbf{\Sigma}_{\mathcal{A}\mathcal{A}}\mathbf{M}_{h,\mathcal{A}*} = \mathbf{\Delta}_{h,\mathcal{A}*}, \quad \mathbf{\Sigma}_{\mathcal{A}^c\mathcal{A}}\mathbf{M}_{h,\mathcal{A}*} = \mathbf{\Delta}_{h,\mathcal{A}^c*}. \tag{S.8}$$

Therefore,

$$\mathbf{M}_{h,\mathcal{A}*} = (\mathbf{\Sigma}_{\mathcal{A}\mathcal{A}})^{-1}\mathbf{\Delta}_{h,\mathcal{A}*}. \tag{S.9}$$

By (S.8), it follows that

$$\mathbf{\Sigma}_{\mathcal{A}^c\mathcal{A}}\mathbf{M}_{h,\mathcal{A}\mathcal{A}} = \mathbf{\Delta}_{h,\mathcal{A}^c\mathcal{A}}.$$

According to (S.9), we have

$$\mathbf{M}_{h,\mathcal{A}\mathcal{A}} = (\mathbf{\Sigma}_{\mathcal{A}\mathcal{A}})^{-1}\mathbf{\Delta}_{h,\mathcal{A}\mathcal{A}}.$$

Since $\mathbf{\Delta}_h$ and $\mathbf{\Sigma}$ are symmetric, then $\mathbf{\Delta}_{h,\mathcal{A}^c\mathcal{A}} = \mathbf{\Delta}_{h,\mathcal{A}\mathcal{A}^c}^\top$, $\mathbf{\Sigma}_{\mathcal{A}^c\mathcal{A}} = \mathbf{\Sigma}_{\mathcal{A}\mathcal{A}^c}^\top$, and

$$\mathbf{\Delta}_{h,\mathcal{A}\mathcal{A}^c} = \mathbf{M}_{h,\mathcal{A}\mathcal{A}}^\top\mathbf{\Sigma}_{\mathcal{A}\mathcal{A}^c} = \mathbf{\Delta}_{h,\mathcal{A}\mathcal{A}}(\mathbf{\Sigma}_{\mathcal{A}\mathcal{A}})^{-1}\mathbf{\Sigma}_{\mathcal{A}\mathcal{A}^c}. \tag{S.10}$$

Therefore, according to (S.10), we have

$$\operatorname{span}\{(\mathbf{\Sigma}_{\mathcal{A}\mathcal{A}})^{-1}\mathbf{\Delta}_{h,\mathcal{A}\mathcal{A}^c}\} \subseteq \operatorname{span}\{(\mathbf{\Sigma}_{\mathcal{A}\mathcal{A}})^{-1}\mathbf{\Delta}_{h,\mathcal{A}\mathcal{A}}\}. \tag{S.11}$$

By combining (S.9) with (S.11), it follows that

$$\operatorname{span}(\mathbf{M}_{h,\mathcal{A}*}) = \operatorname{span}\{(\mathbf{\Sigma}_{\mathcal{A}\mathcal{A}})^{-1}\mathbf{\Delta}_{h,\mathcal{A}*}\} = \operatorname{span}\{(\mathbf{\Sigma}_{\mathcal{A}\mathcal{A}})^{-1}\mathbf{\Delta}_{h,\mathcal{A}\mathcal{A}}\}.$$

This completes the proof.

□

S.13 Proof of Theorem 2

For easy reference, we recall the necessary assumptions in the following:

(A1) The predictors $\mathbf{X}_1, \ldots, \mathbf{X}_n \in \mathbb{R}^p$ are i.i.d. centered sub-Gaussian random vectors with covariance matrix $\mathbf{\Sigma}$. That is, $\mathrm{E}(\mathbf{X}_i) = \mathbf{0}$ and $\sup_{\mathbf{a}\in\mathbb{R}^p, \|\mathbf{a}\|_2=1} \|\mathbf{a}^\top\mathbf{X}_i\|_{\psi_2} < \infty$, where $\|Z\|_{\psi_2} = \inf\{t > 0 : \mathrm{E}\{\exp(Z^2/t^2)\} \leq 2\}$ denotes the sub-Gaussian norm for any random variable $Z \in \mathbb{R}$. Assume there exists a constant $m > 0$ such that $m^{-1} \leq \varphi_{\min}(\mathbf{\Sigma}) \leq \varphi_{\max}(\mathbf{\Sigma}) \leq m$ and $m^{-1} \leq \varphi_{\min}(\mathrm{Cov}\{\mathbf{X}I(\widetilde{Y} = h)\}) \leq \varphi_{\max}(\mathrm{Cov}\{\mathbf{X}I(\widetilde{Y} = h)\}) \leq m$ for $h = 1, \ldots, H$.

(A2) Assume that $\alpha = 1 - \max_{j\in\mathcal{A}^c} \|\boldsymbol{\Sigma}_{j\mathcal{A}}(\boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}})^{-1}\|_1 > 0$.

(A3) Assume that $\|(\boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}})^{-1}\|_\infty$, $\|\boldsymbol{\Sigma}_{\mathcal{A}^c\mathcal{A}}\|_\infty$, and $\max_h \|\boldsymbol{\Delta}_{h,\mathcal{A}*}\|_{\max}$ are bounded from above.

(A4) There exists a constant $a > 0$ such that $\pi_h = \Pr(\widetilde{Y} = h) \geq a$ for $h = 1, \ldots, H$.

(A5) There exists a constant $c > 0$ that does not depend on s, n and p, such that $\|\mathcal{M}_{[:,i,:]}\|_{2,\infty} > c\sqrt{s^2 \log p/n}$ for all $i \in \mathcal{A}$.

(A6) Assume that $\max_h \|\boldsymbol{\Delta}_{h,\mathcal{A}\mathcal{A}}\|_1$ is bounded from above.

(A7) There exists a constant $c' > 0$ that does not depend on s, n and p, such that $\sigma > c'\sqrt{s^2 \log p/n}$.

Define the following notations,

$$\theta_0 = \|(\boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}})^{-1}\|_\infty, \quad \theta_1 = \|\boldsymbol{\Sigma}_{\mathcal{A}^c\mathcal{A}}\|_\infty, \quad \eta_0 = \max_h \|\boldsymbol{\Delta}_{h,\mathcal{A}*}\|_{\max},$$
$$\varepsilon_\Sigma = \|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{\max}, \quad \varepsilon_\Delta = \max_h \|\widehat{\boldsymbol{\Delta}}_h - \boldsymbol{\Delta}_h\|_{\max}.$$

For some constant $c > 0$, define the event

$$E := \left\{\max\{\varepsilon_\Sigma, \varepsilon_\Delta\} \leq \sqrt{c \log p/n}\right\}. \tag{S.12}$$

We show that the results in (i)–(iii) hold under event E.

(Part I) We verify that the following two Conditions hold under event E.

(C-i) $\theta_0 s\varepsilon_\Sigma \leq 1/3$;

(C-ii) $s\varepsilon_\Sigma\theta_0\left\{1 + \frac{3}{2}\left(\theta_1\theta_0 + \frac{1}{3}\right)\right\} \leq M\alpha \min\{\lambda_1, 1\}$, where $M = (\alpha\lambda_1 + \sqrt{H}(\alpha-2)\varepsilon_\Delta)(\sqrt{H}\eta_0\alpha\lambda_1 + \alpha\lambda_1 + \sqrt{H}\varepsilon_\Delta\alpha)^{-1}$

Since we assume that $n \geq c_1 s^2 \log p$ for some large enough constant $c_1 > 0$ and that θ_0 is bounded from above by Assumption (A3), by taking

$$n \geq 9\theta_0^2 c s^2 \log p,$$

we have

$$\theta_0 s\varepsilon_\Sigma \leq \theta_0 s\sqrt{c \log p/n} \leq \frac{1}{3},$$

which gives Condition (C-i). Furthermore, by Assumption (A3), θ_0, θ_1, and η_0 are bounded, we take

$$\lambda_1 \geq \max\left\{2s\alpha^{-1}(\sqrt{H}\eta_0 + 1)\theta_0\left[1 + \frac{3}{2}\left(\theta_1\theta_0 + \frac{1}{3}\right)\right], 4\sqrt{H}\alpha^{-1}\right\}\sqrt{\frac{c \log p}{n}}.$$

Then

$$s\varepsilon_\Sigma\theta_0\left\{1 + \frac{3}{2}\left(\theta_1\theta_0 + \frac{1}{3}\right)\right\} \leq s\theta_0\left\{1 + \frac{3}{2}\left(\theta_1\theta_0 + \frac{1}{3}\right)\right\}\sqrt{\frac{c \log p}{n}} \leq \frac{1}{2(\sqrt{H}\eta_0 + 1)}\alpha\lambda_1, \tag{S.13}$$

and

$$4\sqrt{H}\varepsilon_\Delta \le 4\sqrt{H}\sqrt{c\log p/n} \le \alpha\lambda_1. \tag{S.14}$$

It is known that $\alpha\lambda_1 + \sqrt{H}\eta_0\alpha\lambda_1 \ge \alpha\lambda_1 - 2\sqrt{H}\varepsilon_\Delta$, then

$$M = \frac{\alpha\lambda_1 - 2\sqrt{H}\varepsilon_\Delta + \sqrt{H}\alpha\varepsilon_\Delta}{\alpha\lambda_1 + \sqrt{H}\alpha\eta_0\lambda_1 + \sqrt{H}\alpha\varepsilon_\Delta} \ge \frac{\alpha\lambda_1 - 2\sqrt{H}\varepsilon_\Delta}{\alpha\lambda_1 + \sqrt{H}\eta_0\alpha\lambda_1}.$$

According to (S.14), we have

$$M \ge \frac{1}{2(\sqrt{H}\eta_0 + 1)}. \tag{S.15}$$

By combining (S.13) and (S.15), we have

$$s\varepsilon_\Sigma\theta_0\left\{1 + \frac{3}{2}\left(\theta_1\theta_0 + \frac{1}{3}\right)\right\} \le M\alpha\lambda_1. \tag{S.16}$$

Meanwhile, by taking

$$n \ge 4\alpha^{-2}(\sqrt{H}\eta_0 + 1)^2\theta_0^2\left\{1 + \frac{3}{2}\left(\theta_1\theta_0 + \frac{1}{3}\right)\right\}^2 cs^2\log p,$$

we obtain

$$s\theta_0\left\{1 + \frac{3}{2}\left(\theta_1\theta_0 + \frac{1}{3}\right)\right\}\sqrt{\frac{c\log p}{n}} \le \frac{1}{2(\sqrt{H}\eta_0 + 1)}\alpha \le M\alpha.$$

Hence,

$$s\varepsilon_\Sigma\theta_0\left\{1 + \frac{3}{2}\left(\theta_1\theta_0 + \frac{1}{3}\right)\right\} \le s\theta_0\left\{1 + \frac{3}{2}\left(\theta_1\theta_0 + \frac{1}{3}\right)\right\}\sqrt{\frac{c\log p}{n}} \le M\alpha. \tag{S.17}$$

By combining (S.16) and (S.17), we obtain

$$s\varepsilon_\Sigma\theta_0\left\{1 + \frac{3}{2}\left(\theta_1\theta_0 + \frac{1}{3}\right)\right\} \le M\alpha\min\{\lambda_1, 1\}.$$

Hence, Condition (C-ii) holds.

(Part II) We show that statements (i)–(iii) holds under Conditions (C-i)(C-ii). Let $\mathbf{m} = (\mathbf{m}_1, \ldots, \mathbf{m}_H) \in \mathbb{R}^{p^2\times H}$, where $\mathbf{m}_h = \mathrm{vec}(\mathbf{M}_h)$ for $h = 1, \ldots, H$, and let $\widehat{\mathbf{m}} \in \mathbb{R}^{p^2\times H}$ denote the solution of (S.6). Define the set $\mathcal{S} = \{i + (j-1)p : i \in \mathcal{A}, j \in \{1, \ldots, p\}\}$. We need the following result.

Lemma S.3. *Assuming Assumption (A2) and Conditions (C-i)(C-ii) hold, we have* $\widehat{\mathbf{m}}_{\mathcal{S}^c} = \mathbf{0}$ *and* $\|\widehat{\mathbf{m}} - \mathbf{m}\|_{2,\infty} \le (\lambda_1 + \sqrt{H}\varepsilon_\Delta)\theta_0 + \frac{3}{2}s\varepsilon_\Sigma\theta_0^2(\lambda_1 + \sqrt{H}\eta_0 + \sqrt{H}\varepsilon_\Delta)$.

According to the definition of $\widehat{\mathcal{A}}$, by the result $\widehat{\mathbf{m}}_{\mathcal{S}^c} = \mathbf{0}$ in Lemma S.3, we have $\widehat{\mathcal{A}} \subseteq \mathcal{A}$ in statement (i). Moreover, by (S.14), we have $\sqrt{H}\varepsilon_\Delta \le \alpha\lambda_1/4 < \lambda_1/4$, and by Condition (C-i), we have $\theta_0 s\varepsilon_\Sigma \le 1/3$, it follows that

$$\begin{aligned}\|\widehat{\mathbf{m}} - \mathbf{m}\|_{2,\infty} \le& (\lambda_1 + \sqrt{H}\varepsilon_\Delta)(\theta_0 + \frac{3}{2}s\varepsilon_\Sigma\theta_0^2) + \frac{3}{2}\theta_0^2\sqrt{H}\eta_0 s\varepsilon_\Sigma \\ \le& \frac{15}{8}\theta_0\lambda_1 + \frac{3}{2}\theta_0^2\sqrt{H}\eta_0 s\varepsilon_\Sigma.\end{aligned}$$

Since we further take $\lambda_1 \leq \kappa_2 s\sqrt{\log p/n}$ for some constant $\kappa_2 > 0$, and under event E, $\varepsilon_\Sigma \lesssim \sqrt{\log p/n}$, then it follows that

$$\|\widehat{\mathbf{m}} - \mathbf{m}\|_{2,\infty} \lesssim \sqrt{\frac{s^2 \log p}{n}}.$$

Since

$$\max_{i=1,\ldots,p} \|\widehat{\mathcal{M}}_{[:,i,:]} - \mathcal{M}_{[:,i,:]}\|_{2,\infty} = \max_{i,j=1,\ldots,p} \|\widehat{\mathcal{M}}_{[:,i,j]} - \mathcal{M}_{[:,i,j]}\|_2 = \|\widehat{\mathbf{m}} - \mathbf{m}\|_{2,\infty},$$

where the last equality comes from the correspondence between $\widehat{\mathcal{M}}$ and $\widehat{\mathbf{m}}$, we have statement (ii) that

$$\max_{i=1,\ldots,p} \|\widehat{\mathcal{M}}_{[:,i,:]} - \mathcal{M}_{[:,i,:]}\|_{2,\infty} \lesssim \sqrt{\frac{s^2 \log p}{n}}.$$

If we further assume Assumption (A5), then by taking large enough constant c, we have

$$\|\widehat{\mathcal{M}}_{[:,i,:]}\|_{2,\infty} \geq \|\mathcal{M}_{[:,i,:]}\|_{2,\infty} - \|\mathcal{M}_{[:,i,:]} - \widehat{\mathcal{M}}_{[:,i,:]}\|_{2,\infty} > 0, \quad \forall i \in \mathcal{A}.$$

Therefore, $\mathcal{A} \subseteq \widehat{\mathcal{A}}$. Combined with the result that $\widehat{\mathcal{A}} \subseteq \mathcal{A}$ we have established in statement (i), we obtain statement (iii) that $\widehat{\mathcal{A}} = \mathcal{A}$.

(Part III) Lastly, we show that event E holds with high probability by the following two lemmas.

Lemma S.4. *Assume that Assumption (A1) holds. There exist some constants $C, C' > 0$ such that, for any $0 < \varepsilon \leq C'$ we have*

$$\Pr(\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_{\max} \geq \varepsilon) \leq Cp^2 \exp(-C' n\varepsilon^2).$$

Lemma S.5. *Assume that Assumptions (A1) & (A4) hold. There exist some constants $C, C' > 0$ such that, for any $0 < \varepsilon \leq C'$ we have*

$$\Pr(\|\widehat{\mathbf{\Sigma}}_h - \mathbf{\Sigma}_h\|_{\max} \geq \varepsilon) \leq Cp^2 \exp(-C' n\varepsilon^2).$$

By Lemmas S.4 and S.5, under Assumption (A1), there exist some constants $C, C', c > 0$ such that, with probability at least $1 - Cp^{-C'}$, we have

$$\max\{\varepsilon_\Sigma, \varepsilon_\Delta\} \leq \sqrt{c \log p/n},$$

where c is a constant depending on $\|X_j\|_{\psi_2}$, $\|X_j I(\widetilde{Y} = h)\|_{\psi_2}$, and π_h, for $j = 1, \ldots, p$, and $h = 1, \ldots, H$. This completes the proof. □

S.13.1 Proof of Lemma S.3

(Part I) Let $\boldsymbol{\Theta} = \mathbf{I} \otimes \boldsymbol{\Sigma}$, $\widehat{\boldsymbol{\Theta}} = \mathbf{I} \otimes \widehat{\boldsymbol{\Sigma}}$, and $\boldsymbol{\delta} = (\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_H) \in \mathbb{R}^{p^2 \times H}$. Recall that $\widehat{\mathbf{m}}$ is the solution of the following optimization problem,

$$\underset{\substack{(\mathbf{a}_1,\ldots,\mathbf{a}_H) \\ \mathbf{a}_h \in \mathbb{R}^{p^2}, h=1,\ldots,H}}{\operatorname{argmin}} \sum_{h=1}^{H} \left\{ \frac{1}{2} \mathbf{a}_h^\top (\mathbf{I} \otimes \widehat{\boldsymbol{\Sigma}}) \mathbf{a}_h - \mathbf{a}_h^\top \widehat{\boldsymbol{\delta}}_h \right\} + \lambda_1 \sum_{j=1}^{p^2} \sqrt{\sum_{h=1}^{H} a_{h,j}^2}, \tag{S.18}$$

We consider the oracle estimator $\widetilde{\mathbf{m}} = (\widetilde{\mathbf{m}}_{\mathcal{S}*}^\top, \mathbf{0}^\top)^\top$, where $\widetilde{\mathbf{m}}_{\mathcal{S}*}$ is the solution of

$$\underset{\mathbf{g} \in R^{s^2 \times H}}{\operatorname{argmin}} \frac{1}{2} \operatorname{tr}\left(\mathbf{g}^\top \widehat{\boldsymbol{\Theta}}_{\mathcal{S}\mathcal{S}} \mathbf{g} \right) - \operatorname{tr}\left(\mathbf{g}^\top \widehat{\boldsymbol{\delta}}_{\mathcal{S}*} \right) + \lambda_1 \sum_{j=1}^{s^2} \|\mathbf{g}_{j*}\|_2. \tag{S.19}$$

We prove $\widehat{\mathbf{m}}_{\mathcal{S}^c} = \mathbf{0}$ by showing that $\widetilde{\mathbf{m}} = \widehat{\mathbf{m}}$. Due to the convexity of (S.18) in the main paper, we only need to show that the derivative of (S.18) is zero at $\widetilde{\mathbf{m}}$. Equivalently, we need to show that

$$\|\widehat{\boldsymbol{\Theta}}_{j*} \widetilde{\mathbf{m}} - \widehat{\boldsymbol{\delta}}_{j*}\|_2 \leq \lambda_1, \quad \forall j = 1, \ldots, p^2.$$

Since $\widetilde{\mathbf{m}}_{\mathcal{S}^c *} = \mathbf{0}$, it is equivalent to show that

$$\|\widehat{\boldsymbol{\Theta}}_{j\mathcal{S}} \widetilde{\mathbf{m}}_{\mathcal{S}*} - \widehat{\boldsymbol{\delta}}_{j*}\|_2 \leq \lambda_1, \quad \forall j = 1, \ldots, p^2.$$

Since (S.19) is convex, by taking the first derivative of (S.19), we obtain

$$\widehat{\boldsymbol{\Theta}}_{\mathcal{S}\mathcal{S}} \widetilde{\mathbf{m}}_{\mathcal{S}*} - \widehat{\boldsymbol{\delta}}_{\mathcal{S}*} + \lambda_1 \mathbf{z} = 0 \quad \Longrightarrow \quad \widetilde{\mathbf{m}}_{\mathcal{S}*} = (\widehat{\boldsymbol{\Theta}}_{\mathcal{S}\mathcal{S}})^{-1} (\widehat{\boldsymbol{\delta}}_{\mathcal{S}*} - \lambda_1 \mathbf{z}), \tag{S.20}$$

where $\mathbf{z} = \partial \|\widetilde{\mathbf{m}}_{\mathcal{S}*}\|_{2,1}$. And it follows that

$$\|\widehat{\boldsymbol{\Theta}}_{j\mathcal{S}} \widetilde{\mathbf{m}}_{\mathcal{S}*} - \widehat{\boldsymbol{\delta}}_{j*}\|_2 \leq \lambda_1, \quad \forall j \in \mathcal{S}.$$

Then, it remains to show that

$$\|\widehat{\boldsymbol{\Theta}}_{j\mathcal{S}} \widetilde{\mathbf{m}}_{\mathcal{S}*} - \widehat{\boldsymbol{\delta}}_{j*}\|_2 \leq \lambda_1, \quad \forall j \in \mathcal{S}^c. \tag{S.21}$$

By (S.20), we have

$$\begin{aligned} &\|\widehat{\boldsymbol{\Theta}}_{j\mathcal{S}} \widetilde{\mathbf{m}}_{\mathcal{S}*} - \widehat{\boldsymbol{\delta}}_{j*}\|_2 = \|\widehat{\boldsymbol{\Theta}}_{j\mathcal{S}} (\widehat{\boldsymbol{\Theta}}_{\mathcal{S}\mathcal{S}})^{-1} (\widehat{\boldsymbol{\delta}}_{\mathcal{S}*} - \lambda_1 \mathbf{z}) - \widehat{\boldsymbol{\delta}}_{j*}\|_2 \\ \leq &\lambda_1 \|\widehat{\boldsymbol{\Theta}}_{j\mathcal{S}} (\widehat{\boldsymbol{\Theta}}_{\mathcal{S}\mathcal{S}})^{-1} \mathbf{z}\|_2 + \|\widehat{\boldsymbol{\Theta}}_{j\mathcal{S}} (\widehat{\boldsymbol{\Theta}}_{\mathcal{S}\mathcal{S}})^{-1} \widehat{\boldsymbol{\delta}}_{\mathcal{S}*} - \boldsymbol{\delta}_{j*}\|_2 + \|\widehat{\boldsymbol{\delta}}_{j*} - \boldsymbol{\delta}_{j*}\|_2. \end{aligned}$$

Since $\mathbf{M}_h = \boldsymbol{\Sigma}^{-1} \boldsymbol{\Delta}_h$, then $\mathbf{m}_h = (\mathbf{I} \otimes \boldsymbol{\Sigma}^{-1}) \boldsymbol{\delta}_h = \boldsymbol{\Theta}^{-1} \boldsymbol{\delta}_h$ and $\boldsymbol{\delta} = \boldsymbol{\Theta} \mathbf{m} = \boldsymbol{\Theta}_{*\mathcal{S}} \mathbf{m}_{\mathcal{S}*}$. It follows that

$$\boldsymbol{\delta}_{j*} = \boldsymbol{\Theta}_{j\mathcal{S}} \mathbf{m}_{\mathcal{S}*}, \quad \boldsymbol{\delta}_{\mathcal{S}*} = \boldsymbol{\Theta}_{\mathcal{S}\mathcal{S}} \mathbf{m}_{\mathcal{S}*}, \tag{S.22}$$

which gives us $\boldsymbol{\delta}_{j*} = \boldsymbol{\Theta}_{j\mathcal{S}} (\boldsymbol{\Theta}_{\mathcal{S}\mathcal{S}})^{-1} \boldsymbol{\delta}_{\mathcal{S}*}$ for all $j \in 1, \ldots, p^2$. For any $j \in \mathcal{S}^c$ and $h \in$

$\{1,\ldots,H\}$, we have

$$\begin{aligned}\|\widehat{\boldsymbol{\Theta}}_{j\mathcal{S}}\widetilde{\mathbf{m}}_{\mathcal{S}*}-\widehat{\boldsymbol{\delta}}_{j*}\|_2 \leq&\lambda_1\|\widehat{\boldsymbol{\Theta}}_{j\mathcal{S}}(\widehat{\boldsymbol{\Theta}}_{\mathcal{S}\mathcal{S}})^{-1}\mathbf{z}\|_2+\|\widehat{\boldsymbol{\Theta}}_{j\mathcal{S}}(\widehat{\boldsymbol{\Theta}}_{\mathcal{S}\mathcal{S}})^{-1}\widehat{\boldsymbol{\delta}}_{\mathcal{S}*}-\boldsymbol{\Theta}_{j\mathcal{S}}(\boldsymbol{\Theta}_{\mathcal{S}\mathcal{S}})^{-1}\boldsymbol{\delta}_{\mathcal{S}*}\|_2+\|\widehat{\boldsymbol{\delta}}_{j*}-\boldsymbol{\delta}_{j*}\|_2\\ \leq&\lambda_1\|\widehat{\boldsymbol{\Theta}}_{j\mathcal{S}}(\widehat{\boldsymbol{\Theta}}_{\mathcal{S}\mathcal{S}})^{-1}\|_1+\|\widehat{\boldsymbol{\Theta}}_{j\mathcal{S}}(\widehat{\boldsymbol{\Theta}}_{\mathcal{S}\mathcal{S}})^{-1}-\boldsymbol{\Theta}_{j\mathcal{S}}(\boldsymbol{\Theta}_{\mathcal{S}\mathcal{S}})^{-1}\|_1\|\boldsymbol{\delta}_{\mathcal{S}*}\|_{2,\infty}\\ &+\|\widehat{\boldsymbol{\Theta}}_{j\mathcal{S}}(\widehat{\boldsymbol{\Theta}}_{\mathcal{S}\mathcal{S}})^{-1}\|_1\|\widehat{\boldsymbol{\delta}}_{\mathcal{S}*}-\boldsymbol{\delta}_{\mathcal{S}*}\|_{2,\infty}+\|\widehat{\boldsymbol{\delta}}_{j}-\boldsymbol{\delta}_{j*}\|_2\\ \leq&\lambda_1\|\widehat{\boldsymbol{\Theta}}_{j\mathcal{S}}(\widehat{\boldsymbol{\Theta}}_{\mathcal{S}\mathcal{S}})^{-1}\|_1+\sqrt{H}\eta_0\|\widehat{\boldsymbol{\Theta}}_{j\mathcal{S}}(\widehat{\boldsymbol{\Theta}}_{\mathcal{S}\mathcal{S}})^{-1}-\boldsymbol{\Theta}_{j\mathcal{S}}(\boldsymbol{\Theta}_{\mathcal{S}\mathcal{S}})^{-1}\|_1\\ &+\sqrt{H}\varepsilon_\Delta\|\widehat{\boldsymbol{\Theta}}_{j\mathcal{S}}(\widehat{\boldsymbol{\Theta}}_{\mathcal{S}\mathcal{S}})^{-1}\|_1+\sqrt{H}\varepsilon_\Delta.\end{aligned}$$

Consequently, (S.21) is true if

$$\begin{aligned}&\max_{j\in S^c}\sqrt{H}\eta_0\|\widehat{\boldsymbol{\Theta}}_{j\mathcal{S}}(\widehat{\boldsymbol{\Theta}}_{\mathcal{S}\mathcal{S}})^{-1}-\boldsymbol{\Theta}_{j\mathcal{S}}(\boldsymbol{\Theta}_{\mathcal{S}\mathcal{S}})^{-1}\|_1+\sqrt{H}\varepsilon_\Delta(1+\|\widehat{\boldsymbol{\Theta}}_{j\mathcal{S}}(\widehat{\boldsymbol{\Theta}}_{\mathcal{S}\mathcal{S}})^{-1}\|_1)\leq(1-M)\,\alpha\lambda_1\\ &\max_{j\in S^c}\|\widehat{\boldsymbol{\Theta}}_{j\mathcal{S}}(\widehat{\boldsymbol{\Theta}}_{\mathcal{S}\mathcal{S}})^{-1}\|_1\leq 1-(1-M)\alpha.\end{aligned}\tag{S.23}$$

We have

$$\|\widehat{\boldsymbol{\Theta}}_{j\mathcal{S}}-\boldsymbol{\Theta}_{j\mathcal{S}}\|_1=\|\{\mathbf{I}\otimes(\widehat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma})\}_{j\mathcal{S}}\|_1\leq s\|\widehat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}\|_{\max}\leq s\varepsilon_\Sigma.$$

To proceed, we need the following lemma.

Lemma S.6. *Assume that $\theta_0 s\varepsilon_\Sigma\leq 1/3$, we have $\|(\widehat{\boldsymbol{\Theta}}_{\mathcal{S}\mathcal{S}})^{-1}-(\boldsymbol{\Theta}_{\mathcal{S}\mathcal{S}})^{-1}\|_\infty\leq\frac{3}{2}s^2\varepsilon_\Sigma^2\theta_0^3+s\varepsilon_\Sigma\theta_0^2$.*

By Lemma S.6, given that $\theta_0 s\varepsilon_\Sigma\leq 1/3$, it follows that

$$\|(\widehat{\boldsymbol{\Theta}}_{\mathcal{S}\mathcal{S}})^{-1}-(\boldsymbol{\Theta}_{\mathcal{S}\mathcal{S}})^{-1}\|_\infty\leq\frac{3}{2}s^2\varepsilon_\Sigma^2\theta_0^3+s\varepsilon_\Sigma\theta_0^2.$$

Then, for any $j\in\mathcal{S}^c$,

$$\begin{aligned}&\|\widehat{\boldsymbol{\Theta}}_{j\mathcal{S}}(\widehat{\boldsymbol{\Theta}}_{\mathcal{S}\mathcal{S}})^{-1}-\boldsymbol{\Theta}_{j\mathcal{S}}(\boldsymbol{\Theta}_{\mathcal{S}\mathcal{S}})^{-1}\|_1\\ \leq&\|(\widehat{\boldsymbol{\Theta}}_{j\mathcal{S}}-\boldsymbol{\Theta}_{j\mathcal{S}})(\boldsymbol{\Theta}_{\mathcal{S}\mathcal{S}})^{-1}\|_1+\|\boldsymbol{\Theta}_{j\mathcal{S}}\{(\widehat{\boldsymbol{\Theta}}_{\mathcal{S}\mathcal{S}})^{-1}-(\boldsymbol{\Theta}_{\mathcal{S}\mathcal{S}})^{-1}\}\|_1+\|(\widehat{\boldsymbol{\Theta}}_{j\mathcal{S}}-\boldsymbol{\Theta}_{j\mathcal{S}})\{(\widehat{\boldsymbol{\Theta}}_{\mathcal{S}\mathcal{S}})^{-1}-(\boldsymbol{\Theta}_{\mathcal{S}\mathcal{S}})^{-1}\}\|_1\\ \leq&\|\widehat{\boldsymbol{\Theta}}_{j\mathcal{S}}-\boldsymbol{\Theta}_{j\mathcal{S}}\|_1\|(\boldsymbol{\Theta}_{\mathcal{S}\mathcal{S}})^{-1}\|_\infty+\|\boldsymbol{\Theta}_{j\mathcal{S}}\|_1\|(\widehat{\boldsymbol{\Theta}}_{\mathcal{S}\mathcal{S}})^{-1}-(\boldsymbol{\Theta}_{\mathcal{S}\mathcal{S}})^{-1}\|_\infty+\|\widehat{\boldsymbol{\Theta}}_{j\mathcal{S}}-\boldsymbol{\Theta}_{j\mathcal{S}}\|_1\|(\widehat{\boldsymbol{\Theta}}_{\mathcal{S}\mathcal{S}})^{-1}-(\boldsymbol{\Theta}_{\mathcal{S}\mathcal{S}})^{-1}\|_\infty\\ \leq&s\varepsilon_\Sigma\theta_0+(\theta_1+s\varepsilon_\Sigma)\left(\frac{3}{2}s^2\varepsilon_\Sigma^2\theta_0^3+s\varepsilon_\Sigma\theta_0^2\right)\\ \leq&s\varepsilon_\Sigma\theta_0+s\varepsilon_\Sigma\theta_0(\theta_1\theta_0+s\varepsilon_\Sigma\theta_0)\left(\frac{3}{2}s\varepsilon_\Sigma\theta_0+1\right)\\ \leq&s\varepsilon_\Sigma\theta_0\left\{1+(\theta_1\theta_0+s\varepsilon_\Sigma\theta_0)\left(\frac{3}{2}s\varepsilon_\Sigma\theta_0+1\right)\right\}\end{aligned}$$

Under Conditions (C-i)(C-ii), it follows that

$$\|\widehat{\boldsymbol{\Theta}}_{j\mathcal{S}}(\widehat{\boldsymbol{\Theta}}_{\mathcal{S}\mathcal{S}})^{-1}-\boldsymbol{\Theta}_{j\mathcal{S}}(\boldsymbol{\Theta}_{\mathcal{S}\mathcal{S}})^{-1}\|_1\leq s\varepsilon_\Sigma\theta_0\left\{1+\frac{3}{2}\left(\theta_1\theta_0+\frac{1}{3}\right)\right\}\leq M\alpha\min\{\lambda_1,1\},$$

where the last inequality follows from (S.16). Consequently, by Assumption (A2), we obtain

$$\begin{aligned}\max_{j\in\mathcal{S}^c}\|\widehat{\boldsymbol{\Theta}}_{j\mathcal{S}}(\widehat{\boldsymbol{\Theta}}_{\mathcal{S}\mathcal{S}})^{-1}\|_1 \leq& \max_{j\in\mathcal{S}^c}\|\widehat{\boldsymbol{\Theta}}_{j\mathcal{S}}(\widehat{\boldsymbol{\Theta}}_{\mathcal{S}\mathcal{S}})^{-1}-\boldsymbol{\Theta}_{j\mathcal{S}}(\boldsymbol{\Theta}_{\mathcal{S}\mathcal{S}})^{-1}\|_1+\max_{j\in\mathcal{S}^c}\|\boldsymbol{\Theta}_{j\mathcal{S}}(\boldsymbol{\Theta}_{\mathcal{S}\mathcal{S}})^{-1}\|_1\\ \leq& \max_{j\in\mathcal{S}^c}\|\widehat{\boldsymbol{\Theta}}_{j\mathcal{S}}(\widehat{\boldsymbol{\Theta}}_{\mathcal{S}\mathcal{S}})^{-1}-\boldsymbol{\Theta}_{j\mathcal{S}}(\boldsymbol{\Theta}_{\mathcal{S}\mathcal{S}})^{-1}\|_1+\max_{j\in\mathcal{A}^c}\|\boldsymbol{\Sigma}_{j\mathcal{A}}(\boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}})^{-1}\|_1\\ \leq& M\alpha\min\{\lambda_1,1\}+(1-\alpha)\leq 1-(1-M)\alpha.\end{aligned}$$

For $M=\frac{\alpha\lambda_1+\sqrt{H}(\alpha-2)\varepsilon_\Delta}{\sqrt{H}\eta_0\alpha\lambda_1+\alpha\lambda_1+\sqrt{H}\varepsilon_\Delta\alpha}$, we have

$$\begin{aligned}&\max_{j\in S^c}\sqrt{H}\eta_0\|\widehat{\boldsymbol{\Theta}}_{j\mathcal{S}}(\widehat{\boldsymbol{\Theta}}_{\mathcal{S}\mathcal{S}})^{-1}-\boldsymbol{\Theta}_{j\mathcal{S}}(\boldsymbol{\Theta}_{\mathcal{S}\mathcal{S}})^{-1}\|_1+\sqrt{H}\varepsilon_\Delta(1+\|\widehat{\boldsymbol{\Theta}}_{j\mathcal{S}}(\widehat{\boldsymbol{\Theta}}_{\mathcal{S}\mathcal{S}})^{-1}\|_1)\\ \leq&\sqrt{H}\eta_0\alpha\lambda_1 M+\sqrt{H}\left\{2-(1-M)\,\alpha\right\}\varepsilon_\Delta=(1-M)\,\alpha\lambda.\end{aligned}$$

Hence, (S.23) is proven and we have $\widehat{\mathbf{m}}_{\mathcal{S}^c}=\mathbf{0}$.

