Supplementary Materials for "Subspace Estimation with Automatic Dimension and Variable Selection in Sufficient Dimension Reduction"

A Additional numerical results

A.1 Meat properties data analysis

We apply SEAS methods and other competitors on the meat properties data that was studied in Sæbø et al. (2008) and Cook and Zhang (2015). This data set contains 100 spectral measurements from infrared transmittance between wavelength 850nm and 1050nm and the measured protein, fat, and water percentages for 54 pork and 49 beef samples. All the variables are log-transformed and standardized in preprocessing step. It is of interest to study two problems: (1) the regression problem of protein content on the other measurements in pork; and (2) the classification problem of pork versus beef based on the spectral measurements and the measured protein, fat, and water percentages. In the following two parts, we focus on these two problems, respectively.

Continuous response

In this part, we focus on the regression problem of the $n = 54$ pork samples, where the response Y is the percentage of protein and the $p = 102$ predictors include water and fat percentages and infrared spectra. We compare the performances of SEAS-SIR, SEAS-Intra, SEAS-PFC, Lasso-SIR, natural SSIR, and refined SSIR. In SEAS-PFC, we used $\boldsymbol{f}(Y) = (Y, Y^2)^\top$ as the fitting function. Similar to Section 6, we constructed 100 bootstrap replicates, the average of subspace distance $\mathcal{D}$, the simple matching coefficient ψ, the sparsity level $\widehat{s}$, and the estimated structural dimension $\widehat{d}$ along with the standard errors are summarized in Table A.1. All the three SEAS methods have outperformed other competitors, while the SEAS-PFC had the best performance.

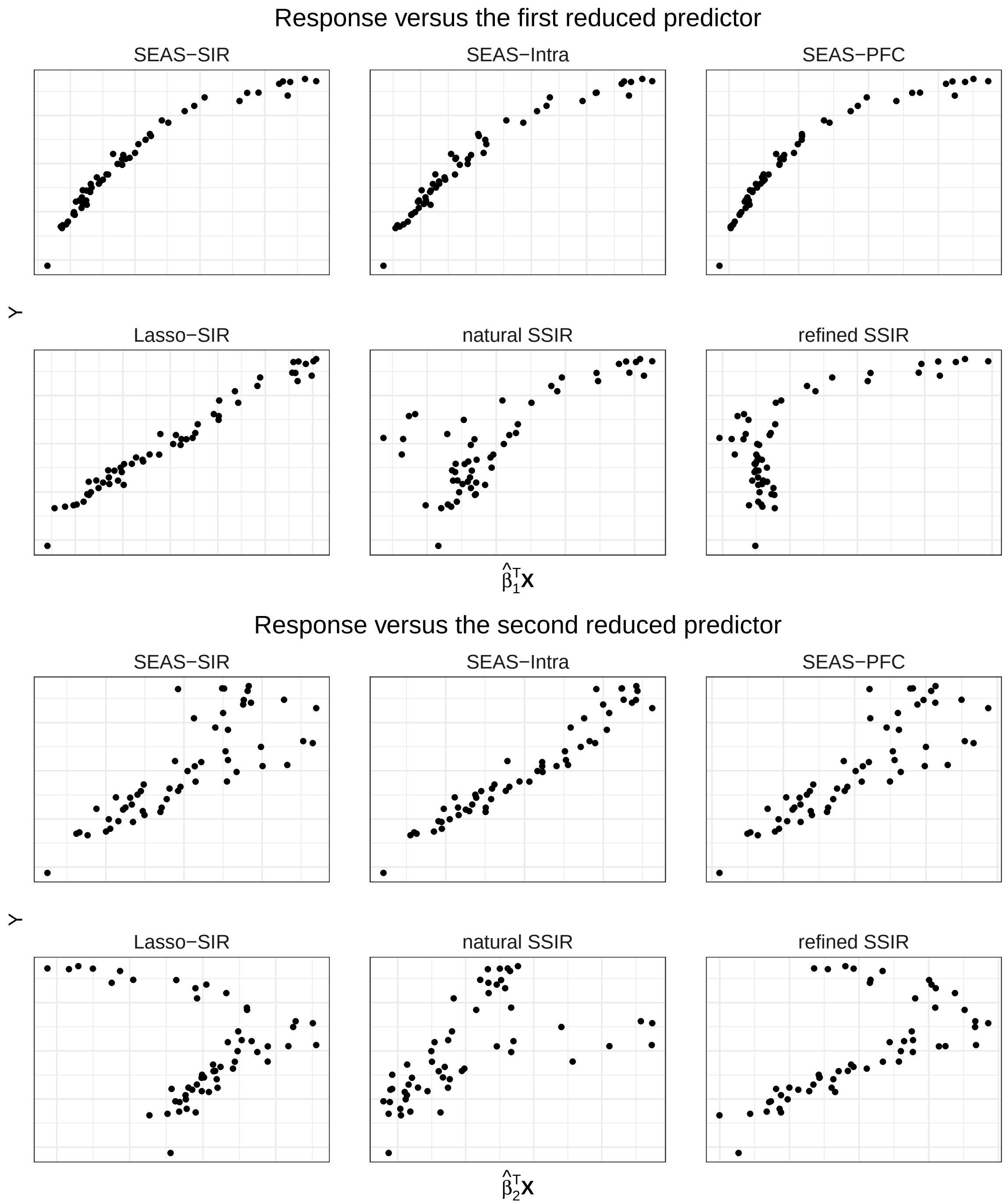

Figure A.1: The scatterplot of the response against the first two reduced predictors in meat properties data set.

The dimension $\widehat{d}$ is selected as 2 for most of the time. For each method, we plotted the response Y versus the reduced predictor $\widehat{\boldsymbol{\beta}}_1^\top \mathbf{X}$ and $\widehat{\boldsymbol{\beta}}_2^\top \mathbf{X}$ in Figure A.1. From the upper panel in Figure A.1, we observed a clear non-linear pattern between the response and the first

Methods	$\mathcal{D} \times 100$	$\psi(\%)$	$\widehat{s}$	$\widehat{d}$	PMSE $\times 100$
SEAS-SIR	58.5 (0.7)	92.8 (0.2)	3.8 (0.3)	2.4	4.5 (0.4)
SEAS-Intra	59.0 (0.8)	90.4 (0.2)	4.5 (0.3)	2.3	3.8 (0.3)
SEAS-PFC	**41.4 (3.0)**	**97.8 (0.2)**	3.5 (0.1)	1.9	**3.3 (0.3)**
Lasso-SIR	93.5 (1.2)	88.5 (0.7)	12.0 (0.7)	1.7	5.4 (0.3)
natural SSIR	76.3 (1.0)	91.0 (0.2)	4.5 (0.2)	2.0	6.2 (0.4)
refined SSIR	64.5 (0.8)	88.1 (0.3)	8.2 (0.3)	2.0	4.8 (0.3)

Table A.1: Pork samples in meat properties data set. The averages of the subspace distance $\mathcal{D} = \mathcal{D}(\mathcal{S}_{\widehat{\boldsymbol{\beta}}}, \mathcal{S}_{\widehat{\boldsymbol{\beta}}_b})$, the simple matching coefficient ψ (%), the sparsity level $\widehat{s}$, the estimated dimension $\widehat{d}$ and the corresponding standard error based on 100 bootstrap data sets. The standard errors for $\widehat{d}$ are all less than 0.1 and are thus omitted. The prediction mean squared error (PMSE) is based on kernel regression.

reduced predictor for SEAS methods. However, we only observed a linear pattern for Lasso-SIR. And the patterns for natural SSIR and refined SSIR are not clear. For the lower panel in Figure A.1, no obvious pattern between the response and the second reduced predictor is indicated. The reason might be that most regression information is obtained by the first reduced predictor. While detecting non-linearity is difficult in high-dimensional data set, the plots suggest that SEAS methods may be more effective in finding non-linear relationships between Y and $\mathbf{X}$ in high dimensions. Because of the non-linear relationship, the prediction mean squared errors (PMSE) based on kernel regression is also included in Table A.1. The SEAS methods, as expected, have demonstrated superior predictive power.

Binary response

In this part, we study the classification problem of the full meat properties data. The pork and beef samples are coded in 1 and 2 in response, and the $p = 103$ predictors consist of 100 spectral measurements from infrared transmittance between wavelength 850nm and 1050nm and the measured protein, fat, and water percentages for meat samples. We aim to distinguish samples between pork and beef using the predictors. Since the response is binary, SEAS-Intra and SEAS-PFC are the same as SEAS-SIR in implementation.

We compare SEAS-SIR, Lasso-SIR, natural SSIR, and refined SSIR on the original and 100 bootstrap data sets. Similar to the previous part, we record the average of subspace distance $\mathcal{D}$, the simple matching coefficient ψ, the sparsity level $\widehat{s}$, and the estimated structural dimension $\widehat{d}$ along with the standard errors in Table A.2.

From Table A.2, it can be seen that SEAS-SIR approaches the smallest $\mathcal{D}$, and the nearly 100% ψ, indicating its superiority in subspace estimation and variable selection consistency.

Methods	Estimation				Classification error (%)			
	$\mathcal{D} \times 100$	$\psi(\%)$	$\widehat{s}$	$\widehat{d}$	Logistic	SVM	LDA	RF
SEAS-SIR	**15.2 (0.6)**	98.4 (0.1)	11.9 (0.1)	1.0	2.5	2.5	1.0	2.5
Lasso-SIR	80.2 (1.9)	55.3 (1.3)	41.9 (3.8)	1.0	1.0	0.5	0.6	1.2
natural SSIR	41.8 (3.2)	99.0 (0.1)	3.4 (0.1)	1.0	2.5	2.1	1.4	2.9
refined SSIR	78.6 (1.2)	63.2 (0.7)	50.7 (1.8)	1.0	2.1	1.2	1.5	2.6

Table A.2: Meat properties data set. The averages of the subspace distance $\mathcal{D} = \mathcal{D}(\mathcal{S}_{\widehat{\boldsymbol{\beta}}}, \mathcal{S}_{\widehat{\boldsymbol{\beta}}_b})$, the simple matching coefficient ψ (%), the sparsity level $\widehat{s}$, the estimated dimension $\widehat{d}$ and the standard errors based on 100 bootstrap data sets. The standard errors for $\widehat{d}$ are all zero and are thus omitted. The classification error rate (%) based on logistic regression model, support vector machine (SVM), linear discriminant analysis (LDA) and random forest (RF) are also reported. The standard errors for the classification error rate are all less than 0.3 (%) and are omitted.

The rank $\widehat{d}$ were selected as one on average by both SEAS-SIR and Lasso-SIR. Then we report the results with rank 1 for natural SSIR and refined SSIR. Meanwhile, SEAS-SIR selects the proper number of variables compared with other competitors. Lasso-SIR and refined SSIR select too many variables, which impedes the interpretation of the model. And natural SSIR selects too few variables, resulting in its unstable subspace estimation.

Next, we compare the classification errors of each SIR-based estimator. Similar to Section 6, we randomly select 20% samples from the original dataset as the testing data set and keep the rest as the training data set in each replicate. Then the CS is estimated on the training data set and used to reduce the predictors in both training and testing data sets. Four classifiers are used, including logistic regression, support vector machine (SVM), linear discriminant analysis (LDA), and random forest (RF). The classification error based on 100 replicates is presented in Table A.2. It can be seen that SEAS-SIR makes an accurate prediction. The classification error of SEAS-SIR is comparable to those of other competitors.

A.2 Computation time

To gain some insights into the computation efficiency of SEAS methods and the competitors, we record the computation time for each method in model (M3a) as a demonstration. Since the computation times of SEAS-SIR, SEAS-Intra, and SEAS-PFC are similar, we only report the result of SEAS-SIR. The most computationally intensive part of the refined SSIR is the computation of its initial value from the natural SSIR, hence the computation times of these two methods are almost the same, and we only report the time of natural SSIR. For a fair comparison, we record the computation time on the selected tuning parameter for each of the three methods (SEAS-SIR, Lasso-SIR, and natural SSIR). We note that SEAS-SIR has more tuning parameters than Lasso-SIR. Specifically, in our simulations, SEAS-SIR runs over 300 tuples of tuning parameters $(\lambda_1, \lambda_2, \gamma)$ in total while Lasso-SIR runs over 100 tuning parameters in total. Although the SEAS-SIR is slightly faster than Lasso-SIR in one run, as shown in Figure A.2, it is slightly slower than Lasso-SIR in our simulation studies as more tuning parameter candidates are considered. In comparison, natural SSIR is hundreds or even thousands of times slower than both SEAS-SIR and Lasso-SIR.

We generate the central subspace basis $\boldsymbol{\beta} \in \mathbb{R}^{1500\times 2}$ from model (M3a), where the sparsity level $s = 6$. In each data replicate, we generate $n = 200$ observations $(\mathbf{X}_i, Y_i)$, $i = 1, \ldots, n$, from model (M3a), where $\mathbf{X}_i \in \mathbb{R}^{1500}$ and $Y \in \mathbb{R}$. For some p, we take the first p variables from each $\mathbf{X}_i$ and generate new observations set $(\mathbf{X}_i^p, Y_i)$, $i = 1, \ldots, n$, where $\mathbf{X}_i^p$ is the subvector containing the first p variables of $\mathbf{X}_i$. We vary p over $\{200, 400, 600, 1000, 1500\}$, and we implement three methods on each set $(\mathbf{X}_i^p, Y_i)$, $i = 1, \ldots, n$. For each of SEAS-SIR and Lasso-SIR, we use the same tuning parameter sequences across all p's. As SEAS-SIR and Lasso-SIR run much faster than natural SSIR, we implement SEAS-SIR and Lasso-SIR on 16 replicates and natural SSIR on only one replicate. For each p, we record the averaged computation time over 16 replicates for SEAS-SIR and Lasso-SIR and the computation time on one replicate for natural SSIR.

The results are displayed in Figure A.2. Each line describes the relationship between the (logarithmic) computation time and the dimension p. It can be seen that SEAS-SIR is the fastest one among the three competitors. Lasso-SIR is slightly slower than SEAS-SIR as p increases since it sequentially estimates each direction in the central subspace. Natural SSIR is slower than other methods in magnitude since it requires the optimization on $p \times p$ matrices. On average, for any p between 200 and 1500, SEAS-SIR finishes one run in 0.5 seconds, and Lasso-SIR finishes one run in 1.3 seconds. For natural SSIR, when $p = 1500$, it takes around 1.5 hours to finish one run. The result in Figure A.2 also shows the scalability of the SEAS method in the high-dimensional setting. As p gets higher, the computation time of SEAS-SIR increases slower than other competitors. Specifically, from $p = 200$ to $p = 600$, the logarithmic computation time of SEAS-SIR, Lasso-SIR, and natural SSIR increases by 1.242, 1.798, and 3.636, respectively. And from $p = 600$ to $p = 1500$, the logarithmic computation time of three methods increases by 1.839, 2.037, and 2.462.

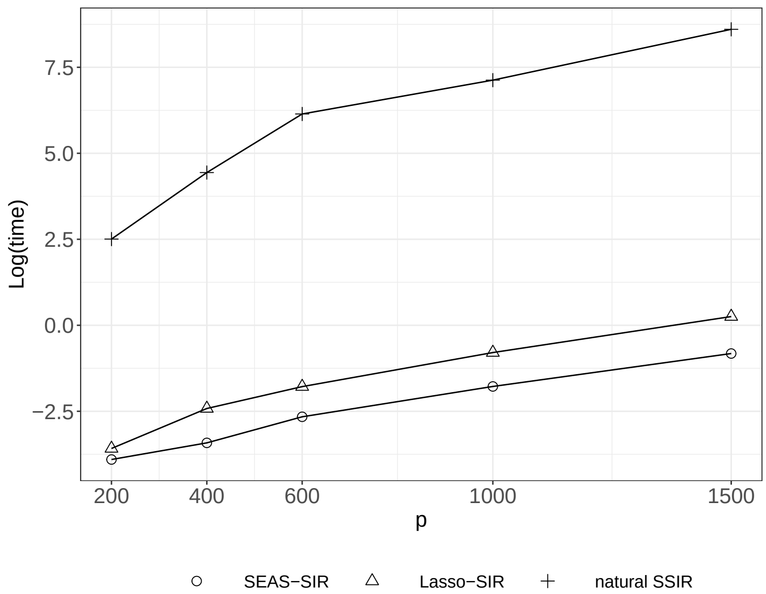

Figure A.2: The mean of log computation time (seconds) under model (M3a). When $p = 200$, SEAS-SIR, Lasso-SIR and natural SSIR finish one run in 0.020, 0.028 and 12.257 seconds; when $p = 1500$, SEAS-SIR, Lasso-SIR and natural SSIR finish one run in 0.4, 1.3 and 5455.9 seconds.

B Additional implementation details

B.1 Tuning parameters selection

There are three tuning parameters involved in Algorithm 1: the sparsity parameter λ_1, the low-rank parameter λ_2, and the ADMM parameter γ. We use cross-validated distance correlation (Székely et al., 2007) to select the tuning parameters. In simulation studies of Section A, the following cross-validation approach works well for all SEAS methods.

For any random vectors $\mathbf{W} \in \mathbb{R}^p$ and $\mathbf{Z} \in \mathbb{R}^q$, the distance correlation, denoted as $\mathrm{dCor}(\mathbf{W}, \mathbf{Z}) \in [0, 1]$ measures the dependence between the two random vectors. In particular, $\mathrm{dCor}(\mathbf{W}, \mathbf{Z}) = 0$ if and only if $\mathbf{W}$ and $\mathbf{Z}$ are independent. For the observed random samples $\mathbf{W}_n \in \mathbb{R}^{n\times p}$ and $\mathbf{Z}_n \in \mathbb{R}^{n\times q}$, if the sample distance correlation $\mathrm{dCor}_n(\mathbf{W}, \mathbf{Z}) = 1$, then there exist some vector $\boldsymbol{a} \in \mathbb{R}^p$, scalar b and orthonormal matrix $\mathbf{C} \in \mathbb{R}^{q\times p}$ such that $\mathbf{W}_n = \mathbf{1}_n \boldsymbol{a}^\top + b\mathbf{Z}_n\mathbf{C}$, where $\mathbf{1}_n$ is the n-dimensional vector with all elements being one. The distance correlation is a natural choice as the evaluation criterion for SDR problems. In recent years, distance correlation has been applied in SDR problems for both estimation (e.g., Sheng and Yin, 2016) and diagnostics (e.g., Chen et al., 2015). In particular, Sheng and Yin (2016) showed that under mild assumptions, the distance covariance between Y and $\boldsymbol{\eta}^\top\mathbf{X}$ is maximized at the CS over all $\boldsymbol{\eta} \in \mathbb{R}^{p\times d^\star}$ such that $\boldsymbol{\eta}^\top\boldsymbol{\Sigma}\boldsymbol{\eta} = \mathbf{I}_{d^\star}$. Motivated by this, we select our tuning parameters by maximizing $\mathrm{dCor}_n(Y, \widehat{\boldsymbol{\beta}}^\top \mathbf{X})$.

Specifically, we choose the optimal triplet of tuning parameters $(\lambda_1, \lambda_2, \gamma)$ by a grid search with K-fold cross-validation. We generate candidate sequence for each of λ_1, λ_2, and γ. As demonstrated in the simulation study in Section A.2, with the given tuning parameter $(\lambda_1, \lambda_2, \gamma)$, SEAS is computationally efficient and scales well to high dimensions. If a faster cross-validation is desired, one can alternatively construct a one-dimensional grid between $(\widetilde{\lambda}_1, \widetilde{\lambda}_2, \widetilde{\gamma})$ and $(a\widetilde{\lambda}_1, a\widetilde{\lambda}_2, a\widetilde{\gamma})$, where $(\widetilde{\lambda}_1, \widetilde{\lambda}_2, \widetilde{\gamma})$ is some pre-specified large-valued triplet and a is a small factor, e.g., 10^{-3}. The optimal triplet is then searched over $(\tau\widetilde{\lambda}_1, \tau\widetilde{\lambda}_2, \tau\widetilde{\gamma})$, where $\tau \in [a, 1]$ is the tuning parameter. The similar strategy is also adopted in other works (see Uematsu et al., 2019). Let $\mathbf{X}_n^{(k)} \in \mathbb{R}^{m_k \times p}$ and $\mathbf{Y}_n^{(k)} \in \mathbb{R}^{m_k}$ denote m_k predictor and response observations in the k-th hold-out fold C_k, $k = 1, \ldots, K$. With each tuning parameter triplet $(\lambda_1, \lambda_2, \gamma)$, we implement the Algorithm 1 using all the observations outside the fold C_k and obtain the estimator $\widehat{\boldsymbol{\beta}}^{(k)}$. Then we transform the design matrix $\mathbf{X}_n^{(k)}$ to the reduced version $\mathbf{X}_n^{(k)}\widehat{\boldsymbol{\beta}}^{(k)}$ and compute the sample distance correlation $\mathrm{dCor}_n(\mathbf{Y}_n^{(k)}, \mathbf{X}_n^{(k)}\widehat{\boldsymbol{\beta}}^{(k)})$. Therefore, the tuning parameter triplet $(\lambda_1, \lambda_2, \gamma)$ is chosen simultaneously to maximize the average sample distance correlation $\{\sum_{k=1}^{K} \mathrm{dCor}_n(\mathbf{Y}_n^{(k)}, \mathbf{X}_n^{(k)}\widehat{\boldsymbol{\beta}}^{(k)})\}/K$.

In our experience, SEAS-Intra and SEAS-PFC are more sensitive to extreme values in the samples $\{Y_i\}_{i=1}^n$ than SEAS-SIR. This is because we use polynomials in the PFC fitting functions $\{f_h(Y_i)\}_{i=1}^n$, $h = 1, \ldots, H$, and need to estimate conditional covariance functions in intraslice covariance method. To reduce the influence of the extremely large values in $|Y_i|$, we implement SEAS-Intra by truncating samples $\{Y_i\}_{i=1}^n$ at 10% and 90% percentiles in the computation of $\widehat{\mathbf{U}}_{\text{Intra}}$; and, for each $h = 1, \ldots, H$, we threshold samples $\{f_h(Y_i)\}_{i=1}^n$ at 10% and 90% percentiles during the computation of $\widehat{\mathbf{U}}_{\text{PFC}}$. In contrast, SEAS-SIR only relies on the order information of samples $\{Y_i\}_{i=1}^n$ and is robust to the extreme values.