(Part II) From Part I, we have $\widehat{\mathbf{m}}=\widetilde{\mathbf{m}}$. By (S.20) and (S.22), we have

$$\begin{aligned}\|\widehat{\mathbf{m}}-\mathbf{m}\|_{2,\infty}=&\|(\widehat{\boldsymbol{\Theta}}_{\mathcal{S}\mathcal{S}})^{-1}(\widehat{\boldsymbol{\delta}}_{\mathcal{S}*}-\lambda_1\mathbf{z})-(\boldsymbol{\Theta}_{\mathcal{S}\mathcal{S}})^{-1}\boldsymbol{\delta}_{\mathcal{S}*}\|_{2,\infty}\\ \leq&\lambda_1\|(\widehat{\boldsymbol{\Theta}}_{\mathcal{S}\mathcal{S}})^{-1}\|_\infty+\|(\widehat{\boldsymbol{\Theta}}_{\mathcal{S}\mathcal{S}})^{-1}(\widehat{\boldsymbol{\delta}}_{\mathcal{S}*}-\boldsymbol{\delta}_{\mathcal{S}*})\|_{2,\infty}+\|\{(\widehat{\boldsymbol{\Theta}}_{\mathcal{S}\mathcal{S}})^{-1}-(\boldsymbol{\Theta}_{\mathcal{S}\mathcal{S}})^{-1}\}\boldsymbol{\delta}_{\mathcal{S}*}\|_{2,\infty}\\ \leq&(\lambda_1+\sqrt{H}\varepsilon_\Delta)\|(\widehat{\boldsymbol{\Theta}}_{\mathcal{S}\mathcal{S}})^{-1}\|_\infty+\sqrt{H}\eta_0\|(\widehat{\boldsymbol{\Theta}}_{\mathcal{S}\mathcal{S}})^{-1}-(\boldsymbol{\Theta}_{\mathcal{S}\mathcal{S}})^{-1}\|_\infty\\ \leq&(\lambda_1+\sqrt{H}\varepsilon_\Delta)\|(\boldsymbol{\Theta}_{\mathcal{S}\mathcal{S}})^{-1}\|_\infty+(\lambda_1+\sqrt{H}\eta_0+\sqrt{H}\varepsilon_\Delta)\|(\widehat{\boldsymbol{\Theta}}_{\mathcal{S}\mathcal{S}})^{-1}-(\boldsymbol{\Theta}_{\mathcal{S}\mathcal{S}})^{-1}\|_\infty.\end{aligned}$$

According to Lemma S.6, since we have $\theta_0 s\varepsilon_\Sigma\leq 1/3$ in Condition (C-i), then

$$\|(\widehat{\boldsymbol{\Theta}}_{\mathcal{S}\mathcal{S}})^{-1}-(\boldsymbol{\Theta}_{\mathcal{S}\mathcal{S}})^{-1}\|_\infty\leq\frac{3}{2}s^2\varepsilon_\Sigma^2\theta_0^3+s\varepsilon_\Sigma\theta_0^2.$$

Consequently, we have

$$\begin{aligned}\|\widehat{\mathbf{m}}-\mathbf{m}\|_{2,\infty}\leq&(\lambda_1+\sqrt{H}\varepsilon_\Delta)\theta_0+\left(\frac{3}{2}s^2\varepsilon_\Sigma^2\theta_0^3+s\varepsilon_\Sigma\theta_0^2\right)(\lambda_1+\sqrt{H}\eta_0+\sqrt{H}\varepsilon_\Delta)\\ \leq&(\lambda_1+\sqrt{H}\varepsilon_\Delta)\theta_0+s\varepsilon_\Sigma\theta_0\left(\frac{3}{2}s\varepsilon_\Sigma\theta_0^2+\theta_0\right)(\lambda_1+\sqrt{H}\eta_0+\sqrt{H}\varepsilon_\Delta)\\ \leq&(\lambda_1+\sqrt{H}\varepsilon_\Delta)\theta_0+\frac{3}{2}s\varepsilon_\Sigma\theta_0^2(\lambda_1+\sqrt{H}\eta_0+\sqrt{H}\varepsilon_\Delta).\end{aligned}$$

□

S.13.2 Proof of Lemma S.4

We cite the following lemma from Cai et al. (2022).

Lemma S.7. *(Cai et al., 2022, Lemma 6) Suppose that $\mathbf{X}_i$, $i=1,\ldots,n$, are i.i.d. centered sub-Gaussian random vectors with parameter $\kappa^2>0$. Denote $\boldsymbol{\Sigma}=\mathrm{Cov}(\mathbf{X}_i)$, we also assume that $m^{-1}\leq\varphi_{\min}(\boldsymbol{\Sigma})\leq\varphi_{\max}(\boldsymbol{\Sigma})\leq m$ for some positive constant $m>1$. For any $\varepsilon>0$ and any fixed non-zero vectors $\mathbf{u},\mathbf{v}\in\mathbb{R}^p$, we have*

$$\Pr\left(\left|\frac{1}{n}\sum_{i=1}^n\mathbf{u}^\top\mathbf{X}_i\mathbf{X}_i^\top\mathbf{v}-\mathbf{u}^\top\boldsymbol{\Sigma}\mathbf{v}\right|\geq\varepsilon\|\mathbf{u}\|_2\|\mathbf{v}\|_2\right)\leq 2\exp\left(-cn\min\left\{\frac{\varepsilon^2}{\kappa^4},\frac{\varepsilon}{\kappa^2}\right\}\right)$$

for some constant $c>0$.

Let $\mathbf{e}_j \in \mathbb{R}^p$ denote the standard basis in $\mathbb{R}^p$ with the j-th element being one and other elements being zero. Then, by Lemma S.7, taking $\mathbf{u} = \mathbf{e}_i$ and $\mathbf{v} = \mathbf{e}_j$, we obtain

$$\Pr(|\widehat{\Sigma}_{ij} - \Sigma_{ij}| \geq \varepsilon) \leq 2\exp\left(-cn\min\left\{\frac{\varepsilon^2}{\kappa^4}, \frac{\varepsilon}{\kappa^2}\right\}\right).$$

Further taking $\varepsilon \leq \kappa^2$, we obtain

$$\Pr(|\widehat{\Sigma}_{ij} - \Sigma_{ij}| \geq \varepsilon) \leq C\exp(-C'n\varepsilon^2).$$

Hence,

$$\Pr(\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{\max} \geq \varepsilon) \leq Cp^2\exp(-C'n\varepsilon^2),$$

and the proof is completed. □

S.13.3 Proof of Lemma S.5

Let $n_h = \sum_{i=1}^n I(\widetilde{Y}_i = h)$ and $\overline{\mathbf{X}}_h = n_h^{-1}\sum_{i=1}^n \mathbf{X}_i I(\widetilde{Y}_i = h)$, for $h = 1, \ldots, H$. And we denote $\boldsymbol{\mu}_h = \mathrm{E}(\mathbf{X} \mid \widetilde{Y} = h)$, for $h = 1, \ldots, H$. $\boldsymbol{\psi}_h = \mathrm{E}\{\mathbf{X}I(\widetilde{Y} = h)\}$ for $h = 1, \ldots, H$. We have

$$\begin{aligned}
\boldsymbol{\Sigma}_h =& \frac{1}{\pi_h}\mathrm{E}\{\mathbf{X}\mathbf{X}^\top I(\widetilde{Y} = h)\} - \boldsymbol{\mu}_h\boldsymbol{\mu}_h^\top \\
\widehat{\boldsymbol{\Sigma}}_h =& \frac{1}{n_h}\sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i^\top I(\widetilde{Y}_i = h) - \overline{\mathbf{X}}_h\overline{\mathbf{X}}_h^\top.
\end{aligned}$$

Rewrite $\boldsymbol{\Sigma}_h$ as

$$\boldsymbol{\Sigma}_h = \frac{1}{\pi_h}\mathrm{Cov}\{\mathbf{X}I(\widetilde{Y} = h)\} + (\pi_h - 1)\boldsymbol{\mu}_h\boldsymbol{\mu}_h^\top.$$

Define the matrix $\boldsymbol{\Xi}_h$ as

$$\boldsymbol{\Xi}_h = \frac{1}{n}\sum_{i=1}^n \left\{\mathbf{X}_i I(\widetilde{Y}_i = h) - \pi_h\boldsymbol{\mu}_h\right\}\left\{\mathbf{X}_i I(\widetilde{Y}_i = h) - \pi_h\boldsymbol{\mu}_h\right\}^\top,$$

and we rewrite $\widehat{\boldsymbol{\Sigma}}_h$ as

$$\widehat{\boldsymbol{\Sigma}}_h = \frac{n}{n_h}\boldsymbol{\Xi}_h + \overline{\mathbf{X}}_h\pi_h\boldsymbol{\mu}_h^\top + \pi_h\boldsymbol{\mu}_h\overline{\mathbf{X}}_h^\top - \frac{n}{n_h}\pi_h^2\boldsymbol{\mu}_h\boldsymbol{\mu}_h^\top - \overline{\mathbf{X}}_h\overline{\mathbf{X}}_h^\top.$$

Then, for any $j, k = 1, \ldots, p$,

$$\Pr(|\widehat{\Sigma}_{h,jk} - \Sigma_{h,jk}| \geq \varepsilon) \leq I_1 + \cdots + I_5,$$

where

$$\begin{aligned}
I_1 &= \Pr\left(\left|\frac{n}{n_h}\mathbf{e}_j^\top\boldsymbol{\Xi}_h\mathbf{e}_k - \frac{1}{\pi_h}\mathbf{e}_j^\top\mathrm{Cov}\{\mathbf{X}I(\widetilde{Y} = h)\}\mathbf{e}_k\right| \geq \frac{\varepsilon}{5}\right), \\
I_2 &= \Pr\left(\left|\overline{X}_{h,j}\pi_h\mu_{h,k} - \pi_h\mu_{h,j}\mu_{h,k}\right| \geq \frac{\varepsilon}{5}\right), \quad I_3 = \Pr\left(\left|\pi_h\mu_{h,j}\overline{X}_{h,k} - \pi_h\mu_{h,j}\mu_{h,k}\right| \geq \frac{\varepsilon}{5}\right), \\
I_4 &= \Pr\left(\left|\frac{n}{n_h}\pi_h^2\mu_{h,j}\mu_{h,k} - \pi_h\mu_{h,j}\mu_{h,k}\right| \geq \frac{\varepsilon}{5}\right), \quad I_5 = \Pr\left(\left|\overline{X}_{h,j}\overline{X}_{h,k} - \mu_{h,j}\mu_{h,k}\right| \geq \frac{\varepsilon}{5}\right),
\end{aligned}$$

and $\mathbf{e}_j$ denote the jth standard basis vector in $\mathbb{R}^p$.

(Part I) We first show the following two useful results,

(i) $\Pr\left(|n/n_h - 1/\pi_h| \geq \varepsilon\right) \leq C\exp(-C'n\varepsilon^2)$, for $h = 1, \ldots, H$.

(ii) $\Pr\left(\left|\overline{X}_{h,j} - \mu_{h,j}\right| \geq \varepsilon\right) \leq C\exp(-C'n\varepsilon^2)$, for $h = 1, \ldots, H$, and $j = 1, \ldots, p$.

According to Hoeffding's inequality, under Assumption (A4), by taking small ε, we have

$$\begin{aligned}\Pr\left(n_k < \frac{n}{\pi_k^{-1} + \varepsilon}\right) &= \Pr\left\{n_k - n\pi_k < -n\left(\pi_k - \frac{1}{\pi_k^{-1} + \varepsilon}\right)\right\} \\ &\leq \exp\left(-2n\left\{\frac{\varepsilon}{\pi_h^{-1}(\pi_h^{-1} + \varepsilon)}\right\}^2\right) \leq \exp(-Cn\varepsilon^2).\end{aligned}$$

for some positive constant C. Similarly,

$$\Pr\left(n_k > \frac{n}{\pi_h^{-1} - \varepsilon}\right) = \Pr\left\{n_k - n\pi_k > n\left(\frac{1}{\pi_k^{-1} - \varepsilon} - \pi_k\right)\right\} \leq \exp(-Cn\varepsilon^2).$$

Therefore, we obtain that for any $h = 1, \ldots, H$,

$$\Pr\left(\left|\frac{n}{n_k} - \frac{1}{\pi_h}\right| \geq \varepsilon\right) \leq C\exp(-C'n\varepsilon^2). \tag{S.24}$$

Meanwhile, we have

$$\begin{aligned}&|\overline{X}_{h,j} - \mu_{h,j}| \\ &= \left|\frac{1}{n_h}\sum_{i=1}^n X_{i,j}I(\widetilde{Y}_i = h) - \frac{1}{\pi_h}\mathrm{E}\{X_jI(\widetilde{Y} = h)\}\right| \\ &= \left|\frac{1}{n_h}\sum_{i=1}^n X_{i,j}I(\widetilde{Y}_i = h) - \frac{n}{n_h}\mathrm{E}\{X_jI(\widetilde{Y} = h)\}\right| + \left|\frac{n}{n_h}\mathrm{E}\{X_jI(\widetilde{Y} = h)\} - \frac{1}{\pi_h}\mathrm{E}\{X_jI(\widetilde{Y} = h)\}\right|.\end{aligned} \tag{S.25}$$

For the first part in (S.25), we have

$$\begin{aligned}&\Pr\left\{\left|\frac{1}{n_h}\sum_{i=1}^n X_{i,j}I(\widetilde{Y}_i = h) - \frac{n}{n_h}\mathrm{E}\{X_jI(\widetilde{Y} = h)\}\right| > \varepsilon\right\} \\ &\leq \Pr\left\{\left|\sum_{i=1}^n X_{i,j}I(\widetilde{Y}_i = h) - n\mathrm{E}\{X_jI(\widetilde{Y} = h)\}\right| > n_h\varepsilon, n_h > \pi_h n/2\right\} + \Pr(n_h \leq \pi_h n/2)\end{aligned}$$

By Hoeffding's inequality stated in Lemma S.16, under Assumption (A4), we have

$$\Pr(n_h \leq \pi_h n/2) = \Pr(n_h - \pi_h n \leq -\pi_h n/2) \leq \exp(-Cn). \tag{S.26}$$

for some positive constant C.

Since $X_{i,j}$ is sub-Gaussian, then

$$\mathrm{E}\exp\left\{\frac{(X_{i,j}I(\widetilde{Y}_i = h))^2}{\|X_{i,j}\|_{\psi_2}^2}\right\} \leq \mathrm{E}\exp\left\{\frac{X_{i,j}^2}{\|X_{i,j}\|_{\psi_2}^2}\right\} \leq 2$$

implying that

$$\|X_{i,j}I(\widetilde{Y}_i = h)\|_{\psi_2} \leq \|X_{i,j}\|_{\psi_2} < \infty.$$

Hence, $X_{i,j}I(\widetilde{Y}_i = h)$ is sub-Gaussian. Therefore, by Lemma S.17,

$$\begin{aligned}
&\Pr\left\{\left|\sum_{i=1}^{n} X_{i,j}I(\widetilde{Y}_i = h) - n\mathrm{E}\{X_jI(\widetilde{Y} = h)\}\right| > n_h\varepsilon, n_h > \pi_h n/2\right\} \\
\leq &\Pr\left\{\left|\sum_{i=1}^{n} X_{i,j}I(\widetilde{Y}_i = h) - n\mathrm{E}\{X_jI(\widetilde{Y} = h)\}\right| > \pi_h n\varepsilon/2\right\} \\
\leq &C\exp(-C'n\varepsilon^2).
\end{aligned}$$

It follows that

$$\Pr\left\{\left|\frac{1}{n_h}\sum_{i=1}^{n} X_{i,j}I(\widetilde{Y}_i = h) - \frac{n}{n_h}\mathrm{E}\{X_jI(\widetilde{Y} = h)\}\right| > \varepsilon\right\} \leq C\exp(-C'n\varepsilon^2). \tag{S.27}$$

We now derive the second part in (S.25). Since $X_{i,j}I(\widetilde{Y}_i = h)$ is sub-Gaussian, then $|\mathrm{E}(X_jI(\widetilde{Y} = h))|$ is bounded from above. It follows that,

$$\Pr\left(\left|\frac{n}{n_h}\mathrm{E}\{X_jI(\widetilde{Y} = h)\} - \frac{1}{\pi_h}\mathrm{E}\{X_jI(\widetilde{Y} = h)\}\right| \geq \varepsilon\right) \leq \Pr\left(\left|\frac{n}{n_h} - \frac{1}{\pi_h}\right| \geq C\varepsilon\right).$$

According to (S.24)

$$\Pr\left(\left|\frac{n}{n_h}\mathrm{E}\{X_jI(\widetilde{Y} = h)\} - \frac{1}{\pi_h}\mathrm{E}\{X_jI(\widetilde{Y} = h)\}\right| \geq \varepsilon\right) \leq C\exp(-C'n\varepsilon^2). \tag{S.28}$$

for some positive constants C and C'. Combining (S.25)–(S.28), it follows that

$$\Pr\left(\left|\overline{X}_{h,j} - \mu_{h,j}\right| \geq \varepsilon\right) \leq C\exp(-C'n\varepsilon^2). \tag{S.29}$$

Therefore, we have proved results (i) and (ii). Next, we derive the upper bounds for I_1–I_5 using these two results.

(Part II) For I_1, we have

$$\begin{aligned}
I_1 \leq &\Pr\left(\left|\frac{n}{n_h}\mathbf{e}_j^\top\boldsymbol{\Xi}_h\mathbf{e}_k - \frac{n}{n_k}\mathbf{e}_j^\top\mathrm{Cov}\{\mathbf{X}I(\widetilde{Y} = h)\}\mathbf{e}_k\right| \geq \varepsilon/10\right) \\
&+ \Pr\left(\left|\left(\frac{n}{n_k} - \frac{1}{\pi_h}\right)\mathbf{e}_j^\top\mathrm{Cov}\{\mathbf{X}I(\widetilde{Y} = h)\}\mathbf{e}_k\right| \geq \varepsilon/10\right).
\end{aligned} \tag{S.30}$$

For the first term on right-hand side of (S.30),

$$\begin{aligned}
&\Pr\left(\left|\frac{n}{n_h}\mathbf{e}_j^\top\boldsymbol{\Xi}_h\mathbf{e}_k - \frac{n}{n_k}\mathbf{e}_j^\top\mathrm{Cov}\{\mathbf{X}I(\widetilde{Y} = h)\}\mathbf{e}_k\right| \geq \varepsilon/10\right) \\
\leq &\Pr\left(\left|\mathbf{e}_j^\top\boldsymbol{\Xi}_h\mathbf{e}_k - \mathbf{e}_j^\top\mathrm{Cov}\{\mathbf{X}I(\widetilde{Y} = h)\}\mathbf{e}_k\right| \geq \frac{n_h\varepsilon}{10n}, n_h \geq \frac{n\pi_h}{2}\right) + \Pr\left(n_h < \frac{n\pi_h}{2}\right) \\
\leq &\Pr\left(\left|\mathbf{e}_j^\top\boldsymbol{\Xi}_h\mathbf{e}_k - \mathbf{e}_j^\top\mathrm{Cov}\{\mathbf{X}I(\widetilde{Y} = h)\}\mathbf{e}_k\right| \geq \frac{\pi_h\varepsilon}{20}\right) + \Pr\left(n_h < \frac{n\pi_h}{2}\right).
\end{aligned}$$

Under Assumption (A1), we are given that $m^{-1} \leq \varphi_i(\mathrm{Cov}\{\mathbf{X}I(\widetilde{Y} = h)\}) \leq m$ for all $i \in [p]$ and $h \in [H]$. Then, according to Lemma S.7 and under Assumption (A4), we obtain

$$\Pr\left(\left|\mathbf{e}_j^\top\boldsymbol{\Xi}_h\mathbf{e}_k - \mathbf{e}_j^\top\mathrm{Cov}\{\mathbf{X}I(\widetilde{Y} = h)\}\mathbf{e}_k\right| \geq \frac{\pi_h\varepsilon}{20}\right) \leq C\exp(-C'n\varepsilon^2)$$

for $0 < \varepsilon \leq C'$. And by (S.26), we have

$$\Pr\left(n_h < \frac{n\pi_h}{2}\right) \leq C\exp(-C'n).$$

Thus,

$$\Pr\left(\left|\frac{n}{n_h}\mathbf{e}_j^\top \mathbf{\Xi}_h \mathbf{e}_k - \frac{n}{n_k}\mathbf{e}_j^\top \mathrm{Cov}\{\mathbf{X}I(\widetilde{Y}=h)\}\mathbf{e}_k\right| \geq \varepsilon/10\right) \leq C\exp(-C'n\varepsilon^2). \tag{S.31}$$

Since

$$\begin{aligned}|\mathbf{e}_j^\top \mathrm{Cov}\{\mathbf{X}I(\widetilde{Y}=h)\}\mathbf{e}_k| \leq& |\mathrm{E}\{X_jX_kI(\widetilde{Y}=h)\}| + |\mathrm{E}\{X_jI(\widetilde{Y}=h)\}||\mathrm{E}\{X_kI(\widetilde{Y}=h)\}| \\ \leq& \mathrm{E}(|X_jX_k|) + \mathrm{E}(|X_j|)E(|X_k|),\end{aligned}$$

then $|\mathbf{e}_j^\top \mathrm{Cov}\{\mathbf{X}I(\widetilde{Y}=h)\}\mathbf{e}_k|$ is bounded from above. The second term on the right-hand side of (S.30) is

$$\begin{aligned}\Pr\left(\left|\left(\frac{n}{n_k}-\frac{1}{\pi_h}\right)\mathbf{e}_j^\top \mathrm{Cov}\{\mathbf{X}I(\widetilde{Y}=h)\}\mathbf{e}_k\right| \geq \varepsilon/10\right) \leq& \Pr\left(\left|\frac{n}{n_k}-\frac{1}{\pi_h}\right| \geq C\varepsilon\right) \\ \leq& C\exp(-C'n\varepsilon^2),\end{aligned} \tag{S.32}$$

where the second inequality follows from statement (i). Combining (S.30)–(S.32), we have

$$I_1 \leq C\exp(-C'n\varepsilon^2).$$

Since $|\mu_{h,k}| = |\pi_h^{-1}\mathrm{E}\{X_kI(\widetilde{Y}=h)\}| \leq \pi_h^{-1}E(|X_k|)$, under Assumption (A4), $|\mu_{h,k}|$ is bounded from above. Then we have

$$I_2 = \Pr\left(\left|\overline{X}_{h,j}\pi_h\mu_{h,k} - \pi_h\mu_{h,j}\mu_{h,k}\right| \geq \frac{\varepsilon}{5}\right) \leq \Pr\left(\left|\overline{X}_{h,j}-\mu_{h,j}\right| \geq C\varepsilon\right).$$

By statement (ii), it follows that

$$I_2 \leq C\exp(-C'n\varepsilon^2).$$

Similarly, we can derive

$$I_3 \leq C\exp(-C'n\varepsilon^2).$$

For I_4, by statement (i), we obtain

$$I_4 = \Pr\left(\left|\frac{n}{n_h}\pi_h^2\mu_{h,j}\mu_{h,k} - \pi_h\mu_{h,j}\mu_{h,k}\right| \geq \frac{\varepsilon}{5}\right) \leq \Pr\left(\left|\frac{n}{n_h}-\frac{1}{\pi_h}\right| \geq C\varepsilon\right) \leq C\exp(-C'n\varepsilon^2).$$

For I_5,

$$\begin{aligned}I_5 =& \Pr\left(\left|\overline{X}_{h,j}\overline{X}_{h,k} - \mu_{h,j}\mu_{h,k}\right| \geq \frac{\varepsilon}{5}\right) \\ \leq& \Pr\left(\left|(\overline{X}_{h,j}-\mu_{h,j})(\overline{X}_{h,k}-\mu_{h,k})\right| \geq \frac{\varepsilon}{15}\right) + \Pr\left(\left|\mu_{h,j}(\overline{X}_{h,k}-\mu_{h,k})\right| \geq \frac{\varepsilon}{15}\right) \\ &+ \Pr\left(\left|\mu_{h,k}(\overline{X}_{h,j}-\mu_{h,j})\right| \geq \frac{\varepsilon}{15}\right) \\ \leq& \Pr\left(\left|\overline{X}_{h,j}-\mu_{h,j}\right| \geq C\sqrt{\varepsilon}\right) + \Pr\left(\left|\overline{X}_{h,k}-\mu_{h,k}\right| \geq C\sqrt{\varepsilon}\right) \\ &+ \Pr\left(\left|\overline{X}_{h,k}-\mu_{h,k}\right| \geq C\varepsilon\right) + \Pr\left(\left|\overline{X}_{h,j}-\mu_{h,j}\right| \geq C\varepsilon\right).\end{aligned}$$

Using statement (ii), it follows that

$$I_5 \leq C \exp(-C' n\varepsilon^2).$$

By combining the upper bounds for I_1–I_5, we obtain

$$\Pr(|\widehat{\Sigma}_{h,jk} - \Sigma_{h,jk}| \geq \varepsilon) \leq C \exp(-C' n\varepsilon^2),$$

and it follows that

$$\Pr(\|\widehat{\boldsymbol{\Sigma}}_h - \boldsymbol{\Sigma}_h\|_{\max} \geq \varepsilon) \leq C p^2 \exp(-C' n\varepsilon^2),$$

which completes the proof. □

S.13.4 Proof of Lemma S.6

Let $\mathbf{E} = \widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta} = \mathbf{I} \otimes (\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})$ and $\mathbf{R} = (\boldsymbol{\Theta}_{\mathcal{SS}} + \mathbf{E}_{\mathcal{SS}})^{-1} - (\boldsymbol{\Theta}_{\mathcal{SS}})^{-1} + (\boldsymbol{\Theta}_{\mathcal{SS}})^{-1}\mathbf{E}_{\mathcal{SS}}(\boldsymbol{\Theta}_{\mathcal{SS}})^{-1}$. We show that

$$\|\mathbf{R}\|_\infty \leq \frac{3}{2} s^2 \varepsilon_\Sigma^2 \theta_0^3 \tag{S.33}$$

and the techniques used in the following arguments are similar to those used in the proof of Lemma 5 of Ravikumar et al. (2011). For any $j \in [p^2]$, we have

$$\|\mathbf{E}_{j\mathcal{S}}\|_1 = \|\{\mathbf{I} \otimes (\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})\}_{j\mathcal{S}}\|_1 \leq s\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{\max} \leq s\varepsilon_\Sigma,$$

and then

$$\|\mathbf{E}_{\mathcal{SS}}\|_\infty \leq s\varepsilon_\Sigma.$$

Consequently, by the assumption that $\theta_0 s\varepsilon_\Sigma \leq 1/3$, we have

$$\|(\boldsymbol{\Theta}_{\mathcal{SS}})^{-1}\mathbf{E}_{\mathcal{SS}}\|_\infty \leq \|(\boldsymbol{\Theta}_{\mathcal{SS}})^{-1}\|_\infty \|\mathbf{E}_{\mathcal{SS}}\|_\infty \leq \theta_0 s\varepsilon_\Sigma \leq \frac{1}{3}. \tag{S.34}$$

Then, by convergent matrix expansion,

$$\begin{aligned}
&(\boldsymbol{\Theta}_{\mathcal{SS}} + \mathbf{E}_{\mathcal{SS}})^{-1} \\
=&[\boldsymbol{\Theta}_{\mathcal{SS}}\{\mathbf{I} + (\boldsymbol{\Theta}_{\mathcal{SS}})^{-1}\mathbf{E}_{\mathcal{SS}}\}]^{-1} = \{\mathbf{I} + (\boldsymbol{\Theta}_{\mathcal{SS}})^{-1}\mathbf{E}_{\mathcal{SS}}\}^{-1}(\boldsymbol{\Theta}_{\mathcal{SS}})^{-1} = \sum_{k=0}^{\infty}(-1)^k\{(\boldsymbol{\Theta}_{\mathcal{SS}})^{-1}\mathbf{E}_{\mathcal{SS}}\}^k(\boldsymbol{\Theta}_{\mathcal{SS}})^{-1} \\
=&(\boldsymbol{\Theta}_{\mathcal{SS}})^{-1} - (\boldsymbol{\Theta}_{\mathcal{SS}})^{-1}\mathbf{E}_{\mathcal{SS}}(\boldsymbol{\Theta}_{\mathcal{SS}})^{-1} + \sum_{k=2}^{\infty}(-1)^k\{(\boldsymbol{\Theta}_{\mathcal{SS}})^{-1}\mathbf{E}_{\mathcal{SS}}\}^k(\boldsymbol{\Theta}_{\mathcal{SS}})^{-1} \\
=&(\boldsymbol{\Theta}_{\mathcal{SS}})^{-1} - (\boldsymbol{\Theta}_{\mathcal{SS}})^{-1}\mathbf{E}_{\mathcal{SS}}(\boldsymbol{\Theta}_{\mathcal{SS}})^{-1} + (\boldsymbol{\Theta}_{\mathcal{SS}})^{-1}\mathbf{E}_{\mathcal{SS}}(\boldsymbol{\Theta}_{\mathcal{SS}})^{-1}\mathbf{E}_{\mathcal{SS}}\boldsymbol{\Psi}(\boldsymbol{\Theta}_{\mathcal{SS}})^{-1},
\end{aligned}$$

where $\boldsymbol{\Psi} = \sum_{k=0}^{\infty}(-1)^k\{(\boldsymbol{\Theta}_{\mathcal{SS}})^{-1}\mathbf{E}_{\mathcal{SS}}\}^k$. Therefore,

$$\mathbf{R} = (\boldsymbol{\Theta}_{\mathcal{SS}})^{-1}\mathbf{E}_{\mathcal{SS}}(\boldsymbol{\Theta}_{\mathcal{SS}})^{-1}\mathbf{E}_{\mathcal{SS}}\boldsymbol{\Psi}(\boldsymbol{\Theta}_{\mathcal{SS}})^{-1}.$$

Then,

$$\|\mathbf{R}\|_\infty \leq \|(\boldsymbol{\Theta}_{\mathcal{SS}})^{-1}\|_\infty^3 \|\mathbf{E}_{\mathcal{SS}}\|_\infty^2 \|\boldsymbol{\Psi}\|_\infty. \tag{S.35}$$

And for $\|\boldsymbol{\Theta}\|_\infty$, by (S.34), we have

$$\|\boldsymbol{\Psi}\|_\infty \leq \sum_{k=0}^{\infty}\|(\boldsymbol{\Theta}_{\mathcal{SS}})^{-1}\mathbf{E}_{\mathcal{SS}}\|_\infty^k \leq \frac{1}{1 - \|(\boldsymbol{\Theta}_{\mathcal{SS}})^{-1}\mathbf{E}_{\mathcal{SS}}\|_\infty} \leq \frac{3}{2}. \tag{S.36}$$

Combining (S.35) and (S.36), we obtain (S.33).

Since $(\widehat{\boldsymbol{\Theta}}_{\mathcal{SS}})^{-1} - (\boldsymbol{\Theta}_{\mathcal{SS}})^{-1} = \mathbf{R} - (\boldsymbol{\Theta}_{\mathcal{SS}})^{-1}\mathbf{E}_{\mathcal{SS}}(\boldsymbol{\Theta}_{\mathcal{SS}})^{-1}$, then

$$\begin{aligned}&\|(\widehat{\boldsymbol{\Theta}}_{\mathcal{SS}})^{-1} - (\boldsymbol{\Theta}_{\mathcal{SS}})^{-1}\|_\infty \le \|\mathbf{R}\|_\infty + \|(\boldsymbol{\Theta}_{\mathcal{SS}})^{-1}\mathbf{E}_{\mathcal{SS}}(\boldsymbol{\Theta}_{\mathcal{SS}})^{-1}\|_\infty \\ \le&\frac{3}{2}s^2\varepsilon_\Sigma^2\theta_0^3 + \|\mathbf{E}_{\mathcal{SS}}\|_\infty\|(\boldsymbol{\Theta}_{\mathcal{SS}})^{-1}\|_\infty^2 \le \frac{3}{2}s^2\varepsilon_\Sigma^2\theta_0^3 + s\varepsilon_\Sigma\theta_0^2.\end{aligned}$$

□

S.14 Proof of Theorem 3

Define the event $F := \{\varphi_{\min}(\widehat{\boldsymbol{\Sigma}}_{\mathcal{AA}}) \ge K^{-1}\}$ for some $K > 0$. Let

$$I = E \cap F,$$

where the event E is defined in (S.12). We show that under event I, we have

$$\widehat{d} = d, \quad D(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}) \lesssim \sqrt{\frac{s^2 \log p}{n\sigma^2}}.$$

(Part I) Recall that $\widehat{\mathbf{B}} = (\widehat{\mathbf{B}}_1, \ldots, \widehat{\mathbf{B}}_H) \in \mathbb{R}^{\widehat{s}\times(H\widehat{s})}$ is the minimizer of the nuclear-norm penalized optimization problem (6). Denote $\mathbf{B} = (\mathbf{M}_{1,\mathcal{AA}}, \ldots, \mathbf{M}_{H,\mathcal{AA}}) \in \mathbb{R}^{s\times(Hs)}$ and let $\eta_1 = \max_h \|\boldsymbol{\Delta}_{h,\mathcal{AA}}\|_1$. We have the following result.

Lemma S.8. *Under conditions that $\widehat{\mathcal{A}} = \mathcal{A}$ and $\varphi_{\min}(\widehat{\boldsymbol{\Sigma}}_{\mathcal{AA}}) \ge K^{-1}$ for some $K > 0$, we have that $\|\widehat{\mathbf{B}} - \mathbf{B}\|_F \le 2K\{(\theta_0\eta_1\varepsilon_\Sigma + \varepsilon_\Delta)\sqrt{s^2H} + \lambda_2\sqrt{s}\}$.*

According to the proof of Theorem 2, under event E and Assumptions (A1)–(A5), we have $\widehat{\mathcal{A}} = \mathcal{A}$. Therefore, under event I and Assumptions (A1)–(A5), the conditions in Lemma S.8 hold. Moreover, by Assumptions (A3) and (A6), θ_0 and η_1 are bounded, and under event E, we have $\max\{\varepsilon_\Sigma, \varepsilon_\Delta\} \le \sqrt{c\log p/n}$. Then, by taking $0 < \lambda_2 \lesssim \sqrt{s\log p/n}$, we obtain from Lemma S.8 that

$$\|\widehat{\mathbf{B}} - \mathbf{B}\|_F \le c'\sqrt{s^2\log p/n}, \tag{S.37}$$

where $c' > 0$ is a constant depending on parameters $m, \theta_0, \eta_1, H, \pi_h, \|X_j\|_{\psi_2}$, and $\|X_j I(\widetilde{Y} = h)\|_{\psi_2}$, for $j = 1, \ldots, p$, and $h = 1, \ldots, H$.

Then, it follows that $\|\widehat{\mathbf{B}} - \mathbf{B}\| \le \|\widehat{\mathbf{B}} - \mathbf{B}\|_F \le c'\sqrt{s^2\log p/n}$. By Assumption (A7), we have $\|\widehat{\mathbf{B}} - \mathbf{B}\| < \sigma/2$. By Weyl's inequality stated in Lemma S.20, we have

$$\widehat{\sigma}_d \ge \sigma_d - \|\widehat{\mathbf{B}} - \mathbf{B}\| > \sigma/2, \quad \widehat{\sigma}_{d+1} \le \sigma_{d+1} + \|\widehat{\mathbf{B}} - \mathbf{B}\| < \sigma/2. \tag{S.38}$$

Recall the definition of $\widehat{d}$, that is,

$$\widehat{d} = |\{j \mid \widehat{\sigma}_j \ge \delta\}| \wedge d_{\max}.$$

Since we take $\delta = \sigma/2$ and let $d_{\max} \ge d$, we obtain that $\widehat{d} = d$.

(Part II) Let $\boldsymbol{\eta} \in \mathbb{R}^{s\times d}$ be the matrix consisting of top-d left singular vectors of $\mathbf{B}$ and construct $\boldsymbol{\beta} = (\boldsymbol{\eta}^\top, \mathbf{0}^\top)^\top$. Then $\boldsymbol{\beta}$ is the basis matrix of $\mathcal{S}_{\text{SAVE}}$. Recall that $\widehat{\boldsymbol{\eta}}_j$

is the left singular vector of $\widehat{\mathbf{B}}$ corresponding to the jth largest singular value and that $\widehat{\boldsymbol{\beta}}_j = (\widehat{\boldsymbol{\eta}}_j^\top, \mathbf{0}^\top)^\top$. Let $\widehat{\boldsymbol{\eta}} = (\widehat{\boldsymbol{\eta}}_1, \ldots, \widehat{\boldsymbol{\eta}}_{\widehat{d}}) \in \mathbb{R}^{\widehat{s} \times \widehat{d}}$. According to Part I, $\widehat{\mathcal{A}} = \mathcal{A}$ and $\widehat{d} = d$ under event I, then $\widehat{\boldsymbol{\eta}} = (\widehat{\boldsymbol{\eta}}_1, \ldots, \widehat{\boldsymbol{\eta}}_d) \in \mathbb{R}^{s \times d}$. Since the size of $\widehat{\boldsymbol{\eta}}$ matches that of $\boldsymbol{\eta}$ under event I, then $D(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta})$ is reduced to $D(\widehat{\boldsymbol{\eta}}, \boldsymbol{\eta})$, that is,

$$D(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}) = \|\mathbf{P}_{\widehat{\boldsymbol{\beta}}} - \mathbf{P}_{\boldsymbol{\beta}}\|_F / \sqrt{2d} = \|\mathbf{P}_{\widehat{\boldsymbol{\eta}}} - \mathbf{P}_{\boldsymbol{\eta}}\|_F / \sqrt{2d}. \tag{S.39}$$

We define the principal angles between the subspaces spanned by $\widehat{\boldsymbol{\eta}}$ and $\boldsymbol{\eta}$ as $(\theta_1, \ldots, \theta_d) = (\cos^{-1}\omega_1, \ldots, \cos^{-1}\omega_d)$, where $\omega_1 \geq \cdots \geq \omega_d$ are the singular values of $\widehat{\boldsymbol{\eta}}^\top \boldsymbol{\eta}$. Define $\sin\Theta(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}) = \mathrm{diag}(\sin\theta_1, \ldots \sin\theta_d) \in \mathbb{R}^{d \times d}$. According to Lemma 1 in Cai and Zhang (2018), $\|\mathbf{P}_{\widehat{\boldsymbol{\eta}}} - \mathbf{P}_{\boldsymbol{\eta}}\|_F / \sqrt{2d} = \|\sin\Theta(\widehat{\boldsymbol{\eta}}, \boldsymbol{\eta})\|_F / \sqrt{d}$, then we have

$$D(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}) = \|\sin\Theta(\widehat{\boldsymbol{\eta}}, \boldsymbol{\eta})\|_F / \sqrt{d}. \tag{S.40}$$

By Wedin's sin-Θ theorem stated in Lemma S.21, we have

$$\|\sin\Theta(\widehat{\boldsymbol{\eta}}, \boldsymbol{\eta})\|_F \leq \frac{\|\widehat{\mathbf{B}} - \mathbf{B}\|_F}{|\widehat{\sigma}_d - \sigma_{d+1}|} = \frac{\|\widehat{\mathbf{B}} - \mathbf{B}\|_F}{\widehat{\sigma}_d}. \tag{S.41}$$

Therefore,

$$D(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}) \leq \frac{\|\widehat{\mathbf{B}} - \mathbf{B}\|_F}{\sqrt{d}\widehat{\sigma}_d}. \tag{S.42}$$

According to (S.37), we have $\|\widehat{\mathbf{B}} - \mathbf{B}\|_F \lesssim \sqrt{s^2 \log p / n}$, and by (S.38), $\widehat{\sigma}_d > \sigma/2$. It follows that

$$D(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}) \lesssim \sqrt{\frac{s^2 \log p}{n\sigma^2}}.$$

(Part III) We show that event I holds with high probability. Since we have shown that event E holds with probability at least $1 - Cp^{-C'}$, it remains to show the high probability of event F.