B.2 Choosing the number of slices H

A practical issue for slicing-based SEAS methods (SEAS-SIR and SEAS-Intra) is the choice of H. Even for low-dimensional settings, choosing the number of slices H is known to be a theoretically challenging but crucial issue in SIR and inverse regression estimators. Existing theoretical studies on the optimal choice of H rely on fix-p asymptotics (Zhu and Ng, 1995; Cook and Zhang, 2014) and may not apply to high-dimensional scenarios. As such, we investigate the empirical performance of slicing-based methods (ours and other existing methods) in high-dimensional simulations where $n = 200$ and $p = 1000$. While we treat H as pre-specified and fixed in our theoretical studies, in this section, we provide some empirical findings and guidance on how to choose H for our SEAS methods in high dimensions. It is noted that SEAS-SIR and SEAS-Intra are robust to H varying from 5 to 20, and also achieve the best results when $H = 5$. Hence, we recommend using $H = 5$ for small sample size scenarios ($n \leq 200$) and $H = 20$ for larger sample size scenarios ($n > 200$).

We compare slicing-based sparse SDR methods in the more challenging models (M3a)–(M4) of Section 5. We set $n = 200$ and $p = 1000$, and, for 100 simulation replicates under each setting, we vary $H \in \{5, 10, 15, 20\}$. Because the result of natural SSIR is worse than that of refined SSIR, and the trends of their performance (over the varying H) are very similar, we summarize the results of SEAS-SIR, SEAS-Intra, Lasso-SIR, and refined SSIR in Figure B.3 but leave out the natural SSIR.

Recall that $\mathcal{D}(\mathcal{S}_{\boldsymbol{\beta}^\star}, \mathcal{S}_{\widehat{\boldsymbol{\beta}}}) = \|\mathbf{P}_{\mathcal{S}_{\boldsymbol{\beta}^\star}} - \mathbf{P}_{\mathcal{S}_{\widehat{\boldsymbol{\beta}}}}\|_F / \sqrt{2d^\star}$ is our subspace estimation criterion, which can be affected by the rank selection results. Also recall that the refined SSIR uses the true dimension while the other methods select $\widehat{d}$ based on cross-validation. Therefore, in Figure B.3, we first report the estimated rank $\widehat{d}$ (the left three plots), except for the refined SSIR that requires the true rank $d^\star$, then report the subspace estimation error $\mathcal{D}$ (the right three plots) when using $d^\star$ in all methods.

From Figure B.3, as H increases, the estimated rank tends to increase. This is expected as the dimension of $\mathbf{B} \in \mathbb{R}^{p \times H}$ increases with H for the SEAS method. For both SEAS-Intra and SEAS-SIR, the rank is estimated accurately for all H and under all three models. The best results are achieved at $H = 5$. However, the estimated rank from Lasso-SIR increases almost linearly as H, leading to a substantial over-estimation of d.

With known true rank $d^\star$, we compare all four estimators with H varying over $5, 10, 15$, and 20. The subspace distance shown in Figure B.3 suggests that both SEAS estimators outperform the other two methods substantially in the latter two models. Overall, all methods (in particular SEAS-SIR and SEAS-Intra) are robust to the choice of $H \in [5, 20]$.

B.3 Technical details for Algorithm 1

We provide the details for the update of $\mathbf{B}^{(t)}$ using the group-wise coordinate descent algorithm proposed by Mai et al. (2019). According to Lemma 3 in Section 3, updating $\mathbf{B}^{(t)}$ is equivalent to solving the $L_{2,1}$ norm penalized optimization problem

$$\underset{\mathbf{B} \in \mathbb{R}^{p \times H}}{\operatorname{argmin}} \; \frac{1}{2}\operatorname{tr}\{\mathbf{B}^\top(\widehat{\boldsymbol{\Sigma}} + \gamma \mathbf{I}_p)\mathbf{B}\} - \operatorname{tr}\{\mathbf{B}^\top(\widehat{\mathbf{U}} - \boldsymbol{\omega}^{(t-1)} + \gamma \mathbf{C}^{(t-1)})\} + \lambda_1 \|\mathbf{B}\|_{2,1}. \tag{B.4}$$

Let $\widehat{\mathbf{Z}}$ and $\widehat{\mathbf{W}}$ denote $\widehat{\boldsymbol{\Sigma}} + \gamma \mathbf{I}_p$ and $\widehat{\mathbf{U}} - \boldsymbol{\omega}^{(t-1)} + \gamma \mathbf{C}^{(t-1)}$. The algorithm is presented in Algorithm B.

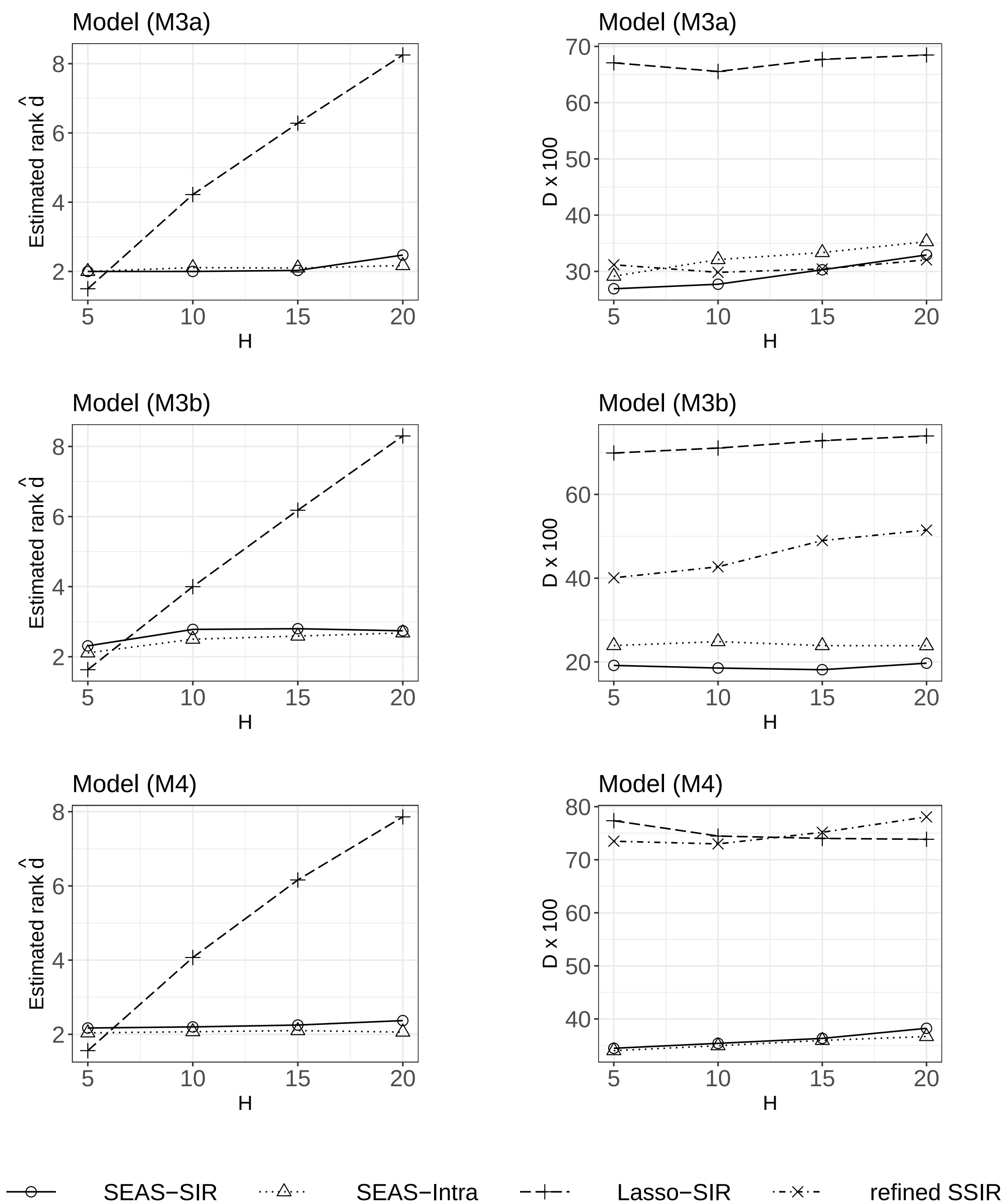

Figure B.3: On the left plots, we report the estimated ranks $\widehat{d}$ over H in models (M3a)–(M4); on the right plots, we report the distance $\mathcal{D} \times 100$ with true $d^{\star}$ for all methods. These results are based on 100 replicates. Note that the refined SSIR does not select d and is not included on the left three plots.

Algorithm B Group-wise coordinate descent algorithm

1. Input $\widehat{\mathbf{Z}} = (\widehat{z}_{ij})$, $\widehat{\mathbf{W}} = (\widehat{w}_{ij})$. For $j = 1, \ldots, p$, initialize each row of $\mathbf{B}^{(0)}$ as $\mathbf{B}_j^{(0)} = (b_{j1}^{(0)}, \ldots, b_{jH}^{(0)})^\top = \mathbf{0} \in \mathbb{R}^H$ and calculate the auxillary vector $\boldsymbol{\theta}_1^{(0)} = (\theta_{11}^{(0)}, \ldots, \theta_{1H}^{(0)})^\top \in \mathbb{R}^H$ as
$$\theta_{1h}^{(0)} = \frac{\widehat{w}_{1h} - \sum_{j' \neq 1} \widehat{z}_{j'1} b_{j'h}^{(0)}}{\widehat{z}_{11}}, \quad h = 1, \ldots, H. \tag{B.1}$$

2. For steps $t = 1, 2, \ldots$, repeat the following until the convergence of $\mathbf{B}^{(t)}$: for each $j = 1, \ldots, p$,

   (a) Based on $\boldsymbol{\theta}_j^{(t-1)}$, update $\mathbf{B}_j^{(t)}$ as
   $$\mathbf{B}_j^{(t)} = \boldsymbol{\theta}_j^{(t-1)} \left(1 - \frac{\lambda_1}{\widehat{z}_{jj} \|\boldsymbol{\theta}_j^{(t-1)}\|_2}\right)_+, \tag{B.2}$$
   where $(\cdot)_+$ is the positive part function.

   (b) Based on $\mathbf{B}_1^{(t)}, \ldots, \mathbf{B}_{j-1}^{(t)}, \mathbf{B}_{j+1}^{(t-1)}, \ldots, \mathbf{B}_p^{(t-1)}$, update the auxillary vector $\boldsymbol{\theta}_j^{(t-1)}$ as
   $$\theta_{jh}^{(t-1)} = \frac{\widehat{w}_{jh} - (\sum_{j'<j} \widehat{z}_{j'j} b_{j'h}^{(t)} + \sum_{j'>j} \widehat{z}_{j'j} b_{j'h}^{(t-1)})}{\widehat{z}_{jj}}, \quad h = 1, \ldots, H. \tag{B.3}$$

3. At termination, output $\mathbf{B}^{(t)}$ as the solution.

C Variable selection consistency

In Section 4, Corollary 2 shows that, with high probability, the important variables are correctly selected, and the false positives are weak. In this section, we push these results a step further and derive the variable selection consistency that the important variables are exactly identified with no false positives. To this end, we make three additional assumptions as follows.

(A8) The predictors $\mathbf{X}_1, \ldots, \mathbf{X}_n \in \mathbb{R}^p$ are i.i.d. normally distributed with mean zero and covariance matrix $\boldsymbol{\Sigma}$. There further exists some constant $m > 0$ such that $m^{-1} \leq \varphi_{\min}(\boldsymbol{\Sigma}) \leq \varphi_{\max}(\boldsymbol{\Sigma}) \leq m$.

(A9) There exists some constant $Q > 0$ such that $\|\boldsymbol{\Sigma}_{\mathcal{A}^c\mathcal{A}} \boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1}\|_\infty \leq Q$.

(A10) There exists some constant $0 < \alpha \leq 1$ such that $\max_{j \in \mathcal{A}^c} \|\boldsymbol{\Sigma}_{j\mathcal{A}} \boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1} \mathbf{R}\|_2 \leq 1 - \alpha$, where $\mathbf{R} \in \partial \|\mathbf{B}_{\mathcal{A}}^\star\|_{2,1}$.

We impose the additional assumptions for variable selection consistency because we need to study the subdifferential of our objective function in (6). The subdifferential involves functionals of $\widehat{\mathbf{U}}$ and $\widehat{\boldsymbol{\Sigma}}$ that are harder to bound than the quantities in Conditions (C1) & (C2). Assumptions (A8)–(A10) assist us in this aspect. All of them are reasonably popular in the literature. Assumption (A8) requires $\mathbf{X}$ to be normal with a well-conditioned $\boldsymbol{\Sigma}$. This is slightly stronger than assuming that $\mathbf{X}$ is sub-Gaussian, but the normality assumption is often made as a technical assumption in many high-dimensional theoretical studies. For example, in the high-dimensional SIR literature, Tan et al. (2020) assumes that $\mathbf{X}$ is normal within each slice, while Lin et al. (2021) assumes $\mathbf{X}$ to be normal with an identity covariance matrix. Assumptions (A9) & (A10) are similar to the irrepresentable condition in the lasso and group lasso literature (Zhao and Yu, 2006; Bach, 2008). Even in easier problems of linear regression and discriminant analysis, such conditions are often required (Zhao and Yu, 2006; Bach, 2008; Mai et al., 2012, 2019). Since regression and discriminant analysis problems can be viewed as sub-problems of sufficient dimension reduction, it makes sense to make such assumptions for SEAS to achieve variable selection consistency as well.

The following theorem gives the variable selection consistency result.

Theorem C.1. *Assume that Conditions (C1)(C2) and Assumptions (A1)(A2)(A4) and (A8)–(A10) hold. Further assume that $C_1 s \log p \leq n$ for some sufficiently large constant $C_1 > 0$. There exist constants $c_1, c_2, C_B, C, C' > 0$ such that, when $2c_1\sqrt{\log p/n} \leq \lambda_1 \leq 3c_1\sqrt{\log p/n}$, $\lambda_2 \leq c_2\lambda_1$ and $C_B\sqrt{s\log p/n} \leq \delta \leq 2C_B\sqrt{s\log p/n}$, we have $\widehat{\mathcal{A}} = \mathcal{A}$ with probability at least $1 - C\exp\{-C'(s \wedge \log p)\}$.*

Theorem C.1 implies that, if we allow s and p to diverge with n at rate of $s \log p = o(n)$, we are able to recover the important variables exactly with probability approaching one provided that Conditions (C1) & (C2) hold for the specific estimator. Just as our main result in Theorem 1, Theorem C.1 allows us to establish variable selection consistency for various SEAS methods such as SEAS-SIR, SEAS-Intra, and SEAS-PFC by verifying Conditions (C1) and (C2). To the best of our knowledge, we are the first to establish variable selection consistency results for SDR methods in high-dimensional settings.

In what follows, we present the variable selection consistency for SEAS-SIR, SEAS-Intra and SEAS-PFC. Let $\mathcal{A}_{\mathrm{SIR}}$, $\mathcal{A}_{\mathrm{Intra}}$, and $\mathcal{A}_{\mathrm{PFC}}$ denote the true active sets for SEAS-SIR, SEAS-Intra and SEAS-PFC. And let $\widehat{\mathcal{A}}_{\mathrm{SIR}}$, $\widehat{\mathcal{A}}_{\mathrm{Intra}}$, and $\widehat{\mathcal{A}}_{\mathrm{PFC}}$ denote the estimated active sets for the three SEAS methods.

Corollary C.1. *Assume that Assumptions (A1), (A2), (A4) and (A8)–(A10) hold. Further assume that $C_1 s \log p \leq n$ for some sufficiently large constant $C_1 > 0$. There exist constants $c_1, c_2, C_B, C, C' > 0$ such that, when $2c_1\sqrt{\log p/n} \leq \lambda_1 \leq 3c_1\sqrt{\log p/n}$, $\lambda_2 \leq c_2\lambda_1$ and*

$C_B\sqrt{s\log p/n} \le \delta \le 2C_B\sqrt{s\log p/n}$, *we have that* $\widehat{\mathcal{A}}_{\mathrm{SIR}} = \mathcal{A}_{\mathrm{SIR}}$ *with probability at least* $1 - C\exp\{-C'(s \wedge \log p)\}$.

Corollary C.2. *Assume that Assumptions (A1), (A2), (A4), (A5) and (A8)–(A10) hold. Further assume that* $C_1 s\log p \le n$ *for some sufficiently large constant* $C_1 > 0$. *There exist constants* $c_1, c_2, C_B, C, C' > 0$ *such that, when* $2c_1\sqrt{\log p/n} \le \lambda_1 \le 3c_1\sqrt{\log p/n}$, $\lambda_2 \le c_2\lambda_1$ *and* $C_B\sqrt{s\log p/n} \le \delta \le 2C_B\sqrt{s\log p/n}$, *we have that* $\widehat{\mathcal{A}}_{\mathrm{Intra}} = \mathcal{A}_{\mathrm{Intra}}$ *with probability at least* $1 - C\exp\{-C'(s \wedge \log p)\}$.

Corollary C.3. *Assume that Assumptions (A1), (A2), (A4), (A6) and (A8)–(A10) hold. Further assume that* $C_1 s\log p \le n$ *for some sufficiently large constant* $C_1 > 0$. *There exist constants* $c_1, c_2, C_B, C, C' > 0$ *such that, when* $2c_1\sqrt{\log p/n} \le \lambda_1 \le 3c_1\sqrt{\log p/n}$, $\lambda_2 \le c_2\lambda_1$ *and* $C_B\sqrt{s\log p/n} \le \delta \le 2C_B\sqrt{s\log p/n}$, *we have that* $\widehat{\mathcal{A}}_{\mathrm{PFC}} = \mathcal{A}_{\mathrm{PFC}}$ *with probability at least* $1 - C\exp\{-C'(s \wedge \log p)\}$.

Corollaries C.1–C.3 show that all the three SEAS estimators we consider achieve variable selection consistency in ultra-high dimensions. These estimators require somewhat different assumptions for reasons parallel to our discussion of Corollaries 3–5. We do not repeat them here to avoid redundancy.

D Further discussion on $\sigma_1/\sigma_{d^\star}^2$

Since σ_1 and $\sigma_{d^\star}$ are the largest and the smallest nonzero singular values of $\mathbf{B}^\star = \boldsymbol{\Sigma}^{-1}\mathbf{U}$, the specific form of $\sigma_1/\sigma_{d^\star}^2$ depends on the SDR method adopted and the model assumption between $\mathbf{X}$ and Y. In the case $d^\star = 1$ or under the assumption that $\sigma_1 \le C\sigma_{d^\star}$ for some constant $C > 0$, we have $\sigma_1/\sigma_{d^\star}^2 \le C/\sigma_{d^\star}$. And under the weaker Assumption (A4) that $\sigma_1 \le T$ for some constant $T > 0$, which is necessary for Corollaries 3–5, we have $\sigma_1/\sigma_{d^\star}^2 \le T/\sigma_{d^\star}^2$. For demonstration purpose, we focus on $\sigma_{\mathrm{SIR},1}/\sigma_{\mathrm{SIR},d^\star}^2$ in SEAS-SIR under the model with $d^\star = 1$. We consider two special models: the linear regression model and the binary linear discriminant model. The bounds for $\sigma_{\mathrm{SIR},1}/\sigma_{\mathrm{SIR},d^\star}^2$ under these two models are derived in the next two lemmas.

Lemma D.1. *Consider the linear regression model* $Y = \boldsymbol{\beta}^\top\mathbf{X} + \epsilon$, *where* $\boldsymbol{\beta} \in \mathbb{R}^p$, $\|\boldsymbol{\beta}\|_2 = \tau > 0$, $\mathbf{X} \sim N(\mathbf{0}, \mathbf{I})$, $\epsilon \sim N(0, \omega^2)$, *and* ϵ *is independent of* $\mathbf{X}$. *Assume that* ω *and* τ *are bounded from below. Let* α *denote the signal-to-noise ratio. In this model,* $\alpha = \tau^2/(\tau^2 + \omega^2)$. *Then, we have* $d^\star = 1$ *and* $\sigma_{\mathrm{SIR},1}/\sigma_{\mathrm{SIR},d^\star}^2 \le C\alpha^{-1/2}$ *for some positive constant* C *that does not depend on any of the parameters.*

Lemma D.2. *For $Y \in \{1, 2\}$, assume that $\mathbb{P}(Y = 1) = \mathbb{P}(Y = 2) = 1/2$, $\mathbf{X} \mid (Y = 1) \sim N(\boldsymbol{\mu}/2, \mathbf{I})$ and $\mathbf{X} \mid (Y = 2) \sim N(-\boldsymbol{\mu}/2, \mathbf{I})$, where $\boldsymbol{\mu} \in \mathbb{R}^p$. Further assume that $\|\boldsymbol{\mu}\|_2$ is bounded from above. Then, we have $d^\star = 1$ and $\sigma_{\mathrm{SIR},1}/\sigma^2_{\mathrm{SIR},d^\star} \leq C\|\boldsymbol{\mu}\|_2^{-1}$ for some positive constant C that does not depend on any of the parameters.*

In Lemmas D.1 and D.2, $d^\star = 1$ and $\sigma_{\mathrm{SIR},1} = \sigma_{\mathrm{SIR},d^\star}$. Then, $\sigma_{\mathrm{SIR},1}/\sigma^2_{\mathrm{SIR},d^\star} = 1/\sigma_{\mathrm{SIR},d^\star}$, which is bounded from above by the model-specific parameter. To be specific, by Lemma D.1, in the linear regression model, with the higher signal-to-noise ratio, $\sigma_{\mathrm{SIR},1}/\sigma^2_{\mathrm{SIR},d^\star}$ gets lower. And by Lemma D.2, in the binary linear discriminant model, easy computation gives the Bayes error rate as $1 - \Phi(\|\boldsymbol{\mu}\|_2/2)$, where Φ is the cumulative distribution function of the standard normal distribution. Therefore, the smaller the Bayes error rate for the classification problem is, the lower $\sigma_{\mathrm{SIR},1}/\sigma^2_{\mathrm{SIR},d^\star}$ is. In general, the term $\sigma_{\mathrm{SIR},1}/\sigma^2_{\mathrm{SIR},d^\star}$ describes the complexity of the model. For the less complicated model, i.e.,the linear regression model with the higher signal-to-noise ratio or the binary linear discriminant model with the smaller Bayes error rate, $\sigma_{\mathrm{SIR},1}/\sigma^2_{\mathrm{SIR},d^\star}$ becomes smaller, resulting in a lower non-asymptotic bound for $\mathcal{D}(\mathcal{S}_{\mathrm{SIR}}, \mathcal{S}_{\widehat{\boldsymbol{\beta}}_{\mathrm{SIR}}})$ in Corollary 3.

E Preparation for proofs

We begin with the introduction of some notations to facilitate the presentation of Sections E–H. For a matrix $\mathbf{A} = (a_{ij}) \in \mathbb{R}^{p\times q}$, define the $L_{2,\infty}$*-norm* as $\|\mathbf{A}\|_{2,\infty} = \max_{i\in\{1,\ldots,p\}} \|\mathbf{A}_i\|_2$. For an index set $\mathcal{I}$, let $\mathcal{I}^c$ denote its complement set. For a vector $\mathbf{v} \in \mathbb{R}^p$, let $\mathbf{v}_{\mathcal{I}} \in \mathbb{R}^{|\mathcal{I}|}$ denote the sub-vector of $\mathbf{v}$ indexed by $\mathcal{I}$. For index sets $\mathcal{I} \subseteq \{1, \ldots, p\}$ and $\mathcal{J} \subseteq \{1, \ldots, q\}$, let $\mathbf{A}_{\mathcal{I}\mathcal{J}} \in \mathbb{R}^{|\mathcal{I}|\times|\mathcal{J}|}$ denote the submatrix of $\mathbf{A}$ with rows and columns indexed by $\mathcal{I}$ and $\mathcal{J}$, and $\mathbf{A}_{\mathcal{I}} \in \mathbb{R}^{|\mathcal{I}|\times q}$ denote the rows of $\mathbf{A}$ indexed by $\mathcal{I}$. Moreover, let the capital letter $\mathbf{A}_i \in \mathbb{R}^q$ denote the i-th row of $\mathbf{A}$, the lowercase letter $\mathbf{a}_j \in \mathbb{R}^p$ denote the j-th column of $\mathbf{A}$, and the non-bold lowercase letter a_{ij} denote the (i, j)-th element of $\mathbf{A}$. For observations $(\mathbf{X}_i, Y_i)$, $i = 1, \ldots, n$, let $\mathbb{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n) \in \mathbb{R}^{p\times n}$ and let $\mathbb{Y} = (Y_1, \ldots, Y_n)^\top \in \mathbb{R}^n$. For any two numbers a and b, let $a \wedge b = \min\{a, b\}$ and $a \vee b = \max\{a, b\}$. For a convex function g, let ∂g denote the subdifferential of g.

In what follows, we introduce some useful propositions and lemmas that were used in Sections F, G, and H, including the spectral inequalities, the definitions and the derivative lemmas for sub-Gaussian random variable and vector.

Proposition E.1. *(Stewart and Sun, 1990, Corollary 4.10) 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*

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

Proposition E.2. *(Yu et al., 2015, Theorem 3) 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. Fix* $1 \leq r \leq s \leq \mathrm{rank}(\mathbf{A})$ *and assume that* $\min\{\sigma_{r-1}^2 - \sigma_r^2, \sigma_s^2 - \sigma_{s+1}^2\} > 0$*, where we define* $\sigma_0^2 = \infty$ *and* $\sigma_{\mathrm{rank}(\mathbf{A})+1}^2 = -\infty$*. Let* $d = s - r + 1$*, and let* $\mathbf{U} = (\mathbf{u}_r, \mathbf{u}_{r+1}, \ldots, \mathbf{u}_s) \in \mathbb{R}^{q\times d}$ *and* $\widehat{\mathbf{U}} = (\widehat{\mathbf{u}}_r, \widehat{\mathbf{u}}_{r+1}, \ldots, \widehat{\mathbf{u}}_s) \in \mathbb{R}^{q\times d}$ *denote the left singular vectors satisfying* $\mathbf{A}^\top\mathbf{u}_j = \sigma_j\mathbf{v}_j$ *and* $\widehat{\mathbf{A}}^\top\widehat{\mathbf{u}}_j = \widehat{\sigma}_j\widehat{\mathbf{v}}_j$*,* $j = r, \ldots, s$*. 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*

$$\|\sin\Theta(\widehat{\mathbf{U}}, \mathbf{U})\|_F \leq \frac{2(2\sigma_1 + \|\widehat{\mathbf{A}} - \mathbf{A}\|)\min\{d^{1/2}\|\widehat{\mathbf{A}} - \mathbf{A}\|, \|\widehat{\mathbf{A}} - \mathbf{A}\|_{\mathrm{F}}\}}{\min\{\sigma_{r-1}^2 - \sigma_r^2, \sigma_s^2 - \sigma_{s+1}^2\}}.$$

Moreover, there exists an orthogonal matrix $\mathbf{O} \in \mathbb{R}^{d\times d}$ *such that*

$$\|\widehat{\mathbf{U}}\mathbf{O} - \mathbf{U}\|_F \leq \frac{2^{3/2}(2\sigma_1 + \|\widehat{\mathbf{A}} - \mathbf{A}\|)\min\{d^{1/2}\|\widehat{\mathbf{A}} - \mathbf{A}\|, \|\widehat{\mathbf{A}} - \mathbf{A}\|_{\mathrm{F}}\}}{\min\{\sigma_{r-1}^2 - \sigma_r^2, \sigma_s^2 - \sigma_{s+1}^2\}}.$$

Proposition E.3. *(Horn and Johnson, 2012, Theorem 4.5.9) Let* $\mathbf{A}, \mathbf{S} \in \mathbb{R}^{p\times p}$ *with* $\mathbf{A}$ *Hermitian and* $\mathbf{S}$ *non-singular. Let* $\varphi_k(\cdot)$ *denote the* k*-th largest eigenvalue of a matrix and let* $\sigma_1 \geq \cdots \geq \sigma_n > 0$ *be the singular values of* $\mathbf{S}$*. For each* $k = 1, \ldots, n$*, there is a positive real number* $\theta_k \in [\sigma_n^2, \sigma_1^2]$ *such that*

$$\varphi_k\left(\mathbf{S}\mathbf{A}\mathbf{S}^\top\right) = \theta_k\varphi_k(\mathbf{A}).$$

Proposition E.4. *(Zhou and Li, 2014, Lemma 1) For any matrix* $\mathbf{B} \in \mathbb{R}^{p_1\times p_2}$*, let* $\mathbf{b} = \sigma(\mathbf{B}) = (\sigma_1(\mathbf{B}), \ldots, \sigma_q(\mathbf{B}))^\top$ *denote the vector of singular values of* $\mathbf{B}$ *arranged in decreasing order, where* $q = \min\{p_1, p_2\}$*. And let* $\mathbf{U}\mathrm{diag}(\mathbf{b})\mathbf{V}^\top$ *denote the singular value decomposition of* $\mathbf{B}$*, where* $\mathbf{U} \in \mathbb{R}^{p_1\times q}$ *and* $\mathbf{V} \in \mathbb{R}^{p_2\times q}$*. For a convex function* f*, the subdifferential of the function* $J(\mathbf{B}) = f \circ \sigma(\mathbf{B})$ *is*

$$\partial J(\mathbf{B}) = \partial(f \circ \sigma)(\mathbf{B}) = \mathbf{U}\mathrm{diag}\{\partial f(\mathbf{b})\}\mathbf{V}^\top.$$

Proposition E.5. *(Laurent and Massart, 2000, Lemma 1) Let* U *be a* χ^2 *statistic with* D *degrees of freedom. For any positive* x*,*

$$\mathbb{P}(U - D \geq 2\sqrt{Dx} + 2x) \leq \exp(-x),$$
$$\mathbb{P}(D - U \geq 2\sqrt{Dx}) \leq \exp(-x).$$

Proposition E.6. *([Cai et al., 2019], 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 $m^{-1} \le \varphi_{\min}(\boldsymbol{\Sigma}) \le \varphi_{\max}(\boldsymbol{\Sigma}) \le m$ for some positive constant $m > 1$, where $\varphi_{\min}(\boldsymbol{\Sigma})$ and $\varphi_{\max}(\boldsymbol{\Sigma})$ denote the smallest and the largest eigenvalues of $\boldsymbol{\Sigma}$. For any $t > 0$ and any fixed non-zero vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^p$, we have*

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} \mathbf{u}^\top \mathbf{X}_i \mathbf{X}_i^\top \mathbf{v} - \mathbf{u}^\top \boldsymbol{\Sigma} \mathbf{v}\right| \ge t \|\mathbf{u}\|_2 \|\mathbf{v}\|_2\right) \le 2\exp\left(-cn \min\left\{\frac{t^2}{\kappa^4}, \frac{t}{\kappa^2}\right\}\right)$$

for some constant $c > 0$.

Proposition E.7. *([Cai et al., 2019], Lemma 7) For $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)^\top \in \mathbb{R}^{n\times p}$, 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 $m^{-1} \le \varphi_{\min}(\boldsymbol{\Sigma}) \le \varphi_{\max}(\boldsymbol{\Sigma}) \le m$ for some positive constant $m > 1$, where $\varphi_{\min}(\boldsymbol{\Sigma})$ and $\varphi_{\max}(\boldsymbol{\Sigma})$ denote the smallest and the largest eigenvalues of $\boldsymbol{\Sigma}$. Suppose $\mathcal{A} \subseteq \{1, \ldots, p\}$ is with cardinality s. Then for any $t > 0$, we have*

$$\mathbb{P}\left(\left\|\frac{1}{n}(\mathbf{X}^\top\mathbf{X})_{\mathcal{A}\mathcal{A}}\boldsymbol{\Sigma}^{-1}_{\mathcal{A}\mathcal{A}} - \mathbf{I}_s\right\| \ge t\right) \le 2\exp\left[(\log 9)s - \frac{n}{2}\left\{\frac{t^2}{(32m^2\kappa^2)^2} \wedge \frac{t}{32m^2\kappa^2}\right\}\right].$$

Lemma E.1. *For $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{p\times q}$ $(p \ge q)$, let $\sigma_1(\mathbf{A}) \ge \ldots \ge \sigma_q(\mathbf{A}) \ge 0$ and $\sigma_1(\mathbf{B}) \ge \ldots \ge \sigma_q(\mathbf{B}) \ge 0$ be the singular values of $\mathbf{A}$ and $\mathbf{B}$ separately, then we have*

$$\sum_{i=1}^{q}\{\sigma_i(\mathbf{A}) - \sigma_i(\mathbf{B})\}^2 \le \|\mathbf{A} - \mathbf{B}\|_F^2.$$

Proof of Lemma E.1. We construct the symmetric matrix $\widetilde{\mathbf{A}} \in \mathbb{R}^{(p+q)\times(p+q)}$ from $\mathbf{A}$ as

$$\widetilde{\mathbf{A}} = \begin{pmatrix} \mathbf{0} & \mathbf{A} \\ \mathbf{A}^\top & \mathbf{0} \end{pmatrix},$$

which satisfies $\widetilde{\mathbf{A}} = \widetilde{\mathbf{A}}^\top$. Similarly, we construct the symmetric matrix $\widetilde{\mathbf{B}} \in \mathbb{R}^{(p+q)\times(p+q)}$ from $\mathbf{B}$. Let $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$ denote the singular value decomposition of $\mathbf{A}$. We split $\mathbf{U}$ as $(\mathbf{U}_1, \mathbf{U}_2)$, where $\mathbf{U}_1 \in \mathbb{R}^{p\times q}$ and $\mathbf{U}_2 \in \mathbb{R}^{p\times(p-q)}$, and $\mathbf{D} = (\boldsymbol{\Lambda}^\top, \mathbf{0}^\top)^\top$, where $\boldsymbol{\Lambda} = \mathrm{diag}\{\sigma_1(\mathbf{A}), \ldots, \sigma_q(\mathbf{A})\} \in \mathbb{R}^{q\times q}$ and $\mathbf{0} \in \mathbb{R}^{(p-q)\times q}$. Then we have the spectral decomposition of $\widetilde{\mathbf{A}}$ as $\widetilde{\mathbf{A}} = \widetilde{\mathbf{U}}\widetilde{\mathbf{D}}\widetilde{\mathbf{U}}^\top$, where

$$\widetilde{\mathbf{U}} = \begin{pmatrix} \mathbf{U}_1/\sqrt{2} & -\mathbf{U}_1/\sqrt{2} & \mathbf{U}_2 \\ \mathbf{V}/\sqrt{2} & \mathbf{V}/\sqrt{2} & \mathbf{0}_1 \end{pmatrix},\ \widetilde{\mathbf{D}} = \mathrm{diag}(\boldsymbol{\Lambda}, -\boldsymbol{\Lambda}, \mathbf{0}_2),$$

where $\mathbf{0}_1 \in \mathbb{R}^{q\times(p-q)}$ and $\mathbf{0}_2 \in \mathbb{R}^{(p-q)\times(p-q)}$. Thus, the eigenvalues of $\widetilde{\mathbf{A}}$ are $\{\sigma_1(\mathbf{A}), \ldots, \sigma_q(\mathbf{A}), -\sigma_1(\mathbf{A}), \ldots, -\sigma_q(\mathbf{A}), 0, \ldots, 0\}$. Similarly, we obtain the eigenvalues of $\widetilde{\mathbf{B}}$ from the singular values of $\mathbf{B}$. Then, $\sum_{i=1}^{n}\{\varphi_i(\widetilde{\mathbf{A}})-\varphi_i(\widetilde{\mathbf{B}})\}^2 = 2\sum_{i=1}^{n}\{\sigma_i(\mathbf{A})-\sigma_i(\mathbf{B})\}^2$ and $\|\widetilde{\mathbf{A}}-\widetilde{\mathbf{B}}\|_F^2 = 2\mathrm{tr}\{(\mathbf{A}-\mathbf{B})(\mathbf{A}-\mathbf{B})^\top\} = 2\|\mathbf{A}-\mathbf{B}\|_F^2$.

Hoffman-Wielandt inequality (Hoffman and Wielandt, 1953, Theorem 1) states that for $p\times p$ symmetric matrices $\mathbf{Z}$ and $\widehat{\mathbf{Z}}$, let $\varphi_i(\cdot)$ be the i-th eigenvalue of a matrix, then

$$\sum_{i=1}^{p}\{\varphi_i(\mathbf{Z})-\varphi_i(\widehat{\mathbf{Z}})\}^2 \le \|\mathbf{Z}-\widehat{\mathbf{Z}}\|_F^2.$$

By applying Hoffman-Wielandt inequality, we have

$$\sum_{i=1}^{n}\{\varphi_i(\widetilde{\mathbf{A}})-\varphi_i(\widetilde{\mathbf{B}})\}^2 \le \|\widetilde{\mathbf{A}}-\widetilde{\mathbf{B}}\|_F^2 \Rightarrow \sum_{i=1}^{q}\{\sigma_i(\mathbf{A})-\sigma_i(\mathbf{B})\}^2 \le \|\mathbf{A}-\mathbf{B}\|_F^2.$$

Thus, the proof is complete. □

Lemma E.2. *Let $Z_1,\ldots,Z_n$ be independent random variables satisfying the moments conditions*

$$\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}(|Z_i|^k) \le \frac{k!}{2}v^2 r^{k-2}, \quad \forall k\ge 2$$

for some positive constants v and r. Then, for any $\epsilon>0$,

$$\mathbb{P}\{|S_n-\mathrm{E}(S_n)|\ge n\epsilon\} \le 2\exp\left(\frac{-n\epsilon^2/2}{v^2+r\epsilon}\right),$$

where $S_n=\sum_{i=1}^{n} Z_i$.

Proof of Lemma E.2. According to Lemma 8 in Birgé and Massart (1998), under the conditions (E.2),

$$\mathbb{P}\{S_n-\mathrm{E}(S_n)\ge n\epsilon\} \le \exp\left(\frac{-n\epsilon^2/2}{v^2+r\epsilon}\right).$$

By replacing Z_i with $-Z_i$, $i=1,\cdots,n$, which still satisfies the conditions (E.2), we have

$$\mathbb{P}\{-S_n+\mathrm{E}(S_n)\ge n\epsilon\} \le \exp\left(\frac{-n\epsilon^2/2}{v^2+r\epsilon}\right).$$

Thus,

$$\mathbb{P}\{S_n-\mathrm{E}(S_n)\le -n\epsilon\} \le \exp\left(\frac{-n\epsilon^2/2}{v^2+r\epsilon}\right),$$

and we obtain the desired inequality. □

Lemma E.3. *For random variable $\mathbf{X}\in\mathbb{R}^p$, let $\mathbf{\Sigma}=\mathrm{Cov}(\mathbf{X})\in\mathbb{R}^{p\times p}$. Let $\varphi_{\min}(\mathbf{\Sigma})$ and $\varphi_{\max}(\mathbf{\Sigma})$ denote the smallest and the largest eigenvalues of $\mathbf{\Sigma}$. If $m^{-1}\le\varphi_{\min}(\mathbf{\Sigma})\le\varphi_{\max}(\mathbf{\Sigma})\le m$ for some constant $m>1$, then for any set $\mathcal{A}\subseteq\{1,\ldots,p\}$ and $k\in\mathcal{A}^c$, where $\mathcal{A}^c$ is the complement of $\mathcal{A}$, we have*

$$m^{-1}\le \mathbf{\Sigma}_{kk}-\mathbf{\Sigma}_{k\mathcal{A}}\mathbf{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1}\mathbf{\Sigma}_{\mathcal{A}k}\le m.$$

Proof of Lemma E.3. Without loss of generality, assume that $\mathcal{A} = \{1, \ldots, s\}$ for some $s < p$. Given that $m^{-1} \le \varphi_{\min}(\mathbf{\Sigma}) \le \varphi_{\max}(\mathbf{\Sigma}) \le m$, we have $m^{-1} \le \varphi_{\min}(\mathbf{\Sigma}^{-1}) \le \varphi_{\max}(\mathbf{\Sigma}^{-1}) \le m$. Then $m^{-1} \le \varphi_{\min}\{(\mathbf{\Sigma}^{-1})_{\mathcal{A}^c\mathcal{A}^c}\} \le \varphi_{\max}\{(\mathbf{\Sigma}^{-1})_{\mathcal{A}^c\mathcal{A}^c}\} \le m$. By block matrix inverse formula,

$$(\mathbf{\Sigma}^{-1})_{\mathcal{A}^c\mathcal{A}^c} = (\mathbf{\Sigma}_{\mathcal{A}^c\mathcal{A}^c} - \mathbf{\Sigma}_{\mathcal{A}^c\mathcal{A}}\mathbf{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1}\mathbf{\Sigma}_{\mathcal{A}\mathcal{A}^c})^{-1}.$$

Let $\mathbf{\Sigma}_{\mathcal{A}^c\circ\mathcal{A}} = \mathbf{\Sigma}_{\mathcal{A}^c\mathcal{A}^c} - \mathbf{\Sigma}_{\mathcal{A}^c\mathcal{A}}\mathbf{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1}\mathbf{\Sigma}_{\mathcal{A}\mathcal{A}^c}$, then $m^{-1} \le \varphi_{\min}(\mathbf{\Sigma}_{\mathcal{A}^c\circ\mathcal{A}}) \le \varphi_{\max}(\mathbf{\Sigma}_{\mathcal{A}^c\circ\mathcal{A}}) \le m$. In addition, for any $k \in \mathcal{A}^c$, $\varphi_{\min}(\mathbf{\Sigma}_{\mathcal{A}^c\circ\mathcal{A}}) \le (\mathbf{\Sigma}_{\mathcal{A}^c\circ\mathcal{A}})_{kk} \le \varphi_{\max}(\mathbf{\Sigma}_{\mathcal{A}^c\circ\mathcal{A}})$. Hence, the proof is finished. $\square$

Next, we present the definitions of (centered) sub-Gaussian random variable and (centered) sub-Gaussian random vector along with some well-known lemmas. The proofs of Lemmas E.4–E.5 are omitted here. For more details, see, for example, Wainwright (2019).

Definition E.1. *A random variable X is (centered) sub-Gaussian with parameter σ^2 if* $\mathrm{E}(X) = 0$ *and its moment generating function satisfies*

$$\mathrm{E}(e^{sX}) \le \exp(\sigma^2 s^2/2), \quad \forall s \in \mathbb{R}.$$

The sub-Gaussian random variable X is denoted as $X \sim \mathrm{subG}(\sigma^2)$.

Definition E.2. *A random vector $\mathbf{X} \in \mathbb{R}^p$ is (centered) sub-Gaussian random vector with parameters $\sigma^2 > 0$ if $\mathrm{E}(\mathbf{X}) = \mathbf{0}$ and*

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

Lemma E.4. *Let $X \sim \mathrm{subG}(\sigma^2)$, then for any positive integer $k \ge 1$,*

$$\mathrm{E}(|X|^k) \le (2\sigma^2)^{k/2} k\Gamma(k/2).$$

Lemma E.5 (Hoeffding's inequality)**.** *Let $X_1, \ldots, X_n$ be n independent random variables bounded by the interval $[a_i, b_i]$, $i = 1, \ldots, n$. Define the sum $S_n = \sum_{i=1}^n X_i$ and for any $t \ge 0$ we have*

$$\mathbb{P}\{|S_n - \mathrm{E}(S_n)| \ge t\} \le 2\exp\left\{-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right\}$$

Lemma E.6. *Suppose that $\mathbf{X} \in \mathbb{R}^p$ is a sub-Gaussian random vector with parameter $\sigma^2 > 0$ and let $\mathrm{Cov}(\mathbf{X}) = \mathbf{\Sigma}$. Then we have,*

(1) $X_j \sim \mathrm{subG}(\sigma^2)$, $j = 1, \ldots, p$,

(2) $\|\mathbf{\Sigma}\| \le 4\sigma^2$.

Proof of Lemma E.6. According to the definition of sub-Gaussian random vector, for any $s > 0$ and $\mathbf{a} \in \mathbb{R}^p$ such that $\|\mathbf{a}\|_2 = 1$, we have

$$\mathrm{E}\{\exp(s\mathbf{a}^\top \mathbf{X})\} \leq \exp(s^2\sigma^2/2).$$

Therefore, $\mathbf{a}^\top \mathbf{X} \sim \mathrm{subG}(\sigma^2)$. For $\mathbf{a} = \mathbf{e}_j$, where $\mathbf{e}_j$ is the j-th standard basis vector in $\mathbb{R}^p$, we have

$$X_j = \mathbf{e}_j^\top \mathbf{X} \sim \mathrm{subG}(\sigma^2), \quad j = 1, \dots, p.$$

By Lemma E.4, for $\mathbf{a} \in \mathbb{R}^p$ such that $\|\mathbf{a}\|_2 = 1$, we have $\mathrm{Cov}(\mathbf{a}^\top \mathbf{X}) \leq 4\sigma^2$. Then,

$$\mathbf{a}^\top \boldsymbol{\Sigma} \mathbf{a} = \mathrm{Cov}(\mathbf{a}^\top \mathbf{X}) \leq 4\sigma^2, \quad \forall \|\mathbf{a}\|_2 = 1.$$

It follows that

$$\|\boldsymbol{\Sigma}\| = \sup_{\|\mathbf{a}\|_2=1} \mathbf{a}^\top \boldsymbol{\Sigma} \mathbf{a} \leq 4\sigma^2.$$

□

Lemma E.7. *Suppose that $\mathbf{X} \in \mathbb{R}^{p\times p}$ and $\mathbf{X} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, then $\mathbf{X}$ is a sub-Gaussian random vector with parameter $\|\boldsymbol{\Sigma}\|$.*

Proof of Lemma E.7. For any $\mathbf{a} \in \mathbb{R}^p$, $\mathbf{a}^\top \mathbf{X} \sim N(0, \mathbf{a}^\top \boldsymbol{\Sigma} \mathbf{a})$. The moment generating function of $\mathbf{a}^\top \mathbf{X}$ is

$$\mathrm{E}\{\exp(\lambda \mathbf{a}^\top \mathbf{X})\} = \exp(\lambda^2 \mathbf{a}^\top \boldsymbol{\Sigma} \mathbf{a}/2) \leq \exp(\lambda^2 \|\mathbf{a}\|_2^2 \|\boldsymbol{\Sigma}\|/2), \quad \lambda \in \mathbb{R}.$$

Hence, $\mathbf{X}$ is a sub-Gaussian random vector with parameter $\|\boldsymbol{\Sigma}\|$. □

F Proof of Theorem 1

For ease of presentation, throughout the technical proofs, we use C and C' to denote generic positive numbers that can vary from line to line.