By applying Theorem 10.5.11 in Vershynin (2018), we obtain that

$$0.9 \leq \varphi_{\min}^{\widehat{\boldsymbol{\Sigma}}}(s) \leq \varphi_{\max}^{\widehat{\boldsymbol{\Sigma}}}(s) \leq 1.1$$

with probability at least $1 - C\exp(-C'n)$ for some constants $C, C' > 0$, if $cs\log(ep/s) \leq n$ for some sufficiently large constant $c > 0$. Since we assume that $n \geq c_2 s^2 \log p$, by taking large enough constant $c_2 > 0$, we have $cs\log(ep/s) \leq n$ for some constant $c > 0$. Therefore,

$$\varphi_{\min}(\widehat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A}}) \geq \varphi_{\min}^{\widehat{\boldsymbol{\Sigma}}}(s) \geq 0.9$$

with probability at least $1 - C\exp(-C'n)$. Then, by the union bound, event I holds with probability at least $1 - Cp^{-C'}$.

□

S.14.1 An intermediate result for subspace estimation

The following proposition makes explicit the interaction among the subspace estimation error $D(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta})$, the bias induced by the nuclear-regularization $\lambda_2\sqrt{s}$, and the signal gap $\sigma \equiv \sigma_d$.

Proposition S.1. *Under the same conditions as in Theorem 3, with probability at least* $1 - Cp^{-C'}$*, we have*

$$D(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}) \lesssim \frac{2K\{(\theta_0\eta_1\varepsilon_\Sigma + \varepsilon_\Delta)\sqrt{s^2H} + \lambda_2\sqrt{s}\}}{\sigma}.$$

Proof of Proposition S.1. The proof is similar to that of Theorem 3. Let $I = E \cap F$, where the event E is defined in (S.12), and the event $F := \{\varphi_{\min}(\widehat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A}}) \geq K^{-1}\}$ for some $K > 0$. Under event I, by (S.42), we have

$$D(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}) \leq \frac{\|\widehat{\mathbf{B}} - \mathbf{B}\|_F}{\sqrt{d}\widehat{\sigma}_d}.$$

Moreover, by (S.38), $\widehat{\sigma}_d > \sigma/2$. Then, by Lemma S.8, it follows that

$$D(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}) \lesssim \frac{2K\{(\theta_0\eta_1\varepsilon_\Sigma + \varepsilon_\Delta)\sqrt{s^2H} + \lambda_2\sqrt{s}\}}{\sigma},$$

which completes the proof.

□

S.14.2 Proof of Lemma S.8

Under the assumption that $\widehat{\mathcal{A}} = \mathcal{A}$, we have

$$\widehat{\mathbf{B}} = \underset{\mathbf{B}_1,\ldots,\mathbf{B}_H \in \mathbb{R}^{s\times s}}{\operatorname{argmin}} \sum_{h=1}^{H} \left\{ \frac{1}{2}\operatorname{tr}(\mathbf{B}_h^\top \widehat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A}} \mathbf{B}_h) - \operatorname{tr}(\mathbf{B}_h^\top \widehat{\boldsymbol{\Delta}}_{h,\mathcal{A}\mathcal{A}}) \right\} + \lambda_2 \|(\mathbf{B}_1, \ldots, \mathbf{B}_H)\|_*,$$

Also, recall that $\mathbf{M}_{h,\mathcal{A}\mathcal{A}} = (\boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}})^{-1}\boldsymbol{\Delta}_{h,\mathcal{A}\mathcal{A}}$. We define $\mathbf{B} = (\mathbf{M}_{1,\mathcal{A}\mathcal{A}}, \ldots, \mathbf{M}_{H,\mathcal{A}\mathcal{A}}) \in \mathbb{R}^{s\times(Hs)}$, $\boldsymbol{\Delta}^{(\mathcal{A})} = (\boldsymbol{\Delta}_{1,\mathcal{A}\mathcal{A}}, \ldots, \boldsymbol{\Delta}_{H,\mathcal{A}\mathcal{A}})$, and $\widehat{\boldsymbol{\Delta}}^{(\mathcal{A})} = (\widehat{\boldsymbol{\Delta}}_{1,\mathcal{A}\mathcal{A}}, \ldots, \widehat{\boldsymbol{\Delta}}_{H,\mathcal{A}\mathcal{A}})$. Since $\widehat{\mathbf{B}}$ is the minimizer, then

$$\begin{aligned}
&\lambda_2(\|\widehat{\mathbf{B}}\|_* - \|\mathbf{B}\|_*) \\
\leq &\frac{1}{2}\operatorname{tr}(\mathbf{B}^\top \widehat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A}} \mathbf{B}) - \operatorname{tr}(\mathbf{B}^\top \widehat{\boldsymbol{\Delta}}^{(\mathcal{A})}) - \frac{1}{2}\operatorname{tr}(\widehat{\mathbf{B}}^\top \widehat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A}} \widehat{\mathbf{B}}) + \operatorname{tr}(\widehat{\mathbf{B}}^\top \widehat{\boldsymbol{\Delta}}^{(\mathcal{A})}) \\
= &\sum_{k=1}^{sH} \left(\frac{1}{2}\mathbf{B}_{*k}^\top \widehat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A}} \mathbf{B}_{*k} - \mathbf{B}_{*k}^\top \widehat{\boldsymbol{\Delta}}_{*k}^{(\mathcal{A})} - \frac{1}{2}\widehat{\mathbf{B}}_{*k}^\top \widehat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A}} \widehat{\mathbf{B}}_{*k} + \widehat{\mathbf{B}}_{*k}^\top \widehat{\boldsymbol{\Delta}}_{*k}^{(\mathcal{A})} \right) \\
= &\sum_{k=1}^{sH} \left\{ -\frac{1}{2}(\widehat{\mathbf{B}}_{*k} - \mathbf{B}_{*k})^\top \widehat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A}} (\widehat{\mathbf{B}}_{*k} - \mathbf{B}_{*k}) - \langle \widehat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A}} \mathbf{B}_{*k} - \widehat{\boldsymbol{\Delta}}_{*k}^{(\mathcal{A})}, \widehat{\mathbf{B}}_{*k} - \mathbf{B}_{*k} \rangle \right\}. \qquad \text{(S.43)}
\end{aligned}$$

Rearranging (S.43), we obtain

$$\begin{aligned}
&\sum_{k=1}^{sH}(\widehat{\mathbf{B}}_{*k}-\mathbf{B}_{*k})^\top\widehat{\mathbf{\Sigma}}_{\mathcal{A}\mathcal{A}}(\widehat{\mathbf{B}}_{*k}-\mathbf{B}_{*k})\\
\leq&-2\sum_{k=1}^{sH}\langle\widehat{\mathbf{\Sigma}}_{\mathcal{A}\mathcal{A}}\mathbf{B}_{*k}-\widehat{\mathbf{\Delta}}_{*k}^{(\mathcal{A})},\widehat{\mathbf{B}}_{*k}-\mathbf{B}_{*k}\rangle-2\lambda_2(\|\widehat{\mathbf{B}}\|_*-\|\mathbf{B}\|_*)\\
\leq&2\sum_{k=1}^{sH}\|\widehat{\mathbf{\Sigma}}_{\mathcal{A}\mathcal{A}}\mathbf{B}_{*k}-\widehat{\mathbf{\Delta}}_{*k}^{(\mathcal{A})}\|_\infty\|\widehat{\mathbf{B}}_{*k}-\mathbf{B}_{*k}\|_1+2\lambda_2\|\widehat{\mathbf{B}}-\mathbf{B}\|_*, \qquad \text{(S.44)}
\end{aligned}$$

For $\|\widehat{\mathbf{\Sigma}}_{\mathcal{A}\mathcal{A}}\mathbf{B}_{*k}-\widehat{\mathbf{\Delta}}_{*k}^{(\mathcal{A})}\|_\infty$, we have

$$\begin{aligned}
\|\widehat{\mathbf{\Sigma}}_{\mathcal{A}\mathcal{A}}\mathbf{B}_{*k}-\widehat{\mathbf{\Delta}}_{*k}^{(\mathcal{A})}\|_\infty\leq&\|\widehat{\mathbf{\Sigma}}_{\mathcal{A}\mathcal{A}}\mathbf{B}_{*k}-\mathbf{\Sigma}_{\mathcal{A}\mathcal{A}}\mathbf{B}_{*k}+\mathbf{\Sigma}_{\mathcal{A}\mathcal{A}}\mathbf{B}_{*k}-\widehat{\mathbf{\Delta}}_{*k}^{(\mathcal{A})}\|_\infty\\
\leq&\|\widehat{\mathbf{\Sigma}}_{\mathcal{A}\mathcal{A}}\mathbf{B}_{*k}-\mathbf{\Sigma}_{\mathcal{A}\mathcal{A}}\mathbf{B}_{*k}\|_\infty+\|\mathbf{\Delta}_{*k}^{(\mathcal{A})}-\widehat{\mathbf{\Delta}}_{*k}^{(\mathcal{A})}\|_\infty\\
\leq&\|\widehat{\mathbf{\Sigma}}-\mathbf{\Sigma}\|_{\max}\|\mathbf{B}_{*k}\|_1+\max_h\|\mathbf{\Delta}_h-\widehat{\mathbf{\Delta}}_h\|_{\max},
\end{aligned}$$

where we used the fact that $\mathbf{\Delta}^{(\mathcal{A})}=\mathbf{\Sigma}_{\mathcal{A}\mathcal{A}}\mathbf{B}$ in the second inequality. Moreover,

$$\|\mathbf{B}_{*k}\|_1=\|(\mathbf{\Sigma}_{\mathcal{A}\mathcal{A}})^{-1}\mathbf{\Delta}_{*k}^{(\mathcal{A})}\|_1\leq\|(\mathbf{\Sigma}_{\mathcal{A}\mathcal{A}})^{-1}\|_1\|\mathbf{\Delta}_{*k}^{(\mathcal{A})}\|_1\leq\|(\mathbf{\Sigma}_{\mathcal{A}\mathcal{A}})^{-1}\|_1\|\mathbf{\Delta}^{(\mathcal{A})}\|_1\leq\theta_0\eta_1.$$

Hence, we obtain

$$\|\widehat{\mathbf{\Sigma}}_{\mathcal{A}\mathcal{A}}\mathbf{B}_{*k}-\widehat{\mathbf{\Delta}}_{*k}^{(\mathcal{A})}\|_\infty\leq\theta_0\eta_1\varepsilon_\Sigma+\varepsilon_\Delta. \qquad \text{(S.45)}$$

By plugging (S.45) into (S.44), we obtain

$$\begin{aligned}
&\sum_{k=1}^{sH}(\widehat{\mathbf{B}}_{*k}-\mathbf{B}_{*k})^\top\widehat{\mathbf{\Sigma}}_{\mathcal{A}\mathcal{A}}(\widehat{\mathbf{B}}_{*k}-\mathbf{B}_{*k})\\
\leq&2(\theta_0\eta_1\varepsilon_\Sigma+\varepsilon_\Delta)\sqrt{s^2H}\|\widehat{\mathbf{B}}-\mathbf{B}\|_F+2\lambda_2\sqrt{s}\|\widehat{\mathbf{B}}-\mathbf{B}\|_F. \qquad \text{(S.46)}
\end{aligned}$$

On the other hand, we assume that

$$\varphi_{\min}(\widehat{\mathbf{\Sigma}}_{\mathcal{A}\mathcal{A}})\geq K^{-1}.$$

Then,

$$\begin{aligned}
\sum_{k=1}^{sH}(\widehat{\mathbf{B}}_{*k}-\mathbf{B}_{*k})^\top\widehat{\mathbf{\Sigma}}_{\mathcal{A}\mathcal{A}}(\widehat{\mathbf{B}}_{*k}-\mathbf{B}_{*k})\geq&\varphi_{\min}(\widehat{\mathbf{\Sigma}}_{\mathcal{A}\mathcal{A}})\sum_{k=1}^{sH}(\widehat{\mathbf{B}}_{*k}-\mathbf{B}_{*k})^\top(\widehat{\mathbf{B}}_{*k}-\mathbf{B}_{*k})\\
\geq&K^{-1}\|\widehat{\mathbf{B}}-\mathbf{B}\|_F^2. \qquad \text{(S.47)}
\end{aligned}$$

By combining (S.46) and (S.47), we have

$$\|\widehat{\mathbf{B}}-\mathbf{B}\|_F\leq2K\{(\theta_0\eta_1\varepsilon_\Sigma+\varepsilon_\Delta)\sqrt{s^2H}+\lambda_2\sqrt{s}\}. \qquad \text{(S.48)}$$

□

S.15 Proof of Theorem 4

Additional assumptions are listed in the following:

(A8) There exists a constant $m > 0$ such that $m^{-1} \leq \varphi_{\min}(\boldsymbol{\Sigma}_k) \leq \varphi_{\max}(\boldsymbol{\Sigma}_k) \leq m$, $\|\boldsymbol{\mu}_k - \boldsymbol{\mu}_1\|_2 \leq m$ for $k = 1, \ldots, K$.

(A9) There exists some constant $T > 0$ such that $\|\boldsymbol{\Sigma}_k^{-1} - \boldsymbol{\Sigma}_1^{-1}\|_F \leq T$, for $k = 2, \ldots, K$.

(A10) Denote $Q_{k,k'}(\mathbf{X}) = Q_k(\mathbf{X}) - Q_{k'}(\mathbf{X})$, $1 \leq k < k' \leq K$, and let $f_{Q_{k,k'}}$ denote the density function of $Q_{k,k'}(\mathbf{X})$. Assume that for some constants $\delta, M > 0$, we have $\sup_{|x| \leq \delta} f_{Q_{k,k'}}(x) < M$.

For the new observation $\mathbf{X}^*$, denote the SAVE variates given the true $\boldsymbol{\beta}$ as $\mathbf{Z}^* = \boldsymbol{\beta}^\top \mathbf{X}^*$. We rewrite the kth discriminant function $Q_k(\mathbf{Z}^*) = Q_k(\mathbf{X}^*)$ as

$$\begin{aligned} Q_k(\mathbf{Z}^*) =& \frac{1}{2}(\mathbf{X}^* - \boldsymbol{\mu}_1)^\top \boldsymbol{\beta}\{(\boldsymbol{\beta}^\top \boldsymbol{\Sigma}_k \boldsymbol{\beta})^{-1} - (\boldsymbol{\beta}^\top \boldsymbol{\Sigma}_1 \boldsymbol{\beta})^{-1}\}\boldsymbol{\beta}^\top (\mathbf{X}^* - \boldsymbol{\mu}_1) \\ & - (\boldsymbol{\mu}_k - \boldsymbol{\mu}_1)^\top \boldsymbol{\beta}(\boldsymbol{\beta}^\top \boldsymbol{\Sigma}_k \boldsymbol{\beta})^{-1}\boldsymbol{\beta}^\top (\mathbf{X}^* - \boldsymbol{\mu}_1) + \frac{1}{2}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_1)^\top \boldsymbol{\beta}(\boldsymbol{\beta}^\top \boldsymbol{\Sigma}_k \boldsymbol{\beta})^{-1}\boldsymbol{\beta}^\top (\boldsymbol{\mu}_k - \boldsymbol{\mu}_1) \\ & - \frac{1}{2}\log\left(\frac{|\boldsymbol{\beta}^\top \boldsymbol{\Sigma}_1 \boldsymbol{\beta}|}{|\boldsymbol{\beta}^\top \boldsymbol{\Sigma}_k \boldsymbol{\beta}|}\right) + \log\left(\frac{\pi_1}{\pi_k}\right). \end{aligned}$$

The estimated discriminant function $\widehat{Q}_k(\widehat{\mathbf{Z}}^*)$ can be expressed as

$$\begin{aligned} \widehat{Q}_k(\widehat{\mathbf{Z}}^*) =& \frac{1}{2}(\mathbf{X}^* - \widehat{\boldsymbol{\mu}}_1)^\top \widehat{\boldsymbol{\beta}}\{(\widehat{\boldsymbol{\beta}}^\top \widehat{\boldsymbol{\Sigma}}_k \widehat{\boldsymbol{\beta}})^{-1} - (\widehat{\boldsymbol{\beta}}^\top \widehat{\boldsymbol{\Sigma}}_1 \widehat{\boldsymbol{\beta}})^{-1}\}\widehat{\boldsymbol{\beta}}^\top (\mathbf{X}^* - \widehat{\boldsymbol{\mu}}_1) \\ & - (\widehat{\boldsymbol{\mu}}_k - \widehat{\boldsymbol{\mu}}_1)^\top \widehat{\boldsymbol{\beta}}(\widehat{\boldsymbol{\beta}}^\top \widehat{\boldsymbol{\Sigma}}_k \widehat{\boldsymbol{\beta}})^{-1}\widehat{\boldsymbol{\beta}}^\top (\mathbf{X}^* - \widehat{\boldsymbol{\mu}}_1) + \frac{1}{2}(\widehat{\boldsymbol{\mu}}_k - \widehat{\boldsymbol{\mu}}_1)^\top \widehat{\boldsymbol{\beta}}(\widehat{\boldsymbol{\beta}}^\top \widehat{\boldsymbol{\Sigma}}_k \widehat{\boldsymbol{\beta}})^{-1}\widehat{\boldsymbol{\beta}}^\top (\widehat{\boldsymbol{\mu}}_k - \widehat{\boldsymbol{\mu}}_1) \\ & - \frac{1}{2}\log\left(\frac{|\widehat{\boldsymbol{\beta}}^\top \widehat{\boldsymbol{\Sigma}}_1 \widehat{\boldsymbol{\beta}}|}{|\widehat{\boldsymbol{\beta}}^\top \widehat{\boldsymbol{\Sigma}}_k \widehat{\boldsymbol{\beta}}|}\right) + \log\left(\frac{\widehat{\pi}_1}{\widehat{\pi}_k}\right). \end{aligned}$$

We prove the consistency for the classification error by recognizing that for some $\varepsilon_0 > 0$,

$$\begin{aligned} R(\widehat{\phi}) - R(\phi^*) &\leq \Pr(\widehat{\phi}(\mathbf{X}^*) \neq \phi^*(\mathbf{X}^*)) \\ &\leq 1 - \Pr(|\widehat{Q}_k(\widehat{\mathbf{Z}}^*) - Q_k(\mathbf{Z}^*)| < \varepsilon_0, |Q_k(\mathbf{Z}^*) - Q_{k'}(\mathbf{Z}^*)| \geq 2\varepsilon_0, 1 \leq k < k' \leq K) \\ &\leq \bigcup_{k=1}^{K} \Pr(|\widehat{Q}_k(\widehat{\mathbf{Z}}^*) - Q_k(\mathbf{Z}^*)| \geq \varepsilon_0) + \bigcup_{k=1}^{K} \bigcup_{k'=k+1}^{K} \Pr(|Q_k(\mathbf{Z}^*) - Q_{k'}(\mathbf{Z}^*)| < 2\varepsilon_0) \end{aligned} \tag{S.49}$$

Since $Q_k(\mathbf{Z}^*) - Q_{k'}(\mathbf{Z}^*) = Q_k(\mathbf{X}^*) - Q_{k'}(\mathbf{X}^*) = Q_{k,k'}(\mathbf{X}^*)$, according to Assumption (A10), by taking $\varepsilon_0 \leq \delta$, we have

$$\Pr(|Q_k(\mathbf{Z}^*) - Q_{k'}(\mathbf{Z}^*)| < 2\varepsilon_0) = \Pr(|Q_{k,k'}(\mathbf{X}^*)| < 2\varepsilon_0) \leq 4M\varepsilon_0. \tag{S.50}$$

To show the high-probability upper bound of $\Pr(|\widehat{Q}_k(\widehat{\mathbf{Z}}^*) - Q_k(\mathbf{Z}^*)| \geq \varepsilon_0)$, we require the following conditions:

(D-i) $\widehat{\mathcal{A}} = \mathcal{A}$;

(D-ii) $\|\mathbf{P}_{\widehat{\boldsymbol{\beta}}} - \mathbf{P}_{\boldsymbol{\beta}}\|_F \lesssim \sqrt{s^2 \log p/(n\sigma^2)}$;

(D-iii) $\|\widehat{\boldsymbol{\Sigma}}_{k,\mathcal{A}\mathcal{A}} - \boldsymbol{\Sigma}_{k,\mathcal{A}\mathcal{A}}\| \lesssim \sqrt{s^2 \log p/n}$, $k = 1, \ldots, K$;

(D-iv) $\|\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_k\|_\infty \lesssim \sqrt{\log p/n}$, $k = 1, \ldots, K$;

(D-v) $|\widehat{\pi}_k - \pi_k| \lesssim \sqrt{\log p/n}$, $k = 1, \ldots, K$.

The following lemma is needed.

Lemma S.9. *Under Conditions (D-i)–(D-v) and Assumptions (A4)(A8)(A9), we have*

$$\begin{aligned}&\Pr(|\widehat{Q}_k(\widehat{\mathbf{Z}}^*) - Q_k(\mathbf{Z}^*)| \geq \varepsilon_0)\\ \lesssim &\exp\left\{-C\min\left(\frac{\varepsilon_0^2}{s^2\log p/(n\sigma^2)}, \frac{\varepsilon_0}{\sqrt{s^2\log p/(n\sigma^2)}}\right)\right\} + \frac{s^2\log p/(n\sigma^2)}{\{\varepsilon_0 - \sqrt{s^2\log p/(n\sigma^2)}\}^2}.\end{aligned}$$

By Lemma S.9 and the results in (S.49) and (S.50), we have

$$\begin{aligned}&R(\widehat{\phi}) - R(\phi^*)\\ \lesssim &\exp\left\{-C\min\left(\frac{\varepsilon_0^2}{s^2\log p/(n\sigma^2)}, \frac{\varepsilon_0}{\sqrt{s^2\log p/(n\sigma^2)}}\right)\right\} + \frac{s^2\log p/(n\sigma^2)}{\{\varepsilon_0 - \sqrt{s^2\log p/(n\sigma^2)}\}^2} + \varepsilon_0.\end{aligned}$$

By taking $\varepsilon_0 = C\{s^2\log p/(n\sigma^2)\}^{1/3}$ for some constant $C > 0$ such that $\varepsilon_0 \geq 2\sqrt{s^2\log p/(n\sigma^2)}$, it follows that

$$R(\widehat{\phi}) - R(\phi^*) \lesssim \left(\frac{s^2\log p}{n\sigma^2}\right)^{1/3} + \exp\left\{-C\left(\frac{s^2\log p}{n\sigma^2}\right)^{-1/6}\right\} \lesssim \left(\frac{s^2\log p}{n\sigma^2}\right)^{1/3}.$$

It remains to prove that Conditions (D-i)–(D-v) hold with high probability.

We first show that under QDA model, Assumption (A1) is implied by Assumption (A8). Under QDA model, $\mathbf{X}$ follows the gaussian mixture model, which is clearly the sub-Gaussian distribution. And according to Lemma S.22, under Assumption (A8), the eigenvalues of $\mathrm{Cov}\{\mathbf{X}I(\widetilde{Y} = h)\}$ are bounded. For $\boldsymbol{\Sigma}$, since

$$\boldsymbol{\Sigma} = \mathrm{E}\{\mathrm{Cov}(\mathbf{X} \mid Y)\} + \mathrm{Cov}\{\mathrm{E}(\mathbf{X} \mid Y)\} = \sum_{k=1}^K \pi_k\boldsymbol{\Sigma}_k + \sum_{k=1}^K \pi_k(\boldsymbol{\mu}_k - \boldsymbol{\mu})(\boldsymbol{\mu}_k - \boldsymbol{\mu})^\top.$$

Therefore,

$$\varphi_{\min}(\boldsymbol{\Sigma}) \geq \sum_{k=1}^K \pi_k\varphi_{\min}(\boldsymbol{\Sigma}_k) \geq m^{-1},$$

and

$$\varphi_{\max}(\boldsymbol{\Sigma}) \leq \sum_{k=1}^K \pi_k\varphi_{\max}(\boldsymbol{\Sigma}_k) + \sum_{k=1}^K \pi_k\|\boldsymbol{\mu}_k - \boldsymbol{\mu}\|_2^2 \leq m + 4m^2.$$

Therefore, Assumption (A1) holds.

Under Assumptions (A2)–(A8), by Theorems 2 and 3, we have that $\widehat{\mathcal{A}} = \mathcal{A}$ and that

$$\|\mathbf{P}_{\widehat{\boldsymbol{\beta}}} - \mathbf{P}_{\boldsymbol{\beta}}\|_F \lesssim \sqrt{\frac{s^2\log p}{n\sigma^2}}$$

with probability at least $1 - Cp^{-C'}$. For $\|\widehat{\boldsymbol{\Sigma}}_{k,\mathcal{A}\mathcal{A}} - \boldsymbol{\Sigma}_{k,\mathcal{A}\mathcal{A}}\|$, since

$$\|\widehat{\boldsymbol{\Sigma}}_{k,\mathcal{A}\mathcal{A}} - \boldsymbol{\Sigma}_{k,\mathcal{A}\mathcal{A}}\| \leq s\|\widehat{\boldsymbol{\Sigma}}_k - \boldsymbol{\Sigma}_k\|_{\max}.$$

By Lemma S.5, we obtain that

$$\|\widehat{\boldsymbol{\Sigma}}_{k,\mathcal{A}\mathcal{A}} - \boldsymbol{\Sigma}_{k,\mathcal{A}\mathcal{A}}\| \lesssim \sqrt{\frac{s^2 \log p}{n}}$$

with probability at least $1 - Cp^{-C'}$. Similar to the proof of (S.29), we derive

$$\Pr(|\widehat{\mu}_{k,j} - \mu_{k,j}| \geq \varepsilon) \leq C \exp(-C' n\varepsilon^2).$$

Then

$$\|\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_k\|_\infty \lesssim \sqrt{\frac{\log p}{n}}$$

with probability at least $1 - Cp^{-C'}$. Moreover, under Assumption (A4), by Hoeffding's inequality stated in Lemma S.16, we have that

$$\Pr(|\widehat{\pi}_k - \pi_k| \geq \varepsilon) \leq 2 \exp(-Cn\varepsilon^2).$$

Then, we have

$$|\widehat{\pi}_k - \pi_k| \lesssim \sqrt{\frac{\log p}{n}}$$

with probability at least $1 - Cp^{-C'}$.

Therefore, Conditions (D-i)–(D-v) hold with probability at least $1 - Cp^{-C'}$, which completes the proof. □

S.15.1 Proof of Lemma S.9

Recall that $\widehat{\pi}_k = n_k/n$, $\widehat{\boldsymbol{\Psi}}_k = \widehat{\boldsymbol{\beta}}^\top \widehat{\boldsymbol{\Sigma}}_k \widehat{\boldsymbol{\beta}}$, and $\widehat{\boldsymbol{\eta}}_k = \widehat{\boldsymbol{\beta}}^\top \overline{\mathbf{X}}_k$. Define following notations,

$$\boldsymbol{\Psi}_k = \boldsymbol{\beta}^\top \boldsymbol{\Sigma}_k \boldsymbol{\beta}, \quad \mathbf{D}_k = \boldsymbol{\Psi}_k^{-1} - \boldsymbol{\Psi}_1^{-1}, \quad \widehat{\mathbf{D}}_k = \widehat{\boldsymbol{\Psi}}_k^{-1} - \widehat{\boldsymbol{\Psi}}_1^{-1}, \quad \mathbf{O} = \mathbf{O}_2 \mathbf{O}_1^\top,$$

where $\mathbf{O}_1, \mathbf{O}_2 \in \mathbb{R}^{d \times d}$ are matrices composed of left and right singular value vectors of $\widehat{\boldsymbol{\beta}}^\top \boldsymbol{\beta}$. We first introduce the following helpful lemmas to facilitate the proof.

Lemma S.10. *Let $\mathbf{O}_1, \mathbf{O}_2$ denote the matrices composed of left and right singular vectors of $\widehat{\boldsymbol{\beta}}^\top \boldsymbol{\beta}$, and $\mathbf{O} = \mathbf{O}_2 \mathbf{O}_1^\top$, then we have*

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\mathbf{O}\|_F \leq \|\mathbf{P}_{\widehat{\boldsymbol{\beta}}} - \mathbf{P}_{\boldsymbol{\beta}}\|_F.$$

Lemma S.11. *Under Assumption (A8), we have*

$$\|\boldsymbol{\Psi}_k\| \leq m, \quad \|\boldsymbol{\Psi}_k^{-1}\| \leq m, \quad k = 1, \ldots, K.$$

Moreover, by further assuming Assumption (A9), we have

$$\|\mathbf{D}_k\|_F \leq m^4 M, \quad k = 1, \ldots, K.$$

Lemma S.12. *Under Conditions (D-i)–(D-iii) and Assumption (A8), assuming that $c_1 s^2 \log p \leq n\sigma^2$ for some sufficiently large constant $c_1 > 0$, we have*

$$\|\widehat{\boldsymbol{\beta}} \widehat{\boldsymbol{\Psi}}_k^{-1} \widehat{\boldsymbol{\beta}}^\top - \boldsymbol{\beta} \boldsymbol{\Psi}_k^{-1} \boldsymbol{\beta}^\top\|_F \lesssim \sqrt{\frac{s^2 \log p}{n\sigma^2}}, \quad k = 1, \ldots, K.$$

Lemma S.13. *Under Conditions (D-i)–(D-iii), assuming that $c_1 s^2 \log p \le n\sigma^2$ for some sufficiently large constant $c_1 > 0$, we have*

$$\|\widehat{\mathbf{\Psi}}_k - \mathbf{O}^\top \mathbf{\Psi}_k \mathbf{O}\|_F \lesssim \sqrt{\frac{s^2 \log p}{n\sigma^2}}, \quad k = 1, \ldots, K.$$

Moreover, by further assuming Assumption (A8), we have

$$\|\widehat{\mathbf{\Psi}}_k^{-1} - \mathbf{O}^\top \mathbf{\Psi}_k^{-1} \mathbf{O}\|_F \lesssim \sqrt{\frac{s^2 \log p}{n\sigma^2}}, \quad k = 1, \ldots, K.$$

Since

$$\Pr(|\widehat{Q}_k(\widehat{\mathbf{Z}}^*) - Q_k(\mathbf{Z}^*)| \ge \varepsilon_0) = \sum_{k'=1}^{K} \pi_{k'} \Pr(|\widehat{Q}_k(\widehat{\mathbf{Z}}^*) - Q_k(\mathbf{Z}^*)| \ge \varepsilon_0 \mid Y^* = k'), \tag{S.51}$$

we decompose $\widehat{Q}_k(\widehat{\mathbf{Z}}^*) - Q_k(\mathbf{Z}^*)$ as

$$\widehat{Q}_k(\widehat{\mathbf{Z}}^*) - Q_k(\mathbf{Z}^*) = h_1(\mathbf{X}^*) + h_2(\mathbf{X}^*) + g_1 + g_2 + g_3 + g_4 + g_5 + g_6 + g_7,$$

where

$$\begin{aligned}
h_1(\mathbf{X}^*) =& \frac{1}{2}\left[(\mathbf{X}^* - \boldsymbol{\mu}_{k'})^\top (\widehat{\boldsymbol{\beta}}\widehat{\mathbf{D}}_k\widehat{\boldsymbol{\beta}}^\top - \boldsymbol{\beta}\mathbf{D}_k\boldsymbol{\beta}^\top)(\mathbf{X}^* - \boldsymbol{\mu}_{k'}) - \mathrm{tr}\{\mathbf{\Sigma}_{k'}^{1/2}(\widehat{\boldsymbol{\beta}}\widehat{\mathbf{D}}_k\widehat{\boldsymbol{\beta}}^\top - \boldsymbol{\beta}\mathbf{D}_k\boldsymbol{\beta}^\top)\mathbf{\Sigma}_{k'}^{1/2}\}\right], \\
h_2(\mathbf{X}^*) =& (\mathbf{X}^* - \boldsymbol{\mu}_{k'})^\top \{(\boldsymbol{\mu}_{k'} - \widehat{\boldsymbol{\mu}}_1)^\top \widehat{\boldsymbol{\beta}}\widehat{\mathbf{D}}_k\widehat{\boldsymbol{\beta}}^\top - (\boldsymbol{\mu}_{k'} - \boldsymbol{\mu}_1)^\top \boldsymbol{\beta}\mathbf{D}_k\boldsymbol{\beta}^\top - (\widehat{\boldsymbol{\mu}}_k - \widehat{\boldsymbol{\mu}}_1)^\top \widehat{\boldsymbol{\beta}}\widehat{\mathbf{\Psi}}_k^{-1}\widehat{\boldsymbol{\beta}}^\top \\
& + (\boldsymbol{\mu}_k - \boldsymbol{\mu}_1)^\top \boldsymbol{\beta}\mathbf{\Psi}_k^{-1}\boldsymbol{\beta}^\top\}, \\
g_1 =& \frac{1}{2}\left[\mathrm{tr}\{\mathbf{\Sigma}_{k'}^{1/2}(\widehat{\boldsymbol{\beta}}\widehat{\mathbf{D}}_k\widehat{\boldsymbol{\beta}}^\top - \boldsymbol{\beta}\mathbf{D}_k\boldsymbol{\beta}^\top)\mathbf{\Sigma}_{k'}^{1/2}\} - \mathrm{tr}\{(\widehat{\mathbf{D}}_k - \mathbf{O}^\top\mathbf{D}_k\mathbf{O})(\mathbf{O}^\top\mathbf{\Psi}_1\mathbf{O})\}\right], \\
g_2 =& \frac{1}{2}\{(\boldsymbol{\mu}_{k'} - \widehat{\boldsymbol{\mu}}_1)^\top \widehat{\boldsymbol{\beta}}\widehat{\mathbf{D}}_k\widehat{\boldsymbol{\beta}}^\top(\boldsymbol{\mu}_{k'} - \widehat{\boldsymbol{\mu}}_1) - (\boldsymbol{\mu}_{k'} - \boldsymbol{\mu}_1)^\top \boldsymbol{\beta}\mathbf{D}_k\boldsymbol{\beta}^\top(\boldsymbol{\mu}_{k'} - \boldsymbol{\mu}_1)\}, \\
g_3 =& \frac{1}{2}\{(\widehat{\boldsymbol{\mu}}_k - \widehat{\boldsymbol{\mu}}_1)^\top \widehat{\boldsymbol{\beta}}\widehat{\mathbf{\Psi}}_k^{-1}\widehat{\boldsymbol{\beta}}^\top(\widehat{\boldsymbol{\mu}}_k - \widehat{\boldsymbol{\mu}}_1) - (\boldsymbol{\mu}_k - \boldsymbol{\mu}_1)^\top \boldsymbol{\beta}\mathbf{\Psi}_k^{-1}\boldsymbol{\beta}^\top(\boldsymbol{\mu}_k - \boldsymbol{\mu}_1)\}, \\
g_4 =& -(\widehat{\boldsymbol{\mu}}_k - \widehat{\boldsymbol{\mu}}_1)^\top \widehat{\boldsymbol{\beta}}\widehat{\mathbf{\Psi}}_k^{-1}\widehat{\boldsymbol{\beta}}^\top(\boldsymbol{\mu}_1 - \widehat{\boldsymbol{\mu}}_1), \\
g_5 =& -\{(\widehat{\boldsymbol{\mu}}_k - \widehat{\boldsymbol{\mu}}_1)^\top \widehat{\boldsymbol{\beta}}\widehat{\mathbf{\Psi}}_k^{-1}\widehat{\boldsymbol{\beta}}^\top - (\boldsymbol{\mu}_k - \boldsymbol{\mu}_1)^\top \boldsymbol{\beta}\mathbf{\Psi}_k^{-1}\boldsymbol{\beta}^\top\}(\boldsymbol{\mu}_{k'} - \boldsymbol{\mu}_1), \\
g_6 =& -\frac{1}{2}\left[\log(|\widehat{\mathbf{\Psi}}_1|/|\widehat{\mathbf{\Psi}}_k|) - \log(|\mathbf{\Psi}_1|/|\mathbf{\Psi}_k|) - \mathrm{tr}\{(\widehat{\mathbf{D}}_k - \mathbf{O}^\top\mathbf{D}_k\mathbf{O})(\mathbf{O}^\top\mathbf{\Psi}_1\mathbf{O})\}\right], \\
g_7 =& \log(\widehat{\pi}_1/\widehat{\pi}_k) - \log(\pi_1/\pi_k).
\end{aligned}$$

It follows that

$$\begin{aligned}
& \Pr(|\widehat{Q}_k(\widehat{\mathbf{Z}}^*) - Q_k(\mathbf{Z}^*)| \ge \varepsilon_0 \mid Y^* = k') \\
\le & \Pr(|h_1(\mathbf{X}^*)| \ge \varepsilon_0/2 \mid Y^* = k') + \Pr\left(\left|h_2(\mathbf{X}^*) + \sum_{j=1}^{7} g_j\right| \ge \varepsilon_0/2 \middle| Y^* = k'\right) \\
=: & L_1 + L_2.
\end{aligned} \tag{S.52}$$

Part I: Upper bound of L_1.