Recall that for a positive semi-definite matrix $\mathbf{M} \in \mathbb{R}^{p\times p}$ and any $k \in \{1, \dots, p\}$, we define the smallest and the largest restricted eigenvalues as

$$\varphi_{\min}^{\mathbf{M}}(k) = \min_{\|\mathbf{u}\|_0 \leq k, \mathbf{u}\neq 0} \frac{\mathbf{u}^\top \mathbf{M}\mathbf{u}}{\mathbf{u}^\top \mathbf{u}}, \quad \varphi_{\max}^{\mathbf{M}}(k) = \max_{\|\mathbf{u}\|_0 \leq k, \mathbf{u}\neq 0} \frac{\mathbf{u}^\top \mathbf{M}\mathbf{u}}{\mathbf{u}^\top \mathbf{u}}.$$

Also recall Conditions (C1)(C2) and Assumptions (A1)(A2) in Section 4:

(C1) There exist constants $C, C' > 0$ such that $\mathbb{P}(\|\widehat{\mathbf{U}} - \mathbf{U}\|_{\max} \leq \epsilon, \|(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})\mathbf{b}_h^\star\|_\infty \leq \epsilon, \forall h) \geq 1 - Cp\exp(-C'n\epsilon^2)$ for $0 < \epsilon \leq C'$, where $\mathbf{b}_h^\star$ is the h-th column of $\mathbf{B}^\star$.

(C2) There exist constants $K > 1$ and $C, C' > 0$ such that

$$\mathbb{P}\{K^{-1} \leq \varphi_{\min}^{\widehat{\boldsymbol{\Sigma}}}(k) \leq \varphi_{\max}^{\widehat{\boldsymbol{\Sigma}}}(k) \leq K\} \geq 1 - C\exp(-C'k)$$

for any $k < p$ satisfying $ck\log(ep/k) \leq n$ for some constant $c > 1$.

(A1) There exists some sufficiently large constant $M > 0$ which does not scale with n, p, s, such that $\sigma_{d^\star} \geq M\sqrt{s\log p/n}$.

(A2) There exits some sufficiently large constant $N > 0$ which does not scale with n, p, s, such that $\min_{j\in\mathcal{A}} \|\mathbf{B}_j^\star\|_2 \geq N(\sigma_1^2/\sigma_{d^\star}^2)\sqrt{s\log p/n}$.

Also recall that $\mathbf{B}^\star = \boldsymbol{\Sigma}^{-1}\mathbf{U}$ and the estimator $\widehat{\mathbf{B}}$ is the solution of the following doubly penalized optimization problem

$$\underset{\mathbf{B}\in\mathbb{R}^{p\times H}}{\operatorname{argmin}} \frac{1}{2}\operatorname{tr}(\mathbf{B}^\top\widehat{\boldsymbol{\Sigma}}\mathbf{B}) - \operatorname{tr}(\mathbf{B}^\top\widehat{\mathbf{U}}) + \lambda_1\|\mathbf{B}\|_{2,1} + \lambda_2\|\mathbf{B}\|_*. \tag{F.1}$$

Let $L(\mathbf{B}) = \frac{1}{2}\operatorname{tr}(\mathbf{B}^\top\widehat{\boldsymbol{\Sigma}}\mathbf{B}) - \operatorname{tr}(\mathbf{B}^\top\widehat{\mathbf{U}}) + \lambda_1\|\mathbf{B}\|_{2,1} + \lambda_2\|\mathbf{B}\|_*$, the following lemma gives the upper bound of $\|\widehat{\mathbf{B}} - \mathbf{B}^\star\|_F$.

Lemma F.1. *Assume that Conditions (C1) and (C2) hold and that $C_1 s\log p \leq n$ for some sufficiently large constant $C_1 > 0$. There exist constants $c_1, c_2, C_B, C, C' > 0$ such that, when $2c_1\sqrt{\log p/n} \leq \lambda_1 \leq 3c_1\sqrt{\log p/n}$ and $\lambda_2 \leq c_2\lambda_1$, we have that*

$$\|\widehat{\mathbf{B}} - \mathbf{B}^\star\|_F \leq C_B\sqrt{s\log p/n}$$

with probability at least $1 - C\exp\{-C'(s\wedge\log p)\}$.

Proof of Lemma F.1. For simplicity, we assume that $\mathrm{E}(\mathbf{X}_i) = 0$, $i = 1, \ldots, n$ and take $\widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i^\top = \frac{1}{n}\mathbb{X}\mathbb{X}^\top$.

Since $\widehat{\mathbf{B}}$ is the minimizer of the function $L(\mathbf{B})$, we have $L(\widehat{\mathbf{B}}) \leq L(\mathbf{B}^\star)$, and it follows that

$$\begin{aligned}
&\lambda_1(\|\widehat{\mathbf{B}}\|_{2,1} - \|\mathbf{B}^\star\|_{2,1}) + \lambda_2(\|\widehat{\mathbf{B}}\|_* - \|\mathbf{B}^\star\|_*)\\
\leq&\frac{1}{2}\operatorname{tr}\{(\mathbf{B}^\star)^\top\widehat{\boldsymbol{\Sigma}}\mathbf{B}^\star\} - \operatorname{tr}\{(\mathbf{B}^\star)^\top\widehat{\mathbf{U}}\} - \frac{1}{2}\operatorname{tr}(\widehat{\mathbf{B}}^\top\widehat{\boldsymbol{\Sigma}}\widehat{\mathbf{B}}) + \operatorname{tr}(\widehat{\mathbf{B}}^\top\widehat{\mathbf{U}})\\
=&\sum_{h=1}^H\{\frac{1}{2}(\mathbf{b}_h^\star)^\top\widehat{\boldsymbol{\Sigma}}\mathbf{b}_h^\star - (\mathbf{b}_h^\star)^\top\widehat{\mathbf{u}}_h - \frac{1}{2}\widehat{\mathbf{b}}_h^\top\widehat{\boldsymbol{\Sigma}}\widehat{\mathbf{b}}_h + \widehat{\mathbf{b}}_h^\top\widehat{\mathbf{u}}_h\}\\
=&\sum_{h=1}^H\{-\frac{1}{2}(\widehat{\mathbf{b}}_h - \mathbf{b}_h^\star)^\top\widehat{\boldsymbol{\Sigma}}(\widehat{\mathbf{b}}_h - \mathbf{b}_h^\star) - \langle\widehat{\boldsymbol{\Sigma}}\mathbf{b}_h^\star - \widehat{\mathbf{u}}_h, \widehat{\mathbf{b}}_h - \mathbf{b}_h^\star\rangle\} \qquad \text{(F.2)}\\
\leq&\sum_{h=1}^H|\langle\widehat{\boldsymbol{\Sigma}}\mathbf{b}_h^\star - \widehat{\mathbf{u}}_h, \widehat{\mathbf{b}}_h - \mathbf{b}_h^\star\rangle| \leq \sum_{h=1}^H\|\widehat{\boldsymbol{\Sigma}}\mathbf{b}_h^\star - \widehat{\mathbf{u}}_h\|_\infty\|\widehat{\mathbf{b}}_h - \mathbf{b}_h^\star\|_1,
\end{aligned}$$

where we use the fact that $\widehat{\boldsymbol{\Sigma}}$ is positive semi-definite in the second inequality. We take $\epsilon = \sqrt{C_0 \log p/n}$ with the constant $C_0 = 2/C'$, where C' comes from Condition (C1). We also take $C_1 \geq 2/(C')^3$. Since we assume that $C_1 s \log p \leq n$, then $\log p/n \leq 1/C_1$ and $\epsilon \leq \sqrt{C_0/C_1} \leq C'$. Meanwhile, since $C_0 C' > 1$, by Condition (C1), with probability at least $1 - Cp^{-C'}$ for some constants $C, C' > 0$, we have

$$\begin{aligned} &\|\widehat{\boldsymbol{\Sigma}}\mathbf{b}_h^\star - \widehat{\mathbf{u}}_h\|_\infty \leq \|\widehat{\boldsymbol{\Sigma}}\mathbf{b}_h^\star - \boldsymbol{\Sigma}\mathbf{b}_h^\star + \boldsymbol{\Sigma}\mathbf{b}_h^\star - \widehat{\mathbf{u}}_h\|_\infty \\ \leq &\|(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})\mathbf{b}_h^\star\|_\infty + \|\mathbf{u}_h - \widehat{\mathbf{u}}_h\|_\infty \leq 2\sqrt{C_0 \log p/n}, \end{aligned} \tag{F.3}$$

where we use the fact that $\mathbf{b}_h^\star = \boldsymbol{\Sigma}^{-1}\mathbf{u}_h$. Since we assume that $H \leq A$ for some constant $A > 0$, then

$$\begin{aligned} &\lambda_1(\|\widehat{\mathbf{B}}\|_{2,1} - \|\mathbf{B}^\star\|_{2,1}) + \lambda_2(\|\widehat{\mathbf{B}}\|_* - \|\mathbf{B}^\star\|_*) \leq 2\sqrt{C_0 \log p/n}\sum_{h=1}^H \|\widehat{\mathbf{b}}_h - \mathbf{b}_h^\star\|_1 \\ \leq &2\sqrt{H}\sqrt{C_0 \log p/n}\sum_{j=1}^p \|\widehat{\mathbf{B}}_j - \mathbf{B}_j^\star\|_2 \leq 2\sqrt{A}\sqrt{C_0 \log p/n}\sum_{j=1}^p \|\widehat{\mathbf{B}}_j - \mathbf{B}_j^\star\|_2 \\ = &c_1\sqrt{\log p/n}\sum_{j=1}^p \|\widehat{\mathbf{B}}_j - \mathbf{B}_j^\star\|_2, \end{aligned} \tag{F.4}$$

where $c_1 = 2\sqrt{AC_0} = 2\sqrt{2A/C'}$. On the other hand,

$$\begin{aligned} &\lambda_1(\|\widehat{\mathbf{B}}\|_{2,1} - \|\mathbf{B}^\star\|_{2,1}) + \lambda_2(\|\widehat{\mathbf{B}}\|_* - \|\mathbf{B}^\star\|_*) \\ = &\lambda_1\sum_{j\in\mathcal{A}^c} \|\widehat{\mathbf{B}}_j\|_2 + \lambda_1\sum_{j\in\mathcal{A}}(\|\widehat{\mathbf{B}}_j\|_2 - \|\mathbf{B}_j^\star\|_2) + \lambda_2(\|\widehat{\mathbf{B}}\|_* - \|\mathbf{B}^\star\|_*) \\ \geq &\lambda_1\sum_{j\in\mathcal{A}^c} \|\widehat{\mathbf{B}}_j - \mathbf{B}_j^\star\|_2 - \lambda_1\sum_{j\in\mathcal{A}} \|\widehat{\mathbf{B}}_j - \mathbf{B}_j^\star\|_2 - \lambda_2\|\widehat{\mathbf{B}} - \mathbf{B}^\star\|_*. \end{aligned} \tag{F.5}$$

Let $\boldsymbol{\Delta} = \widehat{\mathbf{B}} - \mathbf{B}^\star \in \mathbb{R}^{p\times H}$. Since we assume that $\lambda_1 \geq 2c_1\sqrt{\log p/n}$ and $\lambda_2 \leq c_2\lambda_1$ for some positive constant c_2, by combining (F.4) and (F.5), we have that

$$\sum_{j\in\mathcal{A}^c} \|\boldsymbol{\Delta}_j\|_2 \leq 3\sum_{j\in\mathcal{A}} \|\boldsymbol{\Delta}_j\|_2 + 2c_2\|\boldsymbol{\Delta}\|_*. \tag{F.6}$$

By (F.2), we have that

$$\begin{aligned} &\sum_{h=1}^H (\widehat{\mathbf{b}}_h - \mathbf{b}_h^\star)^\top \widehat{\boldsymbol{\Sigma}}(\widehat{\mathbf{b}}_h - \mathbf{b}_h^\star) \\ \leq &-2\sum_{h=1}^H \langle \widehat{\boldsymbol{\Sigma}}\mathbf{b}_h^\star - \widehat{\mathbf{u}}_h, \widehat{\mathbf{b}}_h - \mathbf{b}_h^\star \rangle - 2\lambda_1(\|\widehat{\mathbf{B}}\|_{2,1} - \|\mathbf{B}^\star\|_{2,1}) - 2\lambda_2(\|\widehat{\mathbf{B}}\|_* - \|\mathbf{B}^\star\|_*) \\ \leq &2\sum_{h=1}^H \|\widehat{\boldsymbol{\Sigma}}\mathbf{b}_h^\star - \widehat{\mathbf{u}}_h\|_\infty \|\widehat{\mathbf{b}}_h - \mathbf{b}_h^\star\|_1 + 2\lambda_1\|\boldsymbol{\Delta}\|_{2,1} + 2\lambda_2\|\boldsymbol{\Delta}\|_*. \end{aligned} \tag{F.7}$$

Moreover, since we assume that $\lambda_1 \le 3c_1\sqrt{\log p/n}$, by combining (F.3) with (F.7), we have

$$\begin{aligned}
&\sum_{h=1}^{H}(\widehat{\mathbf{b}}_h - \mathbf{b}_h^\star)^\top \widehat{\boldsymbol{\Sigma}}(\widehat{\mathbf{b}}_h - \mathbf{b}_h^\star) && \text{(F.8)}\\
\le& 4\sqrt{C_0 \log p/n}\sum_{h=1}^{H}\|\widehat{\mathbf{b}}_h - \mathbf{b}_h^\star\|_1 + 6c_1\sqrt{\log p/n}\|\boldsymbol{\Delta}\|_{2,1} + 6c_1c_2\sqrt{\log p/n}\|\boldsymbol{\Delta}\|_* \\
\le& (4\sqrt{C_0} + 6c_1 + 6c_1c_2)\sqrt{\log p/n}\left(\sum_{h=1}^{H}\|\widehat{\mathbf{b}}_h - \mathbf{b}_h^\star\|_1 + \|\boldsymbol{\Delta}\|_{2,1} + \|\boldsymbol{\Delta}\|_*\right). && \text{(F.9)}
\end{aligned}$$

For $\|\boldsymbol{\Delta}\|_*$, since we assume that $H \le s \le p$, we have

$$\|\boldsymbol{\Delta}\|_* \le \sqrt{\min\{p, H\}}\|\boldsymbol{\Delta}\|_F \le \sqrt{H}\|\boldsymbol{\Delta}\|_F \le \sqrt{s}\|\boldsymbol{\Delta}\|_F. \tag{F.10}$$

For $\|\boldsymbol{\Delta}\|_{2,1}$, since we assume that $H \le s$, by (F.6) and (F.10),

$$\begin{aligned}
\|\boldsymbol{\Delta}\|_{2,1} = \sum_{j\in\mathcal{A}^c}\|\boldsymbol{\Delta}_j\|_2 + \sum_{j\in\mathcal{A}}\|\boldsymbol{\Delta}_j\|_2 \le 4\sum_{j\in\mathcal{A}}\|\boldsymbol{\Delta}_j\|_2 + 2c_2\|\boldsymbol{\Delta}\|_* \le 4\sqrt{s}\sqrt{\sum_{j\in\mathcal{A}}\|\boldsymbol{\Delta}_j\|_2^2} + 2c_2\sqrt{s}\|\boldsymbol{\Delta}\|_F \\
\le 4\sqrt{s}\|\boldsymbol{\Delta}\|_F + 2c_2\sqrt{s}\|\boldsymbol{\Delta}\|_F \le (4 + 2c_2)\sqrt{s}\|\boldsymbol{\Delta}\|_F. && \text{(F.11)}
\end{aligned}$$

And it follows that

$$\sum_{h=1}^{H}\|\widehat{\mathbf{b}}_h - \mathbf{b}_h^\star\|_1 \le \sqrt{H}\|\boldsymbol{\Delta}\|_{2,1} \le (4+2c_2)\sqrt{A}\sqrt{s}\|\boldsymbol{\Delta}\|_F. \tag{F.12}$$

Combining (F.9)–(F.12), we have

$$\sum_{h=1}^{H}(\widehat{\mathbf{b}}_h - \mathbf{b}_h^\star)^\top \widehat{\boldsymbol{\Sigma}}(\widehat{\mathbf{b}}_h - \mathbf{b}_h^\star) = \mathrm{tr}\{\boldsymbol{\Delta}^\top\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Delta}\} \le \widetilde{C}\sqrt{s\log p/n}\|\boldsymbol{\Delta}\|_F, \tag{F.13}$$

where the constant $\widetilde{C} = \{4\sqrt{2/C'} + 6c_1(1+c_2)\}\{(4+2c_2)(\sqrt{A}+1)+1\}$.

On the other hand, we re-express

$$\mathrm{tr}(\boldsymbol{\Delta}^\top\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Delta}) = \|n^{-1/2}\mathbb{X}^\top\boldsymbol{\Delta}\|_F^2.$$

The matrix $\boldsymbol{\Delta} \in \mathbb{R}^{p\times H}$ is decomposed in the following way. Let $J_0 = \mathcal{A}$, and let $J_1 = \{j_1, \dots, j_k\} \subset \mathcal{A}^c$ correspond to the k rows with the largest ℓ_2-norm in $\boldsymbol{\Delta}_{\mathcal{A}^c}$. We define $\widetilde{\mathcal{A}} = \mathcal{A} \cup J_1$. Then we partition $\widetilde{\mathcal{A}}^c$ into disjoint subsets $J_2, \dots, J_L$ of size k (with $|J_L| \le k$) such that $L > 2$. The subset J_l is the set of indices corresponding to the k rows with the largest ℓ_2 norm in $\boldsymbol{\Delta}$ outside $\mathcal{A} \cup \bigcup_{j=1}^{l-1} J_j$. For a matrix $\mathbf{A} = (a_{ij}) \in \mathbb{R}^{p\times q}$ and index set

$\mathcal{I} \subseteq \{1, \ldots, p\}$, let $\mathbf{A}^0_{\mathcal{I}} = (a_{ij} 1_{\{i \in \mathcal{I}\}}) \in \mathbb{R}^{p \times q}$ denote the sparsified $\mathbf{A}$ with the rows outside the index $\mathcal{I}$ replaced with zero. Then by triangle inequality, we have

$$\|n^{-1/2}\mathbb{X}^\top \boldsymbol{\Delta}\|_F \geq \|n^{-1/2}\mathbb{X}^\top \boldsymbol{\Delta}^0_{\widetilde{\mathcal{A}}}\|_F - \sum_{l \geq 2} \|n^{-1/2}\mathbb{X}^\top \boldsymbol{\Delta}^0_{J_l}\|_F. \tag{F.14}$$

Let $\mathbf{v}_h$ denotes the h-th column of $\boldsymbol{\Delta}^0_{\widetilde{\mathcal{A}}}$, then we have

$$\begin{aligned} \|n^{-1/2}\mathbb{X}^\top \boldsymbol{\Delta}^0_{\widetilde{\mathcal{A}}}\|_F &= \sqrt{\operatorname{tr}\{(\boldsymbol{\Delta}^0_{\widetilde{\mathcal{A}}})^\top \widehat{\boldsymbol{\Sigma}} \boldsymbol{\Delta}^0_{\widetilde{\mathcal{A}}}\}} \geq \sqrt{\varphi^{\widehat{\boldsymbol{\Sigma}}}_{\min}(s+k) \sum_{h=1}^{H} \|\mathbf{v}_h\|_2^2} \\ &= \sqrt{\varphi^{\widehat{\boldsymbol{\Sigma}}}_{\min}(s+k)} \|\boldsymbol{\Delta}^0_{\widetilde{\mathcal{A}}}\|_F = \sqrt{\varphi^{\widehat{\boldsymbol{\Sigma}}}_{\min}(s+k)} \|\boldsymbol{\Delta}_{\widetilde{\mathcal{A}}}\|_F. \end{aligned} \tag{F.15}$$

Similarly, we have

$$\|n^{-1/2}\mathbb{X}^\top \boldsymbol{\Delta}^0_{J_l}\|_F \leq \sqrt{\varphi^{\widehat{\boldsymbol{\Sigma}}}_{\max}(k)} \|\boldsymbol{\Delta}_{J_l}\|_F, \quad l = 2, \ldots, L. \tag{F.16}$$

Thus, combining (F.14)–(F.16), we have

$$\|n^{-1/2}\mathbb{X}^\top \boldsymbol{\Delta}\|_F \geq \sqrt{\varphi^{\widehat{\boldsymbol{\Sigma}}}_{\min}(s+k)} \|\boldsymbol{\Delta}_{\widetilde{\mathcal{A}}}\|_F - \sqrt{\varphi^{\widehat{\boldsymbol{\Sigma}}}_{\max}(k)} \sum_{l \geq 2} \|\boldsymbol{\Delta}_{J_l}\|_F. \tag{F.17}$$

We assume that $H \leq A$ for some constant $A > 0$. According to the definition of $J_1, \ldots, J_L$, we have

$$\begin{aligned} &\sum_{l \geq 2} \|\boldsymbol{\Delta}_{J_l}\|_F \leq \sqrt{k} \sum_{l \geq 2} \max_{j \in J_l} \|\boldsymbol{\Delta}_j\|_2 \leq \sqrt{k} \sum_{l \geq 2} k^{-1} \sum_{j \in J_{l-1}} \|\boldsymbol{\Delta}_j\|_2 \leq k^{-1/2} \sum_{j \in \mathcal{A}^c} \|\boldsymbol{\Delta}_j\|_2 \\ \leq& 3k^{-1/2} \sum_{j \in \mathcal{A}} \|\boldsymbol{\Delta}_j\|_2 + 2c_2 k^{-1/2} \|\boldsymbol{\Delta}\|_* \leq 3\sqrt{s/k} \|\boldsymbol{\Delta}_{\widetilde{\mathcal{A}}}\|_F + 2c_2 \sqrt{A/k} \|\boldsymbol{\Delta}\|_F, \end{aligned} \tag{F.18}$$

where we used (F.6) in the second-last inequality. By (F.17) and (F.18), we have

$$\|n^{-1/2}\mathbb{X}^\top \boldsymbol{\Delta}\|_F \geq \left\{ \sqrt{\varphi^{\widehat{\boldsymbol{\Sigma}}}_{\min}(s+k)} - 3\sqrt{s/k} \sqrt{\varphi^{\widehat{\boldsymbol{\Sigma}}}_{\max}(k)} \right\} \|\boldsymbol{\Delta}_{\widetilde{\mathcal{A}}}\|_F - 2c_2 \sqrt{A/k} \sqrt{\varphi^{\widehat{\boldsymbol{\Sigma}}}_{\max}(k)} \|\boldsymbol{\Delta}\|_F.$$

Under Condition (C2), there exist constants $K > 1$ and $C, C' > 0$ such that

$$\mathbb{P}\{K^{-1} \leq \varphi^{\widehat{\boldsymbol{\Sigma}}}_{\min}(k) \leq \varphi^{\widehat{\boldsymbol{\Sigma}}}_{\max}(k) \leq K\} \geq 1 - C\exp\{-C'k\}$$

for any $k < p$ satisfying $ck \log(ep/k) \leq n$ for some constant $c > 1$. Let $k = c_0 s$ with the constant $c_0 = \max\{\lceil 36K^2 \rceil, \lceil 256Ac_2^2K^2 \rceil\}$, where $\lceil a \rceil$ denotes the smallest integer larger than the number $a \in \mathbb{R}$. Since we assume that $C_1 s \log p \leq n$, by taking $C_1 \geq c(c_0 + 1)$, we have

$$\frac{(s+k)\log\{ep/(s+k)\}}{n} \leq (c_0 + 1)\frac{s \log p}{n} \leq \frac{c_0 + 1}{C_1} = \frac{1}{c},$$

where we used the fact that $s+k=(c_0+1)s\geq e$. Hence, $c(s+k)\log\{ep/(s+k)\}\leq n$. Since $g(x)=x\log(ep/x)$ is monotonically increasing for $0<x\leq p$, then we have $ck\log(ep/k)\leq n$. Therefore, by Condition (C2), we have

$$\varphi_{\min}^{\widehat{\boldsymbol{\Sigma}}}(s+k)\geq K^{-1},\quad \varphi_{\max}^{\widehat{\boldsymbol{\Sigma}}}(k)\leq K.$$

with probability at least $1-C\exp(-C's)$. As we select $c_0=\max\{\lceil 36K^2\rceil,\lceil 256Ac_2^2K^2\rceil\}$, then

$$\begin{aligned}\|n^{-1/2}\mathbb{X}^\top\boldsymbol{\Delta}\|_F\geq&\left\{K^{-1/2}-3\sqrt{K/c_0}\right\}\|\boldsymbol{\Delta}_{\widetilde{\mathcal{A}}}\|_F-2c_2\sqrt{AK/(c_0s)}\|\boldsymbol{\Delta}\|_F\\ \geq&2^{-1}K^{-1/2}\|\boldsymbol{\Delta}_{\widetilde{\mathcal{A}}}\|_F-8^{-1}K^{-1/2}\|\boldsymbol{\Delta}\|_F.\end{aligned}\tag{F.19}$$

In addition, by (F.18),

$$\|\boldsymbol{\Delta}\|_F\leq\|\boldsymbol{\Delta}_{\widetilde{\mathcal{A}}}\|_F+\sum_{l\geq 2}\|\boldsymbol{\Delta}_{J_l}\|_F\leq(1+3/\sqrt{c_0})\|\boldsymbol{\Delta}_{\widetilde{\mathcal{A}}}\|_F+2c_2\sqrt{A/c_0s}\|\boldsymbol{\Delta}\|_F.\tag{F.20}$$

Since $c_0\geq 36K^2$ and $c_0\geq 256Ac_2^2K^2$, then $1+3/\sqrt{c_0}\leq 3/2$ and $2c_2\sqrt{A/c_0s}\leq 8^{-1}$. And it follows from (F.20) that

$$\|\boldsymbol{\Delta}_{\widetilde{\mathcal{A}}}\|_F\geq(7/12)\|\boldsymbol{\Delta}\|_F.\tag{F.21}$$

Combining (F.19) and (F.21), it follows that

$$\|n^{-1/2}\mathbb{X}^\top\boldsymbol{\Delta}\|_F\geq 6^{-1}K^{-1/2}\|\boldsymbol{\Delta}\|_F.$$

Thus,

$$\mathrm{tr}\{\boldsymbol{\Delta}^\top\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Delta}\}=\|n^{-1/2}\mathbb{X}^\top\boldsymbol{\Delta}\|_F^2\geq(36K)^{-1}\|\boldsymbol{\Delta}\|_F^2.\tag{F.22}$$

By combining (F.13) and (F.22), we establish that

$$\|\boldsymbol{\Delta}\|_F=\|\widehat{\mathbf{B}}-\mathbf{B}^\star\|_F\leq C_B\sqrt{s\log p/n},$$

for the constant $C_B=36K\widetilde{C}$ with probability at least $1-C\exp\{-C'(s\wedge\log p)\}$. This completes the proof. $\square$

Next, we prove the rank selection consistency in statement (ii) of Theorem 1. Let $\widehat{\sigma}_i$ denote the ith singular value of $\widehat{\mathbf{B}}$. According to Lemma E.1, we have

$$\sum_{i=1}^H(\widehat{\sigma}_i-\sigma_i)^2\leq\|\widehat{\mathbf{B}}-\mathbf{B}^\star\|_F^2.\tag{F.23}$$

Specially, $|\widehat{\sigma}_i-\sigma_i|\leq\|\widehat{\mathbf{B}}-\mathbf{B}^\star\|_F$ for each $i=1,\ldots,H$. Therefore, by Lemma F.1, with probability at least $1-C\exp\{-C'(s\wedge\log p)\}$, we have

$$|\widehat{\sigma}_i-\sigma_i|\leq C_B\sqrt{s\log p/n},\quad i=1,\ldots,H.\tag{F.24}$$

In Assumption (A1), we assume that $\sigma_{d^\star} \geq M\sqrt{s\log p/n}$. By taking $M \geq 3C_B$, for $1 \leq i \leq d^\star$, we have $\widehat{\sigma}_i \geq \sigma_i - C_B\sqrt{s\log p/n} \geq (M - C_B)\sqrt{s\log p/n} \geq 2C_B\sqrt{s\log p/n}$. And for $d^\star < i \leq H$, we have $\widehat{\sigma}_i \leq C_B\sqrt{s\log p/n}$. By definition, $\widehat{d} = \#\{i \mid \widehat{\sigma}_i \geq \delta\}$. Since we select δ such that $C_B\sqrt{s\log p/n} < \delta \leq 2C_B\sqrt{s\log p/n}$, we have the rank selection consistency, i.e., $\widehat{d} = d^\star$.

With the results in statements (i) and (ii) of Theorem 1, we now prove the statment (iii) in the rest of this section. Recall the estimate $\widehat{\boldsymbol{\beta}}$ is the matrix composed of top-$\widehat{d}$ left singular vectors of $\widehat{\mathbf{B}}$. Since we have proved that $\widehat{d} = d^\star$ in statement (ii), the estimator $\widehat{\boldsymbol{\beta}}$ is a $p \times d^\star$ matrix. Let $\boldsymbol{\alpha} \in \mathbb{R}^{p\times d^\star}$ denote the matrix composed of the top-$d^\star$ left singular vectors of $\mathbf{B}^\star$. We define the principal angles between the subspaces spanned by $\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{\alpha}$ as $(\theta_1, \ldots, \theta_{d^\star}) = (\cos^{-1}\omega_1, \ldots, \cos^{-1}\omega_{d^\star})$, where $\omega_1 \geq \cdots \geq \omega_{d^\star}$ are the singular values of $\widehat{\boldsymbol{\beta}}^\top\boldsymbol{\alpha}$. Let $\widehat{\boldsymbol{\beta}}^\top\boldsymbol{\alpha} = \mathbf{M}\mathbf{D}\mathbf{N}^\top$ denote the singular value decomposition of $\widehat{\boldsymbol{\beta}}^\top\boldsymbol{\alpha}$, where $\mathbf{M}, \mathbf{N} \in \mathbb{R}^{d^\star\times d^\star}$ and $\mathbf{D} = \mathrm{diag}(\omega_1, \ldots, \omega_{d^\star}) \in \mathbb{R}^{d^\star\times d^\star}$. And we define $\sin\Theta(\widehat{\boldsymbol{\beta}}, \boldsymbol{\alpha}) = \mathrm{diag}(\sin\theta_1, \ldots \sin\theta_{d^\star}) \in \mathbb{R}^{d^\star\times d^\star}$. By Lemma F.1, we have $\|\widehat{\mathbf{B}} - \mathbf{B}^\star\| \leq \|\widehat{\mathbf{B}} - \mathbf{B}^\star\|_F \leq C_B\sqrt{s\log p/n}$. And by Assumption (A1), we further have $\|\widehat{\mathbf{B}} - \mathbf{B}^\star\| \leq (C_B/M)\sigma_{d^\star} \leq \sigma_1/3$. Then by Proposition E.2, we have

$$
\begin{aligned}
\|\sin\Theta(\widehat{\boldsymbol{\beta}}, \boldsymbol{\alpha})\|_F \leq& \frac{2(2\sigma_1 + \|\widehat{\mathbf{B}} - \mathbf{B}^\star\|)\min\{\sqrt{d^\star}\|\widehat{\mathbf{B}} - \mathbf{B}^\star\|, \|\widehat{\mathbf{B}} - \mathbf{B}^\star\|_{\mathrm{F}}\}}{\sigma_{d^\star}^2} \\
\leq& (14/3)(\sigma_1/\sigma_{d^\star}^2)\sqrt{d^\star}\|\widehat{\mathbf{B}} - \mathbf{B}^\star\|_F \leq (14C_B/3)(\sigma_1/\sigma_{d^\star}^2)\sqrt{d^\star}\sqrt{s\log p/n}.
\end{aligned}
$$

Since $\boldsymbol{\beta}^\star$ and $\boldsymbol{\alpha}$ span the same subspace, we have $\mathbf{P}_{\boldsymbol{\alpha}} = \mathbf{P}_{\boldsymbol{\beta}^\star}$. The subspace distance $\mathcal{D}(\mathcal{S}_{\boldsymbol{\beta}^\star}, \mathcal{S}_{\widehat{\boldsymbol{\beta}}})$ is then connected with $\|\sin\Theta(\widehat{\boldsymbol{\beta}}, \boldsymbol{\alpha})\|_F$ via

$$
\begin{aligned}
\mathcal{D}(\mathcal{S}_{\boldsymbol{\beta}^\star}, \mathcal{S}_{\widehat{\boldsymbol{\beta}}}) &= \|\mathbf{P}_{\boldsymbol{\beta}^\star} - \mathbf{P}_{\widehat{\boldsymbol{\beta}}}\|_F/\sqrt{2d^\star} = \|\mathbf{P}_{\boldsymbol{\alpha}} - \mathbf{P}_{\widehat{\boldsymbol{\beta}}}\|_F/\sqrt{2d^\star} = \|\boldsymbol{\alpha}\boldsymbol{\alpha}^\top - \widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\beta}}^\top\|_F/\sqrt{2d^\star} \\
&= \sqrt{\mathrm{tr}\{\boldsymbol{\alpha}\boldsymbol{\alpha}^\top + \widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\beta}}^\top - 2\widehat{\boldsymbol{\beta}}^\top\boldsymbol{\alpha}(\widehat{\boldsymbol{\beta}}^\top\boldsymbol{\alpha})^\top\}}/\sqrt{2d^\star} = \left(d^\star - \sum_{i=1}^{d^\star}\omega_i^2\right)^{1/2}/\sqrt{d^\star} \\
&= \left(\sum_{i=1}^{d^\star}\sin\theta_i^2\right)^{1/2}/\sqrt{d^\star} = \|\sin\Theta(\widehat{\boldsymbol{\beta}}, \boldsymbol{\alpha})\|_F/\sqrt{d^\star},
\end{aligned}
$$

where we used the fact that $\mathrm{tr}\{\widehat{\boldsymbol{\beta}}^\top\boldsymbol{\alpha}(\widehat{\boldsymbol{\beta}}^\top\boldsymbol{\alpha})^\top\} = \mathrm{tr}(\mathbf{M}\mathbf{D}^2\mathbf{M}^\top) = \mathrm{tr}(\mathbf{D}^2) = \sum_{i=1}^{d^\star}\omega_i^2$. Therefore, we have

$$
\mathcal{D}(\mathcal{S}_{\boldsymbol{\beta}^\star}, \mathcal{S}_{\widehat{\boldsymbol{\beta}}}) \leq (14C_B/3)(\sigma_1/\sigma_{d^\star}^2)\sqrt{s\log p/n} = C(\sigma_1/\sigma_{d^\star}^2)\sqrt{s\log p/n}
$$

for some constant $C > 0$, which completes the proof.

G Proof of Theorem C.1

Recall that our target optimization problem is

$$\begin{aligned} &\widehat{\mathbf{B}} = \operatorname*{argmin}_{\mathbf{B}\in\mathbb{R}^{p\times H}} \widehat{L}(\mathbf{B}), \\ &\text{where } \widehat{L}(\mathbf{B}) = \frac{1}{2}\mathrm{tr}(\mathbf{B}^\top\widehat{\boldsymbol{\Sigma}}\mathbf{B}) - \mathrm{tr}(\mathbf{B}^\top\widehat{\mathbf{U}}) + \lambda_1\|\mathbf{B}\|_{2,1} + \lambda_2\|\mathbf{B}\|_*, \end{aligned} \tag{G.1}$$

To prove Theorem C.1, we define an oracle estimator $\widetilde{\mathbf{G}} \in \mathbb{R}^{s\times H}$ in the following:

$$\begin{aligned} &\widetilde{\mathbf{G}} = \operatorname*{argmin}_{\mathbf{G}\in\mathbb{R}^{s\times H}} \widetilde{L}(\mathbf{G}), \\ &\text{where } \widetilde{L}(\mathbf{G}) = \frac{1}{2}\mathrm{tr}(\mathbf{G}^\top\widehat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A}}\mathbf{G}) - \mathrm{tr}(\mathbf{G}^\top\widehat{\mathbf{U}}_{\mathcal{A}}) + \lambda_1\|\mathbf{G}\|_{2,1} + \lambda_2\|\mathbf{G}\|_*. \end{aligned} \tag{G.2}$$

The oracle estimator $\widetilde{\mathbf{G}}$ is not a legitimate estimator since it requires the oracle knowledge of the active set $\mathcal{A}$. It is only an intermediate quantity to facilitate our proof. Since the oracle estimator $\widetilde{\mathbf{G}}$ is the minimizer of $\widetilde{L}(\mathbf{G})$, we have

$$\mathbf{0}_{s\times H} \in \partial\widetilde{L}(\widetilde{\mathbf{G}}) = \widehat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A}}\widetilde{\mathbf{G}} - \widehat{\mathbf{U}}_{\mathcal{A}} + \lambda_1\partial\|\widetilde{\mathbf{G}}\|_{2,1} + \lambda_2\partial\|\widetilde{\mathbf{G}}\|_*, \tag{G.3}$$

where $\partial\|\widetilde{\mathbf{G}}\|_{2,1}$ and $\partial\|\widetilde{\mathbf{G}}\|_*$ are the subdifferentials of the $L_{2,1}$-norm and nuclear norm evaluated at $\widetilde{\mathbf{G}}$. According to Condition (C2), since we assume that $C_1 s\log p \le n$ for some sufficiently large constant $C_1 > 0$, we have that $\varphi_{\min}(\widehat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A}}) \ge \varphi_{\min}^{\widehat{\boldsymbol{\Sigma}}}(s) \ge K^{-1}$ holds with probability at least $1 - C\exp(-C's)$, for some constant $K > 1$. Hence, $\widehat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A}}$ is invertible with high probability. By (G.3), the oracle estimator $\widetilde{\mathbf{G}}$ can be expressed as

$$\widetilde{\mathbf{G}} = \widehat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A}}^{-1}\widehat{\mathbf{U}}_{\mathcal{A}} - \lambda_1\widehat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A}}^{-1}\widetilde{\mathbf{R}} - \lambda_2\widehat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A}}^{-1}\widetilde{\mathbf{S}},$$

where $\widetilde{\mathbf{R}} \in \partial\|\widetilde{\mathbf{G}}\|_{2,1}$ and $\widetilde{\mathbf{S}} \in \partial\|\widetilde{\mathbf{G}}\|_*$. For the population counterpart of $\widetilde{\mathbf{G}}$, $\mathbf{B}^\star_{\mathcal{A}}$, since $\boldsymbol{\Sigma}\mathbf{B}^\star = \mathbf{U}$ and $\mathbf{B}^\star_{\mathcal{A}^c} = \mathbf{0}_{(p-s)\times H}$, we have $\mathbf{B}^\star_{\mathcal{A}} = \boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1}\mathbf{U}_{\mathcal{A}}$. For consistent notations, in the rest of the proof, we take

$$\mathbf{G}^\star := \mathbf{B}^\star_{\mathcal{A}} = \boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1}\mathbf{U}_{\mathcal{A}}. \tag{G.4}$$

In Corollary 2, we have established $\mathcal{A} \subseteq \widehat{\mathcal{A}}$. For exact recovery, it remains to show that $\widehat{\mathcal{A}} \subseteq \mathcal{A}$. We prove this fact by showing that $\widehat{\mathbf{B}} = (\widetilde{\mathbf{G}}^\top, \mathbf{0}_{(p-s)\times H}^\top)^\top$. Without loss of generality, we assume that $\mathcal{A} = \{1, \ldots, s\}$. In the following, Lemma G.1 provides an equivalent condition for the equality $\widehat{\mathbf{B}} = (\widetilde{\mathbf{G}}^\top, \mathbf{0}_{(p-s)\times H}^\top)^\top$. Lemmas G.2–G.8 give useful technical results for verifying this condition, which eventually leads to Theorem C.1.

Lemma G.1. *For the oracle estimator* $\widetilde{\mathbf{G}}$ *defined as in* (G.2)*, the estimator* $(\widetilde{\mathbf{G}}^\top, \mathbf{0}_{(p-s)\times H}^\top)^\top$ *is the solution to* (G.1) *if and only if*

$$\|\widehat{\boldsymbol{\Sigma}}_{\mathcal{A}^c\mathcal{A}}\widetilde{\mathbf{G}} - \widehat{\mathbf{U}}_{\mathcal{A}^c}\|_{2,\infty} \le \lambda_1. \tag{G.5}$$

Proof of Lemma G.1. For $\widetilde{\mathbf{G}}$, the subdifferential $\partial\|\widetilde{\mathbf{G}}\|_{2,1} = (\partial\|\widetilde{\mathbf{G}}_1\|_2, \ldots, \partial\|\widetilde{\mathbf{G}}_s\|_2)^\top$. Let $\widetilde{\mathbf{M}}\mathrm{diag}(\widetilde{\boldsymbol{\sigma}})\widetilde{\mathbf{N}}^\top$ denote the singular value decomposition of $\widetilde{\mathbf{G}}$, where $\widetilde{\mathbf{M}} \in \mathbb{R}^{s\times H}$, $\widetilde{\mathbf{N}} \in \mathbb{R}^{H\times H}$ and $\mathrm{diag}(\widetilde{\boldsymbol{\sigma}}) \in \mathbb{R}^{H\times H}$ is a diagonal matrix with the singular values on the diagonal positions. By Proposition E.4, the subdifferential $\partial\|\widetilde{\mathbf{G}}\|_* = \widetilde{\mathbf{M}}\mathrm{diag}(\partial\|\widetilde{\boldsymbol{\sigma}}\|_1)\widetilde{\mathbf{N}}^\top$.

Let $\mathbf{A} = (\widetilde{\mathbf{G}}^\top, \mathbf{0}_{(p-s)\times H}^\top)^\top$, then $\partial\|\mathbf{A}\|_{2,1} = (\partial\|\widetilde{\mathbf{G}}\|_{2,1}^\top, \partial\|\mathbf{0}_{(p-s)\times H}\|_{2,1}^\top)^\top$. And the singular value decomposition of $\mathbf{A}$ is

$$\mathbf{A} = \begin{pmatrix} \widetilde{\mathbf{M}} \\ \mathbf{0}_{(p-s)\times H} \end{pmatrix} \mathrm{diag}(\widetilde{\boldsymbol{\sigma}})\widetilde{\mathbf{N}}^\top,$$

then by Proposition E.4, the subdifferential $\partial\|\mathbf{A}\|_*$ is

$$\partial\|\mathbf{A}\|_* = \begin{pmatrix} \widetilde{\mathbf{M}} \\ \mathbf{0}_{(p-s)\times H} \end{pmatrix} \mathrm{diag}(\partial\|\widetilde{\boldsymbol{\sigma}}\|_1)\widetilde{\mathbf{N}}^\top.$$

Since $\widehat{L}(\mathbf{B})$ is convex, the matrix $\mathbf{A} = (\widetilde{\mathbf{G}}^\top, \mathbf{0}_{(p-s)\times H}^\top)^\top$ is the solution to (G.1), or equivalently, $\mathbf{A}$ equals to $\widehat{\mathbf{B}}$, if and only if

$$\begin{aligned} \mathbf{0}_{p\times H} \in & \begin{pmatrix} \widehat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A}} \\ \widehat{\boldsymbol{\Sigma}}_{\mathcal{A}^c\mathcal{A}} \end{pmatrix} \widetilde{\mathbf{G}} - \begin{pmatrix} \widehat{\mathbf{U}}_{\mathcal{A}} \\ \widehat{\mathbf{U}}_{\mathcal{A}^c} \end{pmatrix} + \lambda_1 \begin{pmatrix} \partial\|\widetilde{\mathbf{G}}\|_{2,1} \\ \partial\|\mathbf{0}_{(p-s)\times H}\|_{2,1} \end{pmatrix} \\ & + \lambda_2 \begin{pmatrix} \widetilde{\mathbf{M}} \\ \mathbf{0}_{(p-s)\times H} \end{pmatrix} \mathrm{diag}(\partial\|\widetilde{\boldsymbol{\sigma}}\|_1)\widetilde{\mathbf{N}}^\top. \end{aligned} \tag{G.6}$$

Since the oracle estimator $\widetilde{\mathbf{G}}$ is the minimizer of $\widetilde{L}(\mathbf{G})$, we have

$$\mathbf{0}_{s\times H} \in \partial\widetilde{L}(\widetilde{\mathbf{G}}) = \widehat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A}}\widetilde{\mathbf{G}} - \widehat{\mathbf{U}}_{\mathcal{A}} + \lambda_1\partial\|\widetilde{\mathbf{G}}\|_{2,1} + \lambda_2\partial\|\widetilde{\mathbf{G}}\|_*.$$

Then (G.6) holds if and only if

$$\mathbf{0}_{(p-s)\times H} \in \widehat{\boldsymbol{\Sigma}}_{\mathcal{A}^c\mathcal{A}}\widetilde{\mathbf{G}} - \widehat{\mathbf{U}}_{\mathcal{A}^c} + \lambda_1\partial\|\mathbf{0}_{(p-s)\times H}\|_{2,1},$$

which is equivalent to

$$\|\widehat{\boldsymbol{\Sigma}}_{\mathcal{A}^c\mathcal{A}}\widetilde{\mathbf{G}} - \widehat{\mathbf{U}}_{\mathcal{A}^c}\|_{2,\infty} \le \lambda_1. \tag{G.7}$$

Thus, the desired conclusion is obtained. □

The following lemma provides the concentration result of the oracle estimator $\widetilde{\mathbf{G}}$.