Rewrite $\mathbf{X}^* \mid (Y^* = k') = \boldsymbol{\mu}_{k'} + \boldsymbol{\Sigma}_{k'}^{1/2}\mathbf{U}$, where $\mathbf{U} \sim N(\mathbf{0}, \mathbf{I}_p)$, and let

$$\mathbf{M} = \boldsymbol{\Sigma}_{k'}^{1/2}(\widehat{\boldsymbol{\beta}}\widehat{\mathbf{D}}_k\widehat{\boldsymbol{\beta}}^\top - \boldsymbol{\beta}\mathbf{D}_k\boldsymbol{\beta}^\top)\boldsymbol{\Sigma}_{k'}^{1/2},$$

then

$$(\mathbf{X}^* - \boldsymbol{\mu}_{k'})^\top(\widehat{\boldsymbol{\beta}}\widehat{\mathbf{D}}_k\widehat{\boldsymbol{\beta}}^\top - \boldsymbol{\beta}\mathbf{D}_k\boldsymbol{\beta}^\top)(\mathbf{X}^* - \boldsymbol{\mu}_{k'}) \mid (Y^* = k') \sim \mathbf{U}^\top\mathbf{M}\mathbf{U}.$$

Moreover,

$$\mathrm{E}(\mathbf{U}^\top\mathbf{M}\mathbf{U}) = \mathrm{tr}(\mathbf{M}) = \mathrm{tr}\{\boldsymbol{\Sigma}_{k'}^{1/2}(\widehat{\boldsymbol{\beta}}\widehat{\mathbf{D}}_k\widehat{\boldsymbol{\beta}}^\top - \boldsymbol{\beta}\mathbf{D}_k\boldsymbol{\beta}^\top)\boldsymbol{\Sigma}_{k'}^{1/2}\}.$$

Therefore, by Hanson-Wright inequality stated in Lemma S.19, we have

$$\begin{aligned} L_1 =& \Pr(|h_1(\mathbf{X}^*)| \geq \varepsilon_0/2 \mid Y^* = k') = \Pr(|\mathbf{U}^\top\mathbf{M}\mathbf{U} - \mathrm{E}(\mathbf{U}^\top\mathbf{M}\mathbf{U})| \geq \varepsilon_0) \\ &\leq 2\exp\left\{-C\min\left(\frac{\varepsilon_0^2}{\|\mathbf{M}\|_F^2}, \frac{\varepsilon_0}{\|\mathbf{M}\|}\right)\right\}. \end{aligned} \tag{S.53}$$

In addition,

$$\|\mathbf{M}\|_F^2 \lesssim \|\widehat{\boldsymbol{\beta}}\widehat{\mathbf{D}}_k\widehat{\boldsymbol{\beta}}^\top - \boldsymbol{\beta}\mathbf{D}_k\boldsymbol{\beta}^\top\|_F^2$$

and

$$\|\mathbf{M}\| \lesssim \|\widehat{\boldsymbol{\beta}}\widehat{\mathbf{D}}_k\widehat{\boldsymbol{\beta}}^\top - \boldsymbol{\beta}\mathbf{D}_k\boldsymbol{\beta}^\top\|.$$

According to Lemma S.12, it follows that

$$\|\widehat{\boldsymbol{\beta}}\widehat{\mathbf{D}}_k\widehat{\boldsymbol{\beta}}^\top - \boldsymbol{\beta}\mathbf{D}_k\boldsymbol{\beta}^\top\|_F \leq \|\widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\Psi}}_k^{-1}\widehat{\boldsymbol{\beta}}^\top - \boldsymbol{\beta}\boldsymbol{\Psi}_k^{-1}\boldsymbol{\beta}^\top\|_F + \|\widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\Psi}}_1^{-1}\widehat{\boldsymbol{\beta}}^\top - \boldsymbol{\beta}\boldsymbol{\Psi}_1^{-1}\boldsymbol{\beta}^\top\|_F \lesssim \sqrt{\frac{s^2\log p}{n\sigma^2}}.$$

Then,

$$L_1 \leq 2\exp\left\{-C\min\left(\frac{\varepsilon_0^2}{s^2\log p/(n\sigma^2)}, \frac{\varepsilon_0}{\sqrt{s^2\log p/(n\sigma^2)}}\right)\right\}. \tag{S.54}$$

Part II: Upper bound of L_2.

Given $Y^* = k'$, we have

$$h_2(\mathbf{X}^*) \mid (Y^* = k') \sim N(0, \boldsymbol{\theta}_{k',k}^\top\boldsymbol{\Sigma}_{k'}\boldsymbol{\theta}_{k',k}),$$

where

$$\boldsymbol{\theta}_{k',k} = \widehat{\boldsymbol{\beta}}\widehat{\mathbf{D}}_k\widehat{\boldsymbol{\beta}}^\top(\boldsymbol{\mu}_{k'} - \widehat{\boldsymbol{\mu}}_1) - \boldsymbol{\beta}\mathbf{D}_k\boldsymbol{\beta}^\top(\boldsymbol{\mu}_{k'} - \boldsymbol{\mu}_1) + \widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\Psi}}_k^{-1}\widehat{\boldsymbol{\beta}}^\top(\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_k) - \boldsymbol{\beta}\boldsymbol{\Psi}_k^{-1}\boldsymbol{\beta}^\top(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_k).$$

Then, by Markov inequality, when $|\sum_{j=1}^7 g_j| < \varepsilon_0/2$, which will be shown later, we obtain that

$$\Pr\left(\left|h_2(\mathbf{X}^*) + \sum_{j=1}^7 g_j\right| \geq \varepsilon_0/2 \middle| Y^* = k'\right) \leq \frac{\boldsymbol{\theta}_{k',k}^\top\boldsymbol{\Sigma}_h\boldsymbol{\theta}_{k',k}}{(\varepsilon_0/2 - |\sum_{j=1}^7 g_j|)^2} \leq \frac{\|\boldsymbol{\theta}_{k',k}\|_2^2\|\boldsymbol{\Sigma}_{k'}\|}{(\varepsilon_0/2 - |\sum_{j=1}^7 g_j|)^2} \tag{S.55}$$

(i) Upper bound of $\|\boldsymbol{\theta}_{k',k}\|_2$. Under Condition (D-i) that $\widehat{\mathcal{A}} = \mathcal{A}$, we have

$$\begin{aligned}
&\|\widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\Psi}}_k^{-1}\widehat{\boldsymbol{\beta}}^\top(\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_k) - \boldsymbol{\beta}\boldsymbol{\Psi}_k^{-1}\boldsymbol{\beta}^\top(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_k)\|_2 \\
=&\|\widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\Psi}}_k^{-1}\widehat{\boldsymbol{\beta}}^\top\{(\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_k) - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_k)\} + (\widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\Psi}}_k^{-1}\widehat{\boldsymbol{\beta}}^\top - \boldsymbol{\beta}\boldsymbol{\Psi}_k^{-1}\boldsymbol{\beta}^\top)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_k)\|_2 \\
\leq&\|(\widehat{\boldsymbol{\mu}}_{1,\mathcal{A}*} - \boldsymbol{\mu}_{1,\mathcal{A}*}) - (\widehat{\boldsymbol{\mu}}_{k,\mathcal{A}*} - \boldsymbol{\mu}_{k,\mathcal{A}*})\|_2\|(\widehat{\boldsymbol{\beta}}_{\mathcal{A}*}\widehat{\boldsymbol{\Psi}}_k^{-1}\widehat{\boldsymbol{\beta}}_{\mathcal{A}*}^\top - \boldsymbol{\beta}_{\mathcal{A}*}\boldsymbol{\Psi}_k^{-1}\boldsymbol{\beta}_{\mathcal{A}*}^\top) + \boldsymbol{\beta}_{\mathcal{A}*}\boldsymbol{\Psi}_k^{-1}\boldsymbol{\beta}_{\mathcal{A}*}^\top\| \\
&+ \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_k\|_2\|\widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\Psi}}_k^{-1}\widehat{\boldsymbol{\beta}}^\top - \boldsymbol{\beta}\boldsymbol{\Psi}_k^{-1}\boldsymbol{\beta}^\top\|.
\end{aligned}$$

Since $\|\widehat{\boldsymbol{\mu}}_{k,\mathcal{A}*} - \boldsymbol{\mu}_{k,\mathcal{A}*}\|_2 \leq \sqrt{s}\|\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_k\|_\infty$, then by Condition (D-iv), we have

$$\|\widehat{\boldsymbol{\mu}}_{k,\mathcal{A}*} - \boldsymbol{\mu}_{k,\mathcal{A}*}\|_2 \lesssim \sqrt{s\log p/n}.$$

By Assumption (A8), we have $\|\boldsymbol{\mu}_{k,\mathcal{A}*} - \boldsymbol{\mu}_{1,\mathcal{A}*}\|_2 \leq \|\boldsymbol{\mu}_k - \boldsymbol{\mu}_1\|_2 \leq m$. And by Lemma S.11, under Assumption (A8), we have $\|\boldsymbol{\beta}_{\mathcal{A}*}\boldsymbol{\Psi}_k^{-1}\boldsymbol{\beta}_{\mathcal{A}*}^\top\| \leq \|\boldsymbol{\Psi}_k^{-1}\| \leq m$. Moreover, according to Lemma S.12, we obtain

$$\|\widehat{\boldsymbol{\beta}}_{\mathcal{A}*}\widehat{\boldsymbol{\Psi}}_k^{-1}\widehat{\boldsymbol{\beta}}_{\mathcal{A}*}^\top - \boldsymbol{\beta}_{\mathcal{A}*}\boldsymbol{\Psi}_k^{-1}\boldsymbol{\beta}_{\mathcal{A}*}^\top\| \leq \|\widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\Psi}}_k^{-1}\widehat{\boldsymbol{\beta}}^\top - \boldsymbol{\beta}\boldsymbol{\Psi}_k^{-1}\boldsymbol{\beta}^\top\|_F \lesssim \sqrt{\frac{s^2\log p}{n\sigma^2}}.$$

Thus,

$$\|\widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\Psi}}_k^{-1}\widehat{\boldsymbol{\beta}}^\top(\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_k) - \boldsymbol{\beta}\boldsymbol{\Psi}_k^{-1}\boldsymbol{\beta}^\top(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_k)\|_2 \lesssim \sqrt{\frac{s^2\log p}{n\sigma^2}}. \tag{S.56}$$

In addition,

$$\begin{aligned}
&\|\widehat{\boldsymbol{\beta}}\widehat{\mathbf{D}}_k\widehat{\boldsymbol{\beta}}^\top(\boldsymbol{\mu}_{k'} - \widehat{\boldsymbol{\mu}}_1) - \boldsymbol{\beta}\mathbf{D}_k\boldsymbol{\beta}^\top(\boldsymbol{\mu}_{k'} - \boldsymbol{\mu}_1)\|_2 \\
\leq&\|\widehat{\boldsymbol{\beta}}\widehat{\mathbf{D}}_k\widehat{\boldsymbol{\beta}}^\top - \boldsymbol{\beta}\mathbf{D}_k\boldsymbol{\beta}^\top\|\|\boldsymbol{\mu}_{k'} - \boldsymbol{\mu}_1\|_2 + (\|\widehat{\boldsymbol{\beta}}_{\mathcal{A}*}\widehat{\mathbf{D}}_k\widehat{\boldsymbol{\beta}}_{\mathcal{A}*}^\top - \boldsymbol{\beta}_{\mathcal{A}*}\mathbf{D}_k\boldsymbol{\beta}_{\mathcal{A}*}^\top\| + \|\boldsymbol{\beta}_{\mathcal{A}*}\mathbf{D}_k\boldsymbol{\beta}_{\mathcal{A}*}^\top\|)\|\boldsymbol{\mu}_{1,\mathcal{A}*} - \widehat{\boldsymbol{\mu}}_{1,\mathcal{A}*}\|_2.
\end{aligned}$$

By Lemma S.12, we have

$$\|\widehat{\boldsymbol{\beta}}\widehat{\mathbf{D}}_k\widehat{\boldsymbol{\beta}}^\top - \boldsymbol{\beta}\mathbf{D}_k\boldsymbol{\beta}^\top\|_F \leq \|\widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\Psi}}_k^{-1}\widehat{\boldsymbol{\beta}}^\top - \boldsymbol{\beta}\boldsymbol{\Psi}_k^{-1}\boldsymbol{\beta}^\top\|_F + \|\widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\Psi}}_1^{-1}\widehat{\boldsymbol{\beta}}^\top - \boldsymbol{\beta}\boldsymbol{\Psi}_1^{-1}\boldsymbol{\beta}^\top\|_F \lesssim \sqrt{\frac{s^2\log p}{n\sigma^2}}.$$

By Lemma S.11, under Assumption (A9), we have

$$\|\boldsymbol{\beta}_{\mathcal{A}*}\mathbf{D}_k\boldsymbol{\beta}_{\mathcal{A}*}^\top\| \leq \|\mathbf{D}_k\| \leq \|\mathbf{D}_k\|_F \leq m^4M.$$

Then, it follows that

$$\|\widehat{\boldsymbol{\beta}}\widehat{\mathbf{D}}_k\widehat{\boldsymbol{\beta}}^\top(\boldsymbol{\mu}_{k'} - \widehat{\boldsymbol{\mu}}_1) - \boldsymbol{\beta}\mathbf{D}_k\boldsymbol{\beta}^\top(\boldsymbol{\mu}_{k'} - \boldsymbol{\mu}_1)\|_2 \lesssim \sqrt{\frac{s^2\log p}{n\sigma^2}}. \tag{S.57}$$

Therefore, by combining (S.56) and (S.57), it follows that

$$\|\boldsymbol{\theta}_{k',k}\|_2 \lesssim \sqrt{\frac{s^2\log p}{n\sigma^2}}.$$

(ii) Upper bound of $|g_1|$. Since

$$\begin{aligned}
&\mathrm{tr}\{\boldsymbol{\Sigma}_{k'}^{1/2}(\widehat{\boldsymbol{\beta}}\widehat{\mathbf{D}}_k\widehat{\boldsymbol{\beta}}^\top - \boldsymbol{\beta}\mathbf{D}_k\boldsymbol{\beta}^\top)\boldsymbol{\Sigma}_{k'}^{1/2}\} = \mathrm{tr}(\widehat{\mathbf{D}}_k\widehat{\boldsymbol{\beta}}^\top\boldsymbol{\Sigma}_{k'}\widehat{\boldsymbol{\beta}}) - \mathrm{tr}\{(\mathbf{O}^\top\mathbf{D}_k\mathbf{O})(\mathbf{O}^\top\boldsymbol{\Psi}_{k'}\mathbf{O})\} \\
=&\mathrm{tr}\{\widehat{\mathbf{D}}_k(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\mathbf{O})^\top\boldsymbol{\Sigma}_{k'}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\mathbf{O})\} + 2\mathrm{tr}\{\widehat{\mathbf{D}}_k(\boldsymbol{\beta}\mathbf{O})^\top\boldsymbol{\Sigma}_{k'}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\mathbf{O})\} \\
&+ \mathrm{tr}\{(\widehat{\mathbf{D}}_k - \mathbf{O}^\top\mathbf{D}_k\mathbf{O})(\mathbf{O}^\top\boldsymbol{\Psi}_{k'}\mathbf{O})\}.
\end{aligned}$$

Then,

$$\begin{aligned}
|g_1| =& \frac{1}{2}\left|\mathrm{tr}\left\{\boldsymbol{\Sigma}_{k'}^{1/2}(\widehat{\boldsymbol{\beta}}\widehat{\mathbf{D}}_k\widehat{\boldsymbol{\beta}}^\top - \boldsymbol{\beta}\mathbf{D}_k\boldsymbol{\beta}^\top)\boldsymbol{\Sigma}_{k'}^{1/2}\right\} - \mathrm{tr}\{(\widehat{\mathbf{D}}_k - \mathbf{O}^\top\mathbf{D}_k\mathbf{O})(\mathbf{O}^\top\boldsymbol{\Psi}_1\mathbf{O})\}\right| \\
\leq& \frac{1}{2}\Big|\mathrm{tr}\{\widehat{\mathbf{D}}_k(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\mathbf{O})^\top\boldsymbol{\Sigma}_{k'}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\mathbf{O})\} + 2\mathrm{tr}\{\widehat{\mathbf{D}}_k(\boldsymbol{\beta}\mathbf{O})^\top\boldsymbol{\Sigma}_{k'}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\mathbf{O})\} \\
&+\mathrm{tr}\left[(\widehat{\mathbf{D}}_k - \mathbf{O}^\top\mathbf{D}_k\mathbf{O})\{\mathbf{O}^\top(\boldsymbol{\Psi}_{k'} - \boldsymbol{\Psi}_1)\mathbf{O}\}\right]\Big| \\
\leq& \frac{1}{2}\|(\widehat{\mathbf{D}}_k - \mathbf{O}^\top\mathbf{D}_k\mathbf{O}) + \mathbf{O}^\top\mathbf{D}_k\mathbf{O}\|_F\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\mathbf{O}\|_F\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\mathbf{O}\|\|\boldsymbol{\Sigma}_{k'}\| \\
&+ \|(\widehat{\mathbf{D}}_k - \mathbf{O}^\top\mathbf{D}_k\mathbf{O}) + \mathbf{O}^\top\mathbf{D}_k\mathbf{O}\|_F\|\boldsymbol{\beta}\mathbf{O}\|\|\boldsymbol{\Sigma}_{k'}\|\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\mathbf{O}\|_F \\
&+ \frac{1}{2}\|\widehat{\mathbf{D}}_k - \mathbf{O}^\top\mathbf{D}_k\mathbf{O}\|_F\|\boldsymbol{\Psi}_{k'} - \boldsymbol{\Psi}_1\|_F.
\end{aligned}$$

By Lemma S.11,

$$\|\mathbf{O}^\top\mathbf{D}_k\mathbf{O}\|_F \leq \|\mathbf{D}_k\|_F \leq m^4M, \quad \|\boldsymbol{\Sigma}_{k'}\| \leq m,$$

and

$$\|\boldsymbol{\Psi}_{k'} - \boldsymbol{\Psi}_1\|_F \leq \|\boldsymbol{\Psi}_{k'}\mathbf{D}_{k'}\boldsymbol{\Psi}_1\|_F \leq \|\boldsymbol{\Psi}_{k'}\|\|\mathbf{D}_{k'}\|_F\|\boldsymbol{\Psi}_1\| \leq m^6M,$$

According to Lemma S.10, under Condition (D-ii), we have

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\mathbf{O}\|_F \leq \|\mathbf{P}_{\widehat{\boldsymbol{\beta}}} - \mathbf{P}_{\boldsymbol{\beta}}\|_F \lesssim \sqrt{\frac{s^2\log p}{n\sigma^2}}.$$

And by Lemma S.13, under Assumption (A8), we have

$$\|\widehat{\mathbf{D}}_k - \mathbf{O}^\top\mathbf{D}_k\mathbf{O}\|_F \leq \|\widehat{\boldsymbol{\Psi}}_k^{-1} - \mathbf{O}^\top\boldsymbol{\Psi}_k^{-1}\mathbf{O}\|_F + \|\widehat{\boldsymbol{\Psi}}_1^{-1} - \mathbf{O}^\top\boldsymbol{\Psi}_1^{-1}\mathbf{O}\|_F \lesssim \sqrt{\frac{s^2\log p}{n\sigma^2}}.$$

Collecting all pieces above, we have

$$|g_1| \lesssim \sqrt{\frac{s^2\log p}{n\sigma^2}}.$$

(iii) Upper bound of $|g_2|$. Under Condition (D-i) that $\widehat{\mathcal{A}} = \mathcal{A}$, we have

$$\begin{aligned}
|g_2| =& \frac{1}{2}|(\boldsymbol{\mu}_{k'} - \widehat{\boldsymbol{\mu}}_1)^\top\widehat{\boldsymbol{\beta}}\widehat{\mathbf{D}}_k\widehat{\boldsymbol{\beta}}^\top(\boldsymbol{\mu}_{k'} - \widehat{\boldsymbol{\mu}}_1) - (\boldsymbol{\mu}_{k'} - \boldsymbol{\mu}_1)^\top\boldsymbol{\beta}\mathbf{D}_k\boldsymbol{\beta}^\top(\boldsymbol{\mu}_{k'} - \boldsymbol{\mu}_1)| \\
\leq& \frac{1}{2}|(\boldsymbol{\mu}_{k'} - \boldsymbol{\mu}_1)^\top(\widehat{\boldsymbol{\beta}}\widehat{\mathbf{D}}_k\widehat{\boldsymbol{\beta}}^\top - \boldsymbol{\beta}\mathbf{D}_k\boldsymbol{\beta}^\top)(\boldsymbol{\mu}_{k'} - \boldsymbol{\mu}_1)| + |(\boldsymbol{\mu}_1 - \widehat{\boldsymbol{\mu}}_1)^\top\widehat{\boldsymbol{\beta}}\widehat{\mathbf{D}}_k\widehat{\boldsymbol{\beta}}^\top(\boldsymbol{\mu}_{k'} - \boldsymbol{\mu}_1)| \\
&+ \frac{1}{2}|(\boldsymbol{\mu}_1 - \widehat{\boldsymbol{\mu}}_1)^\top\widehat{\boldsymbol{\beta}}\widehat{\mathbf{D}}_k\widehat{\boldsymbol{\beta}}^\top(\boldsymbol{\mu}_1 - \widehat{\boldsymbol{\mu}}_1)| \\
\leq& \frac{1}{2}\|\boldsymbol{\mu}_{k'} - \boldsymbol{\mu}_1\|_2^2\|\widehat{\boldsymbol{\beta}}\widehat{\mathbf{D}}_k\widehat{\boldsymbol{\beta}}^\top - \boldsymbol{\beta}\mathbf{D}_k\boldsymbol{\beta}^\top\| \\
&+ \|\boldsymbol{\mu}_{1,\mathcal{A}*} - \widehat{\boldsymbol{\mu}}_{1,\mathcal{A}*}\|_2\|\boldsymbol{\mu}_{k',\mathcal{A}*} - \boldsymbol{\mu}_{1,\mathcal{A}*}\|_2(\|\widehat{\boldsymbol{\beta}}_{\mathcal{A}*}\widehat{\mathbf{D}}_k\widehat{\boldsymbol{\beta}}_{\mathcal{A}*}^\top - \boldsymbol{\beta}_{\mathcal{A}*}\mathbf{D}_k\boldsymbol{\beta}_{\mathcal{A}*}^\top\| + \|\boldsymbol{\beta}_{\mathcal{A}*}\mathbf{D}_k\boldsymbol{\beta}_{\mathcal{A}*}^\top\|) \\
&+ \frac{1}{2}\|\boldsymbol{\mu}_{1,\mathcal{A}*} - \widehat{\boldsymbol{\mu}}_{1,\mathcal{A}*}\|_2^2(\|\widehat{\boldsymbol{\beta}}_{\mathcal{A}*}\widehat{\mathbf{D}}_k\widehat{\boldsymbol{\beta}}_{\mathcal{A}*}^\top - \boldsymbol{\beta}_{\mathcal{A}*}\mathbf{D}_k\boldsymbol{\beta}_{\mathcal{A}*}^\top\| + \|\boldsymbol{\beta}_{\mathcal{A}*}\mathbf{D}_k\boldsymbol{\beta}_{\mathcal{A}*}^\top\|).
\end{aligned}$$

According to Lemma S.12,

$$\|\widehat{\boldsymbol{\beta}}_{\mathcal{A}*}\widehat{\mathbf{D}}_k\widehat{\boldsymbol{\beta}}_{\mathcal{A}*}^\top - \boldsymbol{\beta}_{\mathcal{A}*}\mathbf{D}_k\boldsymbol{\beta}_{\mathcal{A}*}^\top\| \le \|\widehat{\boldsymbol{\beta}}\widehat{\mathbf{D}}_k\widehat{\boldsymbol{\beta}}^\top - \boldsymbol{\beta}\mathbf{D}_k\boldsymbol{\beta}^\top\|_F$$
$$\le \|\widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\Psi}}_k^{-1}\widehat{\boldsymbol{\beta}}^\top - \boldsymbol{\beta}\boldsymbol{\Psi}_k^{-1}\boldsymbol{\beta}^\top\|_F + \|\widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\Psi}}_1^{-1}\widehat{\boldsymbol{\beta}}^\top - \boldsymbol{\beta}\boldsymbol{\Psi}_1^{-1}\boldsymbol{\beta}^\top\|_F \lesssim \sqrt{\frac{s^2\log p}{n\sigma^2}}.$$

By Condition (D-iv), we have

$$\|\widehat{\boldsymbol{\mu}}_{k,\mathcal{A}*} - \boldsymbol{\mu}_{k,\mathcal{A}*}\|_2 \le \sqrt{s}\|\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_k\|_\infty, \lesssim \sqrt{s\log p/n},$$

and by Assumption (A8), $\|\boldsymbol{\mu}_{k,\mathcal{A}*} - \boldsymbol{\mu}_{1,\mathcal{A}*}\|_2 \le \|\boldsymbol{\mu}_k - \boldsymbol{\mu}_1\|_2 \le m$. According to Lemma S.11, under Assumption (A9), $\|\boldsymbol{\beta}_{\mathcal{A}*}\mathbf{D}_k\boldsymbol{\beta}_{\mathcal{A}*}^\top\| \le \|\mathbf{D}_k\| \le m^4M$. Therefore,

$$|g_2| \lesssim \sqrt{\frac{s^2\log p}{n\sigma^2}}.$$

(iv) Upper bound of $|g_3|$. Under Condition (D-i) that $\widehat{\mathcal{A}} = \mathcal{A}$

$$\begin{aligned}
&(\widehat{\boldsymbol{\mu}}_k - \widehat{\boldsymbol{\mu}}_1)^\top\widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\Psi}}_k^{-1}\widehat{\boldsymbol{\beta}}^\top(\widehat{\boldsymbol{\mu}}_k - \widehat{\boldsymbol{\mu}}_1) - (\boldsymbol{\mu}_k - \boldsymbol{\mu}_1)^\top\boldsymbol{\beta}\boldsymbol{\Psi}_k^{-1}\boldsymbol{\beta}^\top(\boldsymbol{\mu}_k - \boldsymbol{\mu}_1)\\
=&\{(\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_k) - (\widehat{\boldsymbol{\mu}}_1 - \boldsymbol{\mu}_1)\}^\top\{(\widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\Psi}}_k^{-1}\widehat{\boldsymbol{\beta}}^\top - \boldsymbol{\beta}\boldsymbol{\Psi}_k^{-1}\boldsymbol{\beta}^\top) + \boldsymbol{\beta}\boldsymbol{\Psi}_k^{-1}\boldsymbol{\beta}^\top\}(\widehat{\boldsymbol{\mu}}_k - \widehat{\boldsymbol{\mu}}_1)\\
&+ (\boldsymbol{\mu}_k - \boldsymbol{\mu}_1)^\top\{(\widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\Psi}}_k^{-1}\widehat{\boldsymbol{\beta}}^\top - \boldsymbol{\beta}\boldsymbol{\Psi}_k^{-1}\boldsymbol{\beta}^\top)(\widehat{\boldsymbol{\mu}}_k - \widehat{\boldsymbol{\mu}}_1) + \boldsymbol{\beta}\boldsymbol{\Psi}_k^{-1}\boldsymbol{\beta}^\top\{(\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_k) - (\widehat{\boldsymbol{\mu}}_1 - \boldsymbol{\mu}_1)\}\}\\
=&\{(\widehat{\boldsymbol{\mu}}_{k,\mathcal{A}*} - \boldsymbol{\mu}_{k,\mathcal{A}*}) - (\widehat{\boldsymbol{\mu}}_{1,\mathcal{A}*} - \boldsymbol{\mu}_{1,\mathcal{A}*})\}^\top\{(\widehat{\boldsymbol{\beta}}_{\mathcal{A}*}\widehat{\boldsymbol{\Psi}}_k^{-1}\widehat{\boldsymbol{\beta}}_{\mathcal{A}*}^\top - \boldsymbol{\beta}_{\mathcal{A}*}\boldsymbol{\Psi}_k^{-1}\boldsymbol{\beta}_{\mathcal{A}*}^\top) + \boldsymbol{\beta}_{\mathcal{A}*}\boldsymbol{\Psi}_k^{-1}\boldsymbol{\beta}_{\mathcal{A}*}^\top\}(\widehat{\boldsymbol{\mu}}_{k,\mathcal{A}*} - \widehat{\boldsymbol{\mu}}_{1,\mathcal{A}*})\\
&+ (\boldsymbol{\mu}_{k,\mathcal{A}*} - \boldsymbol{\mu}_{1,\mathcal{A}*})^\top(\widehat{\boldsymbol{\beta}}_{\mathcal{A}*}\widehat{\boldsymbol{\Psi}}_k^{-1}\widehat{\boldsymbol{\beta}}_{\mathcal{A}*}^\top - \boldsymbol{\beta}_{\mathcal{A}*}\boldsymbol{\Psi}_k^{-1}\boldsymbol{\beta}_{\mathcal{A}*}^\top)\{(\widehat{\boldsymbol{\mu}}_{k,\mathcal{A}*} - \widehat{\boldsymbol{\mu}}_{1,\mathcal{A}*}) - (\boldsymbol{\mu}_{k,\mathcal{A}*} - \boldsymbol{\mu}_{1,\mathcal{A}*}) + (\boldsymbol{\mu}_{k,\mathcal{A}*} - \boldsymbol{\mu}_{1,\mathcal{A}*})\}\\
&+ (\boldsymbol{\mu}_{k,\mathcal{A}*} - \boldsymbol{\mu}_{1,\mathcal{A}*})^\top\boldsymbol{\beta}_{\mathcal{A}*}\boldsymbol{\Psi}_k^{-1}\boldsymbol{\beta}_{\mathcal{A}*}^\top\{(\widehat{\boldsymbol{\mu}}_{k,\mathcal{A}*} - \boldsymbol{\mu}_{k,\mathcal{A}*}) - (\widehat{\boldsymbol{\mu}}_{1,\mathcal{A}*} - \boldsymbol{\mu}_{1,\mathcal{A}*})\}.
\end{aligned}$$

Then,

$$\begin{aligned}
&|(\widehat{\boldsymbol{\mu}}_k - \widehat{\boldsymbol{\mu}}_1)^\top\widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\Psi}}_k^{-1}\widehat{\boldsymbol{\beta}}^\top(\widehat{\boldsymbol{\mu}}_k - \widehat{\boldsymbol{\mu}}_1) - (\boldsymbol{\mu}_k - \boldsymbol{\mu}_1)^\top\boldsymbol{\beta}\boldsymbol{\Psi}_k^{-1}\boldsymbol{\beta}^\top(\boldsymbol{\mu}_k - \boldsymbol{\mu}_1)|\\
\le&\|(\widehat{\boldsymbol{\mu}}_{k,\mathcal{A}*} - \boldsymbol{\mu}_{k,\mathcal{A}*}) - (\widehat{\boldsymbol{\mu}}_{1,\mathcal{A}*} - \boldsymbol{\mu}_{1,\mathcal{A}*})\|_2\,(\|(\widehat{\boldsymbol{\mu}}_{k,\mathcal{A}*} - \widehat{\boldsymbol{\mu}}_{1,\mathcal{A}*}) - (\boldsymbol{\mu}_{k,\mathcal{A}*} - \boldsymbol{\mu}_{1,\mathcal{A}*})\|_2 + \|\boldsymbol{\mu}_{k,\mathcal{A}*} - \boldsymbol{\mu}_{1,\mathcal{A}*}\|_2)\\
&\cdot(\|\widehat{\boldsymbol{\beta}}_{\mathcal{A}*}\widehat{\boldsymbol{\Psi}}_k^{-1}\widehat{\boldsymbol{\beta}}_{\mathcal{A}*}^\top - \boldsymbol{\beta}_{\mathcal{A}*}\boldsymbol{\Psi}_k^{-1}\boldsymbol{\beta}_{\mathcal{A}*}^\top\| + \|\boldsymbol{\beta}_{\mathcal{A}*}\boldsymbol{\Psi}_k^{-1}\boldsymbol{\beta}_{\mathcal{A}*}^\top\|)\\
&+\|\boldsymbol{\mu}_{k,\mathcal{A}*} - \boldsymbol{\mu}_{1,\mathcal{A}*}\|_2\|\widehat{\boldsymbol{\beta}}_{\mathcal{A}*}\widehat{\boldsymbol{\Psi}}_k^{-1}\widehat{\boldsymbol{\beta}}_{\mathcal{A}*}^\top - \boldsymbol{\beta}_{\mathcal{A}*}\boldsymbol{\Psi}_k^{-1}\boldsymbol{\beta}_{\mathcal{A}*}^\top\|\\
&\cdot(\|(\widehat{\boldsymbol{\mu}}_{k,\mathcal{A}*} - \widehat{\boldsymbol{\mu}}_{1,\mathcal{A}*}) - (\boldsymbol{\mu}_{k,\mathcal{A}*} - \boldsymbol{\mu}_{1,\mathcal{A}*})\|_2 + \|\boldsymbol{\mu}_{k,\mathcal{A}*} - \boldsymbol{\mu}_{1,\mathcal{A}*}\|_2)\\
&+\|\boldsymbol{\mu}_{k,\mathcal{A}*} - \boldsymbol{\mu}_{1,\mathcal{A}*}\|_2\|\boldsymbol{\beta}_{\mathcal{A}*}\boldsymbol{\Psi}_k^{-1}\boldsymbol{\beta}_{\mathcal{A}*}^\top\|\|(\widehat{\boldsymbol{\mu}}_{k,\mathcal{A}*} - \boldsymbol{\mu}_{k,\mathcal{A}*}) - (\widehat{\boldsymbol{\mu}}_{1,\mathcal{A}*} - \boldsymbol{\mu}_{1,\mathcal{A}*})\|_2.
\end{aligned}$$

Similar to arguments in Part (iii), we have

$$|g_3| \lesssim \sqrt{\frac{s^2\log p}{n\sigma^2}}.$$

(v) Upper bound of $|g_4|$. We have

$$\begin{aligned}
&\left|(\widehat{\boldsymbol{\mu}}_1-\widehat{\boldsymbol{\mu}}_k)^\top\widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\Psi}}_k^{-1}\widehat{\boldsymbol{\beta}}^\top(\boldsymbol{\mu}_1-\widehat{\boldsymbol{\mu}}_1)\right| \\
=&\Big|\{(\widehat{\boldsymbol{\mu}}_{1,\mathcal{A}*}-\widehat{\boldsymbol{\mu}}_{k,\mathcal{A}*})-(\boldsymbol{\mu}_{1,\mathcal{A}*}-\boldsymbol{\mu}_{k,\mathcal{A}*})+(\boldsymbol{\mu}_{1,\mathcal{A}*}-\boldsymbol{\mu}_{k,\mathcal{A}*})\}^\top \\
&\cdot\{(\widehat{\boldsymbol{\beta}}_{\mathcal{A}*}\widehat{\boldsymbol{\Psi}}_k^{-1}\widehat{\boldsymbol{\beta}}_{\mathcal{A}*}^\top-\boldsymbol{\beta}_{\mathcal{A}*}\boldsymbol{\Psi}_k^{-1}\boldsymbol{\beta}_{\mathcal{A}*}^\top)+\boldsymbol{\beta}_{\mathcal{A}*}\boldsymbol{\Psi}_k^{-1}\boldsymbol{\beta}_{\mathcal{A}*}^\top\}(\boldsymbol{\mu}_{1,\mathcal{A}*}-\widehat{\boldsymbol{\mu}}_{1,\mathcal{A}*})\Big| \\
\le&\|\boldsymbol{\mu}_{1,\mathcal{A}*}-\widehat{\boldsymbol{\mu}}_{1,\mathcal{A}*}\|_2\left(\|(\widehat{\boldsymbol{\mu}}_{1,\mathcal{A}*}-\widehat{\boldsymbol{\mu}}_{k,\mathcal{A}*})-(\boldsymbol{\mu}_{1,\mathcal{A}*}-\boldsymbol{\mu}_{k,\mathcal{A}*})\|_2+\|\boldsymbol{\mu}_{1,\mathcal{A}*}-\boldsymbol{\mu}_{k,\mathcal{A}*}\|_2\right) \\
&\cdot\left(\|\widehat{\boldsymbol{\beta}}_{\mathcal{A}*}\widehat{\boldsymbol{\Psi}}_k^{-1}\widehat{\boldsymbol{\beta}}_{\mathcal{A}*}^\top-\boldsymbol{\beta}_{\mathcal{A}*}\boldsymbol{\Psi}_k^{-1}\boldsymbol{\beta}_{\mathcal{A}*}^\top\|+\|\boldsymbol{\beta}_{\mathcal{A}*}\boldsymbol{\Psi}_k^{-1}\boldsymbol{\beta}_{\mathcal{A}*}^\top\|\right)
\end{aligned}$$

Similar to arguments in Part (iii), we obtain that

$$|g_4|\lesssim\sqrt{\frac{s\log p}{n}}.$$

(vi) Upper bound of $|g_5|$. Similar to arguments in part (iv), under Condition (D-i) that $\widehat{\mathcal{A}}=\mathcal{A}$, we have

$$\begin{aligned}
&(\widehat{\boldsymbol{\mu}}_k-\widehat{\boldsymbol{\mu}}_1)^\top\widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\Psi}}_k^{-1}\widehat{\boldsymbol{\beta}}^\top-(\boldsymbol{\mu}_k-\boldsymbol{\mu}_1)^\top\boldsymbol{\beta}\boldsymbol{\Psi}_k^{-1}\boldsymbol{\beta}^\top \\
=&\{(\widehat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_k)-(\widehat{\boldsymbol{\mu}}_1-\boldsymbol{\mu}_1)\}^\top\{(\widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\Psi}}_k^{-1}\widehat{\boldsymbol{\beta}}^\top-\boldsymbol{\beta}\boldsymbol{\Psi}_k^{-1}\boldsymbol{\beta}^\top)+\boldsymbol{\beta}\boldsymbol{\Psi}_k^{-1}\boldsymbol{\beta}^\top\} \\
&+(\boldsymbol{\mu}_k-\boldsymbol{\mu}_1)^\top(\widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\Psi}}_k^{-1}\widehat{\boldsymbol{\beta}}^\top-\boldsymbol{\beta}\boldsymbol{\Psi}_k^{-1}\boldsymbol{\beta}^\top) \\
=&\{(\widehat{\boldsymbol{\mu}}_{k,\mathcal{A}*}-\boldsymbol{\mu}_{k,\mathcal{A}*})-(\widehat{\boldsymbol{\mu}}_{1,\mathcal{A}*}-\boldsymbol{\mu}_{k,\mathcal{A}*})\}^\top\{(\widehat{\boldsymbol{\beta}}_{\mathcal{A}*}\widehat{\boldsymbol{\Psi}}_k^{-1}\widehat{\boldsymbol{\beta}}_{\mathcal{A}*}^\top-\boldsymbol{\beta}_{\mathcal{A}*}\boldsymbol{\Psi}_k^{-1}\boldsymbol{\beta}_{\mathcal{A}*}^\top)+\boldsymbol{\beta}_{\mathcal{A}*}\boldsymbol{\Psi}_k^{-1}\boldsymbol{\beta}_{\mathcal{A}*}^\top\} \\
&+(\boldsymbol{\mu}_k-\boldsymbol{\mu}_1)^\top(\widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\Psi}}_k^{-1}\widehat{\boldsymbol{\beta}}^\top-\boldsymbol{\beta}\boldsymbol{\Psi}_k^{-1}\boldsymbol{\beta}^\top).
\end{aligned}$$

We obtain that

$$|g_5|\lesssim\sqrt{\frac{s^2\log p}{n\sigma^2}}.$$

(vii) Upper bound of $|g_6|$. Since

$$\begin{aligned}
&\{(\mathbf{O}^\top\mathbf{D}_k\mathbf{O})(\mathbf{O}^\top\boldsymbol{\Psi}_1\mathbf{O})+\mathbf{I}\}^{-1}=\mathbf{O}^\top(\mathbf{D}_k\boldsymbol{\Psi}_1+\mathbf{I})^{-1}\mathbf{O}=\mathbf{O}^\top(\mathbf{I}-\mathbf{D}_k\boldsymbol{\Psi}_k)\mathbf{O} \\
=&\mathbf{I}-(\mathbf{O}^\top\mathbf{D}_k\mathbf{O})(\mathbf{O}^\top\boldsymbol{\Psi}_k\mathbf{O}),
\end{aligned}$$

then according to Lemma 8.7 in Cai and Zhang (2021), we have

$$\begin{aligned}
&\log(|\widehat{\boldsymbol{\Psi}}_1|/|\widehat{\boldsymbol{\Psi}}_k|)-\log(|\boldsymbol{\Psi}_1|/|\boldsymbol{\Psi}_k|)=\log|\widehat{\mathbf{D}}_k\widehat{\boldsymbol{\Psi}}_1+\mathbf{I}|-\log|(\mathbf{O}^\top\mathbf{D}_k\mathbf{O})(\mathbf{O}^\top\boldsymbol{\Psi}_1\mathbf{O})+\mathbf{I}| \\
\le&\mathrm{tr}\left[\{(\mathbf{O}^\top\mathbf{D}_k\mathbf{O})(\mathbf{O}^\top\boldsymbol{\Psi}_1\mathbf{O})+\mathbf{I}\}^{-1}\{\widehat{\mathbf{D}}_k\widehat{\boldsymbol{\Psi}}_1-(\mathbf{O}^\top\mathbf{D}_k\mathbf{O})(\mathbf{O}^\top\boldsymbol{\Psi}_1\mathbf{O})\}\right] \\
=&\mathrm{tr}\left[\{\mathbf{I}-(\mathbf{O}^\top\mathbf{D}_k\mathbf{O})(\mathbf{O}^\top\boldsymbol{\Psi}_k\mathbf{O})\}\{\widehat{\mathbf{D}}_k\widehat{\boldsymbol{\Psi}}_1-(\mathbf{O}^\top\mathbf{D}_k\mathbf{O})(\mathbf{O}^\top\boldsymbol{\Psi}_1\mathbf{O})\}\right] \\
=&\mathrm{tr}\{\widehat{\mathbf{D}}_k\widehat{\boldsymbol{\Psi}}_1-(\mathbf{O}^\top\mathbf{D}_k\mathbf{O})(\mathbf{O}^\top\boldsymbol{\Psi}_1\mathbf{O})\}-\mathrm{tr}\left[(\mathbf{O}^\top\mathbf{D}_k\mathbf{O})(\mathbf{O}^\top\boldsymbol{\Psi}_k\mathbf{O})\{\widehat{\mathbf{D}}_k\widehat{\boldsymbol{\Psi}}_1-(\mathbf{O}^\top\mathbf{D}_k\mathbf{O})(\mathbf{O}^\top\boldsymbol{\Psi}_1\mathbf{O})\}\right] \\
=&\mathrm{tr}\{\widehat{\mathbf{D}}_k\widehat{\boldsymbol{\Psi}}_1-\widehat{\mathbf{D}}_k(\mathbf{O}^\top\boldsymbol{\Psi}_1\mathbf{O})\}+\mathrm{tr}\{(\widehat{\mathbf{D}}_k-\mathbf{O}^\top\mathbf{D}_k\mathbf{O})(\mathbf{O}^\top\boldsymbol{\Psi}_1\mathbf{O})\} \\
&-\mathrm{tr}\left[(\mathbf{O}^\top\mathbf{D}_k\mathbf{O})(\mathbf{O}^\top\boldsymbol{\Psi}_k\mathbf{O})\{\widehat{\mathbf{D}}_k\widehat{\boldsymbol{\Psi}}_1-(\mathbf{O}^\top\mathbf{D}_k\mathbf{O})(\mathbf{O}^\top\boldsymbol{\Psi}_1\mathbf{O})\}\right].
\end{aligned}$$

Hence,

$$\begin{aligned}
&\log(|\widehat{\mathbf{\Psi}}_1|/|\widehat{\mathbf{\Psi}}_k|) - \log(|\mathbf{\Psi}_1|/|\mathbf{\Psi}_k|) - \mathrm{tr}\{(\widehat{\mathbf{D}}_k - \mathbf{O}^\top \mathbf{D}_k \mathbf{O})(\mathbf{O}^\top \mathbf{\Psi}_1 \mathbf{O})\} \\
\leq & \mathrm{tr}\{\widehat{\mathbf{D}}_k \widehat{\mathbf{\Psi}}_1 - \widehat{\mathbf{D}}_k (\mathbf{O}^\top \mathbf{\Psi}_1 \mathbf{O})\} - \mathrm{tr}\left[(\mathbf{O}^\top \mathbf{D}_k \mathbf{O})(\mathbf{O}^\top \mathbf{\Psi}_k \mathbf{O})\{\widehat{\mathbf{D}}_k \widehat{\mathbf{\Psi}}_1 - (\mathbf{O}^\top \mathbf{D}_k \mathbf{O})(\mathbf{O}^\top \mathbf{\Psi}_1 \mathbf{O})\}\right] \\
\leq & \mathrm{tr}\{\widehat{\mathbf{D}}_k \widehat{\mathbf{\Psi}}_1 - \widehat{\mathbf{D}}_k (\mathbf{O}^\top \mathbf{\Psi}_1 \mathbf{O})\} + \|\mathbf{D}_k\|_F \|\mathbf{\Psi}_k\| \|\widehat{\mathbf{D}}_k \widehat{\mathbf{\Psi}}_1 - (\mathbf{O}^\top \mathbf{D}_k \mathbf{O})(\mathbf{O}^\top \mathbf{\Psi}_1 \mathbf{O})\|_F. \qquad \text{(S.58)}
\end{aligned}$$

For $\|\widehat{\mathbf{D}}_k \widehat{\mathbf{\Psi}}_1 - (\mathbf{O}^\top \mathbf{D}_k \mathbf{O})(\mathbf{O}^\top \mathbf{\Psi}_1 \mathbf{O})\|_F$, we have

$$\begin{aligned}
&\|\widehat{\mathbf{D}}_k \widehat{\mathbf{\Psi}}_1 - (\mathbf{O}^\top \mathbf{D}_k \mathbf{O})(\mathbf{O}^\top \mathbf{\Psi}_1 \mathbf{O})\|_F \\
\leq & \|\widehat{\mathbf{D}}_k \widehat{\mathbf{\Psi}}_1 - \widehat{\mathbf{D}}_k (\mathbf{O}^\top \mathbf{\Psi}_1 \mathbf{O})\|_F + \|(\widehat{\mathbf{D}}_k - \mathbf{O}^\top \mathbf{D}_k \mathbf{O})(\mathbf{O}^\top \mathbf{\Psi}_1 \mathbf{O})\|_F \\
\leq & (\|\widehat{\mathbf{D}}_k - \mathbf{O}^\top \mathbf{D}_k \mathbf{O}\|_F + \|\mathbf{O}^\top \mathbf{D}_k \mathbf{O}\|_F) \|\widehat{\mathbf{\Psi}}_1 - \mathbf{O}^\top \mathbf{\Psi}_1 \mathbf{O}\| + \|\widehat{\mathbf{D}}_k - \mathbf{O}^\top \mathbf{D}_k \mathbf{O}\|_F \|\mathbf{O}^\top \mathbf{\Psi}_1 \mathbf{O}\|
\end{aligned}$$

According to Lemma S.13, under Assumption (A8), we have

$$\begin{aligned}
&\|\widehat{\mathbf{\Psi}}_1 - \mathbf{O}^\top \mathbf{\Psi}_1 \mathbf{O}\| \leq \|\widehat{\mathbf{\Psi}}_1 - \mathbf{O}^\top \mathbf{\Psi}_1 \mathbf{O}\|_F \lesssim \sqrt{\frac{s^2 \log p}{n\sigma^2}}, \\
&\|\widehat{\mathbf{D}}_k - \mathbf{O}^\top \mathbf{D}_k \mathbf{O}\|_F \leq \|\widehat{\mathbf{\Psi}}_k^{-1} - \mathbf{O}^\top \mathbf{\Psi}_k^{-1} \mathbf{O}\|_F + \|\widehat{\mathbf{\Psi}}_1^{-1} - \mathbf{O}^\top \mathbf{\Psi}_1^{-1} \mathbf{O}\|_F \lesssim \sqrt{\frac{s^2 \log p}{n\sigma^2}},
\end{aligned}$$

And by Lemma S.11, we have

$$\|\mathbf{O}^\top \mathbf{D}_k \mathbf{O}\|_F \leq \|\mathbf{D}_k\|_F \leq m^4 M, \quad \|\mathbf{O}^\top \mathbf{\Psi}_1 \mathbf{O}\| \leq \|\mathbf{\Psi}_1\| \leq m.$$

Therefore,

$$\|\widehat{\mathbf{D}}_k \widehat{\mathbf{\Psi}}_1 - (\mathbf{O}^\top \mathbf{D}_k \mathbf{O})(\mathbf{O}^\top \mathbf{\Psi}_1 \mathbf{O})\|_F \lesssim \sqrt{\frac{s^2 \log p}{n\sigma^2}}. \qquad \text{(S.59)}$$

For $\mathrm{tr}\{\widehat{\mathbf{D}}_k \widehat{\mathbf{\Psi}}_1 - \widehat{\mathbf{D}}_k (\mathbf{O}^\top \mathbf{\Psi}_1 \mathbf{O})\}$, we have

$$\begin{aligned}
&\mathrm{tr}\{\widehat{\mathbf{D}}_k \widehat{\mathbf{\Psi}}_1 - \widehat{\mathbf{D}}_k (\mathbf{O}^\top \mathbf{\Psi}_1 \mathbf{O})\} \\
\leq & (\|\widehat{\mathbf{D}}_k - \mathbf{O}^\top \mathbf{D}_k \mathbf{O}\|_F + \|\mathbf{O}^\top \mathbf{D}_k \mathbf{O}\|_F) \|\widehat{\mathbf{\Psi}}_1 - \mathbf{O}^\top \mathbf{\Psi}_1 \mathbf{O}\|_F \lesssim \sqrt{\frac{s^2 \log p}{n\sigma^2}} \qquad \text{(S.60)}
\end{aligned}$$

By combining (S.58)–(S.60), we have

$$\log(|\widehat{\mathbf{\Psi}}_1|/|\widehat{\mathbf{\Psi}}_k|) - \log(|\mathbf{\Psi}_1|/|\mathbf{\Psi}_k|) - \mathrm{tr}\{(\widehat{\mathbf{D}}_k - \mathbf{O}^\top \mathbf{D}_k \mathbf{O})(\mathbf{O}^\top \mathbf{\Psi}_1 \mathbf{O})\} \lesssim \sqrt{\frac{s^2 \log p}{n\sigma^2}}.$$

Following the similar arguments, we obtain that

$$\begin{aligned}
&\mathrm{tr}\{(\widehat{\mathbf{D}}_k - \mathbf{O}^\top \mathbf{D}_k \mathbf{O})(\mathbf{O}^\top \mathbf{\Psi}_1 \mathbf{O})\} - \{\log(|\widehat{\mathbf{\Psi}}_1|/|\widehat{\mathbf{\Psi}}_k|) - \log(|\mathbf{\Psi}_1|/|\mathbf{\Psi}_k|)\} \\
\leq & \mathrm{tr}\{\widehat{\mathbf{D}}_k (\mathbf{O}^\top \mathbf{\Psi}_1 \mathbf{O}) - \widehat{\mathbf{D}}_k \widehat{\mathbf{\Psi}}_1\} - \mathrm{tr}\left[\widehat{\mathbf{D}}_k \widehat{\mathbf{\Psi}}_k \{(\mathbf{O}^\top \mathbf{D}_k \mathbf{O})(\mathbf{O}^\top \mathbf{\Psi}_1 \mathbf{O}) - \widehat{\mathbf{D}}_k \widehat{\mathbf{\Psi}}_1\}\right] \\
\lesssim & \sqrt{\frac{s^2 \log p}{n\sigma^2}}.
\end{aligned}$$

Therefore,

$$|g_6| \lesssim \sqrt{\frac{s^2 \log p}{n\sigma^2}}.$$

(viii) Upper bound of $|g_7|$. We have

$$|\log(\widehat{\pi}_1/\widehat{\pi}_k) - \log(\pi_1/\pi_k)| \leq |\log\widehat{\pi}_1 - \log\pi_1| + |\log\widehat{\pi}_k - \log\pi_k|,$$

and

$$|\log\widehat{\pi}_1 - \log\pi_1| \leq \max\left\{\frac{1}{\pi_1}, \frac{1}{\widehat{\pi}_1}\right\}|\widehat{\pi}_1 - \pi_1|.$$

Under Condition (D-v),

$$\pi_1 - C\sqrt{\log p/n} \leq \widehat{\pi}_1 \leq \pi_1 + C\sqrt{\log p/n},$$

and

$$\max\left\{\frac{1}{\pi_1}, \frac{1}{\widehat{\pi}_1}\right\} \leq \frac{1}{\pi_1 - C\sqrt{\log p/n}}.$$

Then, under Assumption (A4), we have that

$$|\log\widehat{\pi}_1 - \log\pi_1| \lesssim |\widehat{\pi}_1 - \pi_1|.$$

Similarly, we obtain

$$|\log\widehat{\pi}_k - \log\pi_k| \lesssim |\widehat{\pi}_k - \pi_k|,$$

and it follows that

$$|\log(\widehat{\pi}_1/\widehat{\pi}_k) - \log(\pi_1/\pi_k)| \lesssim |\widehat{\pi}_1 - \pi_1| + |\widehat{\pi}_k - \pi_k| \lesssim \sqrt{\log p/n}.$$

Then,

$$|g_7| \lesssim \sqrt{\log p/n}.$$

According to (S.55), we have

$$\Pr\left(\left|h_2(\mathbf{X}^*) + \sum_{l=1}^{6} g_l\right| \geq \varepsilon_0/2 \middle| Y^* = h\right) \lesssim \frac{s^2\log p/(n\sigma^2)}{\{\varepsilon_0 - \sqrt{s^2\log p/(n\sigma^2)}\}^2}. \tag{S.61}$$

By combining (S.52)(S.54)(S.61), we have

$$\begin{aligned}
&\Pr(|\widehat{Q}_k(\widehat{\mathbf{Z}}^*) - Q_k(\mathbf{Z}^*)| \geq \varepsilon_0 \mid Y^* = h) \\
\lesssim &\exp\left\{-C\min\left(\frac{\varepsilon_0^2}{s^2\log p/(n\sigma^2)}, \frac{\varepsilon_0}{\sqrt{s^2\log p/(n\sigma^2)}}\right)\right\} + \frac{s^2\log p/(n\sigma^2)}{\{\varepsilon_0 - \sqrt{s^2\log p/(n\sigma^2)}\}^2}.
\end{aligned}$$

Thus, by (S.51),

$$\begin{aligned}
&\Pr(|\widehat{Q}_k(\widehat{\mathbf{Z}}^*) - Q_k(\mathbf{Z}^*)| \geq \varepsilon_0) \\
\lesssim &\exp\left\{-C\min\left(\frac{\varepsilon_0^2}{s^2\log p/(n\sigma^2)}, \frac{\varepsilon_0}{\sqrt{s^2\log p/(n\sigma^2)}}\right)\right\} + \frac{s^2\log p/(n\sigma^2)}{\{\varepsilon_0 - \sqrt{s^2\log p/(n\sigma^2)}\}^2}.
\end{aligned}$$

□

S.15.2 Proof of Lemma S.10

Let $\mathbf{O}_1\mathbf{\Lambda}\mathbf{O}_2^\top$ denote the singular value decomposition of $\widehat{\boldsymbol{\beta}}^\top\boldsymbol{\beta}$, where $\mathbf{\Lambda} = \mathrm{diag}(\cos\theta_1, \ldots, \cos\theta_d)$, and θ_j, $j = 1, \ldots, d$, are principal angles between $\mathrm{span}(\widehat{\boldsymbol{\beta}})$ and $\mathrm{span}(\boldsymbol{\beta})$. Let $\mathbf{O} = \mathbf{O}_2\mathbf{O}_1^\top$ and $\sin\Theta(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}) = \mathrm{diag}(\sin\theta_1, \ldots, \sin\theta_d)$, we have

$$\begin{aligned}\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\mathbf{O}\|_F^2 &= \mathrm{tr}\left\{(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\mathbf{O})^\top(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\mathbf{O})\right\} = 2d - 2\mathrm{tr}(\widehat{\boldsymbol{\beta}}^\top\boldsymbol{\beta}\mathbf{O}) \\ &= 2d - 2\sum_{j=1}^{d}\cos\theta_j \leq 2d - 2\sum_{j=1}^{d}\cos^2\theta_j \\ &= 2\|\sin\Theta(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta})\|_F^2.\end{aligned}$$

Therefore, according to Proposition 1 in Cai and Zhang (2018), we have

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\mathbf{O}\|_F \leq \sqrt{2}\|\sin\Theta(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta})\|_F = \|\mathbf{P}_{\widehat{\boldsymbol{\beta}}} - \mathbf{P}_{\boldsymbol{\beta}}\|_F.$$

□

S.15.3 Proof of Lemma S.11

For $\|\mathbf{\Psi}_k\|$, by Assumption (A8), we have

$$\|\mathbf{\Psi}_k\| \leq \|\boldsymbol{\beta}\|^2\|\mathbf{\Sigma}_k\| = \|\mathbf{\Sigma}_k\| \leq m$$

for some constant $m > 0$. For $\|\mathbf{\Psi}_k^{-1}\|$, since

$$\|\mathbf{\Psi}_k^{-1}\| = \frac{1}{\varphi_{\min}(\mathbf{\Psi}_k)},$$

and

$$\varphi_{\min}(\mathbf{\Psi}_k) = \varphi_{\min}(\boldsymbol{\beta}^\top\mathbf{\Sigma}_k\boldsymbol{\beta}) \geq \varphi_{\min}(\mathbf{\Sigma}_k),$$

then by Assumption (A8), we have $\varphi_{\min}(\mathbf{\Sigma}_k) \geq m^{-1}$. Thus,

$$\varphi_{\min}(\mathbf{\Psi}_k) \geq \frac{1}{m}, \quad \|\mathbf{\Psi}_k^{-1}\| \leq m, \quad k = 1, \ldots, K. \tag{S.62}$$

For $\|\mathbf{D}_k\|_F$, we have

$$\begin{aligned}\|\mathbf{D}_k\|_F &= \|\mathbf{\Psi}_k^{-1}(\mathbf{\Psi}_1 - \mathbf{\Psi}_k)\mathbf{\Psi}_1^{-1}\|_F \leq \|\mathbf{\Psi}_k^{-1}\|\|\mathbf{\Psi}_1^{-1}\|\|\mathbf{\Psi}_1 - \mathbf{\Psi}_k\|_F \\ &= \|\mathbf{\Psi}_k^{-1}\|\|\mathbf{\Psi}_1^{-1}\|\|\mathbf{\Sigma}_1 - \mathbf{\Sigma}_k\|_F.\end{aligned}$$

Moreover,

$$\|\mathbf{\Sigma}_1 - \mathbf{\Sigma}_k\|_F = \|\mathbf{\Sigma}_k(\mathbf{\Sigma}_k^{-1} - \mathbf{\Sigma}_1^{-1})\mathbf{\Sigma}_1\|_F \leq \|\mathbf{\Sigma}_k\|\|\mathbf{\Sigma}_1\|\|\mathbf{\Sigma}_k^{-1} - \mathbf{\Sigma}_1^{-1}\|_F \leq m^2\|\mathbf{\Sigma}_k^{-1} - \mathbf{\Sigma}_1^{-1}\|_F.$$

According to Assumption (A9), $\|\mathbf{\Sigma}_k^{-1} - \mathbf{\Sigma}_1^{-1}\|_F \leq M$ for some constant $M > 0$ and $k = 2, \ldots, K$. Therefore,

$$\|\mathbf{D}_k\|_F \leq m^4 M, \quad k = 1, \ldots, K. \tag{S.63}$$

□

S.15.4 Proof of Lemma S.12

We have

$$\begin{aligned}
&\widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\Psi}}_k^{-1}\widehat{\boldsymbol{\beta}}^\top - \boldsymbol{\beta}\boldsymbol{\Psi}_k^{-1}\boldsymbol{\beta}^\top \\
=&\widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\Psi}}_k^{-1}\widehat{\boldsymbol{\beta}}^\top - \boldsymbol{\beta}\mathbf{O}(\mathbf{O}^\top\boldsymbol{\Psi}_k^{-1}\mathbf{O})\mathbf{O}^\top\boldsymbol{\beta}^\top \\
=&(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\mathbf{O})\{(\widehat{\boldsymbol{\Psi}}_k^{-1} - \mathbf{O}^\top\boldsymbol{\Psi}_k^{-1}\mathbf{O}) + \mathbf{O}^\top\boldsymbol{\Psi}_k^{-1}\mathbf{O}\}\widehat{\boldsymbol{\beta}}^\top \\
&+ \boldsymbol{\beta}\mathbf{O}\{(\widehat{\boldsymbol{\Psi}}_k^{-1} - \mathbf{O}^\top\boldsymbol{\Psi}_k^{-1}\mathbf{O})\widehat{\boldsymbol{\beta}}^\top + \mathbf{O}^\top\boldsymbol{\Psi}_k^{-1}\mathbf{O}(\widehat{\boldsymbol{\beta}}^\top - \mathbf{O}^\top\boldsymbol{\beta}^\top)\}.
\end{aligned}$$

Hence,

$$\begin{aligned}
&\|\widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\Psi}}_k^{-1}\widehat{\boldsymbol{\beta}}^\top - \boldsymbol{\beta}\boldsymbol{\Psi}_k^{-1}\boldsymbol{\beta}^\top\|_F \\
\le&\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\mathbf{O}\|_F(\|\widehat{\boldsymbol{\Psi}}_k^{-1} - \mathbf{O}^\top\boldsymbol{\Psi}_k^{-1}\mathbf{O}\| + \|\mathbf{O}^\top\boldsymbol{\Psi}_k^{-1}\mathbf{O}\|)\|\widehat{\boldsymbol{\beta}}\| \\
&+ \|\boldsymbol{\beta}\mathbf{O}\|\|\widehat{\boldsymbol{\Psi}}_k^{-1} - \mathbf{O}^\top\boldsymbol{\Psi}_k^{-1}\mathbf{O}\|_F(\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\mathbf{O}\| + \|\boldsymbol{\beta}\mathbf{O}\|) + \|\boldsymbol{\beta}\mathbf{O}\|\|\mathbf{O}^\top\boldsymbol{\Psi}_k^{-1}\mathbf{O}\|\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\mathbf{O}\|_F.
\end{aligned} \tag{S.64}$$

According to Lemma S.10, under Condition (D-ii), we have

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\mathbf{O}\|_F \le \|\mathbf{P}_{\widehat{\boldsymbol{\beta}}} - \mathbf{P}_{\boldsymbol{\beta}}\|_F \lesssim \sqrt{\frac{s^2\log p}{n\sigma^2}}. \tag{S.65}$$

Under Assumption (A8), by Lemma S.11, we have

$$\|\mathbf{O}^\top\boldsymbol{\Psi}_k^{-1}\mathbf{O}\| \le \|\boldsymbol{\Psi}_k^{-1}\| \le m. \tag{S.66}$$

And by Lemma S.13,

$$\|\widehat{\boldsymbol{\Psi}}_k^{-1} - \mathbf{O}^\top\boldsymbol{\Psi}_k^{-1}\mathbf{O}\|_F \lesssim \sqrt{\frac{s^2\log p}{n\sigma^2}} \tag{S.67}$$

By combining (S.64)–(S.70), it follows that

$$\|\widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\Psi}}_k^{-1}\widehat{\boldsymbol{\beta}}^\top - \boldsymbol{\beta}\boldsymbol{\Psi}_k^{-1}\boldsymbol{\beta}^\top\|_F \lesssim \sqrt{\frac{s^2\log p}{n\sigma^2}}.$$

□

S.15.5 Proof of Lemma S.13

Under Condition (D-i) that $\widehat{\mathcal{A}} = \mathcal{A}$, we have $\widehat{\boldsymbol{\beta}}_{\mathcal{A}^c*} = \mathbf{0}$ and

$$\begin{aligned}
\widehat{\boldsymbol{\Psi}}_k - \mathbf{O}^\top\boldsymbol{\Psi}_k\mathbf{O} =&\widehat{\boldsymbol{\beta}}_{\mathcal{A}*}^\top\widehat{\boldsymbol{\Sigma}}_{k,\mathcal{A}\mathcal{A}}\widehat{\boldsymbol{\beta}}_{\mathcal{A}*} - \mathbf{O}^\top\boldsymbol{\beta}_{\mathcal{A}*}^\top\boldsymbol{\Sigma}_{k,\mathcal{A}\mathcal{A}}\boldsymbol{\beta}_{\mathcal{A}*}\mathbf{O} \\
=&(\widehat{\boldsymbol{\beta}}_{\mathcal{A}*} - \boldsymbol{\beta}_{\mathcal{A}*}\mathbf{O})^\top(\widehat{\boldsymbol{\Sigma}}_{k,\mathcal{A}\mathcal{A}}\widehat{\boldsymbol{\beta}}_{\mathcal{A}*}) + (\boldsymbol{\beta}_{\mathcal{A}*}\mathbf{O})^\top(\widehat{\boldsymbol{\Sigma}}_{k,\mathcal{A}\mathcal{A}}\widehat{\boldsymbol{\beta}}_{\mathcal{A}*} - \boldsymbol{\Sigma}_{k,\mathcal{A}\mathcal{A}}\boldsymbol{\beta}_{\mathcal{A}*}\mathbf{O}).
\end{aligned}$$

Then,

$$\begin{aligned}
\|\widehat{\boldsymbol{\Psi}}_k - \mathbf{O}^\top\boldsymbol{\Psi}_k\mathbf{O}\|_F \le&\|\widehat{\boldsymbol{\beta}}_{\mathcal{A}*} - \boldsymbol{\beta}_{\mathcal{A}*}\mathbf{O}\|_F\left(\|\widehat{\boldsymbol{\Sigma}}_{k,\mathcal{A}\mathcal{A}}\widehat{\boldsymbol{\beta}}_{\mathcal{A}*} - \boldsymbol{\Sigma}_{k,\mathcal{A}\mathcal{A}}\boldsymbol{\beta}_{\mathcal{A}*}\mathbf{O}\| + \|\boldsymbol{\Sigma}_{k,\mathcal{A}\mathcal{A}}\boldsymbol{\beta}_{\mathcal{A}*}\mathbf{O}\|\right) \\
&+ \|\boldsymbol{\beta}_{\mathcal{A}*}\mathbf{O}\|\|\widehat{\boldsymbol{\Sigma}}_{k,\mathcal{A}\mathcal{A}}\widehat{\boldsymbol{\beta}}_{\mathcal{A}*} - \boldsymbol{\Sigma}_{k,\mathcal{A}\mathcal{A}}\boldsymbol{\beta}_{\mathcal{A}*}\mathbf{O}\|_F.
\end{aligned}$$

By Lemma S.10, under Condition (D-ii), we have

$$\|\widehat{\boldsymbol{\beta}}_{\mathcal{A}*} - \boldsymbol{\beta}_{\mathcal{A}*}\mathbf{O}\|_F \le \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\mathbf{O}\|_F \le \|\mathbf{P}_{\widehat{\boldsymbol{\beta}}} - \mathbf{P}_{\boldsymbol{\beta}}\|_F \lesssim \sqrt{\frac{s^2 \log p}{\sigma^2 n}}. \tag{S.68}$$

Moreover,

$$\begin{aligned}
&\|\widehat{\boldsymbol{\Sigma}}_{k,\mathcal{A}\mathcal{A}}\widehat{\boldsymbol{\beta}}_{\mathcal{A}*} - \boldsymbol{\Sigma}_{k,\mathcal{A}\mathcal{A}}\boldsymbol{\beta}_{\mathcal{A}*}\mathbf{O}\|_F \\
\le &(\|\widehat{\boldsymbol{\Sigma}}_{k,\mathcal{A}\mathcal{A}} - \boldsymbol{\Sigma}_{k,\mathcal{A}\mathcal{A}}\| + \|\boldsymbol{\Sigma}_{k,\mathcal{A}\mathcal{A}}\|)\|\widehat{\boldsymbol{\beta}}_{\mathcal{A}*} - \boldsymbol{\beta}_{\mathcal{A}*}\mathbf{O}\|_F + \|\widehat{\boldsymbol{\Sigma}}_{k,\mathcal{A}\mathcal{A}} - \boldsymbol{\Sigma}_{k,\mathcal{A}\mathcal{A}}\|_F\|\boldsymbol{\beta}_{\mathcal{A}*}\mathbf{O}\|.
\end{aligned}$$

By Conditions (D-ii) and (D-iii), we have

$$\|\widehat{\boldsymbol{\Sigma}}_{k,\mathcal{A}\mathcal{A}}\widehat{\boldsymbol{\beta}}_{\mathcal{A}*} - \boldsymbol{\Sigma}_{k,\mathcal{A}\mathcal{A}}\boldsymbol{\beta}_{\mathcal{A}*}\|_F \lesssim \sqrt{\frac{s^2 \log p}{n\sigma^2}}.$$

It follows that

$$\|\widehat{\boldsymbol{\Psi}}_k - \mathbf{O}^\top \boldsymbol{\Psi}_k \mathbf{O}\|_F \lesssim \sqrt{\frac{s^2 \log p}{n\sigma^2}}. \tag{S.69}$$

For $\|\widehat{\boldsymbol{\Psi}}_k^{-1} - \mathbf{O}^\top \boldsymbol{\Psi}_k^{-1}\mathbf{O}\|_F$, since

$$\widehat{\boldsymbol{\Psi}}_k^{-1} - \mathbf{O}^\top \boldsymbol{\Psi}_k^{-1}\mathbf{O} = \widehat{\boldsymbol{\Psi}}_k^{-1} - (\mathbf{O}^\top \boldsymbol{\Psi}_k \mathbf{O})^{-1} = \widehat{\boldsymbol{\Psi}}_k^{-1}(\mathbf{O}^\top \boldsymbol{\Psi}_k \mathbf{O} - \widehat{\boldsymbol{\Psi}}_k)(\mathbf{O}^\top \boldsymbol{\Psi}_k \mathbf{O})^{-1},$$

then

$$\|\widehat{\boldsymbol{\Psi}}_k^{-1} - \mathbf{O}^\top \boldsymbol{\Psi}_k^{-1}\mathbf{O}\|_F \le \|\widehat{\boldsymbol{\Psi}}_k^{-1}\|\|\mathbf{O}^\top \boldsymbol{\Psi}_k \mathbf{O} - \widehat{\boldsymbol{\Psi}}_k\|_F\|\mathbf{O}^\top \boldsymbol{\Psi}_k^{-1}\mathbf{O}\|.$$

According to Lemma S.11, under Assumption (A8), $\|\boldsymbol{\Psi}_k^{-1}\| \le m$, and it follows that $\sigma_{\min}(\boldsymbol{\Psi}_k) \ge m^{-1}$. Then by Weyl's inequality, we have

$$\sigma_{\min}(\widehat{\boldsymbol{\Psi}}_k) \ge \sigma_{\min}(\mathbf{O}^\top \boldsymbol{\Psi}_k \mathbf{O}) - \|\widehat{\boldsymbol{\Psi}}_k - \mathbf{O}^\top \boldsymbol{\Psi}_k \mathbf{O}\| \ge \frac{1}{m} - C\sqrt{\frac{s^2 \log p}{n\sigma^2}}$$

for some constant $C > 0$. It follows that

$$\|\widehat{\boldsymbol{\Psi}}_k^{-1}\| = \frac{1}{\sigma_{\min}(\widehat{\boldsymbol{\Psi}}_k)} \le \left(\frac{1}{m} - C\sqrt{\frac{s^2 \log p}{n\sigma^2}}\right)^{-1}.$$

Since we assume that $c_1 s^2 \log p \le n\sigma^2$ for some sufficiently large constant $c_1 > 0$, then $\|\widehat{\boldsymbol{\Psi}}_k^{-1}\|$ is bounded from above. By collecting all the pieces above, we obtain that

$$\|\widehat{\boldsymbol{\Psi}}_k^{-1} - \mathbf{O}^\top \boldsymbol{\Psi}_k^{-1}\mathbf{O}\|_F \lesssim \sqrt{\frac{s^2 \log p}{n\sigma^2}}. \tag{S.70}$$

□

S.16 Proof of Lemma S.1

Let $\mathbf{U} = \mathbf{X}_{\mathcal{A}} = (X_1, X_2)^\top$ and $\boldsymbol{\Gamma} = (\ell, \ell)^\top$. Conditional on $\mathbf{U}$, the variable X_3 has distribution $N(\boldsymbol{\Gamma}^\top \mathbf{U}, 1)$, which does not depend on Y. Therefore, $Y \perp\!\!\!\perp X_3 \mid \mathbf{U}$, and the active set is $\mathcal{A} = \{1, 2\}$.