Lemma G.2. *Assume that Conditions (C1) and (C2) hold and that $C_1 s\log p \le n$ for some sufficiently large constant $C_1 > 0$. There exist constants $c_1, c_2, \widetilde{C}, C, C' > 0$ such that, when $2c_1\sqrt{\log p/n} \le \lambda_1 \le 3c_1\sqrt{\log p/n}$ and $\lambda_2 \le c_2\lambda_1$, we have that*

$$\|\widetilde{\mathbf{G}} - \mathbf{G}^\star\|_F \le \widetilde{C}\sqrt{s\log p/n}$$

with probability at least $1 - C\exp\{-C'(s\wedge\log p)\}$.

Proof of Lemma G.2. The proof is similar to that of Lemma F.1. Since

$$\|\widehat{\mathbf{U}}_{\mathcal{A}} - \mathbf{U}_{\mathcal{A}}\|_{\max} \le \|\widehat{\mathbf{U}} - \mathbf{U}\|_{\max}, \quad \|(\widehat{\mathbf{\Sigma}}_{\mathcal{A}\mathcal{A}} - \mathbf{\Sigma}_{\mathcal{A}\mathcal{A}})\mathbf{g}_h^\star\|_\infty \le \|(\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma})\mathbf{b}_h^\star\|_\infty,$$

then under Condition (C1), we have

$$\mathbb{P}(\|\widehat{\mathbf{U}}_{\mathcal{A}} - \mathbf{U}_{\mathcal{A}}\|_{\max} \le \epsilon, \|(\widehat{\mathbf{\Sigma}}_{\mathcal{A}\mathcal{A}} - \mathbf{\Sigma}_{\mathcal{A}\mathcal{A}})\mathbf{g}_h^\star\|_\infty \le \epsilon, \forall h) \ge 1 - Cp\exp(-C'n\epsilon^2), \quad 0 < \epsilon \le C'. \tag{G.8}$$

Since $\widetilde{\mathbf{G}}$ is the minimizer of the function $\widetilde{L}(\mathbf{G})$, we have $\widetilde{L}(\widetilde{\mathbf{G}}) \le \widetilde{L}(\mathbf{G}^\star)$. Then it follows that

$$\begin{aligned}
&\lambda_1(\|\widetilde{\mathbf{G}}\|_{2,1} - \|\mathbf{G}^\star\|_{2,1}) + \lambda_2(\|\widetilde{\mathbf{G}}\|_* - \|\mathbf{G}^\star\|_*) \\
\le& \frac{1}{2}\mathrm{tr}\{(\mathbf{G}^\star)^\top \widehat{\mathbf{\Sigma}}_{\mathcal{A}\mathcal{A}}\mathbf{G}^\star\} - \mathrm{tr}\{(\mathbf{G}^\star)^\top \widehat{\mathbf{U}}_{\mathcal{A}}\} - \frac{1}{2}\mathrm{tr}(\widetilde{\mathbf{G}}^\top \widehat{\mathbf{\Sigma}}_{\mathcal{A}\mathcal{A}}\widetilde{\mathbf{G}}) + \mathrm{tr}(\widetilde{\mathbf{G}}^\top \widehat{\mathbf{U}}_{\mathcal{A}}) \\
=& \sum_{h=1}^H \{\frac{1}{2}(\mathbf{g}_h^\star)^\top \widehat{\mathbf{\Sigma}}_{\mathcal{A}\mathcal{A}}\mathbf{g}_h^\star - (\mathbf{g}_h^\star)^\top \widehat{\mathbf{u}}_{\mathcal{A},h} - \frac{1}{2}\widetilde{\mathbf{g}}_h^\top \widehat{\mathbf{\Sigma}}_{\mathcal{A}\mathcal{A}}\widetilde{\mathbf{g}}_h + \widetilde{\mathbf{g}}_h^\top \widehat{\mathbf{u}}_{\mathcal{A},h}\} \\
=& \sum_{h=1}^H \{-\frac{1}{2}(\widetilde{\mathbf{g}}_h - \mathbf{g}_h^\star)^\top \widehat{\mathbf{\Sigma}}_{\mathcal{A}\mathcal{A}}(\widetilde{\mathbf{g}}_h - \mathbf{g}_h^\star) - \langle \widehat{\mathbf{\Sigma}}_{\mathcal{A}\mathcal{A}}\mathbf{g}_h^\star - \widehat{\mathbf{u}}_{\mathcal{A},h}, \widetilde{\mathbf{g}}_h - \mathbf{g}_h^\star \rangle\}.
\end{aligned} \tag{G.9}$$

Rearranging (G.9), we obtain

$$\begin{aligned}
&\sum_{h=1}^H (\widetilde{\mathbf{g}}_h - \mathbf{g}_h^\star)^\top \widehat{\mathbf{\Sigma}}_{\mathcal{A}\mathcal{A}}(\widetilde{\mathbf{g}}_h - \mathbf{g}_h^\star) \\
\le& -2\sum_{h=1}^H \langle \widehat{\mathbf{\Sigma}}_{\mathcal{A}\mathcal{A}}\mathbf{g}_h^\star - \widehat{\mathbf{u}}_{\mathcal{A},h}, \widetilde{\mathbf{g}}_h - \mathbf{g}_h^\star \rangle - 2\lambda_1(\|\widetilde{\mathbf{G}}\|_{2,1} - \|\mathbf{G}^\star\|_{2,1}) - 2\lambda_2(\|\widetilde{\mathbf{G}}\|_* - \|\mathbf{G}^\star\|_*) \\
\le& 2\sum_{h=1}^H \|\widehat{\mathbf{\Sigma}}_{\mathcal{A}\mathcal{A}}\mathbf{g}_h^\star - \widehat{\mathbf{u}}_{\mathcal{A},h}\|_\infty \|\widetilde{\mathbf{g}}_h - \mathbf{g}_h^\star\|_1 + 2\lambda_1\|\widetilde{\mathbf{G}} - \mathbf{G}^\star\|_{2,1} + 2\lambda_2\|\widetilde{\mathbf{G}} - \mathbf{G}^\star\|_*.
\end{aligned} \tag{G.10}$$

We take $\epsilon = \sqrt{C_0 \log p/n}$ with the constant $C_0 = 2/C'$ where C' comes from Condition (C1). Then $C_0C' > 1$. Since we assume that $C_1 s \log p \le n$, by taking $C_1 \ge 2/(C')^3$, we have $\epsilon \le \sqrt{C_0/C_1} = C'$. Then, by (G.8), with probability at least $1 - Cp^{-C'}$ for some constants $C, C' > 0$, we have

$$\begin{aligned}
&\|\widehat{\mathbf{\Sigma}}_{\mathcal{A}\mathcal{A}}\mathbf{g}_h^\star - \widehat{\mathbf{u}}_{\mathcal{A},h}\|_\infty \le \|\widehat{\mathbf{\Sigma}}_{\mathcal{A}\mathcal{A}}\mathbf{g}_h^\star - \mathbf{\Sigma}_{\mathcal{A}\mathcal{A}}\mathbf{g}_h^\star + \mathbf{\Sigma}_{\mathcal{A}\mathcal{A}}\mathbf{g}_h^\star - \widehat{\mathbf{u}}_{\mathcal{A},h}\|_\infty \\
\le& \|(\widehat{\mathbf{\Sigma}}_{\mathcal{A}\mathcal{A}} - \mathbf{\Sigma}_{\mathcal{A}\mathcal{A}})\mathbf{g}_h^\star\|_\infty + \|\mathbf{u}_{\mathcal{A},h} - \widehat{\mathbf{u}}_{\mathcal{A},h}\|_\infty \le 2\sqrt{C_0 \log p/n},
\end{aligned} \tag{G.11}$$

where we use the fact that $\mathbf{g}_h^\star = \mathbf{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1}\mathbf{u}_{\mathcal{A},h}$ from (G.4). Since we assume that $\lambda_1 \le$

$3c_1\sqrt{\log p/n}$ and $\lambda_2 \le c_2\lambda_1$, combined with (G.10) and (G.11), we have

$$
\begin{aligned}
&\sum_{h=1}^{H}(\widetilde{\mathbf{g}}_h - \mathbf{g}_h^\star)^\top \widehat{\mathbf{\Sigma}}_{\mathcal{A}\mathcal{A}}(\widetilde{\mathbf{g}}_h - \mathbf{g}_h^\star) \\
\le& 4\sqrt{C_0 \log p/n}\sum_{h=1}^{H}\|\widetilde{\mathbf{g}}_h - \mathbf{g}_h^\star\|_1 + 6c_1\sqrt{\log p/n}\|\widetilde{\mathbf{G}} - \mathbf{G}^\star\|_{2,1} + 6c_1c_2\sqrt{\log p/n}\|\widetilde{\mathbf{G}} - \mathbf{G}^\star\|_* \\
\le& (4\sqrt{C_0} + 6c_1 + 6c_1c_2)\sqrt{\log p/n}\left(\sum_{h=1}^{H}\|\widetilde{\mathbf{g}}_h - \mathbf{g}_h^\star\|_1 + \|\widetilde{\mathbf{G}} - \mathbf{G}^\star\|_{2,1} + \|\widetilde{\mathbf{G}} - \mathbf{G}^\star\|_*\right). \quad \text{(G.12)}
\end{aligned}
$$

For $\sum_{h=1}^{H}\|\widetilde{\mathbf{g}}_h - \mathbf{g}_h^\star\|_1$, we have

$$
\sum_{h=1}^{H}\|\widetilde{\mathbf{g}}_h - \mathbf{g}_h^\star\|_1 \le \sqrt{sH}\|\widetilde{\mathbf{G}} - \mathbf{G}^\star\|_F \le \sqrt{sA}\|\widetilde{\mathbf{G}} - \mathbf{G}^\star\|_F. \quad \text{(G.13)}
$$

And for $\|\widetilde{\mathbf{G}} - \mathbf{G}^\star\|_{2,1}$ and $\|\widetilde{\mathbf{G}} - \mathbf{G}^\star\|_*$, since we assume that $H \le s \le p$, we have

$$
\|\widetilde{\mathbf{G}} - \mathbf{G}^\star\|_{2,1} \le \sqrt{s}\|\widetilde{\mathbf{G}} - \mathbf{G}^\star\|_F, \quad \|\widetilde{\mathbf{G}} - \mathbf{G}^\star\|_* \le \sqrt{s}\|\widetilde{\mathbf{G}} - \mathbf{G}^\star\|_F. \quad \text{(G.14)}
$$

Combining (G.12)–(G.14), we have

$$
\sum_{h=1}^{H}(\widetilde{\mathbf{g}}_h - \mathbf{g}_h^\star)^\top \widehat{\mathbf{\Sigma}}_{\mathcal{A}\mathcal{A}}(\widetilde{\mathbf{g}}_h - \mathbf{g}_h^\star) \le \{4\sqrt{2/C'} + 6c_1(1+c_2)\}(\sqrt{A}+2)\sqrt{s\log p/n}\|\widetilde{\mathbf{G}} - \mathbf{G}^\star\|_F, \quad \text{(G.15)}
$$

As $p > 1$, then $s\log(ep/s) \le s\log p^3 = 3s\log p$. Since we assume that $C_1 s\log p \le n$, by taking $C_1 \ge 3c$, we have $cs\log(ep/s) \le n$. Thus, according to Condition (C2), with probability at least $1 - C\exp(-C's)$, we have

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

Then, we have that

$$
\begin{aligned}
\sum_{h=1}^{H}(\widetilde{\mathbf{g}}_h - \mathbf{g}_h^\star)^\top \widehat{\mathbf{\Sigma}}_{\mathcal{A}\mathcal{A}}(\widetilde{\mathbf{g}}_h - \mathbf{g}_h^\star) &\ge \varphi_{\min}(\widehat{\mathbf{\Sigma}}_{\mathcal{A}\mathcal{A}})\sum_{h=1}^{H}\|\widetilde{\mathbf{g}}_h - \mathbf{g}_h^\star\|_2^2 \\
&\ge K^{-1}\|\widetilde{\mathbf{G}} - \mathbf{G}^\star\|_F^2. \quad \text{(G.16)}
\end{aligned}
$$

Combining (G.15) and (G.16), we obtain

$$
\|\widetilde{\mathbf{G}} - \mathbf{G}^\star\|_F \le \widetilde{C}\sqrt{s\log p/n}
$$

with probability at least $1 - C\exp\{-C'(s \wedge \log p)\}$, where the constants $\widetilde{C} = \{4\sqrt{2/C'} + 6c_1(1+c_2)\}(\sqrt{A}+2)K$. This completes the proof.

□

In the rest of the section, we verify that condition (G.5) in Lemma G.1 holds with high probability.

Denote

$$\boldsymbol{\Omega} = \widehat{\mathbf{U}}_{\mathcal{A}^c} - \boldsymbol{\Sigma}_{\mathcal{A}^c\mathcal{A}}\boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1}\widehat{\mathbf{U}}_{\mathcal{A}} \in \mathbb{R}^{(p-s)\times H}, \quad \boldsymbol{\Phi} = \mathbb{X}_{\mathcal{A}^c} - \boldsymbol{\Sigma}_{\mathcal{A}^c\mathcal{A}}\boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1}\mathbb{X}_{\mathcal{A}} \in \mathbb{R}^{(p-s)\times n}.$$

Then we have

$$\begin{aligned}
&\widehat{\mathbf{U}}_{\mathcal{A}^c} = \boldsymbol{\Sigma}_{\mathcal{A}^c\mathcal{A}}\boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1}\widehat{\mathbf{U}}_{\mathcal{A}} + \boldsymbol{\Omega},\\
&n\widehat{\boldsymbol{\Sigma}}_{\mathcal{A}^c\mathcal{A}} = \mathbb{X}_{\mathcal{A}^c}\mathbb{X}_{\mathcal{A}}^\top = (\boldsymbol{\Sigma}_{\mathcal{A}^c\mathcal{A}}\boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1}\mathbb{X}_{\mathcal{A}} + \boldsymbol{\Phi})\mathbb{X}_{\mathcal{A}}^\top = n\boldsymbol{\Sigma}_{\mathcal{A}^c\mathcal{A}}\boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1}\widehat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A}} + \boldsymbol{\Phi}\mathbb{X}_{\mathcal{A}}^\top.
\end{aligned}$$

Recall that the oracle estimator $\widetilde{\mathbf{G}}$ can be expressed as

$$\widetilde{\mathbf{G}} = \widehat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A}}^{-1}\widehat{\mathbf{U}}_{\mathcal{A}} - \lambda_1\widehat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A}}^{-1}\widetilde{\mathbf{R}} - \lambda_2\widehat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A}}^{-1}\widetilde{\mathbf{S}}, \tag{G.17}$$

where $\widetilde{\mathbf{R}} \in \partial\|\widetilde{\mathbf{G}}\|_{2,1}$ and $\widetilde{\mathbf{S}} \in \partial\|\widetilde{\mathbf{G}}\|_*$. Therefore,

$$\begin{aligned}
&\widehat{\boldsymbol{\Sigma}}_{\mathcal{A}^c\mathcal{A}}\widetilde{\mathbf{G}} - \widehat{\mathbf{U}}_{\mathcal{A}^c}\\
=&(\boldsymbol{\Sigma}_{\mathcal{A}^c\mathcal{A}}\boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1}\widehat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A}} + n^{-1}\boldsymbol{\Phi}\mathbb{X}_{\mathcal{A}}^\top)(\widehat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A}}^{-1}\widehat{\mathbf{U}}_{\mathcal{A}} - \lambda_1\widehat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A}}^{-1}\widetilde{\mathbf{R}} - \lambda_2\widehat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A}}^{-1}\widetilde{\mathbf{S}}) - \boldsymbol{\Sigma}_{\mathcal{A}^c\mathcal{A}}\boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1}\widehat{\mathbf{U}}_{\mathcal{A}} - \boldsymbol{\Omega}\\
=&\boldsymbol{\Sigma}_{\mathcal{A}^c\mathcal{A}}\boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1}\widehat{\mathbf{U}}_{\mathcal{A}} - \lambda_1\boldsymbol{\Sigma}_{\mathcal{A}^c\mathcal{A}}\boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1}\widetilde{\mathbf{R}} - \lambda_2\boldsymbol{\Sigma}_{\mathcal{A}^c\mathcal{A}}\boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1}\widetilde{\mathbf{S}} + n^{-1}\boldsymbol{\Phi}\mathbb{X}_{\mathcal{A}}^\top\widehat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A}}^{-1}(\widehat{\mathbf{U}}_{\mathcal{A}} - \lambda_1\widetilde{\mathbf{R}} - \lambda_2\widetilde{\mathbf{S}})\\
&- \boldsymbol{\Sigma}_{\mathcal{A}^c\mathcal{A}}\boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1}\widehat{\mathbf{U}}_{\mathcal{A}} - \boldsymbol{\Omega}\\
=&-\lambda_1\boldsymbol{\Sigma}_{\mathcal{A}^c\mathcal{A}}\boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1}\widetilde{\mathbf{R}} - \lambda_2\boldsymbol{\Sigma}_{\mathcal{A}^c\mathcal{A}}\boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1}\widetilde{\mathbf{S}} + n^{-1}\boldsymbol{\Phi}\mathbb{X}_{\mathcal{A}}^\top\widehat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A}}^{-1}\widehat{\mathbf{U}}_{\mathcal{A}} - \lambda_1 n^{-1}\boldsymbol{\Phi}\mathbb{X}_{\mathcal{A}}^\top\widehat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A}}^{-1}\widetilde{\mathbf{R}}\\
&- \lambda_2 n^{-1}\boldsymbol{\Phi}\mathbb{X}_{\mathcal{A}}^\top\widehat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A}}^{-1}\widetilde{\mathbf{S}} - \boldsymbol{\Omega}.
\end{aligned} \tag{G.18}$$

We aim to derive the following inequalities with high probability:

$$\lambda_1\|\boldsymbol{\Sigma}_{\mathcal{A}^c\mathcal{A}}\boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1}\widetilde{\mathbf{R}}\|_{2,\infty} \le \lambda_1(1-\alpha/2), \tag{G.19}$$

$$\lambda_2\|\boldsymbol{\Sigma}_{\mathcal{A}^c\mathcal{A}}\boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1}\widetilde{\mathbf{S}}\|_{2,\infty} \le \lambda_1\alpha/10, \tag{G.20}$$

$$\|n^{-1}\boldsymbol{\Phi}\mathbb{X}_{\mathcal{A}}^\top\widehat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A}}^{-1}\widehat{\mathbf{U}}_{\mathcal{A}}\|_{2,\infty} \le \lambda_1\alpha/10, \tag{G.21}$$

$$\lambda_1\|n^{-1}\boldsymbol{\Phi}\mathbb{X}_{\mathcal{A}}^\top\widehat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A}}^{-1}\widetilde{\mathbf{R}}\|_{2,\infty} \le \lambda_1\alpha/10, \tag{G.22}$$

$$\lambda_2\|n^{-1}\boldsymbol{\Phi}\mathbb{X}_{\mathcal{A}}^\top\widehat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A}}^{-1}\widetilde{\mathbf{S}}\|_{2,\infty} \le \lambda_1\alpha/10, \tag{G.23}$$

$$\|\boldsymbol{\Omega}\|_{2,\infty} \le \lambda_1\alpha/10, \tag{G.24}$$

where the constant α comes from the Assumption (A10). In Lemmas G.3–G.8, we prove that each inequality holds with high probability under the conditions in Theorem C.1. Collecting all the inequalities, by (G.18), we obtain

$$\|\widehat{\boldsymbol{\Sigma}}_{\mathcal{A}^c\mathcal{A}}\widetilde{\mathbf{G}} - \widehat{\mathbf{U}}_{\mathcal{A}^c}\|_{2,\infty} \le \lambda_1$$

with high probability.

Lemma G.3. *Under the same conditions in Theorem C.1, there exist some constants* $C, C' > 0$ *such that*

$$\lambda_1 \|\boldsymbol{\Sigma}_{\mathcal{A}^c\mathcal{A}} \boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1} \widetilde{\mathbf{R}}\|_{2,\infty} \le \lambda_1(1 - \alpha/2)$$

with probability at least $1 - C\exp\{-C'(s \wedge \log p)\}$.