Since Y is uniform on $\{1, 2\}$, we have $\boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}} = \mathrm{Cov}(\mathbf{X}_{\mathcal{A}}) = \frac{1}{2}\boldsymbol{\Sigma}_{1,\mathcal{A}\mathcal{A}} + \frac{1}{2}\boldsymbol{\Sigma}_{2,\mathcal{A}\mathcal{A}} = \mathbf{I}_2$. Hence

$\boldsymbol{\Delta}_{1,\mathcal{A}\mathcal{A}} = \boldsymbol{\Sigma}_{1,\mathcal{A}\mathcal{A}} - \boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}} = \tau \mathbf{J}_2$ and $\boldsymbol{\Delta}_{2,\mathcal{A}\mathcal{A}} = \boldsymbol{\Sigma}_{2,\mathcal{A}\mathcal{A}} - \boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}} = -\tau \mathbf{J}_2$. Next, for each h,

$$\begin{aligned}\boldsymbol{\Sigma}_{h,\mathcal{A}\mathcal{A}^c} &= \mathrm{Cov}(\mathbf{X}_{\mathcal{A}}, X_3 \mid Y = h)\\ &= \mathrm{Cov}(\mathbf{U}, \boldsymbol{\Gamma}^\top \mathbf{U} + \varepsilon \mid Y = h)\\ &= \mathrm{Cov}(\mathbf{U}, \boldsymbol{\Gamma}^\top \mathbf{U} \mid Y = h) + \mathrm{Cov}(\mathbf{U}, \varepsilon \mid Y = h).\end{aligned}$$

Since ε is independent of $(\mathbf{U}, Y)$, the second term is zero. For the first term,

$$\mathrm{Cov}(\mathbf{U}, \boldsymbol{\Gamma}^\top \mathbf{U} \mid Y = h) = \boldsymbol{\Sigma}_{h,\mathcal{A}\mathcal{A}} \boldsymbol{\Gamma}.$$

Therefore,

$$\boldsymbol{\Sigma}_{h,\mathcal{A}\mathcal{A}^c} = \boldsymbol{\Sigma}_{h,\mathcal{A}\mathcal{A}} \boldsymbol{\Gamma}.$$

Similarly, $\boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}^c} = \boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}} \boldsymbol{\Gamma}$, and thus

$$\boldsymbol{\Delta}_{h,\mathcal{A}\mathcal{A}^c} = \boldsymbol{\Sigma}_{h,\mathcal{A}\mathcal{A}^c} - \boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}^c} = (\boldsymbol{\Sigma}_{h,\mathcal{A}\mathcal{A}} - \boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}})\boldsymbol{\Gamma} = \boldsymbol{\Delta}_{h,\mathcal{A}\mathcal{A}} \boldsymbol{\Gamma}.$$

By Theorem 1, $\mathbf{M}_{h,\mathcal{A}\mathcal{A}} = (\boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}})^{-1} \boldsymbol{\Delta}_{h,\mathcal{A}\mathcal{A}}$ and $\mathbf{M}_{h,\mathcal{A}\mathcal{A}^c} = (\boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}})^{-1} \boldsymbol{\Delta}_{h,\mathcal{A}\mathcal{A}^c}$. Since $\boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}} = \mathbf{I}_2$, we obtain

$$\mathbf{M}_{1,\mathcal{A}\mathcal{A}} = \tau \mathbf{J}_2, \quad \mathbf{M}_{2,\mathcal{A}\mathcal{A}} = -\tau \mathbf{J}_2,$$

and

$$\mathbf{M}_{1,\mathcal{A}\mathcal{A}^c} = 2\ell\tau\, \mathbf{1}_2, \quad \mathbf{M}_{2,\mathcal{A}\mathcal{A}^c} = -2\ell\tau\, \mathbf{1}_2.$$

For any $i \in \mathcal{A}$, we have $\mathcal{M}_{[:,i,1]} = (\tau, -\tau)^\top$, $\mathcal{M}_{[:,i,2]} = (\tau, -\tau)^\top$, and $\mathcal{M}_{[:,i,3]} = (2\ell\tau, -2\ell\tau)^\top$. Therefore,

$$\max_{j \in \mathcal{A}} \|\mathcal{M}_{[:,i,j]}\|_2 = \sqrt{2}\tau, \quad \max_{j \in \mathcal{A}^c} \|\mathcal{M}_{[:,i,j]}\|_2 = 2\sqrt{2}\ell\tau.$$

If $0 < \tau < r_n/\sqrt{2}$ and $\ell > r_n/(2\sqrt{2}\tau)$, then $\sqrt{2}\tau < r_n < 2\sqrt{2}\ell\tau$, and $\max_{j \in \mathcal{A}} \|\mathcal{M}_{[:,i,j]}\|_2 < r_n < \max_{j \in \mathcal{A}^c} \|\mathcal{M}_{[:,i,j]}\|_2$ for each $i \in \mathcal{A}$. Taking $r_n = c\sqrt{s^2 \log p/n}$ finishes the proof. □

S.17 Proof of Corollary S.1

Let $\widetilde{\varepsilon}_\Sigma = \|\widehat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A}} - \boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}}\|_{\max}$ and $\widetilde{\varepsilon}_\Delta = \max_h \|\widehat{\boldsymbol{\Delta}}_{h,\mathcal{A}\mathcal{A}} - \boldsymbol{\Delta}_{h,\mathcal{A}\mathcal{A}}\|_{\max}$. For some constant $c_3 > 0$, define the event

$$G := \left\{ \max\{\widetilde{\varepsilon}_\Sigma, \widetilde{\varepsilon}_\Delta\} \le \sqrt{c_3 \log s/n} \right\}.$$

We show that Corollary S.1 holds under event $I \cap G$. First, we prove that under event $I = E \cap F$,

$$\|\widehat{\mathbf{B}} - \mathbf{B}\|_F \le 2K\{(\theta_0 \eta_1 \widetilde{\varepsilon}_\Sigma + \widetilde{\varepsilon}_\Delta)\sqrt{s^2 H} + \lambda_2 \sqrt{s}\}.$$

The proof is similar to that of Lemma S.8. By (S.44), we have

$$\begin{aligned}
&\sum_{k=1}^{sH}(\widehat{\mathbf{B}}_{*k}-\mathbf{B}_{*k})^\top\widehat{\mathbf{\Sigma}}_{\mathcal{A}\mathcal{A}}(\widehat{\mathbf{B}}_{*k}-\mathbf{B}_{*k})\\
\leq&2\sum_{k=1}^{sH}\|\widehat{\mathbf{\Sigma}}_{\mathcal{A}\mathcal{A}}\mathbf{B}_{*k}-\widehat{\mathbf{\Delta}}_{*k}^{(\mathcal{A})}\|_\infty\|\widehat{\mathbf{B}}_{*k}-\mathbf{B}_{*k}\|_1+2\lambda_2\|\widehat{\mathbf{B}}-\mathbf{B}\|_*.
\end{aligned}$$

Then, for $\|\widehat{\mathbf{\Sigma}}_{\mathcal{A}\mathcal{A}}\mathbf{B}_{*k}-\widehat{\mathbf{\Delta}}_{*k}^{(\mathcal{A})}\|_\infty$, given that $\mathbf{\Delta}^{(\mathcal{A})}=\mathbf{\Sigma}_{\mathcal{A}\mathcal{A}}\mathbf{B}$,

$$\begin{aligned}
\|\widehat{\mathbf{\Sigma}}_{\mathcal{A}\mathcal{A}}\mathbf{B}_{*k}-\widehat{\mathbf{\Delta}}_{*k}^{(\mathcal{A})}\|_\infty\leq&\|\widehat{\mathbf{\Sigma}}_{\mathcal{A}\mathcal{A}}\mathbf{B}_{*k}-\mathbf{\Sigma}_{\mathcal{A}\mathcal{A}}\mathbf{B}_{*k}\|_\infty+\|\mathbf{\Delta}_{*k}^{(\mathcal{A})}-\widehat{\mathbf{\Delta}}_{*k}^{(\mathcal{A})}\|_\infty\\
\leq&\|\widehat{\mathbf{\Sigma}}_{\mathcal{A}\mathcal{A}}-\mathbf{\Sigma}_{\mathcal{A}\mathcal{A}}\|_{\max}\|\mathbf{B}_{*k}\|_1+\max_h\|\mathbf{\Delta}_{h,\mathcal{A}\mathcal{A}}-\widehat{\mathbf{\Delta}}_{h,\mathcal{A}\mathcal{A}}\|_{\max}\\
\leq&\theta_0\eta_1\widetilde{\varepsilon}_\Sigma+\widetilde{\varepsilon}_\Delta.
\end{aligned}$$

It follows that

$$\begin{aligned}
&\sum_{k=1}^{sH}(\widehat{\mathbf{B}}_{*k}-\mathbf{B}_{*k})^\top\widehat{\mathbf{\Sigma}}_{\mathcal{A}\mathcal{A}}(\widehat{\mathbf{B}}_{*k}-\mathbf{B}_{*k})\\
\leq&2(\theta_0\eta_1\widetilde{\varepsilon}_\Sigma+\widetilde{\varepsilon}_\Delta)\sqrt{s^2H}\|\widehat{\mathbf{B}}-\mathbf{B}\|_F+2\lambda_2\sqrt{s}\|\widehat{\mathbf{B}}-\mathbf{B}\|_F.
\end{aligned}$$

Moreover, under event I, $\varphi_{\min}(\widehat{\mathbf{\Sigma}}_{\mathcal{A}\mathcal{A}})\geq K^{-1}$. Therefore,

$$\sum_{k=1}^{sH}(\widehat{\mathbf{B}}_{*k}-\mathbf{B}_{*k})^\top\widehat{\mathbf{\Sigma}}_{\mathcal{A}\mathcal{A}}(\widehat{\mathbf{B}}_{*k}-\mathbf{B}_{*k})\geq K^{-1}\|\widehat{\mathbf{B}}-\mathbf{B}\|_F^2.$$

Hence,

$$\|\widehat{\mathbf{B}}-\mathbf{B}\|_F\leq 2K\{(\theta_0\eta_1\widetilde{\varepsilon}_\Sigma+\widetilde{\varepsilon}_\Delta)\sqrt{s^2H}+\lambda_2\sqrt{s}\}.$$

Then, under event G, by taking $0<\lambda_2\lesssim\sqrt{s\log s/n}$, it follows that

$$\|\widehat{\mathbf{B}}-\mathbf{B}\|_F\leq\widetilde{c}\sqrt{s^2\log s/n},$$

where $\widetilde{c}>0$ is a constant depending on parameters $m,\theta_0,\eta_1,H,\pi_h,\|X_j\|_{\psi_2}$, and $\|X_jI(\widetilde{Y}=h)\|_{\psi_2}$, for $j=1,\ldots,p$, and $h=1,\ldots,H$. Moreover, $\|\widehat{\mathbf{B}}-\mathbf{B}\|\leq\|\widehat{\mathbf{B}}-\mathbf{B}\|_F\leq\widetilde{c}\sqrt{s^2\log s/n}$. Then under Assumption (A7′), $\|\widehat{\mathbf{B}}-\mathbf{B}\|<\sigma/2$. Therefore, by Weyl's inequality (cf. Lemma S.20), we have $\widehat{\sigma}_d\geq\sigma_d-\|\widehat{\mathbf{B}}-\mathbf{B}\|>\sigma/2$, and $\widehat{\sigma}_{d+1}\leq\sigma_{d+1}+\|\widehat{\mathbf{B}}-\mathbf{B}\|<\sigma/2$. By taking $\delta=\sigma/2$ and let $d_{\max}\geq d$, we obtain that $\widehat{d}=d$.

Following Part II in the proof of Theorem 3, we have

$$\|\sin\Theta(\widehat{\boldsymbol{\eta}},\boldsymbol{\eta})\|_F\leq\frac{\|\widehat{\mathbf{B}}-\mathbf{B}\|_F}{\widehat{\sigma}_d}\lesssim\sqrt{\frac{s^2\log s}{n\sigma^2}}.$$

Moreover,

$$D(\widehat{\boldsymbol{\beta}},\boldsymbol{\beta})\lesssim\sqrt{\frac{s^2\log s}{n\sigma^2}}.$$

Now, we show the high probability that event $I\cap G$ holds. It has been shown in Part III in the proof of Theorem 3 that event I holds with probability at least $1-Cp^{-C'}$. Following the proofs of Lemma S.4 and S.5, it can be easily derived that under Assumption (A1),

there exist some constants $C, C' > 0$ such that, for any $0 < \varepsilon \le C'$ we have

$$\Pr(\|\widehat{\mathbf{\Sigma}}_{\mathcal{A}\mathcal{A}} - \mathbf{\Sigma}_{\mathcal{A}\mathcal{A}}\|_{\max} \ge \varepsilon) \le Cs^2 \exp(-C'n\varepsilon^2).$$

Furthermore, under Assumptions (A1) and (A4),

$$\Pr(\|\widehat{\mathbf{\Sigma}}_{h,\mathcal{A}\mathcal{A}} - \mathbf{\Sigma}_{h,\mathcal{A}\mathcal{A}}\|_{\max} \ge \varepsilon) \le Cs^2 \exp(-C'n\varepsilon^2).$$

By taking $\varepsilon = C\sqrt{\log s/n}$ for some constant $C > 0$, it gives us that event G holds with probability at least $1 - Cs^{-C'}$. In conclusion, event $I \cap G$ holds with probability at least $1 - Cs^{-C'}$. This completes the proof. □

S.18 Proof of Theorem S.2

Define the following notations,

$$\xi_0 = \max_{\bar{h}} \|\mathbf{\Theta}_{\bar{h},\mathcal{A}*}\|_{\max}, \quad \xi_1 = \max_{\bar{h}} \|\mathbf{\Theta}_{\bar{h},\mathcal{A}\mathcal{A}}\|_1, \quad \varepsilon_\Theta = \max_{\bar{h}} \|\widehat{\mathbf{\Theta}}_{\bar{h}} - \mathbf{\Theta}_{\bar{h}}\|_{\max},$$

and define the event

$$J := \left\{\max\{\varepsilon_\Sigma, \varepsilon_\Theta\} \lesssim \sqrt{\log p/n}\right\}.$$

We prove (i)–(iii) hold under event J following the similar arguments in Section S.13.

First, we can show that under event J, the following two conditions hold.

(C-i′) $\theta_0 s\varepsilon_\Sigma \le 1/3$;

(C-ii′) $s\varepsilon_\Sigma\theta_0\left\{1 + \frac{3}{2}\left(\theta_1\theta_0 + \frac{1}{3}\right)\right\} \le M\alpha\min\{\lambda_1, 1\}$, where $M = (\alpha\lambda_1 + \sqrt{\bar{H}}(\alpha-2)\varepsilon_\Theta)(\sqrt{\bar{H}}\xi_0\alpha\lambda_1 + \alpha\lambda_1 + \sqrt{\bar{H}}\varepsilon_\Theta\alpha)^{-1}$

The proof is similar to that in Part I in Section S.13 and is thus omitted.

Define $\mathbf{m}_{\bar{h}} = \mathrm{vec}(\mathbf{M}_{\bar{h}})$, $\bar{h} = 1, \ldots, \bar{H}$, and let $\mathbf{m} = (\mathbf{m}_1, \ldots, \mathbf{m}_{\bar{H}}) \in \mathbb{R}^{p^2 \times \bar{H}}$. Recall that $(\widehat{\mathbf{M}}_1, \ldots, \widehat{\mathbf{M}}_{\bar{H}})$ is the solution of (S.3). Let $\widehat{\mathbf{m}}_{\bar{h}} = \mathrm{vec}(\widehat{\mathbf{M}}_{\bar{h}})$, $\bar{h} = 1, \ldots, \bar{H}$, and $\widehat{\mathbf{m}} = (\mathbf{m}_1, \ldots, \mathbf{m}_{\bar{H}}) \in \mathbb{R}^{p^2 \times \bar{H}}$. We have the following lemma, similar to Lemma S.3.

Lemma S.14. *Assuming Assumption (A2) and Conditions (C-i′)(C-ii′) hold, we have* $\widehat{\mathbf{m}}_{\mathcal{S}^c} = \mathbf{0}$ *and* $\|\widehat{\mathbf{m}} - \mathbf{m}\|_{2,\infty} \le (\lambda_1 + \sqrt{\bar{H}}\varepsilon_\Theta)\theta_0 + \frac{3}{2}s\varepsilon_\Sigma\theta_0^2(\lambda_1 + \sqrt{\bar{H}}\xi_0 + \sqrt{\bar{H}}\varepsilon_\Theta)$.

The proof of Lemma S.14 directly follows from that of Lemma S.3. Moreover, the rest of proof is quite similar to that of Part II in Section S.13 and is also omitted. Then, it remains to show that event J holds with high probability. Since the high probability inequality for ε_Σ has been developed in Lemma S.4, we only need to show the high probability inequality for ε_Θ, as clarified in the following lemma.

Lemma S.15. *Assume that Assumptions (A1) & (A4) hold. There exist some constants* $C, C' > 0$ *such that, for any* $0 < \varepsilon \le C'$ *and* $j, k = 1, \ldots, p$, *we have*

$$\Pr(\max_{\bar{h}} \|\widehat{\mathbf{\Theta}}_{\bar{h}} - \mathbf{\Theta}_{\bar{h}}\|_{\max} \ge \varepsilon) \le Cp^2 \exp(-C'n\varepsilon^2).$$

Therefore, by Lemmas S.4 and S.15, there exist some constants $C, C', c > 0$ such that, with probability at least $1 - Cp^{-C'}$, we have

$$\max\{\varepsilon_\Sigma, \varepsilon_\Theta\} \le \sqrt{c \log p / n}.$$

This completes the proof. □

S.18.1 Proof of Lemma S.15

Recall that $\mathbf{\Theta}_{\bar{h}} = \mathbf{\Lambda}_{\bar{h}} - 2\mathbf{\Sigma}$ and $\mathbf{\Lambda}_{\bar{h}} = \mathrm{E}\{(\mathbf{X} - \mathbf{X}')(\mathbf{X} - \mathbf{X}')^\top \mid \widetilde{Y} = h, \widetilde{Y}' = h'\}$. We decompose $\mathbf{\Lambda}_{\bar{h}}$ as

$$\mathbf{\Lambda}_{\bar{h}} = \mathbf{K}_h + \mathbf{K}_{h'} - \boldsymbol{\mu}_h \boldsymbol{\mu}_{h'}^\top - \boldsymbol{\mu}_{h'} \boldsymbol{\mu}_h^\top,$$

where $\mathbf{K}_h = \mathrm{E}(\mathbf{X}\mathbf{X}^\top \mid \widetilde{Y} = h)$ and $\boldsymbol{\mu}_h = \mathrm{E}(\mathbf{X} \mid \widetilde{Y} = h)$. Also, the estimate $\widehat{\mathbf{\Theta}}_{\bar{h}} = \widehat{\mathbf{\Lambda}}_{\bar{h}} - 2\widehat{\mathbf{\Sigma}}$ with

$$\widehat{\mathbf{\Lambda}}_{\bar{h}} = \widehat{\mathbf{K}}_h + \widehat{\mathbf{K}}_{h'} - \overline{\mathbf{X}}_h \overline{\mathbf{X}}_{h'}^\top - \overline{\mathbf{X}}_{h'} \overline{\mathbf{X}}_h^\top,$$

where $\widehat{\mathbf{K}}_h = \sum_{i=1}^n \mathbf{X}_i \mathbf{X}_i^\top I(Y_i = h)/n_h$ and $\overline{\mathbf{X}}_h = \sum_{i=1}^n \mathbf{X}_i I(Y_i = h)/n_h$. Therefore, we decompose $\|\widehat{\mathbf{\Theta}}_{\bar{h}} - \mathbf{\Theta}_{\bar{h}}\|_{\max}$ as follows,

$$\begin{aligned}\|\widehat{\mathbf{\Theta}}_{\bar{h}} - \mathbf{\Theta}_{\bar{h}}\|_{\max} \le & \|\widehat{\mathbf{K}}_h - \mathbf{K}_h\|_{\max} + \|\widehat{\mathbf{K}}_{h'} - \mathbf{K}_{h'}\|_{\max} + \|\boldsymbol{\mu}_h \boldsymbol{\mu}_{h'}^\top - \overline{\mathbf{X}}_h \overline{\mathbf{X}}_{h'}^\top\|_{\max} \\ & + \|\boldsymbol{\mu}_{h'} \boldsymbol{\mu}_h^\top - \overline{\mathbf{X}}_{h'} \overline{\mathbf{X}}_h^\top\|_{\max} + 2\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_{\max}.\end{aligned}$$

For $\|\widehat{\mathbf{K}}_h - \mathbf{K}_h\|_{\max}$, we have

$$\begin{aligned}\|\widehat{\mathbf{K}}_h - \mathbf{K}_h\|_{\max} = & \frac{1}{n_h} \left\| \sum_{i=1}^n \mathbf{X}_i \mathbf{X}_i^\top I(Y_i = h) - \mathrm{E}(\mathbf{X}\mathbf{X}^\top \mid \widetilde{Y} = h) \right\|_{\max} \\ \le & \left\| \frac{n}{n_h} \mathbf{\Xi}_h + \overline{\mathbf{X}}_h \pi_h \boldsymbol{\mu}_h^\top + \pi_h \boldsymbol{\mu}_h \overline{\mathbf{X}}_h^\top - \frac{n}{n_h} \pi_h^2 \boldsymbol{\mu}_h \boldsymbol{\mu}_h^\top \right. \\ & \left. - \left(\frac{1}{\pi_h} \mathrm{Cov}\{\mathbf{X} I(\widetilde{Y} = h)\} + \pi_h \boldsymbol{\mu}_h \boldsymbol{\mu}_h^\top + \pi_h \boldsymbol{\mu}_h \boldsymbol{\mu}_h^\top - \pi_h \boldsymbol{\mu}_h \boldsymbol{\mu}_h^\top \right) \right\|_{\max} \\ \le & \|\frac{n}{n_h} \mathbf{\Xi}_h - \frac{1}{\pi_h} \mathrm{Cov}\{\mathbf{X} I(\widetilde{Y} = h)\}\|_{\max} + \|\overline{\mathbf{X}}_h \pi_h \boldsymbol{\mu}_h^\top - \pi_h \boldsymbol{\mu}_h \boldsymbol{\mu}_h^\top\|_{\max} \\ & + \|\pi_h \boldsymbol{\mu}_h \overline{\mathbf{X}}_h^\top - \pi_h \boldsymbol{\mu}_h \boldsymbol{\mu}_h^\top\|_{\max} + \|\frac{n}{n_h} \pi_h^2 \boldsymbol{\mu}_h \boldsymbol{\mu}_h^\top - \pi_h \boldsymbol{\mu}_h \boldsymbol{\mu}_h^\top\|_{\max},\end{aligned}$$

where

$$\mathbf{\Xi}_h = \frac{1}{n} \sum_{i=1}^n \left\{ \mathbf{X}_i I(\widetilde{Y}_i = h) - \pi_h \boldsymbol{\mu}_h \right\} \left\{ \mathbf{X}_i I(\widetilde{Y}_i = h) - \pi_h \boldsymbol{\mu}_h \right\}^\top.$$

By the upper bounds of terms I_1–I_4 in the proof of Lemma S.5, we have that

$$\Pr(\|\widehat{\mathbf{K}}_h - \mathbf{K}_h\|_{\max} \ge \varepsilon) \le Cp^2 \exp(-C' n \varepsilon^2).$$

For $\|\boldsymbol{\mu}_h \boldsymbol{\mu}_{h'}^\top - \overline{\mathbf{X}}_h \overline{\mathbf{X}}_{h'}^\top\|_{\max}$, we directly follow the derivation of the upper bound of I_5 in the proof of Lemma S.5 and have that

$$\Pr(\|\boldsymbol{\mu}_h \boldsymbol{\mu}_{h'}^\top - \overline{\mathbf{X}}_h \overline{\mathbf{X}}_{h'}^\top\|_{\max} \ge \varepsilon) \le Cp^2 \exp(-C' n \varepsilon^2).$$

Similar upper bounds can also be established for $\|\widehat{\mathbf{K}}_{h'} - \mathbf{K}_{h'}\|_{\max}$ and $\|\boldsymbol{\mu}_{h'}\boldsymbol{\mu}_h^\top - \overline{\mathbf{X}}_{h'}\overline{\mathbf{X}}_h^\top\|_{\max}$. Moreover, by Lemma S.4,

$$\Pr(\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{\max} \geq \varepsilon) \leq Cp^2 \exp(-C'n\varepsilon^2).$$

In conclusion, we have

$$\Pr(\max_{\bar{h}} \|\widehat{\boldsymbol{\Theta}}_{\bar{h}} - \boldsymbol{\Theta}_{\bar{h}}\|_{\max} \geq \varepsilon) \leq Cp^2 \exp(-C'n\varepsilon^2).$$

□

S.19 Auxiliary definitions and lemmas

Definition S.1. *A random variable X is sub-Gaussian with parameter σ^2, denoted by $X \sim \mathrm{subG}(\sigma^2)$ if*

$$\mathrm{E}\exp\{s(X - \mathrm{E}X)\} \leq \exp(\sigma^2 s^2/2), \quad \forall s \in \mathbb{R}.$$

Definition S.2. *A random vector $\mathbf{X} \in \mathbb{R}^p$ is sub-Gaussian random vector with parameters $\sigma^2 > 0$ if $\mathbf{u}^\top\mathbf{X} \sim \mathrm{subG}(\sigma^2)$ for all $\mathbf{u} \in \mathbb{R}^p$ such that $\|\mathbf{u}\|_2 = 1$, or equivalently,*

$$\mathrm{E}\exp\{\mathbf{a}^\top(\mathbf{X} - \mathrm{E}\mathbf{X})\} \leq \exp(\sigma^2\|\mathbf{a}\|_2^2/2), \quad \forall \mathbf{a} \in \mathbb{R}^p.$$

Definition S.3. *The sub-Gaussian norm of $X \in \mathbb{R}$, denoted by $\|X\|_{\psi_2}$, is defined as*

$$\|X\|_{\psi_2} = \inf\{t > 0 : \mathrm{E}\exp(X^2/t^2) \leq 2\}.$$

Lemma S.16. *(Vershynin, 2018, Theorem 2.2.6, Hoeffding's inequality) Let $X_1, \ldots, X_n$ be independent random variables. Assume that $X_i \in [m_i, M_i]$ for $i = 1, \ldots, n$. Then, for any $t > 0$, there exists some constant $C > 0$ such that*

$$\Pr\left\{\left|\sum_{i=1}^n (X_i - \mathrm{E}X_i)\right| \geq t\right\} \leq 2\exp\left\{-\frac{Ct^2}{\sum_{i=1}^n (M_i - m_i)^2}\right\}.$$

Lemma S.17. *(Vershynin, 2018, Theorem 2.6.3) Let $X_1, \ldots, X_n$ be independent, mean zero, sub-gaussian random variables, and $\mathbf{a} = (a_1, \ldots, a_n) \in \mathbb{R}^n$. Then, there exists some constant $c > 0$ such that for every $t \geq 0$, we have*

$$\Pr\left(\left|\sum_{i=1}^n a_i X_i\right| \geq t\right) \leq 2\exp\left(-\frac{ct^2}{K^2\|\mathbf{a}\|_2^2}\right),$$

where $K = \max_i \|X_i\|_{\psi_2}^2$.

Lemma S.18. *(Zeng et al., 2024, Lemma E.6) Suppose that $\mathbf{X} \in \mathbb{R}^p$ is a sub-Gaussian random vector with parameter $\sigma^2 > 0$ and $\mathrm{Cov}(\mathbf{X}) = \boldsymbol{\Sigma}$. Then we have $\|\boldsymbol{\Sigma}\| \leq 4\sigma^2$.*

Lemma S.19. *(Rudelson and Vershynin, 2013, Theorem 1.1) Let $\mathbf{X} = (X_1, \ldots, X_n) \in \mathbb{R}^n$ be a random vector with independent components X_i which satisfies $\mathrm{E}(X_i) = 0$ and*

$\|X_i\|_{\psi_2} \leq K$. *Let* $\mathbf{A}$ *be an* $n \times n$ *matrix. Then, for any* $t \geq 0$, *there exists some constant* $C > 0$ *such that*

$$\Pr\left(|\mathbf{X}^\top \mathbf{A}\mathbf{X} - \mathrm{E}(\mathbf{X}^\top \mathbf{A}\mathbf{X})| > t\right) \leq 2\exp\left(-C\min\left\{\frac{t^2}{K^4\|\mathbf{A}\|_F^2}, \frac{t}{K^2\|\mathbf{A}\|}\right\}\right).$$

Lemma S.20. *(Stewart and Sun, 1990, Corollary 4.10, Weyl's inequality) Let* $\mathbf{B} = \mathbf{A} + \mathbf{E} \in \mathbb{R}^{p\times p}$, *where* $\mathbf{A} \in \mathbb{R}^{p\times p}$ *and* $\mathbf{E} \in \mathbb{R}^{p\times p}$ *are symmetric matrices. Let* λ_j *and* μ_j *denote the* j*-th eigenvalues of* $\mathbf{B}$ *and* $\mathbf{A}$ *respectively. Then*

$$|\lambda_j - \mu_j| \leq \|\mathbf{E}\|, \quad j = 1, \ldots, p.$$

Lemma S.21. *(Wedin, 1972, Wedin's sin* Θ *theorem) Suppose that* $\mathbf{A}, \widehat{\mathbf{A}} \in \mathbb{R}^{p\times q}$ *have singular values* $\sigma_1 \geq \cdots \geq \sigma_{\min\{p,q\}}$ *and* $\widehat{\sigma}_1 \geq \cdots \geq \widehat{\sigma}_{\min\{p,q\}}$ *respectively. Let* $\mathbf{U}$ *and* $\mathbf{V}$ *be the matrices composed of top-*d *left and right singular vectors of* $\mathbf{A}$, *and let* $\widehat{\mathbf{U}}$ *and* $\widehat{\mathbf{V}}$ *be the matrices composed of top-*d *left and right singular vectors of* $\widehat{\mathbf{A}}$. *For the subspaces spanned by* $\widehat{\mathbf{U}}$ *and* $\mathbf{U}$, *define the principal angles between them as* $(\theta_1, \ldots, \theta_d) = (\cos^{-1}\omega_1, \ldots, \cos^{-1}\omega_d)$ *where* $\omega_1 \geq \cdots \geq \omega_d$ *are the singular values of* $\widehat{\mathbf{U}}^\top\mathbf{U}$. *And define* $\sin\Theta(\widehat{\mathbf{U}}, \mathbf{U}) = \mathrm{diag}(\sin\theta_1, \ldots \sin\theta_d) \in \mathbb{R}^{d\times d}$. *Then, if* $\sigma_d(\widehat{\mathbf{A}}) - \sigma_{d+1}(\mathbf{A}) > 0$, *we have*

$$\|\sin\Theta(\widehat{\mathbf{U}}, \mathbf{U})\|_F \leq \frac{\max\{\|(\widehat{\mathbf{A}} - \mathbf{A})\widehat{\mathbf{V}}\|_F, \|\widehat{\mathbf{U}}^\top(\widehat{\mathbf{A}} - \mathbf{A})\|_F\}}{\sigma_d(\widehat{\mathbf{A}}) - \sigma_{d+1}(\mathbf{A})}.$$

Lemma S.22. *Assume that* $\max_h \|\mathrm{E}(\mathbf{X} \mid \widetilde{Y} = h)\|_2 \leq Q$ *for some constant* $Q > 0$. *And for all* $i \in [p]$ *and* $h \in [H]$, *assume that* $a^{-1} \leq \pi_h \leq a$ *and* $b^{-1} \leq \varphi_i\{\mathrm{Cov}(\mathbf{X} \mid \widetilde{Y} = h)\} \leq b$ *for some constants* $a, b > 0$. *Then, for all* $h \in [H]$, *the eigenvalues of* $\mathrm{Cov}\{\mathbf{X}I(\widetilde{Y} = h)\}$ *are bounded from both sides.*

Proof of Lemma S.22. Since for each h,

$$\begin{aligned}
\mathrm{Cov}\{\mathbf{X}I(\widetilde{Y} = h)\} &= \mathrm{E}\{\mathbf{X}\mathbf{X}^\top I(\widetilde{Y} = h)\} - \mathrm{E}\{\mathbf{X}I(\widetilde{Y} = h)\}\mathrm{E}\{\mathbf{X}^\top I(\widetilde{Y} = h)\} \\
&= \pi_h \mathrm{E}(\mathbf{X}\mathbf{X}^\top \mid \widetilde{Y} = h) - \pi_h^2 \mathrm{E}(\mathbf{X} \mid \widetilde{Y} = h)\mathrm{E}(\mathbf{X}^\top \mid \widetilde{Y} = h) \\
&= \pi_h \mathrm{Cov}(\mathbf{X} \mid \widetilde{Y} = h) + (\pi_h - \pi_h^2)\mathrm{E}(\mathbf{X} \mid \widetilde{Y} = h)\mathrm{E}(\mathbf{X}^\top \mid \widetilde{Y} = h).
\end{aligned}$$

Therefore,

$$\begin{aligned}
\varphi_{\max}\left[\mathrm{Cov}\{\mathbf{X}I(\widetilde{Y} = h)\}\right] \leq& \pi_h \varphi_{\max}\left[\mathrm{Cov}(\mathbf{X} \mid \widetilde{Y} = h)\right] + (\pi_h - \pi_h^2)\|\mathrm{E}(\mathbf{X} \mid \widetilde{Y} = h)\|_2^2 \\
\leq& ab + a(1 - a^{-1})Q^2,
\end{aligned}$$

and

$$\varphi_{\min}\left[\mathrm{Cov}\{\mathbf{X}I(\widetilde{Y} = h)\}\right] \geq \pi_h \varphi_{\min}\left[\mathrm{Cov}(\mathbf{X} \mid \widetilde{Y} = h)\right] \geq a^{-1}b^{-1}.$$

This completes the proof. □

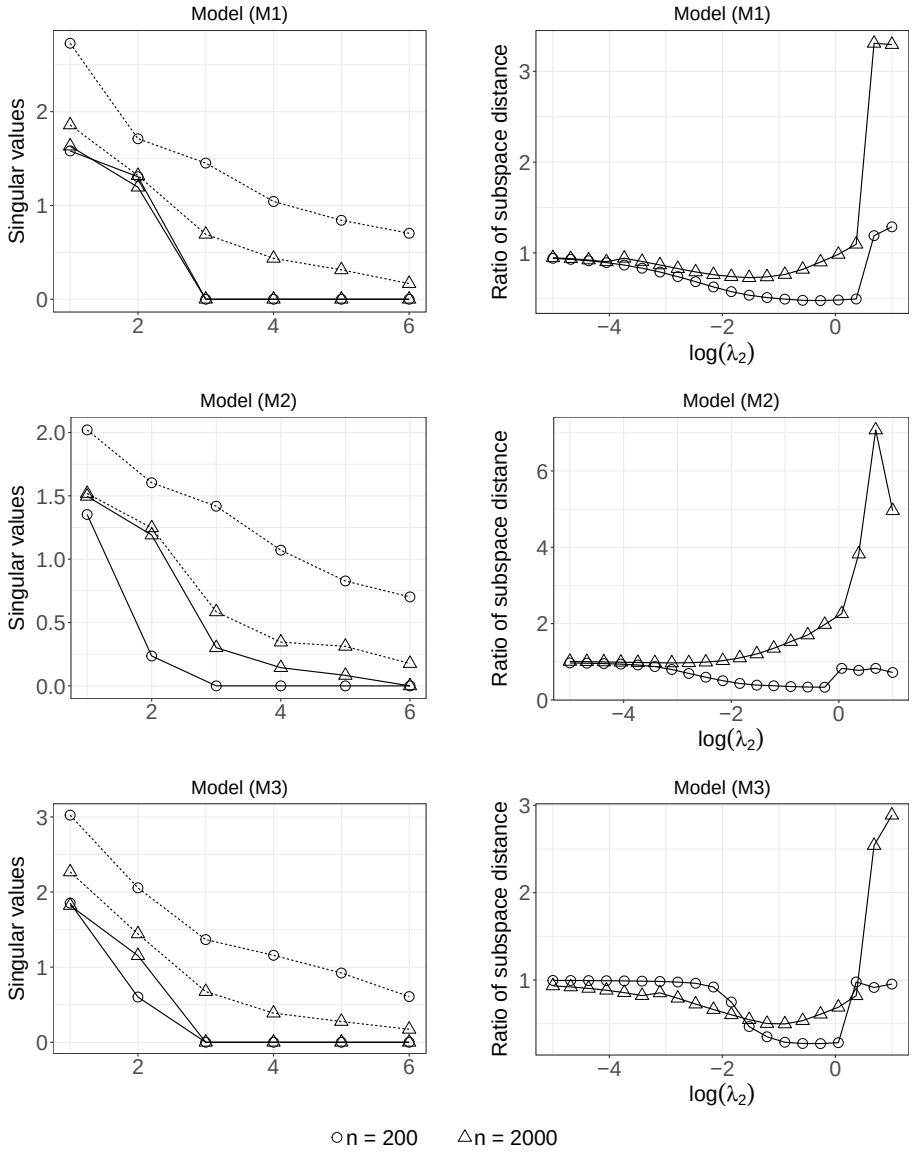

Figure S.3: Models (M1)–(M3). (Left) Singular values of sparse SAVE (solid lines) with the optimal λ_2 and oracle-SAVE (dashed lines); (Right) The ratio of the subspace estimation error of sparse SAVE and oracle-SAVE.

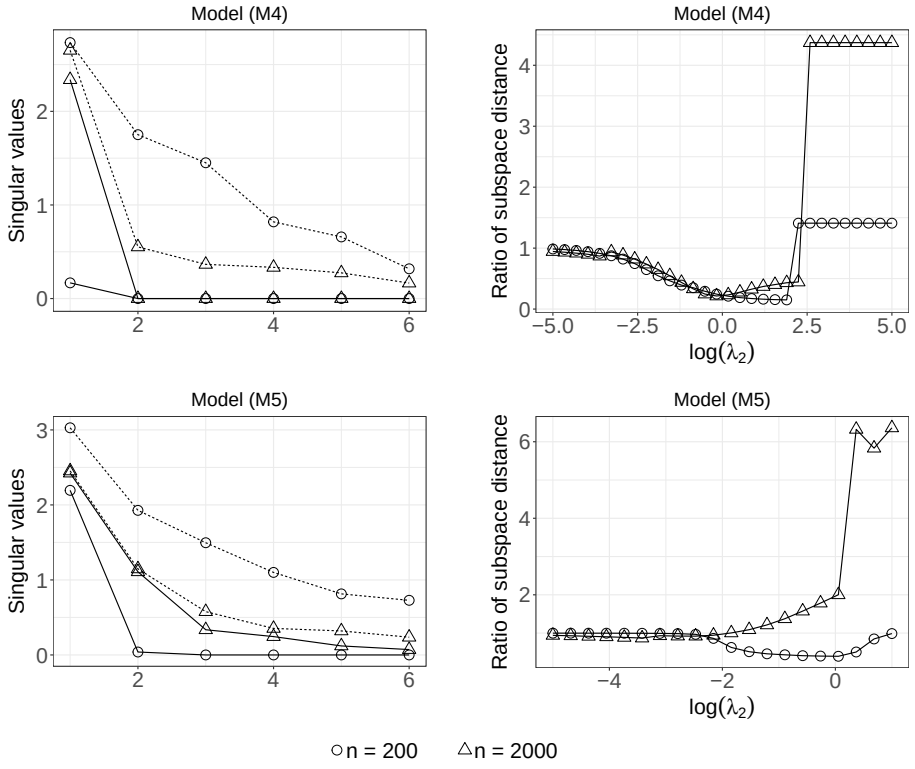

Figure S.4: Models (M4) and (M5). (Left) Singular values of sparse SAVE (solid lines) with the optimal λ_2 and oracle-SAVE (dashed lines); (Right) The ratio of the subspace estimation error of sparse SAVE and oracle-SAVE.

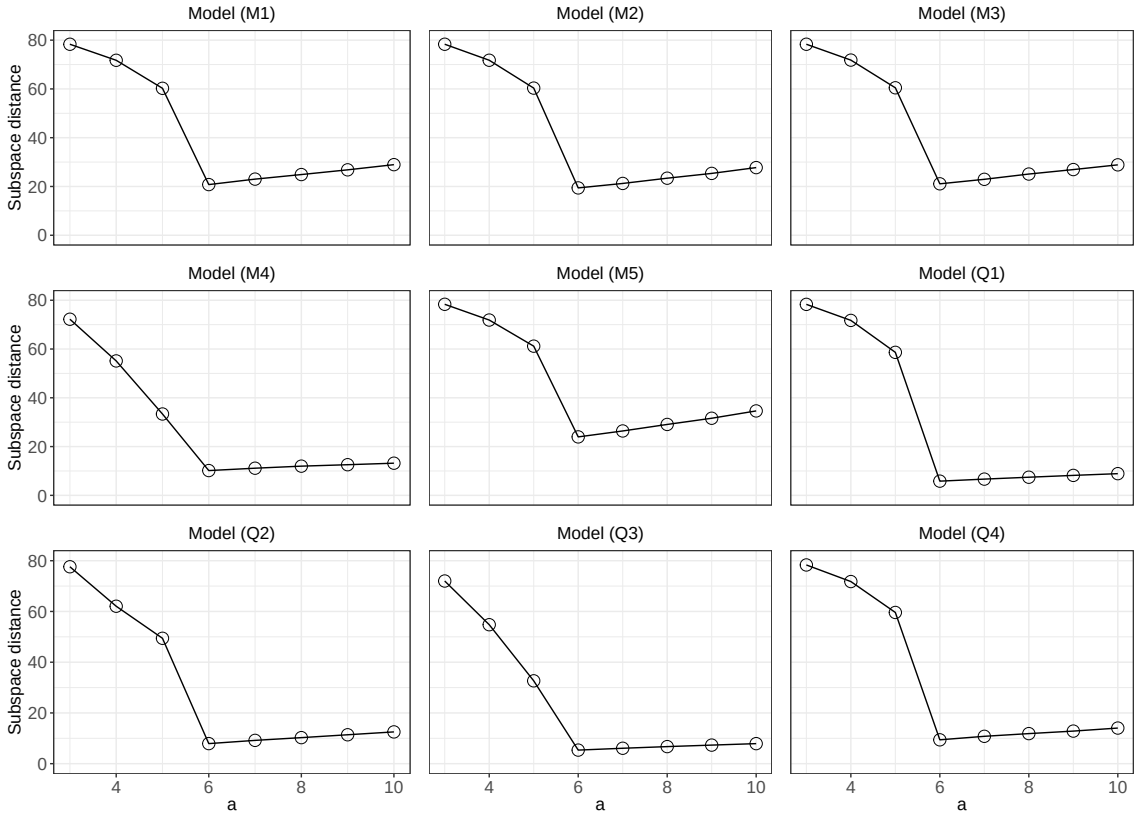

Figure S.5: Subspace estimation error of sparse SAVE given the active set $\{1, \ldots, a\}$ from the first step, where $a = 3, \ldots, 10$. The structural dimension d is given and $n = 300$.

Figure S.6: The word cloud of the top 30 words most frequently selected by sparse SAVE on the news report data.

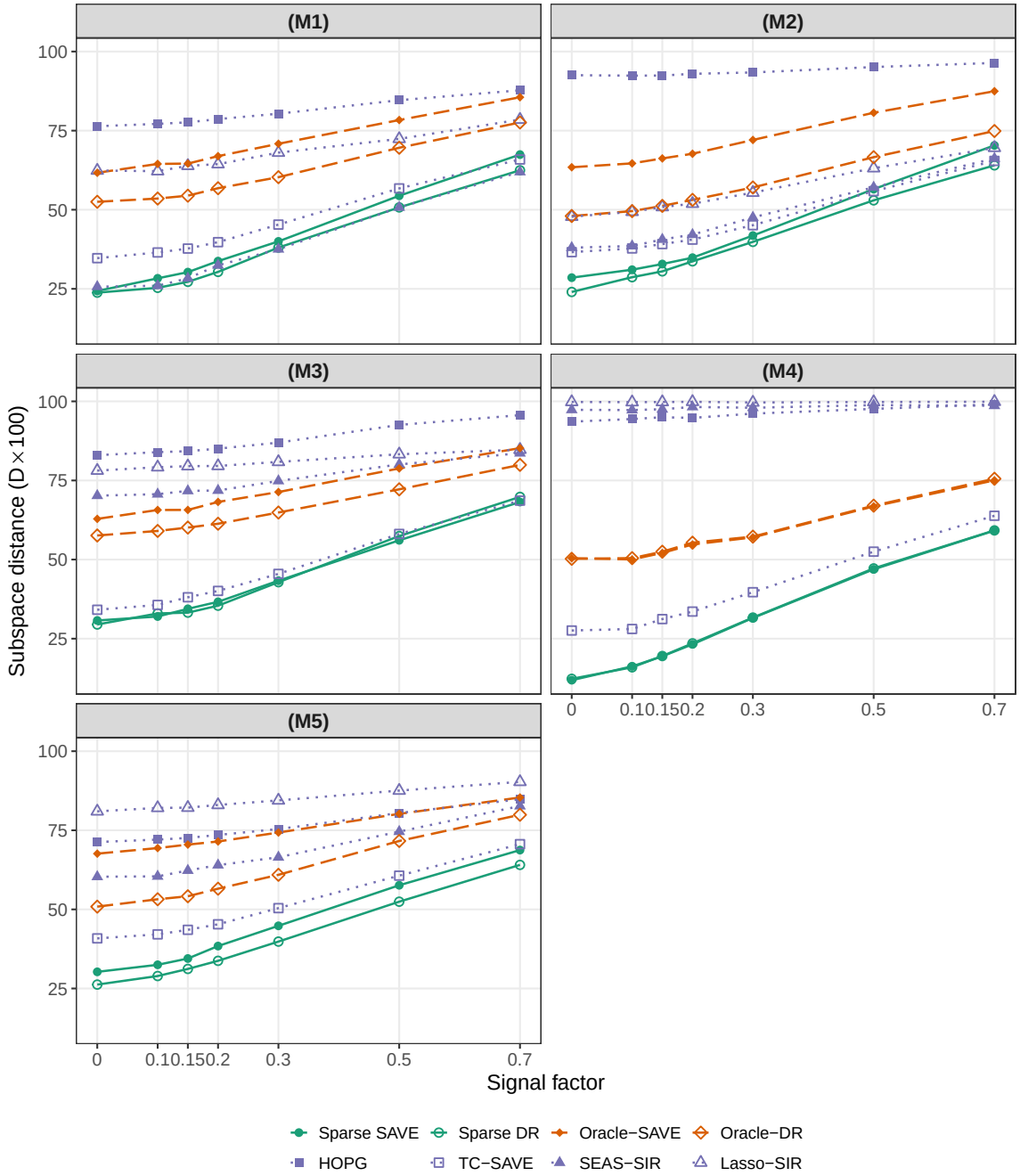

Figure S.7: Subspace estimation error versus signal factor ρ curves under Models (M1)–(M5), where $n = 300$ and $p = 700$.

Model	p	D (×100)							
		Sparse SAVE	Sparse DR	HOPG	TC-SAVE	SEAS-SIR	Lasso-SIR	Oracle-SAVE	Oracle-DR
(M1)	700	**24.3 (0.8)**	**23.8 (0.8)**	76.4 (0.3)	34.7 (0.7)	25.7 (1.1)	62.4 (1.0)	61.7 (1.3)	52.5 (1.3)
	1000	**24.5 (0.9)**	**23.8 (0.8)**	77.3 (0.3)	35.1 (0.6)	25.5 (1.1)	63.3 (1.1)	61.7 (1.3)	52.5 (1.3)
	1500	**25.3 (1.1)**	**23.1 (0.7)**	77.7 (0.3)	35.9 (0.6)	25.8 (1.1)	63.4 (1.1)	61.7 (1.3)	52.5 (1.3)
(M2)	700	**28.5 (1.3)**	**24.0 (0.7)**	92.5 (0.3)	36.6 (0.4)	38.0 (2.1)	47.8 (0.6)	63.4 (1.3)	48.0 (1.4)
	1000	**29.3 (1.4)**	**24.2 (0.7)**	93.0 (0.2)	37.5 (0.6)	38.1 (2.1)	49.3 (0.6)	63.4 (1.3)	48.0 (1.4)
	1500	**30.1 (1.5)**	**24.3 (0.8)**	93.2 (0.3)	37.5 (0.4)	37.7 (2.1)	50.6 (0.7)	63.4 (1.3)	48.0 (1.4)
(M3)	700	**30.7 (1.0)**	**29.5 (1.2)**	83.0 (0.6)	34.1 (0.9)	70.2 (0.9)	78.1 (0.3)	62.9 (1.2)	57.6 (1.0)
	1000	**30.9 (1.0)**	**30.7 (1.3)**	83.6 (0.6)	32.7 (0.8)	71.3 (0.9)	78.5 (0.4)	62.9 (1.2)	57.6 (1.0)
	1500	**31.5 (1.0)**	**31.2 (1.4)**	84.5 (0.7)	34.0 (0.9)	71.3 (0.9)	79.1 (0.3)	62.9 (1.2)	57.6 (1.0)
(M4)	700	**11.9 (0.4)**	**12.3 (0.6)**	93.6 (0.8)	27.6 (0.9)	97.3 (0.9)	99.8 (0.1)	50.4 (1.6)	50.2 (1.6)
	1000	**12.0 (0.4)**	**12.3 (0.6)**	96.8 (0.5)	25.1 (0.8)	98.4 (0.7)	99.8 (0.1)	50.4 (1.6)	50.2 (1.6)
	1500	**12.1 (0.4)**	**12.4 (0.6)**	98.4 (0.2)	25.8 (0.9)	99.3 (0.6)	100.0 (0.0)	50.4 (1.6)	50.2 (1.6)
(M5)	700	**30.3 (1.0)**	**26.2 (0.7)**	71.3 (0.5)	40.8 (1.2)	60.3 (0.8)	81.0 (0.7)	67.6 (1.1)	50.9 (1.2)
	1000	**30.0 (1.0)**	**26.3 (0.7)**	72.2 (0.4)	38.2 (1.1)	61.4 (0.9)	82.6 (0.6)	67.6 (1.1)	50.9 (1.2)
	1500	**30.7 (1.1)**	**26.5 (0.7)**	73.7 (0.5)	39.7 (1.2)	63.3 (0.9)	83.6 (0.6)	67.6 (1.1)	50.9 (1.2)
(Q1)	700	**6.7 (0.3)**	66.1 (1.3)	82.1 (0.5)	78.1 (0.7)	60.5 (0.8)	75.2 (0.5)	**6.8 (0.2)**	27.8 (1.6)
	1000	**7.0 (0.3)**	67.2 (1.2)	83.6 (0.6)	80.0 (0.5)	61.5 (0.8)	76.5 (0.5)	**6.8 (0.2)**	27.8 (1.6)
	1500	**7.4 (0.4)**	68.7 (1.2)	84.1 (0.5)	82.0 (0.4)	62.5 (0.9)	77.9 (0.5)	**6.8 (0.2)**	27.8 (1.6)
(Q2)	700	8.4 (0.2)	**8.2 (0.2)**	82.8 (0.4)	61.3 (0.9)	58.0 (0.5)	73.8 (0.5)	8.9 (0.2)	**8.2 (0.2)**
	1000	8.4 (0.2)	**8.1 (0.2)**	83.6 (0.4)	63.6 (0.7)	58.5 (0.5)	74.6 (0.6)	8.9 (0.2)	**8.2 (0.2)**
	1500	8.5 (0.2)	**8.2 (0.2)**	84.5 (0.4)	64.8 (0.7)	58.8 (0.5)	74.9 (0.6)	8.9 (0.2)	**8.2 (0.2)**
(Q3)	700	**11.9 (1.1)**	14.4 (1.0)	78.4 (0.7)	71.6 (1.9)	17.6 (0.7)	47.5 (0.9)	**7.4 (0.3)**	22.9 (0.8)
	1000	**14.6 (1.1)**	14.7 (1.0)	80.0 (0.6)	79.2 (1.8)	17.9 (0.6)	50.0 (0.9)	**7.4 (0.3)**	22.9 (0.8)
	1500	**16.0 (1.3)**	14.7 (1.0)	81.7 (0.6)	79.9 (1.3)	18.9 (0.7)	52.0 (0.9)	**7.4 (0.3)**	22.9 (0.8)
(Q4)	700	13.5 (0.8)	**10.2 (0.3)**	89.6 (0.3)	83.9 (0.4)	58.3 (0.6)	83.5 (0.6)	10.3 (0.3)	**10.2 (0.3)**
	1000	14.6 (0.9)	**10.3 (0.3)**	90.3 (0.3)	85.0 (0.2)	59.2 (0.7)	85.2 (0.6)	10.3 (0.3)	**10.2 (0.3)**
	1500	15.7 (1.0)	**10.2 (0.4)**	91.1 (0.3)	85.4 (0.2)	59.8 (0.7)	86.2 (0.6)	10.3 (0.3)	**10.2 (0.3)**

Table S.1: Mean (and standard error) of the subspace estimation error D (×100) under Models (M1)–(M5) and (Q1)–(Q4) with $n = 300$ and $p = 700, 1000, 1500$. The best two results in each setting are highlighted.

Model	n	D (×100)							
		Sparse SAVE	Sparse DR	HOPG	TC-SAVE	SEAS-SIR	Lasso-SIR	Oracle-SAVE	Oracle-DR
(M1)	300	**24.3 (0.8)**	**23.8 (0.8)**	76.4 (0.3)	34.7 (0.7)	25.7 (1.1)	62.4 (1.0)	61.7 (1.3)	52.5 (1.3)
	600	**19.8 (0.6)**	**18.5 (0.5)**	70.1 (0.4)	27.9 (0.6)	21.5 (0.4)	49.2 (0.7)	36.9 (1.0)	34.1 (1.2)
(M2)	300	**28.5 (1.3)**	**24.0 (0.7)**	92.5 (0.3)	36.6 (0.4)	38.0 (2.1)	47.8 (0.6)	63.4 (1.3)	48.0 (1.4)
	600	**18.7 (0.6)**	**18.8 (0.7)**	92.3 (0.2)	30.2 (0.4)	35.2 (2.1)	36.7 (0.5)	35.9 (1.3)	30.8 (1.0)
(M3)	300	**30.7 (1.0)**	**29.5 (1.2)**	83.0 (0.6)	34.1 (0.9)	70.2 (0.9)	78.1 (0.3)	62.9 (1.2)	57.6 (1.0)
	600	**23.2 (0.8)**	**22.7 (0.8)**	81.2 (0.4)	24.5 (0.5)	68.2 (0.7)	75.2 (0.5)	43.4 (1.4)	39.7 (1.2)
(M4)	300	**11.9 (0.4)**	**12.3 (0.6)**	93.6 (0.8)	27.6 (0.9)	97.3 (0.9)	99.8 (0.1)	50.4 (1.6)	50.2 (1.6)
	600	**10.4 (0.2)**	**10.3 (0.2)**	71.7 (0.7)	20.5 (0.6)	95.5 (1.1)	99.9 (0.1)	34.3 (1.2)	34.2 (1.2)
(M5)	300	**30.3 (1.0)**	**26.2 (0.7)**	71.3 (0.5)	40.8 (1.2)	60.3 (0.8)	81.0 (0.7)	67.6 (1.1)	50.9 (1.2)
	600	**22.6 (0.8)**	**21.8 (0.7)**	64.1 (0.4)	29.7 (0.7)	49.3 (1.0)	74.8 (0.8)	46.0 (1.5)	34.6 (0.9)
(Q1)	300	**6.7 (0.3)**	66.1 (1.3)	82.1 (0.5)	78.1 (0.7)	60.5 (0.8)	75.2 (0.5)	**6.8 (0.2)**	27.8 (1.6)
	600	**4.7 (0.2)**	32.4 (1.8)	76.4 (0.5)	59.5 (2.3)	53.9 (0.3)	65.2 (0.5)	**4.8 (0.1)**	20.4 (1.3)
(Q2)	300	8.4 (0.2)	**8.2 (0.2)**	82.8 (0.4)	61.3 (0.9)	58.0 (0.5)	73.8 (0.5)	8.9 (0.2)	**8.2 (0.2)**
	600	5.8 (0.2)	**5.5 (0.1)**	79.4 (0.3)	32.0 (0.7)	52.3 (0.6)	66.9 (0.6)	6.1 (0.2)	**5.6 (0.1)**
(Q3)	300	**11.9 (1.1)**	14.4 (1.0)	78.4 (0.7)	71.6 (1.9)	17.6 (0.7)	47.5 (0.9)	**7.4 (0.3)**	22.9 (0.8)
	600	**5.9 (0.2)**	9.2 (0.3)	74.6 (0.6)	38.3 (0.9)	11.6 (0.3)	31.4 (0.7)	**5.3 (0.2)**	16.5 (0.7)
(Q4)	300	13.5 (0.8)	**10.2 (0.3)**	89.6 (0.3)	83.9 (0.4)	58.3 (0.6)	83.5 (0.6)	10.3 (0.3)	**10.2 (0.3)**
	600	**7.2 (0.2)**	**7.1 (0.2)**	85.5 (0.3)	72.6 (1.2)	52.8 (0.4)	70.9 (0.9)	7.4 (0.2)	7.3 (0.2)

Table S.2: Mean (and standard error) of the subspace estimation error D (×100) under Models (M1)–(M5) and (Q1)–(Q4) with $p = 700$ and $n = 300, 600$. The best two results in each setting are highlighted.

Model	p	Classification errors (%)										
		Bayes	Sparse SAVE	Sparse DR	HOPG	TC-SAVE	SEAS-SIR	Lasso-SIR	MSDA	SLR	DA-QDA	SQDA
(Q1)	700	6.0 (0.1)	**6.5 (0.1)**	19.3 (0.7)	35.5 (0.4)	25.4 (0.5)	24.5 (0.7)	37.8 (0.5)	34.9 (0.5)	39.0 (0.5)	24.3 (0.4)	**18.8 (0.4)**
	1000	6.0 (0.1)	**6.4 (0.1)**	19.8 (0.7)	36.4 (0.4)	27.1 (0.5)	25.4 (0.7)	39.0 (0.4)	35.2 (0.5)	39.4 (0.5)	25.4 (0.4)	**19.2 (0.4)**
	1500	6.0 (0.1)	**6.5 (0.1)**	20.5 (0.8)	36.7 (0.4)	29.0 (0.4)	26.4 (0.8)	40.0 (0.4)	35.8 (0.5)	40.1 (0.5)	27.1 (0.4)	**19.5 (0.4)**
(Q2)	700	18.5 (0.2)	**20.0 (0.2)**	**20.0 (0.2)**	50.5 (0.3)	32.6 (0.3)	39.0 (0.6)	48.0 (0.4)	47.2 (0.4)	49.1 (0.3)	46.4 (0.3)	50.0 (0.7)
	1000	18.5 (0.2)	**20.0 (0.2)**	**20.0 (0.2)**	51.1 (0.3)	33.5 (0.3)	39.5 (0.5)	49.2 (0.5)	47.5 (0.4)	49.4 (0.3)	47.6 (0.3)	50.9 (0.7)
	1500	18.5 (0.2)	**20.0 (0.2)**	**20.0 (0.2)**	52.1 (0.3)	34.1 (0.3)	40.0 (0.6)	49.3 (0.5)	47.6 (0.4)	49.7 (0.3)	48.8 (0.3)	51.2 (0.7)
(Q3)	700	9.5 (0.1)	**10.6 (0.3)**	**10.6 (0.3)**	28.3 (0.4)	25.7 (0.8)	10.9 (0.2)	18.0 (0.3)	16.3 (0.2)	17.2 (0.3)	26.9 (0.3)	15.3 (0.2)
	1000	9.5 (0.1)	**10.9 (0.3)**	**10.6 (0.3)**	29.1 (0.4)	28.5 (0.9)	11.0 (0.2)	19.0 (0.4)	16.4 (0.2)	17.6 (0.3)	29.0 (0.3)	15.4 (0.3)
	1500	9.5 (0.1)	**11.0 (0.3)**	**10.7 (0.3)**	30.0 (0.4)	28.8 (0.6)	11.1 (0.2)	19.7 (0.4)	16.5 (0.2)	17.9 (0.3)	32.0 (0.4)	16.3 (0.4)
(Q4)	700	25.0 (0.2)	**27.2 (0.3)**	**26.5 (0.2)**	56.5 (0.3)	50.9 (0.4)	43.7 (0.4)	52.5 (0.4)	52.7 (0.4)	52.5 (0.3)	51.7 (0.3)	52.4 (0.6)
	1000	25.0 (0.2)	**27.4 (0.3)**	**26.5 (0.2)**	57.4 (0.3)	51.8 (0.3)	44.1 (0.5)	53.8 (0.4)	52.9 (0.4)	52.8 (0.3)	52.5 (0.3)	52.4 (0.6)
	1500	25.0 (0.2)	**27.7 (0.3)**	**26.5 (0.2)**	57.9 (0.3)	52.0 (0.3)	44.6 (0.5)	54.5 (0.4)	53.3 (0.4)	53.3 (0.3)	54.0 (0.3)	53.2 (0.6)

Table S.3: Mean (and standard error) of the classification errors under Models (Q1)–(Q4) with $n = 300$ and $p = 700, 1000, 1500$. The best two results excluding Bayes error in each setting are highlighted.

Model	n	Classification errors (%)										
		Bayes	Sparse SAVE	Sparse DR	HOPG	TC-SAVE	SEAS-SIR	Lasso-SIR	MSDA	SLR	DA-QDA	SQDA
(Q1)	300	6.0 (0.1)	**6.5 (0.1)**	19.3 (0.7)	35.5 (0.4)	25.4 (0.5)	24.5 (0.7)	37.8 (0.5)	34.9 (0.5)	39.0 (0.5)	24.3 (0.4)	**18.8 (0.4)**
	600	6.0 (0.1)	**6.2 (0.1)**	**10.4 (0.4)**	31.1 (0.4)	15.6 (0.5)	18.4 (0.3)	28.6 (0.5)	30.1 (0.4)	33.6 (0.4)	24.3 (0.4)	17.6 (0.3)
(Q2)	300	18.5 (0.2)	**20.0 (0.2)**	**20.0 (0.2)**	50.5 (0.3)	32.6 (0.3)	39.0 (0.6)	48.0 (0.4)	47.2 (0.4)	49.1 (0.3)	46.4 (0.3)	50.0 (0.7)
	600	18.5 (0.2)	**19.2 (0.2)**	**19.2 (0.2)**	48.3 (0.3)	24.5 (0.2)	33.5 (0.5)	42.0 (0.4)	43.2 (0.3)	45.6 (0.3)	46.4 (0.3)	54.0 (0.5)
(Q3)	300	9.5 (0.1)	**10.6 (0.3)**	**10.6 (0.3)**	28.3 (0.4)	25.7 (0.8)	10.9 (0.2)	18.0 (0.3)	16.3 (0.2)	17.2 (0.3)	26.9 (0.3)	15.3 (0.2)
	600	9.5 (0.1)	**9.7 (0.1)**	**10.0 (0.1)**	26.4 (0.3)	14.5 (0.2)	10.1 (0.1)	13.2 (0.2)	14.9 (0.2)	14.6 (0.2)	26.9 (0.3)	14.5 (0.2)
(Q4)	300	25.0 (0.2)	**27.2 (0.3)**	**26.5 (0.2)**	56.5 (0.3)	50.9 (0.4)	43.7 (0.4)	52.5 (0.4)	52.7 (0.4)	52.5 (0.3)	51.7 (0.3)	52.4 (0.6)
	600	25.0 (0.2)	**25.8 (0.2)**	**25.9 (0.2)**	54.3 (0.2)	41.8 (0.5)	39.6 (0.3)	44.8 (0.5)	48.4 (0.3)	48.9 (0.3)	51.7 (0.3)	52.4 (0.6)

Table S.4: Mean (and standard error) of the classification errors under Models (Q1)–(Q4) with $p = 700$ and $n = 300, 600$. The best two results excluding Bayes error in each setting are highlighted.

Model	p	Estimated structural dimension						
		Sparse SAVE	Sparse DR	TC-SAVE	SEAS-SIR	Lasso-SIR	Oracle-SAVE	Oracle-DR
(M1) ($d = 2$)	700	1.73 (0.04)	1.67 (0.05)	1.97 (0.02)	1.73 (0.04)	1.50 (0.05)	6.00 (0.00)	6.00 (0.00)
	1000	1.73 (0.04)	1.67 (0.05)	2.00 (0.00)	1.73 (0.04)	1.46 (0.05)	6.00 (0.00)	6.00 (0.00)
	1500	1.73 (0.04)	1.67 (0.05)	1.99 (0.01)	1.73 (0.04)	1.45 (0.05)	6.00 (0.00)	6.00 (0.00)
(M2) ($d = 2$)	700	1.88 (0.06)	1.81 (0.07)	2.00 (0.00)	2.30 (0.09)	1.99 (0.01)	6.00 (0.00)	6.00 (0.00)
	1000	1.91 (0.07)	1.89 (0.11)	2.00 (0.00)	2.34 (0.09)	1.99 (0.01)	6.00 (0.00)	6.00 (0.00)
	1500	1.90 (0.07)	1.76 (0.04)	2.00 (0.00)	2.37 (0.09)	1.99 (0.01)	6.00 (0.00)	6.00 (0.00)
(M3) ($d = 2$)	700	2.08 (0.08)	2.01 (0.01)	2.04 (0.02)	1.11 (0.05)	1.00 (0.00)	6.00 (0.00)	6.00 (0.00)
	1000	2.01 (0.01)	2.08 (0.08)	2.02 (0.01)	1.13 (0.05)	1.01 (0.01)	6.00 (0.00)	6.00 (0.00)
	1500	2.17 (0.11)	2.16 (0.11)	2.03 (0.02)	1.18 (0.06)	1.01 (0.01)	6.00 (0.00)	6.00 (0.00)
(M4) ($d = 1$)	700	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	3.76 (0.07)	1.29 (0.05)	6.00 (0.00)	6.00 (0.00)
	1000	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	3.80 (0.07)	1.20 (0.04)	6.00 (0.00)	6.00 (0.00)
	1500	1.00 (0.00)	1.00 (0.00)	1.01 (0.01)	3.90 (0.05)	1.35 (0.05)	6.00 (0.00)	6.00 (0.00)
(M5) ($d = 2$)	700	1.98 (0.03)	2.00 (0.00)	2.05 (0.03)	1.84 (0.04)	1.06 (0.02)	6.00 (0.00)	6.00 (0.00)
	1000	2.02 (0.03)	2.00 (0.00)	2.04 (0.02)	1.84 (0.04)	1.16 (0.04)	6.00 (0.00)	6.00 (0.00)
	1500	2.02 (0.03)	2.00 (0.00)	2.11 (0.04)	1.84 (0.05)	1.12 (0.03)	6.00 (0.00)	6.00 (0.00)
(Q1) ($d = 2$)	700	2.00 (0.00)	2.11 (0.09)	4.60 (0.11)	1.00 (0.00)	1.00 (0.00)	6.00 (0.00)	5.28 (0.13)
	1000	2.00 (0.00)	2.04 (0.09)	4.65 (0.10)	1.00 (0.00)	1.00 (0.00)	6.00 (0.00)	5.28 (0.13)
	1500	2.00 (0.00)	2.00 (0.08)	4.76 (0.07)	1.00 (0.00)	1.00 (0.00)	6.00 (0.00)	5.28 (0.13)
(Q2) ($d = 3$)	700	3.00 (0.00)	3.00 (0.00)	3.57 (0.12)	1.32 (0.05)	1.00 (0.00)	6.00 (0.00)	6.00 (0.00)
	1000	3.00 (0.00)	3.00 (0.00)	3.64 (0.12)	1.30 (0.05)	1.00 (0.00)	6.00 (0.00)	6.00 (0.00)
	1500	3.00 (0.00)	3.00 (0.00)	3.79 (0.12)	1.31 (0.05)	1.00 (0.00)	6.00 (0.00)	6.00 (0.00)
(Q3) ($d = 1$)	700	1.00 (0.00)	1.00 (0.00)	2.92 (0.16)	1.00 (0.00)	1.00 (0.00)	6.00 (0.00)	5.28 (0.12)
	1000	1.00 (0.00)	1.00 (0.00)	3.33 (0.14)	1.00 (0.00)	1.00 (0.00)	6.00 (0.00)	5.28 (0.12)
	1500	1.00 (0.00)	1.00 (0.00)	3.46 (0.14)	1.00 (0.00)	1.00 (0.00)	6.00 (0.00)	5.28 (0.12)
(Q4) ($d = 2$)	700	2.00 (0.00)	1.97 (0.02)	4.62 (0.10)	1.01 (0.01)	1.00 (0.00)	6.00 (0.00)	6.00 (0.00)
	1000	2.00 (0.00)	1.97 (0.02)	4.67 (0.09)	1.04 (0.02)	1.00 (0.00)	6.00 (0.00)	6.00 (0.00)
	1500	1.99 (0.01)	1.97 (0.02)	4.74 (0.07)	1.02 (0.01)	1.00 (0.00)	6.00 (0.00)	6.00 (0.00)

Table S.5: Mean (and standard error) of the estimated structural dimension under Models (M1)–(M5) and (Q1)–(Q4) with $n = 300$ and $p = 700, 1000, 1500$.

Model	n	Estimated structural dimension						
		Sparse SAVE	Sparse DR	TC-SAVE	SEAS-SIR	Lasso-SIR	Oracle-SAVE	Oracle-DR
(M1) ($d = 2$)	300	1.73 (0.04)	1.67 (0.05)	1.97 (0.02)	1.73 (0.04)	1.50 (0.05)	6.00 (0.00)	6.00 (0.00)
	600	1.71 (0.05)	1.85 (0.04)	2.00 (0.00)	1.82 (0.04)	1.71 (0.05)	6.00 (0.00)	6.00 (0.00)
(M2) ($d = 2$)	300	1.88 (0.06)	1.81 (0.07)	2.00 (0.00)	2.30 (0.09)	1.99 (0.01)	6.00 (0.00)	6.00 (0.00)
	600	1.75 (0.04)	1.74 (0.04)	2.00 (0.00)	1.93 (0.07)	2.00 (0.00)	6.00 (0.00)	6.00 (0.00)
(M3) ($d = 2$)	300	2.08 (0.08)	2.01 (0.01)	2.04 (0.02)	1.11 (0.05)	1.00 (0.00)	6.00 (0.00)	6.00 (0.00)
	600	2.00 (0.00)	2.00 (0.00)	2.00 (0.00)	1.00 (0.00)	1.00 (0.00)	6.00 (0.00)	6.00 (0.00)
(M4) ($d = 1$)	300	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	3.76 (0.07)	1.29 (0.05)	6.00 (0.00)	6.00 (0.00)
	600	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	3.44 (0.11)	1.25 (0.04)	6.00 (0.00)	6.00 (0.00)
(M5) ($d = 2$)	300	1.98 (0.03)	2.00 (0.00)	2.05 (0.03)	1.84 (0.04)	1.06 (0.02)	6.00 (0.00)	6.00 (0.00)
	600	2.00 (0.00)	2.00 (0.00)	2.00 (0.00)	2.03 (0.03)	1.00 (0.00)	6.00 (0.00)	6.00 (0.00)
(Q1) ($d = 2$)	300	2.00 (0.00)	2.11 (0.09)	4.60 (0.11)	1.00 (0.00)	1.00 (0.00)	6.00 (0.00)	5.28 (0.13)
	600	2.00 (0.00)	2.47 (0.08)	3.00 (0.18)	1.00 (0.00)	1.00 (0.00)	6.00 (0.00)	4.92 (0.13)
(Q2) ($d = 3$)	300	3.00 (0.00)	3.00 (0.00)	3.57 (0.12)	1.32 (0.05)	1.00 (0.00)	6.00 (0.00)	6.00 (0.00)
	600	3.00 (0.00)	3.00 (0.00)	3.03 (0.05)	1.60 (0.05)	1.00 (0.00)	6.00 (0.00)	6.00 (0.00)
(Q3) ($d = 1$)	300	1.00 (0.00)	1.00 (0.00)	2.92 (0.16)	1.00 (0.00)	1.00 (0.00)	6.00 (0.00)	5.28 (0.12)
	600	1.00 (0.00)	1.00 (0.00)	1.28 (0.06)	1.00 (0.00)	1.00 (0.00)	6.00 (0.00)	5.05 (0.14)
(Q4) ($d = 2$)	300	2.00 (0.00)	1.97 (0.02)	4.62 (0.10)	1.01 (0.01)	1.00 (0.00)	6.00 (0.00)	6.00 (0.00)
	600	2.00 (0.00)	2.00 (0.00)	3.61 (0.18)	1.01 (0.01)	1.00 (0.00)	6.00 (0.00)	6.00 (0.00)

Table S.6: Mean (and standard error) of the estimated structural dimension under Models (M1)–(M5) and (Q1)–(Q4) with $p = 700$ and $n = 300, 600$.

Model	p	TPR(%) / FPR(%)					
		Sparse SAVE	Sparse DR	HOPG	TC-SAVE	SEAS-SIR	Lasso-SIR
(M1)	700	99.0/0.1	99.0/0.0	100.0/6.6	100.0/0.0	99.0/0.4	93.3/4.5
	1000	98.8/0.0	99.2/0.0	100.0/4.6	100.0/0.0	99.0/0.2	92.7/3.0
	1500	98.5/0.0	99.3/0.0	100.0/3.1	100.0/0.0	99.0/0.2	92.3/2.1
(M2)	700	96.8/0.0	99.7/0.1	99.7/6.6	100.0/0.0	95.0/1.8	99.0/7.1
	1000	96.3/0.0	99.7/0.0	99.7/4.6	99.7/0.0	95.0/1.3	99.2/5.1
	1500	95.7/0.0	99.7/0.0	99.2/3.1	99.8/0.0	95.0/0.8	99.3/3.3
(M3)	700	98.3/0.1	99.8/0.1	91.5/6.7	99.8/0.0	75.2/5.9	71.7/5.4
	1000	97.8/0.0	99.8/0.2	89.0/4.7	100.0/0.0	73.8/4.3	69.8/3.8
	1500	97.8/0.0	99.5/0.0	86.0/3.1	100.0/0.0	71.0/2.4	68.5/2.8
(M4)	700	100.0/0.0	99.7/0.0	100.0/6.6	99.7/0.1	20.2/22.7	1.5/0.8
	1000	100.0/0.0	99.7/0.0	100.0/4.6	99.8/0.0	17.6/16.6	0.8/0.8
	1500	100.0/0.0	99.7/0.0	100.0/3.1	99.7/0.0	10.7/9.4	1.2/0.7
(M5)	700	98.5/0.0	99.5/0.0	100.0/6.6	99.7/0.1	74.2/1.3	64.0/5.3
	1000	98.5/0.0	99.5/0.0	100.0/4.6	100.0/0.0	72.8/1.1	62.5/4.4
	1500	97.8/0.0	99.5/0.0	100.0/3.1	99.5/0.0	69.5/0.7	60.8/3.1
(Q1)	700	100.0/0.1	55.3/0.5	97.3/6.7	58.5/0.0	70.2/1.2	72.0/4.5
	1000	99.8/0.1	53.5/0.3	96.2/4.7	54.7/0.0	67.0/1.0	70.7/3.8
	1500	99.8/0.1	51.0/0.3	94.5/3.1	50.5/0.0	63.8/0.7	67.5/2.9
(Q2)	700	100.0/0.0	100.0/0.0	99.7/6.6	74.2/0.0	73.5/1.3	68.0/4.7
	1000	100.0/0.0	100.0/0.0	99.3/4.6	71.7/0.0	72.5/1.1	66.7/3.6
	1500	100.0/0.0	100.0/0.0	99.3/3.1	70.2/0.0	71.7/0.8	65.0/2.6
(Q3)	700	98.3/0.8	98.5/0.0	100.0/6.6	58.3/0.0	100.0/0.4	98.0/4.3
	1000	97.7/0.8	98.3/0.0	100.0/4.6	46.3/0.0	100.0/0.3	98.0/3.6
	1500	96.8/0.6	98.3/0.0	100.0/3.1	45.0/0.0	100.0/0.2	98.0/2.7
(Q4)	700	99.5/0.4	99.8/0.0	98.3/6.6	44.0/0.0	76.7/1.9	71.0/5.2
	1000	98.3/0.2	99.8/0.0	97.8/4.6	40.0/0.0	74.8/1.3	69.5/4.3
	1500	97.5/0.1	99.8/0.0	96.2/3.1	37.0/0.0	74.0/1.1	66.0/3.1

Table S.7: Mean of TPR (%) and FPR (%) under Models (M1)–(M5) and (Q1)–(Q4) with $n = 300$ and $p = 700, 1000, 1500$. The standard errors of TPR and FPR are less than 2.9% and 1.3% and are omitted.

Model	H	D ($\times 100$)			
		Oracle-SAVE	Oracle-DR	Sparse SAVE	Sparse DR
(M1)	3	50.9 (1.1)	62.5 (1.1)	**19.7 (0.5)**	**21.0 (0.5)**
	5	61.7 (1.3)	52.5 (1.3)	**20.8 (0.6)**	**19.5 (0.5)**
	8	73.8 (1.1)	60.3 (1.2)	**20.5 (0.7)**	**19.3 (0.5)**
(M2)	3	64.3 (1.2)	42.2 (0.8)	**21.5 (0.6)**	**20.1 (0.5)**
	5	63.4 (1.3)	48.0 (1.4)	**19.4 (0.5)**	**17.8 (0.4)**
	8	74.5 (0.9)	63.5 (1.1)	**17.6 (0.4)**	**16.7 (0.5)**
(M3)	3	51.0 (1.2)	47.3 (1.1)	**22.4 (0.5)**	**20.1 (0.5)**
	5	62.9 (1.2)	57.6 (1.0)	**21.1 (0.5)**	**19.8 (0.5)**
	8	73.9 (0.8)	67.7 (0.7)	**20.6 (0.5)**	**19.1 (0.5)**
(M4)	3	41.1 (1.5)	41.2 (1.5)	**10.1 (0.2)**	**10.0 (0.2)**
	5	50.4 (1.6)	50.2 (1.6)	**10.2 (0.3)**	**10.1 (0.3)**
	8	64.7 (1.8)	64.5 (1.7)	**10.7 (0.2)**	**11.1 (0.3)**
(M5)	3	55.0 (1.3)	46.2 (1.0)	**23.4 (0.6)**	**23.0 (0.5)**
	5	67.6 (1.1)	50.9 (1.2)	**24.0 (0.5)**	**22.4 (0.5)**
	8	76.0 (0.8)	65.4 (1.2)	**25.9 (0.6)**	**23.3 (0.5)**

Table S.8: Mean (and standard error) of the subspace estimation error D ($\times 100$) for oracle methods and our proposals with $H = 3, 5, 8$ under Models (M1)–(M5) with $n = 300$. The active set $\mathcal{A}$ and structural dimension d are given. The best two results in each setting are highlighted.