Proof of Lemma G.3. First, we have

$$\lambda_1 \|\boldsymbol{\Sigma}_{\mathcal{A}^c\mathcal{A}} \boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1} \widetilde{\mathbf{R}}\|_{2,\infty} \le \lambda_1 \|\boldsymbol{\Sigma}_{\mathcal{A}^c\mathcal{A}} \boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1} (\widetilde{\mathbf{R}} - \mathbf{R})\|_{2,\infty} + \lambda_1 \|\boldsymbol{\Sigma}_{\mathcal{A}^c\mathcal{A}} \boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1} \mathbf{R}\|_{2,\infty}, \tag{G.25}$$

where $\mathbf{R} \in \partial \|\mathbf{G}^\star\|_{2,1}$. By Assumption (A10), for some constant $0 < \alpha \le 1$, it follows that

$$\lambda_1 \|\boldsymbol{\Sigma}_{\mathcal{A}^c\mathcal{A}} \boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1} \mathbf{R}\|_{2,\infty} \le \lambda_1(1 - \alpha). \tag{G.26}$$

Then, for all $j \in \mathcal{A}$ and $h = 1, \ldots, H$, we have

$$\begin{aligned}
|\widetilde{r}_{jh} - r_{jh}| = \left| \frac{\widetilde{g}_{jh}}{\|\widetilde{\mathbf{G}}_j\|_2} - \frac{g^\star_{jh}}{\|\mathbf{G}^\star_j\|_2} \right| &\le \frac{|\widetilde{g}_{jh} - g^\star_{jh}| \|\mathbf{G}^\star_j\|_2 + |g^\star_{jh}| \|\widetilde{\mathbf{G}}_j - \mathbf{G}^\star_j\|_2}{\|\widetilde{\mathbf{G}}_j\|_2 \|\mathbf{G}^\star_j\|_2} \\
&\le \frac{\|\widetilde{\mathbf{G}}_j - \mathbf{G}^\star_j\|_2 \|\mathbf{G}^\star_j\|_2 + \|\mathbf{G}^\star_j\|_2 \|\widetilde{\mathbf{G}}_j - \mathbf{G}^\star_j\|_2}{\|\widetilde{\mathbf{G}}_j\|_2 \|\mathbf{G}^\star_j\|_2} \le 2 \frac{\|\widetilde{\mathbf{G}}_j - \mathbf{G}^\star_j\|_2}{\|\widetilde{\mathbf{G}}_j\|_2} \\
&\le 2 \frac{\|\widetilde{\mathbf{G}}_j - \mathbf{G}^\star_j\|_2}{|\|\mathbf{G}^\star_j\|_2 - \|\widetilde{\mathbf{G}}_j - \mathbf{G}^\star_j\|_2|}.
\end{aligned}$$

By Lemma G.2, $\|\widetilde{\mathbf{G}}_j - \mathbf{G}^\star_j\|_2 \le \|\widetilde{\mathbf{G}} - \mathbf{G}^\star\|_F \le \widetilde{C}\sqrt{s\log p/n}$ with probability at least $1 - C\exp\{-C'(s \wedge \log p)\}$. And in Assumption (A2), $\min_{j\in\mathcal{A}} \|\mathbf{G}^\star_j\|_2 = \min_{j\in\mathcal{A}} \|\mathbf{B}^\star_j\|_2 \ge N(\sigma_1^2/\sigma_{d^\star}^2)\sqrt{s\log p/n} \ge N\sqrt{s\log p/n}$. We take the constant $N \ge \widetilde{C}(4\alpha^{-1}\sqrt{A}Q + 1)$, where $\widetilde{C} = \{4\sqrt{2/C'} + 6c_1(1 + c_2)\}(\sqrt{A} + 2)K$. Then we have

$$|\widetilde{r}_{jh} - r_{jh}| \le \alpha(2\sqrt{A}Q)^{-1}, \quad \forall j \in \mathcal{A},\ h = 1, \ldots, H.$$

Then, under Assumption (A9), we have

$$\begin{aligned}
|\boldsymbol{\Sigma}_{k\mathcal{A}}^\top \boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1} (\widetilde{\mathbf{r}}_h - \mathbf{r}_h)| \le \|\boldsymbol{\Sigma}_{k\mathcal{A}}^\top \boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1}\|_1 \|\widetilde{\mathbf{r}}_h - \mathbf{r}_h\|_\infty &\le \|\boldsymbol{\Sigma}_{\mathcal{A}^c\mathcal{A}} \boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1}\|_\infty \max_{j\in\mathcal{A}} |\widetilde{r}_{jh} - r_{jh}| \\
&\le \alpha(2\sqrt{A})^{-1}, \quad \forall k \in \mathcal{A}^c,\ h = 1, \ldots, H.
\end{aligned}$$

It follows that

$$\|\boldsymbol{\Sigma}_{k\mathcal{A}}^\top \boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1} (\widetilde{\mathbf{R}} - \mathbf{R})\|_2 \le \sqrt{H} \max_{h=1,\ldots,H} |\boldsymbol{\Sigma}_{k\mathcal{A}}^\top \boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1} (\widetilde{\mathbf{r}}_h - \mathbf{r}_h)| \le \alpha/2, \quad \forall k \in \mathcal{A}^c.$$

Then,

$$\lambda_1 \|\boldsymbol{\Sigma}_{\mathcal{A}^c\mathcal{A}} \boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1} (\widetilde{\mathbf{R}} - \mathbf{R})\|_{2,\infty} \le \lambda_1 \alpha/2. \tag{G.27}$$

Combining (G.25)–(G.27), we obtain

$$\lambda_1 \|\mathbf{\Sigma}_{\mathcal{A}^c\mathcal{A}}\mathbf{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1}\widetilde{\mathbf{R}}\|_{2,\infty} \le \lambda_1(1-\alpha/2)$$

with probability at least $1 - C\exp\{-C'(s \wedge \log p)\}$. $\square$

Lemma G.4. *Under the same conditions in Theorem C.1, we have*

$$\lambda_2 \|\mathbf{\Sigma}_{\mathcal{A}^c\mathcal{A}}\mathbf{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1}\widetilde{\mathbf{S}}\|_{2,\infty} \le \lambda_1\alpha/10.$$

Proof of Lemma G.4. Let $\widetilde{\mathbf{M}}\mathrm{diag}(\widetilde{\boldsymbol{\sigma}})\widetilde{\mathbf{N}}^\top$ denote the singular value decomposition of $\widetilde{\mathbf{G}}$, where $\widetilde{\mathbf{M}} \in \mathbb{R}^{s\times H}$, $\widetilde{\mathbf{N}} \in \mathbb{R}^{H\times H}$ and $\widetilde{\boldsymbol{\sigma}} = (\widetilde{\sigma}_1, \ldots, \widetilde{\sigma}_H) \in \mathbb{R}^H$ consists of singular values in decreasing order. By Proposition E.4, the subdifferential $\partial\|\widetilde{\mathbf{G}}\|_* = \widetilde{\mathbf{M}}\mathrm{diag}(\partial\|\widetilde{\boldsymbol{\sigma}}\|_1)\widetilde{\mathbf{N}}^\top$. Then for each element of $\widetilde{\mathbf{S}}$, we have $\widetilde{s}_{jh} = \widetilde{\mathbf{M}}_j^\top \mathrm{diag}(\partial\|\widetilde{\boldsymbol{\sigma}}\|_1)\widetilde{\mathbf{n}}_h$. Since each element in $\partial\|\widetilde{\boldsymbol{\sigma}}\|_1$, $\widetilde{\mathbf{M}}_j$ and $\widetilde{\mathbf{n}}_h$ is in $[-1, 1]$, then

$$|\widetilde{s}_{jh}| = |\widetilde{\mathbf{M}}_j^\top \mathrm{diag}(\partial\|\widetilde{\boldsymbol{\sigma}}\|_1)\widetilde{\mathbf{n}}_h| \le H, \quad \forall j \in \mathcal{A},\ h = 1, \ldots, H.$$

By Assumption (A9),

$$\begin{aligned}|\mathbf{\Sigma}_{k\mathcal{A}}^\top\mathbf{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1}\widetilde{\mathbf{s}}_h| &\le \|\mathbf{\Sigma}_{k\mathcal{A}}^\top\mathbf{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1}\|_1\|\widetilde{\mathbf{s}}_h\|_\infty \le \|\mathbf{\Sigma}_{\mathcal{A}^c\mathcal{A}}\mathbf{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1}\|_\infty \max_{j\in\mathcal{A}}|\widetilde{s}_{jh}| \\ &\le QH, \quad \forall k \in \mathcal{A}^c,\ h = 1, \ldots, H.\end{aligned}$$

It follows that

$$\|\mathbf{\Sigma}_{k\mathcal{A}}^\top\mathbf{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1}\widetilde{\mathbf{S}}\|_2 \le \sqrt{H}\max_{h=1,\ldots,H}|\mathbf{\Sigma}_{k\mathcal{A}}^\top\mathbf{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1}\widetilde{\mathbf{s}}_h| \le QH^{3/2} \le QA^{3/2}, \quad \forall k \in \mathcal{A}^c.$$

And then we have

$$\|\mathbf{\Sigma}_{\mathcal{A}^c\mathcal{A}}\mathbf{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1}\widetilde{\mathbf{S}}\|_{2,\infty} \le QA^{3/2}. \tag{G.28}$$

Since we assume that $\lambda_2 \le c_2\lambda_1$, by taking $c_2 \le \alpha(10QA^{3/2})^{-1}$, we have

$$\lambda_2 \|\mathbf{\Sigma}_{\mathcal{A}^c\mathcal{A}}\mathbf{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1}\widetilde{\mathbf{S}}\|_{2,\infty} \le \lambda_1\alpha/10.$$

$\square$

Lemma G.5. *Under the same conditions in Theorem C.1, there exist some constants* $C, C' > 0$ *such that*

$$\|n^{-1}\mathbf{\Phi}\mathbb{X}_{\mathcal{A}}^\top\widehat{\mathbf{\Sigma}}_{\mathcal{A}\mathcal{A}}^{-1}\widehat{\mathbf{U}}_{\mathcal{A}}\|_{2,\infty} \le \lambda_1\alpha/10$$

with probability at least $1 - C\exp\{-C'(s \wedge \log p)\}$.

Proof of Lemma G.5. Recall that $\mathbf{\Phi} = \mathbb{X}_{\mathcal{A}^c} - \mathbf{\Sigma}_{\mathcal{A}^c\mathcal{A}}\mathbf{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1}\mathbb{X}_{\mathcal{A}}$. We rewrite the ith column of $\mathbf{\Phi}$ as $\boldsymbol{\varphi}_i = \mathbf{X}_{i,\mathcal{A}^c} - \mathbf{\Sigma}_{\mathcal{A}^c\mathcal{A}}\mathbf{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1}\mathbf{X}_{i,\mathcal{A}} = (\mathbf{I}_{p-s}, -\mathbf{\Sigma}_{\mathcal{A}^c\mathcal{A}}\mathbf{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1})(\mathbf{X}_{i,\mathcal{A}^c}^\top, \mathbf{X}_{i,\mathcal{A}}^\top)^\top$. Then for $k \in \mathcal{A}^c$, $\mathbf{e}_k^\top\boldsymbol{\varphi}_i \sim N(0, \sigma_{kk\cdot\mathcal{A}})$, where $\mathbf{e}_k$ is the k-th standard basis vector in $\mathbb{R}^{p-s}$ and $\sigma_{kk\cdot\mathcal{A}} = \mathbf{\Sigma}_{kk} - \mathbf{\Sigma}_{k\mathcal{A}}\mathbf{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1}\mathbf{\Sigma}_{\mathcal{A}k}$. Since $\boldsymbol{\varphi}_i$'s are mutually independent, then

$$\mathbf{e}_k^\top\mathbf{\Phi} \sim N(\mathbf{0}, \sigma_{kk\cdot\mathcal{A}}\mathbf{I}_n), \quad \forall k \in \mathcal{A}^c.$$

Since $\boldsymbol{\varphi}_i \perp\!\!\!\perp \mathbf{X}_{j\mathcal{A}}$ for $i \neq j$, and

$$\mathrm{Cov}(\boldsymbol{\varphi}_i, \mathbf{X}_{i,\mathcal{A}}) = \mathbf{\Sigma}_{\mathcal{A}^c\mathcal{A}} - \mathbf{\Sigma}_{\mathcal{A}^c\mathcal{A}}\mathbf{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1}\mathbf{\Sigma}_{\mathcal{A}\mathcal{A}} = \mathbf{0}, \quad \forall i = 1, \ldots, n,$$

we have $\mathrm{Cov}\{\mathrm{vec}(\mathbf{\Phi}), \mathrm{vec}(\mathbb{X}_{\mathcal{A}})\} = \mathbf{0}$. In addition, since $\mathrm{vec}(\mathbf{\Phi})$ and $\mathrm{vec}(\mathbb{X}_{\mathcal{A}})$ are normal, then $\mathbf{\Phi} \perp\!\!\!\perp \mathbb{X}_{\mathcal{A}}$. Since $\mathcal{A} = \{j : \|\boldsymbol{\beta}_j^\star\|_2 > 0\}$, then $\mathcal{S}_{\boldsymbol{\beta}^\star} \subseteq \mathrm{span}\{\mathbf{e}_j \in \mathbb{R}^p, j \in \mathcal{A}\}$. And we have $\mathbf{X}_{i,\mathcal{A}^c} \perp\!\!\!\perp Y_i \mid \mathbf{X}_{i,\mathcal{A}}$, $i = 1, \ldots, n$. Given $\mathbb{X}_{\mathcal{A}}$, $\mathbf{\Phi}$ is the function of $\mathbb{X}_{\mathcal{A}^c}$. Thus, we have $\mathbf{\Phi} \perp\!\!\!\perp \mathbb{Y} \mid \mathbb{X}_{\mathcal{A}}$. Combining the results that $\mathbf{\Phi} \perp\!\!\!\perp \mathbb{X}_{\mathcal{A}}$ and $\mathbf{\Phi} \perp\!\!\!\perp \mathbb{Y} \mid \mathbb{X}_{\mathcal{A}}$, we have

$$\mathbf{\Phi} \perp\!\!\!\perp (\mathbb{Y}, \mathbb{X}_{\mathcal{A}}).$$

Since $\widehat{\mathbf{U}}_{\mathcal{A}}$ is a function of $(\mathbb{Y}, \mathbb{X}_{\mathcal{A}})$, then

$$\mathbf{\Phi} \perp\!\!\!\perp (\widehat{\mathbf{U}}_{\mathcal{A}}, \mathbb{X}_{\mathcal{A}}). \tag{G.29}$$

For all $k \in \mathcal{A}^c$, since $\mathbf{\Phi} \perp\!\!\!\perp (\widehat{\mathbf{U}}_{\mathcal{A}}, \mathbb{X}_{\mathcal{A}})$, then we have

$$n^{-1}\mathbf{e}_k^\top\mathbf{\Phi}\mathbb{X}_{\mathcal{A}}^\top\widehat{\mathbf{\Sigma}}_{\mathcal{A}\mathcal{A}}^{-1}\widehat{\mathbf{U}}_{\mathcal{A}} \mid (\widehat{\mathbf{U}}_{\mathcal{A}}, \mathbb{X}_{\mathcal{A}}) \sim N(\mathbf{0}, n^{-1}\sigma_{kk\cdot\mathcal{A}}\widehat{\mathbf{U}}_{\mathcal{A}}^\top\widehat{\mathbf{\Sigma}}_{\mathcal{A}\mathcal{A}}^{-1}\widehat{\mathbf{U}}_{\mathcal{A}}).$$

Let $\mathbf{W}_1 = \widehat{\mathbf{U}}_{\mathcal{A}}^\top\widehat{\mathbf{\Sigma}}_{\mathcal{A}\mathcal{A}}^{-1}\widehat{\mathbf{U}}_{\mathcal{A}}$, then

$$n^{-1/2}\sigma_{kk\cdot\mathcal{A}}^{-1/2}\mathbf{e}_k^\top\mathbf{\Phi}\mathbb{X}_{\mathcal{A}}^\top\widehat{\mathbf{\Sigma}}_{\mathcal{A}\mathcal{A}}^{-1}\widehat{\mathbf{U}}_{\mathcal{A}}\mathbf{W}_1^{-1/2} \mid (\widehat{\mathbf{U}}_{\mathcal{A}}, \mathbb{X}_{\mathcal{A}}) \sim N(\mathbf{0}, \mathbf{I}_H).$$

and

$$\|n^{-1/2}\sigma_{kk\cdot\mathcal{A}}^{-1/2}\mathbf{e}_k^\top\mathbf{\Phi}\mathbb{X}_{\mathcal{A}}^\top\widehat{\mathbf{\Sigma}}_{\mathcal{A}\mathcal{A}}^{-1}\widehat{\mathbf{U}}_{\mathcal{A}}\mathbf{W}_1^{-1/2}\|_2^2 \mid (\widehat{\mathbf{U}}_{\mathcal{A}}, \mathbb{X}_{\mathcal{A}}) \sim \chi_H^2$$

According to Proposition E.5, we have

$$\mathbb{P}\{\|n^{-1/2}\sigma_{kk\cdot\mathcal{A}}^{-1/2}\mathbf{e}_k^\top\mathbf{\Phi}\mathbb{X}_{\mathcal{A}}^\top\widehat{\mathbf{\Sigma}}_{\mathcal{A}\mathcal{A}}^{-1}\widehat{\mathbf{U}}_{\mathcal{A}}\mathbf{W}_1^{-1/2}\|_2^2 \geq H + 2\sqrt{H\epsilon} + 2\epsilon \mid (\widehat{\mathbf{U}}_{\mathcal{A}}, \mathbb{X}_{\mathcal{A}})\} \leq \exp(-\epsilon).$$

By taking the expectation over $(\widehat{\mathbf{U}}_{\mathcal{A}}, \mathbb{X}_{\mathcal{A}})$ both sides, we have

$$\mathbb{P}\{\|n^{-1/2}\sigma_{kk\cdot\mathcal{A}}^{-1/2}\mathbf{e}_k^\top\mathbf{\Phi}\mathbb{X}_{\mathcal{A}}^\top\widehat{\mathbf{\Sigma}}_{\mathcal{A}\mathcal{A}}^{-1}\widehat{\mathbf{U}}_{\mathcal{A}}\mathbf{W}_1^{-1/2}\|_2^2 \geq H + 2\sqrt{H\epsilon} + 2\epsilon\} \leq \exp(-\epsilon).$$

As $p > 1$, we take $\epsilon = c_0 \log p$ with constant $c_0 = (\log 2)^{-1}A > 1$ such that $\epsilon \geq H$, then

$$\mathbb{P}\{\|n^{-1}\mathbf{e}_k^\top\mathbf{\Phi}\mathbb{X}_{\mathcal{A}}^\top\widehat{\mathbf{\Sigma}}_{\mathcal{A}\mathcal{A}}^{-1}\widehat{\mathbf{U}}_{\mathcal{A}}\|_2^2 \geq 5\|\mathbf{W}_1\|\sigma_{kk\cdot\mathcal{A}}c_0 \log p/n\} \leq \exp(-c_0 \log p).$$

Under Assumption (A8), by Lemma E.3, we have $\sigma_{kk\cdot\mathcal{A}} \le m$ for all $k \in \mathcal{A}^c$. Then it follows that

$$\mathbb{P}\{\|n^{-1}\mathbf{e}_k^\top \mathbf{\Phi} \mathbb{X}_{\mathcal{A}}^\top \widehat{\mathbf{\Sigma}}_{\mathcal{A}\mathcal{A}}^{-1} \widehat{\mathbf{U}}_{\mathcal{A}}\|_2 \le \sqrt{5\|\mathbf{W}_1\| m c_0 \log p/n}\} \ge 1 - \exp(-c_0 \log p).$$

By union bound, we have

$$\begin{aligned} &\mathbb{P}\{\|n^{-1}\mathbf{\Phi} \mathbb{X}_{\mathcal{A}}^\top \widehat{\mathbf{\Sigma}}_{\mathcal{A}\mathcal{A}}^{-1} \widehat{\mathbf{U}}_{\mathcal{A}}\|_{2,\infty} \le \sqrt{5\|\mathbf{W}_1\| m c_0 \log p/n}\} \\ \ge& 1 - (p-s)\exp(-c_0 \log p) \ge 1 - \exp\{-(c_0 - 1)\log p\} = 1 - p^{-C} \end{aligned} \tag{G.30}$$

for some constant $C > 0$.

Next, we bound $\|\mathbf{W}_1\|$. By Assumption (A8), $m^{-1} \le \varphi_{\min}(\mathbf{\Sigma}) \le \varphi_{\max}(\mathbf{\Sigma}) \le m$ for some positive constant $m > 1$. By Lemma E.7, $\mathbf{X}_i$ is also a sub-Gaussian random vector with parameter $\|\mathbf{\Sigma}\| \le m$. As $p > 1$, then $s\log(ep/s) \le s \log p^3 = 3s\log p$. Since we assume that $C_1 s \log p \le n$, by taking $C_1 = 3c$, we have $cs\log(ep/s) \le n$. Then by Condition (C2), with probability at least $1 - C\exp(-C's)$, we have

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

According to Condition (C1),

$$\mathbb{P}(\|\widehat{\mathbf{U}} - \mathbf{U}\|_{\max} \le \epsilon) \ge 1 - Cp\exp(-C'n\epsilon^2), \quad 0 < \epsilon \le C'.$$

We select $\epsilon = \sqrt{C_0 \log p/n}$ with the constant $C_0 = 2/C'$ such that $C_0 C' > 1$. Since we assume that $C_1 s\log p \le n$, by taking $C_1 \ge 2/(C')^3$, we have $\epsilon \le \sqrt{C_0/C_1} = C'$. Then we have

$$\|\widehat{\mathbf{U}}_{\mathcal{A}} - \mathbf{U}_{\mathcal{A}}\| \le \sqrt{sH}\|\widehat{\mathbf{U}}_{\mathcal{A}} - \mathbf{U}_{\mathcal{A}}\|_{\max} \le \sqrt{2A/C'}\sqrt{s\log p/n} \le C'\sqrt{A} \tag{G.31}$$

with probability at least $1 - Cp^{-1}$. In addition, since we assume that $\sigma_1 \le T$ for some constant $T > 0$ in Assumption (A4), and that $\|\mathbf{\Sigma}_{\mathcal{A}\mathcal{A}}\| \le \|\mathbf{\Sigma}\| \le m$, then we have

$$\|\mathbf{U}_{\mathcal{A}}\| \le \|\mathbf{\Sigma}_{\mathcal{A}\mathcal{A}}\mathbf{G}^\star\| \le \|\mathbf{\Sigma}_{\mathcal{A}\mathcal{A}}\|\|\mathbf{G}^\star\| \le mT. \tag{G.32}$$

Combining (G.31) and (G.32), we have

$$\|\widehat{\mathbf{U}}_{\mathcal{A}}\| \le \|\widehat{\mathbf{U}}_{\mathcal{A}} - \mathbf{U}_{\mathcal{A}}\| + \|\mathbf{U}_{\mathcal{A}}\| \le C'\sqrt{A} + mT$$

with probability at least $1 - Cp^{-1}$. Therefore,

$$\|\mathbf{W}_1\| \le \|\widehat{\mathbf{U}}_{\mathcal{A}}\|^2 \|\widehat{\mathbf{\Sigma}}_{\mathcal{A}\mathcal{A}}^{-1}\| \le K(C'\sqrt{A} + mT)^2 \tag{G.33}$$

with probability at least $1 - C\exp\{-C'(s \wedge \log p)\}$.