Model	p	Execution time (sec.)						
		Sparse SAVE	Sparse DR	TC-SAVE	SEAS-SIR	Lasso-SIR	DA-QDA	SQDA
(Q1)	700	1.64 (0.05)	**0.97 (0.03)**	158.12 (2.83)	54.87 (1.59)	30.01 (0.75)	17.12 (0.04)	**1.63 (0.04)**
	1500	**11.02 (0.34)**	**7.81 (0.25)**	837.31 (13.48)	651.40 (10.04)	140.02 (2.47)	147.26 (0.74)	11.39 (0.64)
(Q2)	700	7.71 (0.27)	**6.08 (0.22)**	140.71 (2.07)	64.97 (2.23)	53.20 (0.39)	36.49 (0.38)	**4.30 (0.06)**
	1500	**42.52 (1.14)**	52.45 (4.54)	950.96 (10.12)	714.83 (12.49)	220.75 (0.87)	282.99 (1.55)	**46.24 (1.37)**
(Q3)	700	5.52 (0.11)	**2.48 (0.08)**	110.69 (2.22)	54.20 (1.50)	34.58 (1.26)	17.73 (0.05)	**2.10 (0.05)**
	1500	46.65 (1.38)	**19.53 (0.62)**	677.48 (10.85)	687.31 (13.06)	149.84 (5.18)	147.13 (0.94)	**21.29 (0.76)**
(Q4)	700	**6.21 (0.14)**	30.67 (1.08)	175.23 (3.96)	60.18 (2.45)	95.89 (0.32)	33.65 (0.13)	**4.25 (0.07)**
	1500	**40.77 (0.94)**	43.36 (1.08)	971.04 (17.18)	710.56 (11.88)	223.82 (0.79)	295.25 (1.75)	**43.04 (1.38)**

Table S.9: Execution time (in seconds) under Models (Q1)–(Q4), where $n = 300$ and $p = 700, 1500$. The best two results in each setting are highlighted.

Model	ν	D ($\times 100$)							
		Sparse SAVE	Sparse DR	HOPG	TC-SAVE	SEAS-SIR	Lasso-SIR	Oracle-SAVE	Oracle-DR
(M3)	5	80.2 (3.5)	75.6 (2.4)	99.5 (0.1)	**54.0 (1.7)**	77.6 (0.6)	79.6 (0.3)	75.6 (1.0)	**63.5 (0.8)**
	6	80.1 (1.8)	68.9 (1.8)	98.5 (0.3)	**48.3 (1.6)**	75.4 (0.7)	79.1 (0.4)	72.8 (1.0)	**61.2 (0.9)**
	7	79.2 (2.0)	65.3 (1.9)	97.8 (0.4)	**47.8 (1.5)**	74.7 (0.6)	78.5 (0.3)	70.2 (1.1)	**60.6 (0.9)**
	8	73.8 (2.3)	63.5 (2.0)	96.6 (0.4)	**44.3 (1.4)**	75.0 (0.7)	78.7 (0.4)	69.4 (1.0)	**57.0 (1.1)**
	∞	**30.7 (1.0)**	**29.5 (1.2)**	83.0 (0.6)	34.1 (0.9)	70.2 (0.9)	78.1 (0.3)	62.9 (1.2)	57.6 (1.0)
(M4)	5	57.1 (4.4)	56.1 (4.3)	89.5 (0.9)	**52.1 (2.0)**	99.5 (0.2)	99.9 (0.1)	**54.4 (1.6)**	54.6 (1.5)
	6	**33.4 (2.9)**	**33.3 (2.9)**	88.9 (1.0)	42.1 (1.6)	98.5 (0.6)	99.8 (0.1)	51.3 (1.5)	50.9 (1.5)
	7	**25.3 (2.2)**	**25.3 (2.3)**	91.2 (1.0)	38.1 (1.2)	97.5 (0.9)	99.7 (0.1)	49.4 (1.6)	49.6 (1.6)
	8	**19.8 (1.8)**	**20.6 (1.8)**	90.6 (1.0)	36.6 (1.3)	98.3 (0.5)	99.7 (0.1)	50.0 (1.6)	49.8 (1.6)
	∞	**11.9 (0.4)**	**12.3 (0.6)**	93.6 (0.8)	27.6 (0.9)	97.3 (0.9)	99.8 (0.1)	50.4 (1.6)	50.2 (1.6)
(M5)	5	**58.3 (3.7)**	**58.6 (3.8)**	72.7 (0.6)	60.2 (1.6)	74.4 (1.0)	86.1 (0.5)	70.7 (0.9)	63.5 (1.3)
	6	53.9 (2.2)	**48.0 (2.5)**	71.9 (0.6)	**51.9 (1.5)**	70.9 (1.1)	85.9 (0.6)	69.4 (1.0)	58.3 (1.4)
	7	**49.8 (2.0)**	**43.2 (2.0)**	72.4 (0.5)	53.5 (1.4)	68.0 (1.0)	85.4 (0.6)	69.4 (0.9)	57.8 (1.2)
	8	**45.0 (2.1)**	**37.6 (1.9)**	71.7 (0.5)	48.7 (1.3)	66.0 (0.9)	84.0 (0.6)	69.1 (0.9)	56.5 (1.3)
	∞	**30.3 (1.0)**	**26.2 (0.7)**	71.3 (0.5)	40.8 (1.2)	60.3 (0.8)	81.0 (0.7)	67.6 (1.1)	50.9 (1.2)

Table S.10: Mean (and standard error) of the subspace estimation error D ($\times 100$) under Models (M3)–(M5), where $n = 300$, $p = 700$, and predictors follow the multivariate t and normal distributions ($\nu = \infty$). The best two results in each setting are highlighted.

Model	ν	TPR(%) / FPR(%)					
		Sparse SAVE	Sparse DR	HOPG	TC-SAVE	SEAS-SIR	Lasso-SIR
(M3)	5	3.0/0.0	13.7/0.0	29.3/7.2	92.5/0.0	77.0/13.1	70.7/7.0
	6	25.3/0.9	53.0/0.4	41.5/7.1	95.0/0.0	74.8/11.4	71.2/6.8
	7	50.0/2.7	75.7/1.0	47.0/7.1	96.7/0.0	77.2/10.8	72.3/7.1
	8	65.2/3.2	86.7/1.8	53.0/7.0	96.7/0.0	77.7/11.6	71.8/6.4
	∞	98.3/0.1	99.8/0.1	91.5/6.7	99.8/0.0	75.2/5.9	71.7/5.4
(M4)	5	26.5/0.0	26.0/0.0	100.0/6.6	87.8/0.0	20.7/22.7	2.3/1.4
	6	66.3/0.0	66.7/0.0	100.0/6.6	91.0/0.0	21.0/22.5	2.0/1.2
	7	87.3/0.1	87.0/0.1	100.0/6.6	95.3/0.0	25.2/23.7	2.7/1.4
	8	92.5/0.1	92.8/0.1	100.0/6.6	93.8/0.0	21.2/24.1	3.0/1.2
	∞	100.0/0.0	99.7/0.0	100.0/6.6	99.7/0.1	20.2/22.7	1.5/0.8
(M5)	5	21.5/0.0	23.3/0.0	100.0/6.6	92.0/0.1	55.3/1.9	58.2/5.7
	6	53.3/0.0	57.8/0.0	100.0/6.6	96.2/0.0	60.2/1.7	56.8/5.7
	7	73.8/0.0	80.2/0.1	100.0/6.6	95.7/0.0	62.5/1.6	55.7/6.4
	8	81.7/0.1	88.5/0.1	100.0/6.6	97.5/0.0	65.8/1.4	56.5/6.0
	∞	98.5/0.0	99.5/0.0	100.0/6.6	99.7/0.1	74.2/1.3	64.0/5.3

Table S.11: Mean (and standard error) of TPR(%) / FPR(%) under Models (M3)–(M5), where $n = 300$, $p = 700$, and predictors follow the multivariate t and normal distributions ($\nu = \infty$). Over all settings, the maximal standard errors of TPR(%) / FPR(%) are 4.1/0.8, 4.5/0.5, 2.6/0.0, 1.7/0.0, 3.3/1.3, 1.6/0.4 for Sparse SAVE, Sparse DR, HOPG, TC-SAVE, SEAS-SIR, Lasso-SIR, respectively.

Model	D ($\times 100$)							
	Sparse SAVE	Sparse DR	HOPG	TC-SAVE	SEAS-SIR	Lasso-SIR	Oracle-SAVE	Oracle-DR
(Q5-i)	**29.1 (0.9)**	**23.4 (0.8)**	86.5 (0.4)	72.8 (0.4)	60.5 (0.6)	76.8 (0.4)	14.1 (0.2)	13.0 (0.2)
(Q5-ii)	**17.6 (0.7)**	**13.0 (0.5)**	82.5 (0.4)	68.7 (0.5)	62.8 (0.5)	75.0 (0.3)	12.7 (0.2)	11.7 (0.2)
(Q5-iii)	**16.4 (0.6)**	**12.1 (0.4)**	76.9 (0.3)	58.9 (0.5)	63.8 (0.3)	74.5 (0.2)	11.1 (0.2)	10.8 (0.2)
(Q5-iv)	**10.4 (0.4)**	**9.4 (0.3)**	76.9 (0.3)	51.3 (0.4)	63.8 (0.3)	73.2 (0.2)	9.5 (0.2)	9.3 (0.1)

Table S.12: Mean (and standard error) of the subspace estimation error D ($\times 100$) under Models (Q5i)–(Q5iv), where $n = 300$ and $p = 700$. The best two results excluding oracle methods in each setting are highlighted.

Model	Classification errors (%)										
	Bayes	Sparse SAVE	Sparse DR	HOPG	TC-SAVE	SEAS-SIR	Lasso-SIR	MSDA	SLR	DA-QDA	SQDA
(Q5-i)	18.5 (0.2)	**24.6 (0.3)**	**23.3 (0.3)**	52.3 (0.4)	39.0 (0.5)	38.2 (0.5)	50.1 (0.4)	48.2 (0.4)	50.0 (0.4)	48.4 (0.4)	51.6 (0.7)
(Q5-ii)	14.0 (0.2)	**17.5 (0.2)**	**16.5 (0.2)**	48.6 (0.4)	33.7 (0.5)	37.4 (0.5)	49.6 (0.5)	48.3 (0.4)	49.5 (0.4)	45.5 (0.4)	48.9 (0.7)
(Q5-iii)	11.3 (0.2)	**14.4 (0.2)**	**13.9 (0.2)**	43.1 (0.4)	25.9 (0.3)	35.5 (0.5)	50.2 (0.4)	48.0 (0.4)	49.9 (0.4)	43.0 (0.4)	48.3 (0.9)
(Q5-iv)	9.2 (0.2)	**11.5 (0.2)**	**11.4 (0.2)**	43.1 (0.4)	19.9 (0.3)	35.5 (0.5)	49.0 (0.5)	47.6 (0.4)	49.6 (0.3)	41.2 (0.3)	43.6 (0.7)

Table S.13: Mean (and standard error) of the classification errors under Models (Q5i)–(Q5iv), where $n = 300$ and $p = 700$. The best two results excluding Bayes error in each setting are highlighted.

Model	TPR(%) / FPR(%)					
	Sparse SAVE	Sparse DR	HOPG	TC-SAVE	SEAS-SIR	Lasso-SIR
(Q5-i)	86.6/0.5	93.7/0.5	74.9/6.5	35.7/0.0	48.6/0.8	51.5/5.4
(Q5-ii)	96.5/0.2	99.4/0.1	89.6/6.2	52.8/0.0	49.1/0.9	48.6/5.2
(Q5-iii)	95.1/0.1	99.2/0.1	99.1/6.1	68.1/0.0	48.4/0.7	51.8/5.5
(Q5-iv)	99.3/0.0	99.7/0.0	99.1/6.1	79.7/0.0	48.4/0.7	49.5/5.1

Table S.14: Mean (and standard error) of TPR(%) / FPR(%) under Models (Q5-i)–(Q5-iv), where $n = 300$ and $p = 700$. The maximal standard errors of TPR/FPR are 0.9%/0.1%, 0.9%/0.1%, 1.1%/0.0%, 1.5%/0.0%, 1.1%/0.2%, 1.8%/0.4%, corresponding to each method.

Model	Execution time (sec.)						
	Sparse SAVE	Sparse DR	TC-SAVE	SEAS-SIR	Lasso-SIR	DA-QDA	SQDA
(Q5-i)	**9.84 (0.46)**	13.29 (0.31)	102.66 (2.02)	48.86 (0.29)	56.42 (0.34)	35.43 (0.22)	**4.22 (0.06)**
(Q5-ii)	17.25 (0.90)	**11.60 (0.31)**	119.38 (2.60)	46.28 (0.23)	109.30 (1.61)	32.24 (0.04)	**4.04 (0.05)**
(Q5-iii)	7.55 (0.32)	**4.48 (0.16)**	145.14 (3.32)	51.24 (0.17)	53.90 (0.24)	34.64 (0.14)	**4.09 (0.05)**
(Q5-iv)	**3.31 (0.12)**	4.87 (0.17)	133.37 (3.24)	74.87 (0.24)	91.70 (0.44)	32.98 (0.06)	**3.58 (0.05)**

Table S.15: Execution time (in seconds) under Models (Q5i)–(Q5iv), where $n = 300$ and $p = 700$. The best two results in each setting are highlighted.

Model	Setting	Sparse SAVE	Sparse DR	HOPG	TC-SAVE	SEAS-SIR	Lasso-SIR	Oracle-SAVE	Oracle-DR
	0	**24.3 (0.8)**	**23.8 (0.8)**	76.4 (0.3)	34.7 (0.7)	25.7 (1.1)	62.4 (1.0)	61.7 (1.3)	52.5 (1.3)
	0.1	28.3 (0.9)	**25.3 (0.6)**	77.1 (0.4)	36.5 (0.8)	**26.0 (0.8)**	62.2 (1.0)	64.5 (1.2)	53.5 (1.3)
	0.15	30.3 (0.8)	**27.2 (0.5)**	77.6 (0.3)	37.7 (0.7)	**28.3 (0.8)**	63.9 (1.0)	64.6 (1.2)	54.4 (1.3)
(M1)	0.2	33.8 (0.8)	**30.3 (0.6)**	78.6 (0.4)	39.8 (0.7)	**32.4 (0.9)**	64.4 (0.9)	66.9 (1.1)	56.8 (1.2)
	0.3	40.0 (0.6)	**38.0 (0.7)**	80.3 (0.4)	45.3 (0.6)	**37.5 (0.6)**	68.0 (0.9)	70.8 (1.0)	60.3 (1.1)
	0.5	54.4 (0.7)	**50.7 (0.4)**	84.6 (0.3)	56.7 (0.4)	**50.7 (0.3)**	72.4 (0.7)	78.3 (0.9)	69.6 (0.8)
	0.7	67.4 (0.7)	**62.5 (0.3)**	87.7 (0.2)	65.9 (0.3)	**61.9 (0.3)**	78.5 (0.6)	85.6 (0.6)	77.6 (0.5)
	0	**28.5 (1.3)**	**24.0 (0.7)**	92.5 (0.3)	36.6 (0.4)	38.0 (2.1)	47.8 (0.6)	63.4 (1.3)	48.0 (1.4)
	0.1	**31.0 (1.4)**	**28.7 (1.2)**	92.4 (0.3)	37.8 (0.4)	38.5 (1.9)	49.3 (0.7)	64.7 (1.2)	49.6 (1.4)
	0.15	**32.8 (1.4)**	**30.5 (1.0)**	92.4 (0.3)	39.1 (0.4)	40.6 (1.9)	50.8 (0.6)	66.2 (1.1)	51.2 (1.3)
(M2)	0.2	**34.8 (1.1)**	**33.7 (1.0)**	93.0 (0.2)	40.6 (0.3)	42.1 (1.7)	51.9 (0.6)	67.7 (1.1)	53.1 (1.3)
	0.3	**41.8 (1.1)**	**39.8 (0.9)**	93.4 (0.2)	45.1 (0.3)	47.6 (1.5)	55.4 (0.5)	72.1 (1.0)	57.0 (1.1)
	0.5	56.5 (1.1)	**53.0 (0.6)**	95.1 (0.2)	**55.9 (0.3)**	57.1 (1.1)	63.1 (0.4)	80.6 (0.7)	66.6 (0.8)
	0.7	70.3 (0.9)	**64.0 (0.5)**	96.4 (0.2)	**65.4 (0.2)**	66.2 (1.0)	69.6 (0.3)	87.5 (0.5)	74.9 (0.6)
	0	**30.7 (1.0)**	**29.5 (1.2)**	83.0 (0.6)	34.1 (0.9)	70.2 (0.9)	78.1 (0.3)	62.9 (1.2)	57.6 (1.0)
	0.1	**32.0 (0.9)**	**33.0 (1.3)**	84.0 (0.6)	35.7 (0.8)	70.7 (0.9)	79.2 (0.2)	65.7 (1.0)	59.0 (0.9)
	0.15	**34.4 (0.9)**	**33.2 (1.0)**	84.2 (0.6)	38.1 (0.8)	71.7 (0.8)	79.5 (0.2)	65.7 (1.0)	60.1 (0.9)
(M3)	0.2	**36.6 (0.9)**	**35.4 (1.0)**	85.1 (0.6)	40.1 (0.8)	71.8 (0.9)	79.6 (0.2)	68.2 (1.0)	61.3 (0.8)
	0.3	**43.3 (0.8)**	**42.8 (1.0)**	87.0 (0.5)	45.5 (0.7)	74.9 (0.8)	80.9 (0.2)	71.3 (0.9)	64.9 (0.8)
	0.5	**56.1 (0.7)**	**57.4 (0.9)**	92.6 (0.5)	58.1 (0.6)	80.1 (0.5)	83.3 (0.2)	78.8 (0.7)	72.2 (0.6)
	0.7	**68.2 (0.8)**	69.8 (0.9)	95.6 (0.4)	**68.6 (0.5)**	83.6 (0.2)	84.7 (0.2)	85.2 (0.5)	79.9 (0.4)
	0	**11.9 (0.4)**	**12.3 (0.6)**	93.6 (0.8)	27.6 (0.9)	97.3 (0.9)	99.8 (0.1)	50.4 (1.6)	50.2 (1.6)
	0.1	**16.2 (0.5)**	**15.9 (0.4)**	94.4 (0.8)	28.1 (0.8)	97.3 (1.0)	99.8 (0.1)	50.0 (1.5)	50.4 (1.5)
	0.15	**19.4 (0.3)**	**19.5 (0.4)**	95.0 (0.7)	31.2 (0.8)	97.5 (0.9)	99.8 (0.1)	52.1 (1.5)	52.5 (1.4)
(M4)	0.2	**23.2 (0.3)**	**23.5 (0.4)**	94.8 (0.7)	33.6 (0.7)	98.2 (0.7)	99.9 (0.1)	54.8 (1.4)	55.3 (1.4)
	0.3	**31.5 (0.3)**	**31.7 (0.4)**	96.1 (0.6)	39.7 (0.6)	98.0 (0.6)	99.7 (0.2)	57.0 (1.2)	57.2 (1.2)
	0.5	**46.9 (0.3)**	**47.2 (0.4)**	97.6 (0.4)	52.5 (0.5)	98.8 (0.4)	99.8 (0.1)	66.8 (1.0)	67.0 (1.0)
	0.7	**59.2 (0.3)**	**59.2 (0.3)**	99.1 (0.1)	63.8 (0.4)	98.6 (0.5)	99.9 (0.1)	75.0 (0.8)	75.5 (0.9)
	0	**30.3 (1.0)**	**26.2 (0.7)**	71.3 (0.5)	40.8 (1.2)	60.3 (0.8)	81.0 (0.7)	67.6 (1.1)	50.9 (1.2)
	0.1	**32.5 (1.1)**	**28.9 (0.8)**	72.1 (0.5)	42.1 (1.2)	60.4 (0.9)	82.1 (0.6)	69.4 (1.0)	53.2 (1.1)
	0.15	**34.5 (1.1)**	**31.2 (0.7)**	72.6 (0.5)	43.5 (1.2)	62.3 (0.9)	82.2 (0.6)	70.5 (0.9)	54.1 (1.1)
(M5)	0.2	**38.4 (1.3)**	**33.7 (0.7)**	73.5 (0.5)	45.3 (1.1)	64.0 (0.8)	83.0 (0.6)	71.5 (0.8)	56.5 (1.1)
	0.3	**44.8 (1.3)**	**39.8 (0.6)**	75.4 (0.5)	50.4 (1.0)	66.5 (0.8)	84.4 (0.4)	74.3 (0.7)	60.9 (0.9)
	0.5	**57.6 (0.9)**	**52.4 (0.5)**	80.5 (0.4)	60.7 (0.8)	74.6 (0.7)	87.6 (0.4)	80.2 (0.5)	71.6 (0.9)
	0.7	**68.7 (0.7)**	**64.1 (0.4)**	84.8 (0.3)	70.6 (0.7)	82.7 (0.5)	90.3 (0.3)	85.4 (0.4)	79.9 (0.7)

Table S.16: Mean (and standard error) of the subspace estimation error D ($\times 100$) of competing methods under Models (M1)–(M5) and their non-sparse counterparts, where $n = 300$ and $p = 700$. The best two results in each setting are highlighted.

Model	Method	$D(\times 100)$	TPR(%)	FPR(%)	Time (sec.)	$D(\times 100)$	TPR(%)	FPR(%)	Time (sec.)
		Sparse SAVE				Sparse DR			
(M1)	Sequential	24.7 (0.8)	99.0 (0.6)	0.1 (0.0)	4776.4 (115.5)	23.4 (0.8)	99.0 (0.6)	0.0 (0.0)	18228.1 (1735.0)
	Grid-search	24.7 (0.9)	98.7 (0.7)	0.1 (0.0)	19530.5 (336.9)	23.7 (0.8)	98.8 (0.6)	0.0 (0.0)	17634.0 (1569.1)
(M2)	Sequential	28.5 (1.3)	96.8 (1.2)	0.0 (0.0)	4860.4 (90.1)	24.0 (0.7)	99.7 (0.2)	0.1 (0.0)	9288.4 (231.5)
	Grid-search	29.1 (1.1)	98.5 (0.5)	0.1 (0.0)	19694.1 (266.7)	24.9 (0.8)	99.5 (0.3)	0.1 (0.0)	12197.1 (999.3)
(M3)	Sequential	30.7 (1.0)	98.3 (0.5)	0.1 (0.0)	4946.4 (123.4)	29.5 (1.2)	99.8 (0.2)	0.1 (0.0)	12270.7 (441.8)
	Grid-search	31.2 (1.0)	98.7 (0.5)	0.1 (0.0)	29898.3 (2475.5)	30.4 (1.3)	100.0 (0.0)	0.1 (0.0)	42069.0 (3426.6)
(M4)	Sequential	11.9 (0.4)	100.0 (0.0)	0.0 (0.0)	4384.5 (125.6)	12.3 (0.6)	99.7 (0.3)	0.0 (0.0)	10068.7 (275.9)
	Grid-search	12.3 (0.4)	100.0 (0.0)	0.2 (0.0)	28014.2 (3037.7)	12.1 (0.4)	100.0 (0.0)	0.1 (0.0)	26314.7 (2684.6)
(M5)	Sequential	30.3 (1.0)	98.5 (0.6)	0.0 (0.0)	4400.3 (103.9)	26.2 (0.7)	99.5 (0.3)	0.0 (0.0)	13242.4 (811.8)
	Grid-search	28.8 (0.8)	99.0 (0.4)	0.1 (0.0)	15940.7 (779.9)	25.9 (0.7)	99.7 (0.2)	0.0 (0.0)	33025.6 (1893.2)

Table S.17: Comparison results of sparse SAVE/DR using sequential tuning and grid-search tuning strategies, where $n = 300$ and $p = 700$.

Model	Method	δ				
		10^{-4}	5×10^{-4}	10^{-3}	5×10^{-3}	10^{-2}
(M1)	sparse SAVE	1.73 (0.04)	1.73 (0.04)	1.73 (0.04)	1.73 (0.04)	1.73 (0.04)
	sparse DR	1.67 (0.05)	1.67 (0.05)	1.67 (0.05)	1.67 (0.05)	1.67 (0.05)
(M2)	sparse SAVE	1.88 (0.06)	1.88 (0.06)	1.88 (0.06)	1.87 (0.06)	1.86 (0.06)
	sparse DR	1.81 (0.07)	1.81 (0.07)	1.81 (0.07)	1.81 (0.07)	1.81 (0.07)
(M3)	sparse SAVE	2.08 (0.08)	2.08 (0.08)	2.08 (0.08)	2.08 (0.08)	2.08 (0.08)
	sparse DR	2.01 (0.01)	2.01 (0.01)	2.01 (0.01)	2.01 (0.01)	2.01 (0.01)
(M4)	sparse SAVE	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)
	sparse DR	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)
(M5)	sparse SAVE	1.98 (0.03)	1.98 (0.03)	1.98 (0.03)	2.00 (0.03)	1.98 (0.02)
	sparse DR	2.00 (0.00)	2.00 (0.00)	2.00 (0.00)	2.00 (0.00)	2.00 (0.00)

Table S.18: Mean (and standard error) of estimated structural dimension of sparse SAVE and sparse DR under Models (M1)–(M5), where $n = 300$ and $p = 700$. The threshold δ takes value across $\{10^{-4}, 5 \times 10^{-4}, 10^{-3}, 5 \times 10^{-3}, 10^{-2}\}$.

Model	Sparse SAVE					Sparse DR				
	$\widetilde{\lambda}_2 = 0.01$	$\widetilde{\lambda}_2 = 0.1$	$\widetilde{\lambda}_2 = 0.5$	$\widetilde{\lambda}_2 = 1$	$\widetilde{\lambda}_2 = 5$	$\widetilde{\lambda}_2 = 0.01$	$\widetilde{\lambda}_2 = 0.1$	$\widetilde{\lambda}_2 = 0.5$	$\widetilde{\lambda}_2 = 1$	$\widetilde{\lambda}_2 = 5$
(M1)	35.0 (1.7)	27.3 (1.2)	**24.8 (0.9)**	**24.7 (0.8)**	45.9 (1.8)	29.4 (1.4)	27.2 (1.2)	**23.8 (0.8)**	**23.4 (0.8)**	25.2 (1.0)
(M2)	39.1 (1.8)	29.7 (1.2)	**28.7 (1.2)**	**28.5 (1.3)**	49.5 (1.9)	30.1 (1.4)	28.2 (1.2)	**24.7 (0.8)**	**24.0 (0.7)**	24.9 (0.8)
(M3)	40.4 (1.4)	34.9 (1.2)	**30.6 (1.0)**	**30.7 (1.0)**	46.1 (1.7)	30.9 (1.3)	30.6 (1.3)	**29.9 (1.3)**	**29.5 (1.2)**	31.8 (1.3)
(M4)	19.5 (1.6)	15.5 (1.2)	12.3 (0.6)	**11.9 (0.4)**	**12.0 (0.4)**	20.1 (1.6)	16.9 (1.4)	13.0 (0.8)	**12.3 (0.6)**	**12.2 (0.4)**
(M5)	45.2 (1.9)	33.7 (1.5)	**28.9 (0.8)**	**30.3 (1.0)**	48.0 (1.4)	30.3 (1.3)	27.8 (1.0)	**25.9 (0.6)**	**26.2 (0.7)**	39.8 (1.3)
(Q1)	6.8 (0.3)	6.8 (0.3)	**6.7 (0.3)**	**6.7 (0.3)**	69.3 (0.2)	**54.6 (1.8)**	**56.9 (1.9)**	60.0 (1.7)	66.1 (1.3)	70.4 (0.8)
(Q2)	8.1 (0.2)	8.1 (0.2)	**8.1 (0.2)**	**8.1 (0.2)**	42.9 (1.8)	8.0 (0.2)	8.0 (0.2)	**8.0 (0.2)**	**7.9 (0.2)**	44.4 (2.0)
(Q3)	12.2 (1.2)	**12.0 (1.1)**	12.4 (1.1)	**11.9 (1.1)**	76.4 (1.4)	14.7 (1.0)	14.7 (1.0)	**14.5 (1.0)**	**14.4 (1.0)**	78.7 (1.6)
(Q4)	13.6 (0.8)	**13.6 (0.8)**	13.7 (0.8)	**13.5 (0.8)**	43.1 (2.1)	10.4 (0.3)	10.3 (0.3)	**10.3 (0.3)**	**10.2 (0.3)**	56.1 (2.0)

Table S.19: Means (and standard errors) of the subspace estimation error D ($\times 100$) of sparse SAVE/DR with different choices of $\widetilde{\lambda}_2$ under Models (M1)–(M5) and (Q1)–(Q4), where $n = 300$ and $p = 700$. In each model, for each method, the best two results are highlighted.

Model	Sparse SAVE					Sparse DR				
	$\widetilde{\lambda}_2 = 0.01$	$\widetilde{\lambda}_2 = 0.1$	$\widetilde{\lambda}_2 = 0.5$	$\widetilde{\lambda}_2 = 1$	$\widetilde{\lambda}_2 = 5$	$\widetilde{\lambda}_2 = 0.01$	$\widetilde{\lambda}_2 = 0.1$	$\widetilde{\lambda}_2 = 0.5$	$\widetilde{\lambda}_2 = 1$	$\widetilde{\lambda}_2 = 5$
(M1)	0.7 (0.1)	0.6 (0.1)	0.6 (0.1)	0.6 (0.1)	2.2 (0.2)	1.3 (0.1)	1.3 (0.1)	1.4 (0.1)	1.5 (0.1)	1.7 (0.2)
(M2)	0.4 (0.1)	0.5 (0.1)	0.5 (0.1)	0.5 (0.1)	2.5 (0.2)	2.1 (0.3)	2.2 (0.3)	2.1 (0.2)	2.0 (0.2)	2.2 (0.2)
(M3)	0.3 (0.1)	0.5 (0.1)	0.3 (0.1)	0.3 (0.1)	2.0 (0.3)	0.8 (0.1)	0.8 (0.1)	0.8 (0.1)	0.8 (0.1)	0.8 (0.1)
(M4)	5.3 (0.4)	5.8 (0.4)	6.4 (0.4)	6.5 (0.4)	6.7 (0.4)	8.6 (0.7)	9.5 (0.7)	10.6 (0.7)	10.8 (0.7)	11.2 (0.6)
(M5)	0.7 (0.1)	0.8 (0.2)	0.6 (0.1)	0.8 (0.2)	2.7 (0.3)	0.9 (0.1)	0.9 (0.1)	1.0 (0.1)	1.0 (0.1)	2.2 (0.3)
(Q1)	0.5 (0.1)	0.5 (0.1)	0.5 (0.1)	0.5 (0.1)	0.2 (0.0)	0.1 (0.0)	0.1 (0.0)	0.1 (0.0)	0.1 (0.0)	0.1 (0.0)
(Q2)	0.6 (0.1)	0.6 (0.1)	0.6 (0.1)	0.6 (0.1)	0.5 (0.1)	0.5 (0.1)	0.5 (0.1)	0.5 (0.1)	0.6 (0.1)	0.5 (0.1)
(Q3)	0.7 (0.1)	0.7 (0.1)	0.8 (0.1)	0.9 (0.1)	1.4 (0.2)	1.6 (0.1)	1.6 (0.1)	1.6 (0.1)	1.6 (0.1)	0.7 (0.1)
(Q4)	0.5 (0.1)	0.6 (0.1)	0.6 (0.1)	0.6 (0.1)	0.8 (0.1)	0.5 (0.1)	0.5 (0.1)	0.6 (0.1)	0.6 (0.1)	2.2 (0.2)

Table S.20: Means (and standard errors) of the selected optimal λ_2^* of sparse SAVE/DR with different choices of the predetermined $\widetilde{\lambda}_2$ under Models (M1)–(M5) and (Q1)–(Q4), where $n = 300$ and $p = 700$.

References

Cai, T. T. and Zhang, A. (2018). Rate-optimal perturbation bounds for singular subspaces with applications to high-dimensional statistics. *The Annals of Statistics*, 46(1):60 – 89.

Cai, T. T., Zhang, A. R., and Zhou, Y. (2022). Sparse group lasso: Optimal sample complexity, convergence rate, and statistical inference. *IEEE Transactions on Information Theory*.

Cai, T. T. and Zhang, L. (2021). A convex optimization approach to high-dimensional sparse quadratic discriminant analysis. *The Annals of Statistics*, 49(3):1537–1568.

Ceccarelli, D. F., Tang, X., Pelletier, B., Orlicky, S., Xie, W., Plantevin, V., Neculai, D., Chou, Y.-C., Ogunjimi, A., Al-Hakim, A., et al. (2011). An allosteric inhibitor of the human cdc34 ubiquitin-conjugating enzyme. *Cell*, 145(7):1075–1087.

Choi, E.-H. and Park, S.-J. (2023). Txnip: A key protein in the cellular stress response pathway and a potential therapeutic target. *Experimental & Molecular Medicine*, 55(7):1348–1356.

Jaccard, P. (1901). Etude de la distribution florale dans une portion des alpes et du jura. *Bulletin de la Societe Vaudoise des Sciences Naturelles*, 37:547–579.

Jiang, X., Liu, B., Nie, Z., Duan, L., Xiong, Q., Jin, Z., Yang, C., and Chen, Y. (2021). The role of m6a modification in the biological functions and diseases. *Signal transduction and targeted therapy*, 6(1):74.

Jin, Y. and Luo, W. (2025). A unified generalization of the inverse regression methods via column selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, page qkaf038.

Lebaron, S., O'Donohue, M.-F., Smith, S. C., Engleman, K. L., Juusola, J., Safina, N. P., Thiffault, I., Saunders, C. J., and Gleizes, P.-E. (2022). Functionally impaired rpl8 variants associated with diamond–blackfan anemia and a diamond–blackfan anemia-like phenotype. *Human mutation*, 43(3):389–402.

Li, B. and Wang, S. (2007). On directional regression for dimension reduction. *Journal of the American Statistical Association*, 102(479):997–1008.

Li, R., Zhong, W., and Zhu, L. (2012). Feature screening via distance correlation learning. *Journal of the American Statistical Association*, 107(499):1129–1139.

Nesterov, Y. (2013). Gradient methods for minimizing composite functions. *Mathematical Programming*, 140(1):125–161.

Pan, Y. and Mai, Q. (2020). Efficient computation for differential network analysis with applications to quadratic discriminant analysis. *Computational Statistics & Data Analysis*, 144:106884.

Qian, W., Ding, S., and Cook, R. D. (2019). Sparse minimum discrepancy approach to sufficient dimension reduction with simultaneous variable selection in ultrahigh dimension. *Journal of the American Statistical Association*, 114(527):1277–1290.

Ravikumar, P., Wainwright, M. J., Raskutti, G., and Yu, B. (2011). High-dimensional covariance estimation by minimizing $\ell 1$-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980.

Rudelson, M. and Vershynin, R. (2013). Hanson-wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18:1–9.

Sheng, W. and Yin, X. (2016). Sufficient dimension reduction via distance covariance. *Journal of Computational and Graphical Statistics*, 25(1):91–104.

Stewart, G. and Sun, J. (1990). *Matrix Perturbation Theory*. Computer Science and Scientific Computing. ACADEMIC PressINC.

Székely, G. J., Rizzo, M. L., and Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6):2769–2794.

Vershynin, R. (2018). *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Wedin, P.-Å. (1972). Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12(1):99–111.

Xiao, F., Wang, H.-W., Hu, J.-J., Tao, R., Weng, X.-X., Wang, P., Wu, D., Wang, X.-J., Yan, W.-M., Xi, D., et al. (2022). Fibrinogen-like protein 2 deficiency inhibits virus-induced fulminant hepatitis through abrogating inflammatory macrophage activation. *World Journal of Gastroenterology*, 28(4):479.

Yang, X., Zhou, W., Zhou, J., Li, A., Zhang, C., Fang, Z., Wang, C., Liu, S., Hao, A., and Zhang, M. (2024). Pcgf5: An important regulatory factor in early embryonic neural induction. *Heliyon*, 10(6).

Yang, Y., Hsu, P. J., Chen, Y.-S., and Yang, Y.-G. (2018). Dynamic transcriptomic m6a decoration: writers, erasers, readers and functions in rna metabolism. *Cell research*, 28(6):616–624.

Yin, X. and Cook, R. D. (2003). Estimating central subspaces via inverse third moments. *Biometrika*, 90(1):113–125.

Yu, Z., Dong, Y., and Zhu, L.-X. (2016). Trace pursuit: A general framework for model-free variable selection. *Journal of the American Statistical Association*, 111(514):813–821.

Zeng, J., Mai, Q., and Zhang, X. (2024). Subspace estimation with automatic dimension and variable selection in sufficient dimension reduction. *Journal of the American Statistical Association*, 119(545):343–355.

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