Combining (G.30) and (G.33), we have

$$
\begin{aligned}
&\mathbb{P}\{\|n^{-1}\mathbf{\Phi}\mathbb{X}_{\mathcal{A}}^{\top}\widehat{\mathbf{\Sigma}}_{\mathcal{A}\mathcal{A}}^{-1}\widehat{\mathbf{U}}_{\mathcal{A}}\|_{2,\infty} \le \sqrt{5\|\mathbf{W}_1\|mc_0\log p/n},\ \|\mathbf{W}_1\| \le K(C'\sqrt{A}+mT)^2\} \\
\ge& 1 - C\exp\{-C'(s\wedge\log p)\}.
\end{aligned}
$$

Hence,

$$
\mathbb{P}\{\|n^{-1}\mathbf{\Phi}\mathbb{X}_{\mathcal{A}}^{\top}\widehat{\mathbf{\Sigma}}_{\mathcal{A}\mathcal{A}}^{-1}\widehat{\mathbf{U}}_{\mathcal{A}}\|_{2,\infty} \le C_3\sqrt{\log p/n}\} \ge 1 - C\exp\{-C'(s\wedge\log p)\},
$$

where $C_3 = \sqrt{5KAm(\log 2)^{-1}}(C'\sqrt{A}+mT)$. Since we assume that $\lambda_1 \ge 2c_1\sqrt{\log p/n}$, by taking $c_1 \ge 5C_3/\alpha$, we have

$$
\|n^{-1}\mathbf{\Phi}\mathbb{X}_{\mathcal{A}}^{\top}\widehat{\mathbf{\Sigma}}_{\mathcal{A}\mathcal{A}}^{-1}\widehat{\mathbf{U}}_{\mathcal{A}}\|_{2,\infty} \le C_3\frac{\lambda_1}{2c_1} \le \lambda_1\alpha/10
$$

with probability at least $1 - C\exp\{-C'(s\wedge\log p)\}$. □

Lemma G.6. *Under the same conditions in Theorem C.1, there exist some constants $C, C' > 0$ such that*

$$
\lambda_1\|n^{-1}\mathbf{\Phi}\mathbb{X}_{\mathcal{A}}^{\top}\widehat{\mathbf{\Sigma}}_{\mathcal{A}\mathcal{A}}^{-1}\widetilde{\mathbf{R}}\|_{2,\infty} \le \lambda_1\alpha/10
$$

with probability at least $1 - C\exp\{-C'(s\wedge\log p)\}$.

Proof of Lemma G.6. Recall that $\widetilde{\mathbf{R}} \in \partial\|\widetilde{\mathbf{G}}\|_{2,1}$ and $\mathbf{\Phi} = \mathbb{X}_{\mathcal{A}^c} - \mathbf{\Sigma}_{\mathcal{A}^c\mathcal{A}}\mathbf{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1}\mathbb{X}_{\mathcal{A}}$. Since $(\widetilde{\mathbf{R}}, \mathbb{X}_{\mathcal{A}})$ is the function of $(\widehat{\mathbf{U}}_{\mathcal{A}}, \mathbb{X}_{\mathcal{A}})$, by (G.29) that $\mathbf{\Phi} \perp\!\!\!\perp (\widehat{\mathbf{U}}_{\mathcal{A}}, \mathbb{X}_{\mathcal{A}})$, we have

$$
\mathbf{\Phi} \perp\!\!\!\perp (\widetilde{\mathbf{R}}, \mathbb{X}_{\mathcal{A}}).
$$

Then for all $k \in \mathcal{A}^c$, we have

$$
n^{-1}\mathbf{e}_k^{\top}\mathbf{\Phi}\mathbb{X}_{\mathcal{A}}^{\top}\widehat{\mathbf{\Sigma}}_{\mathcal{A}\mathcal{A}}^{-1}\widetilde{\mathbf{R}} \mid (\widetilde{\mathbf{R}}, \mathbb{X}_{\mathcal{A}}) \sim N(\mathbf{0}, n^{-1}\sigma_{kk\cdot\mathcal{A}}\widetilde{\mathbf{R}}^{\top}\widehat{\mathbf{\Sigma}}_{\mathcal{A}\mathcal{A}}^{-1}\widetilde{\mathbf{R}}),
$$

where $\mathbf{e}_k$ is the k-th standard basis vector in $\mathbb{R}^{p-s}$ and $\sigma_{kk\cdot\mathcal{A}} = \mathbf{\Sigma}_{kk} - \mathbf{\Sigma}_{k\mathcal{A}}\mathbf{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1}\mathbf{\Sigma}_{\mathcal{A}k}$. Let $\mathbf{W}_2 = \widetilde{\mathbf{R}}^{\top}\widehat{\mathbf{\Sigma}}_{\mathcal{A}\mathcal{A}}^{-1}\widetilde{\mathbf{R}}$, then similar to the proof of Lemma G.5, we obtain

$$
\mathbb{P}\{\|n^{-1/2}\sigma_{kk\cdot\mathcal{A}}^{-1/2}\mathbf{e}_k^{\top}\mathbf{\Phi}\mathbb{X}_{\mathcal{A}}^{\top}\widehat{\mathbf{\Sigma}}_{\mathcal{A}\mathcal{A}}^{-1}\widetilde{\mathbf{R}}\mathbf{W}_2^{-1/2}\|_2^2 \ge H + 2\sqrt{H\epsilon} + 2\epsilon\} \le \exp(-\epsilon).
$$

As $p > 1$, we take $\epsilon = c_0\log p$ with constant $c_0 = (\log 2)^{-1}A > 1$ such that $\epsilon \ge H$, then

$$
\mathbb{P}\{\|n^{-1}\mathbf{e}_k^{\top}\mathbf{\Phi}\mathbb{X}_{\mathcal{A}}^{\top}\widehat{\mathbf{\Sigma}}_{\mathcal{A}\mathcal{A}}^{-1}\widetilde{\mathbf{R}}\|_2^2 \ge 5\|\mathbf{W}_2\|\sigma_{kk\cdot\mathcal{A}}c_0\log p/n\} \le \exp(-c_0\log p).
$$

Under Assumption (A8), by Lemma E.3, we have $\sigma_{kk\cdot\mathcal{A}} \le m$ for all $k \in \mathcal{A}^c$. Then it follows that

$$
\mathbb{P}\{\|n^{-1}\mathbf{e}_k^{\top}\mathbf{\Phi}\mathbb{X}_{\mathcal{A}}^{\top}\widehat{\mathbf{\Sigma}}_{\mathcal{A}\mathcal{A}}^{-1}\widetilde{\mathbf{R}}\|_2 \le \sqrt{5\|\mathbf{W}_2\|mc_0\log p/n}\} \ge 1 - \exp(-c_0\log p).
$$

By union bound, we have

$$\begin{aligned}&\mathbb{P}\{\|n^{-1}\boldsymbol{\Phi}\mathbb{X}_{\mathcal{A}}^{\top}\widehat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A}}^{-1}\widetilde{\mathbf{R}}\|_{2,\infty}\leq\sqrt{5\|\mathbf{W}_2\|mc_0\log p/n}\}\\ \geq&1-(p-s)\exp(-c_0\log p)\geq 1-\exp\{-(c_0-1)\log p\}=1-p^{-C}\end{aligned} \tag{G.34}$$

for some constant $C>0$.

Since $\varphi_{\min}(\widehat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A}})\geq K^{-1}$ with probability at least $1-C\exp\{-C'k\}$ by Condition (C2) and $\|\widetilde{\mathbf{R}}\|\leq\|\widetilde{\mathbf{R}}\|_F\leq\sqrt{s}$, then

$$\|\mathbf{W}_2\|\leq\|\widetilde{\mathbf{R}}\|^2\|\widehat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A}}^{-1}\|\leq sK \tag{G.35}$$

with probability at least $1-C\exp\{-C'k\}$.

Combining (G.34) and (G.35), we have

$$\begin{aligned}&\mathbb{P}\{\|n^{-1}\boldsymbol{\Phi}\mathbb{X}_{\mathcal{A}}^{\top}\widehat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A}}^{-1}\widetilde{\mathbf{R}}\|_{2,\infty}\leq\sqrt{5\|\mathbf{W}_2\|mc_0\log p/n},\ \|\mathbf{W}_2\|\leq sK\}\\ \geq&1-C\exp\{-C'(s\wedge\log p)\}.\end{aligned}$$

Thus,

$$\mathbb{P}\{\|n^{-1}\boldsymbol{\Phi}\mathbb{X}_{\mathcal{A}}^{\top}\widehat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A}}^{-1}\widetilde{\mathbf{R}}\|_{2,\infty}\leq C_4\sqrt{s\log p/n}\}\geq 1-C\exp\{-C'(s\wedge\log p)\},$$

where $C_4=\sqrt{5KAm(\log 2)^{-1}}$. By taking $C_1\geq 100C_4^2\alpha^{-2}$, we have

$$\lambda_1\|n^{-1}\boldsymbol{\Phi}\mathbb{X}_{\mathcal{A}}^{\top}\widehat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A}}^{-1}\widetilde{\mathbf{R}}\|_{2,\infty}\leq\lambda_1C_4/\sqrt{C_1}\leq\lambda_1\alpha/10$$

with probability at least $1-C\exp\{-C'(s\wedge\log p)\}$. □

Lemma G.7. *Under the same conditions in Theorem C.1, there exist some constants $C,C'>0$ such that*

$$\lambda_2\|n^{-1}\boldsymbol{\Phi}\mathbb{X}_{\mathcal{A}}^{\top}\widehat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A}}^{-1}\widetilde{\mathbf{S}}\|_{2,\infty}\leq\lambda_1\alpha/10$$

with probability at least $1-C\exp\{-C'(s\wedge\log p)\}$.

Proof of Lemma G.7. Recall that $\widetilde{\mathbf{S}}\in\partial\|\widetilde{\mathbf{G}}\|_*$ and $\boldsymbol{\Phi}=\mathbb{X}_{\mathcal{A}^c}-\boldsymbol{\Sigma}_{\mathcal{A}^c\mathcal{A}}\boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1}\mathbb{X}_{\mathcal{A}}$. Since $(\widetilde{\mathbf{S}},\mathbb{X}_{\mathcal{A}})$ is the function of $(\widehat{\mathbf{U}}_{\mathcal{A}},\mathbb{X}_{\mathcal{A}})$, by (G.29) that $\boldsymbol{\Phi}\perp\!\!\!\perp(\widehat{\mathbf{U}}_{\mathcal{A}},\mathbb{X}_{\mathcal{A}})$, we have

$$\boldsymbol{\Phi}\perp\!\!\!\perp(\widetilde{\mathbf{S}},\mathbb{X}_{\mathcal{A}}).$$

Since we select $\lambda_2\leq c_2\lambda_1$ for some constant $c_2>0$, it gives us

$$\lambda_2\|n^{-1}\boldsymbol{\Phi}\mathbb{X}_{\mathcal{A}}^{\top}\widehat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A}}^{-1}\widetilde{\mathbf{S}}\|_{2,\infty}\leq c_2\lambda_1\|n^{-1}\boldsymbol{\Phi}\mathbb{X}_{\mathcal{A}}^{\top}\widehat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A}}^{-1}\widetilde{\mathbf{S}}\|_{2,\infty}.$$

We have

$$n^{-1}\mathbf{e}_k^{\top}\boldsymbol{\Phi}\mathbb{X}_{\mathcal{A}}^{\top}\widehat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A}}^{-1}\widetilde{\mathbf{S}}\mid(\widetilde{\mathbf{S}},\mathbb{X}_{\mathcal{A}})\sim N(\mathbf{0},n^{-1}\sigma_{kk\cdot\mathcal{A}}\widetilde{\mathbf{S}}^{\top}\widehat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A}}^{-1}\widetilde{\mathbf{S}}),$$

where $\mathbf{e}_k$ is the k-th standard basis vector in $\mathbb{R}^{p-s}$ and $\sigma_{kk\cdot\mathcal{A}} = \boldsymbol{\Sigma}_{kk} - \boldsymbol{\Sigma}_{k\mathcal{A}}\boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1}\boldsymbol{\Sigma}_{\mathcal{A}k}$. Let $\mathbf{W}_3 = \widetilde{\mathbf{S}}^\top\widehat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A}}^{-1}\widetilde{\mathbf{S}}$, then similar to the proof of Lemma G.5, we obtain

$$\mathbb{P}\{\|n^{-1/2}\sigma_{kk\cdot\mathcal{A}}^{-1/2}\mathbf{e}_k^\top\boldsymbol{\Phi}\mathbb{X}_{\mathcal{A}}^\top\widehat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A}}^{-1}\widetilde{\mathbf{S}}\mathbf{W}_3^{-1/2}\|_2^2 \geq H + 2\sqrt{H\epsilon} + 2\epsilon\} \leq \exp(-\epsilon).$$

As $p > 1$, we take $\epsilon = c_0 \log p$ with constant $c_0 = (\log 2)^{-1}A$ such that $\epsilon \geq H$, then

$$\mathbb{P}\{\|n^{-1}\mathbf{e}_k^\top\boldsymbol{\Phi}\mathbb{X}_{\mathcal{A}}^\top\widehat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A}}^{-1}\widetilde{\mathbf{S}}\|_2^2 \geq 5\|\mathbf{W}_3\|\sigma_{kk\cdot\mathcal{A}}c_0\log p/n\} \leq \exp(-c_0\log p).$$

Under Assumption (A8), by Lemma E.3, we have $\sigma_{kk\cdot\mathcal{A}} \leq m$ for all $k \in \mathcal{A}^c$. Then it follows that

$$\mathbb{P}\{\|n^{-1}\mathbf{e}_k^\top\boldsymbol{\Phi}\mathbb{X}_{\mathcal{A}}^\top\widehat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A}}^{-1}\widetilde{\mathbf{S}}\|_2 \leq \sqrt{5\|\mathbf{W}_3\|mc_0\log p/n}\} \geq 1 - \exp(-c_0\log p).$$

By union bound, we have

$$\begin{aligned}&\mathbb{P}\{\|n^{-1}\boldsymbol{\Phi}\mathbb{X}_{\mathcal{A}}^\top\widehat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A}}^{-1}\widetilde{\mathbf{S}}\|_{2,\infty} \leq \sqrt{5\|\mathbf{W}_3\|mc_0\log p/n}\}\\ \geq& 1 - (p-s)\exp(-c_0\log p) \geq 1 - \exp\{-(c_0-1)\log p\} = 1 - p^{-C}\end{aligned} \tag{G.36}$$

for some constant $C > 0$.

Let $\widetilde{\mathbf{M}}\text{diag}(\widetilde{\boldsymbol{\sigma}})\widetilde{\mathbf{N}}^\top$ be the singular value decomposition of $\widetilde{\mathbf{G}}$. Then by Proposition E.4, we have $\widetilde{\mathbf{S}} \in \partial\|\widetilde{\mathbf{G}}\|_\star = \widetilde{\mathbf{M}}\text{diag}(\partial\|\widetilde{\boldsymbol{\sigma}}\|_1)\widetilde{\mathbf{N}}^\top$. Since each element in $\partial\|\widetilde{\boldsymbol{\sigma}}\|_1$ is bounded in $[-1,1]$, we have $\|\widetilde{\mathbf{S}}\| \leq 1$. And by Condition (C2), $\varphi_{\min}(\widehat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A}}) \geq K^{-1}$ with probability at least $1 - C\exp\{-C'k\}$. Thus,

$$\|\mathbf{W}_3\| \leq \|\widetilde{\mathbf{S}}\|^2\|\widehat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A}}^{-1}\| \leq K \tag{G.37}$$

with probability at least $1 - C\exp\{-C'k\}$.

Combining (G.36) and (G.37), we obtain

$$\begin{aligned}&\mathbb{P}\{\|n^{-1}\boldsymbol{\Phi}\mathbb{X}_{\mathcal{A}}^\top\widehat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A}}^{-1}\widetilde{\mathbf{S}}\|_{2,\infty} \leq \sqrt{5\|\mathbf{W}_3\|mc_0\log p/n},\ \|\mathbf{W}_3\| \leq K\}\\ \geq& 1 - C\exp\{-C'(s\wedge\log p)\}.\end{aligned}$$

Thus,

$$\mathbb{P}\{\|n^{-1}\boldsymbol{\Phi}\mathbb{X}_{\mathcal{A}}^\top\widehat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A}}^{-1}\widetilde{\mathbf{S}}\|_{2,\infty} \leq C_5\sqrt{\log p/n}\} \geq 1 - C\exp\{-C'(s\wedge\log p)\},$$

where $C_5 = \sqrt{5KAm(\log 2)^{-1}}$. Since we assume that $s\log p \leq n$, by taking $C_1 \geq 100C_5^2c_2^2\alpha^{-2}$, we have

$$\lambda_2\|n^{-1}\boldsymbol{\Phi}\mathbb{X}_{\mathcal{A}}^\top\widehat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A}}^{-1}\widetilde{\mathbf{S}}\|_{2,\infty} \leq c_2\lambda_1\|n^{-1}\boldsymbol{\Phi}\mathbb{X}_{\mathcal{A}}^\top\widehat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A}}^{-1}\widetilde{\mathbf{S}}\|_{2,\infty} \leq \lambda_1\alpha/10$$

with probability at least $1 - C\exp\{-C'(s\wedge\log p)\}$. □

Lemma G.8. *Under the same conditions in Theorem C.1, there exist some constants $C, C' > 0$ such that*

$$\|\boldsymbol{\Omega}\|_{2,\infty} \leq \lambda_1\alpha/10$$

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

Proof of Lemma G.8. Recall that $\mathbf{\Omega} = \widehat{\mathbf{U}}_{\mathcal{A}^c} - \mathbf{\Sigma}_{\mathcal{A}^c\mathcal{A}}\mathbf{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1}\widehat{\mathbf{U}}_{\mathcal{A}}$. Since $\mathbf{\Sigma}\mathbf{B}^\star = \mathbf{U}$ and $\mathbf{B}^\star_{\mathcal{A}^c} = \mathbf{0}$, then $\mathbf{\Sigma}_{\mathcal{A}\mathcal{A}}\mathbf{B}^\star_{\mathcal{A}} = \mathbf{U}_{\mathcal{A}}$ and $\mathbf{\Sigma}_{\mathcal{A}^c\mathcal{A}}\mathbf{B}^\star_{\mathcal{A}} = \mathbf{U}_{\mathcal{A}^c}$, which gives us

$$\mathbf{U}_{\mathcal{A}^c} - \mathbf{\Sigma}_{\mathcal{A}^c\mathcal{A}}\mathbf{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1}\mathbf{U}_{\mathcal{A}} = \mathbf{0}.$$

Hence,

$$\mathbf{\Omega} = (\widehat{\mathbf{U}}_{\mathcal{A}^c} - \mathbf{U}_{\mathcal{A}^c}) - \mathbf{\Sigma}_{\mathcal{A}^c\mathcal{A}}\mathbf{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1}(\widehat{\mathbf{U}}_{\mathcal{A}} - \mathbf{U}_{\mathcal{A}}).$$

For each element of $\mathbf{\Omega}$, we have

$$|\omega_{kh}| \le \|\widehat{\mathbf{U}} - \mathbf{U}\|_{\max} + \|\mathbf{\Sigma}_{k\mathcal{A}}\mathbf{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1}\|_1 \|\widehat{\mathbf{u}}_{\mathcal{A},h} - \mathbf{u}_{\mathcal{A},h}\|_\infty \le (1 + \|\mathbf{\Sigma}_{\mathcal{A}^c\mathcal{A}}\mathbf{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1}\|_\infty)\|\widehat{\mathbf{U}} - \mathbf{U}\|_{\max}.$$

By Condition (C1) and Assumption (A9), for $k \in \mathcal{A}^c$ and $h = 1, \dots, H$, we obtain

$$\mathbb{P}\{|\omega_{kh}| \le (1+Q)\epsilon\} \ge \mathbb{P}(\|\widehat{\mathbf{U}} - \mathbf{U}\|_{\max} \le \epsilon) \ge 1 - Cp\exp(-C'n\epsilon^2), \quad 0 < \epsilon \le C'.$$

Hence, for $k \in \mathcal{A}^c$,

$$\begin{aligned}&\mathbb{P}\left\{(\sum_{h=1}^H \omega_{kh}^2)^{1/2} \le \sqrt{H}(1+Q)\epsilon\right\} \ge \mathbb{P}\{|\omega_{kh}| \le (1+Q)\epsilon,\ h = 1, \dots, H\}\\ \ge& 1 - CHp\exp(-C'n\epsilon^2), \quad 0 < \epsilon \le C'.\end{aligned}$$

Since we assume that $H \le A$ for some constant $A > 0$, it follows that

$$\begin{aligned}&\mathbb{P}\left\{(\sum_{h=1}^H \omega_{kh}^2)^{1/2} \le \sqrt{A}(1+Q)\epsilon\right\} \ge \mathbb{P}\left\{(\sum_{h=1}^H \omega_{kh}^2)^{1/2} \le \sqrt{H}(1+Q)\epsilon\right\}\\ \ge& 1 - CAp\exp(-C'n\epsilon^2), \quad 0 < \epsilon \le C'.\end{aligned}$$

Then we have

$$\mathbb{P}\{\|\mathbf{\Omega}\|_{2,\infty} \ge \sqrt{A}(1+Q)\epsilon\} \le CAp(p-s)\exp(-C'n\epsilon^2), \quad 0 < \epsilon \le C'. \tag{G.38}$$

We select $\epsilon = \sqrt{C_0 \log p/n}$ with the constant $C_0 = 3/C'$. We also take $C_1 \ge 3/(C')^3$. Since we assume that $C_1 s \log p \le n$, then $\log p/n \le 1/C_1$ and $\epsilon \le \sqrt{C_0/C_1} \le C'$. Meanwhile, since $C_0C' > 2$, by (G.38), with probability at least $1 - Cp^{-C'}$ for some constants $C, C' > 0$, we have

$$\|\mathbf{\Omega}\|_{2,\infty} \le (1+Q)\sqrt{3A/C'}\sqrt{\log p/n}$$

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

Since we select $\lambda_1 \ge 2c_1\sqrt{\log p/n}$, by taking $c_1 \ge \{5(1+Q)/\alpha\}\sqrt{3A/C'}$, we have

$$\|\mathbf{\Omega}\|_{2,\infty} \le (1+Q)\sqrt{\frac{3A}{C'}}\frac{\lambda_1}{2c_1} \le \lambda_1\alpha/10$$

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

Now we are ready for the proof of Theorem C.1.

Proof of Theorem C.1. By Lemmas G.3–G.8, we have that $\|\widehat{\boldsymbol{\Sigma}}_{\mathcal{A}^c\mathcal{A}}\widetilde{\mathbf{G}} - \widehat{\mathbf{U}}_{\mathcal{A}^c}\|_{2,\infty} \le \lambda_1$ with probability at least $1 - C\exp\{-C'(s \wedge \log p)\}$. Then by Lemma G.1, the solution to (G.1) is $\widehat{\mathbf{B}} = (\widetilde{\mathbf{G}}^\top, \mathbf{0}^\top_{(p-s)\times H})^\top$. Let $\widehat{\mathcal{B}} = \{j : \|\widehat{\mathbf{B}}_j\|_2 > 0\}$, then $\widehat{\mathcal{B}} \subseteq \mathcal{A}$. Let $\widehat{\mathbf{M}}\widehat{\mathbf{D}}\widehat{\mathbf{N}}^\top$ denote the singular value decomposition of $\widehat{\mathbf{B}}$, where $\widehat{\mathbf{D}} = \mathrm{diag}(\widehat{\sigma}_1, \ldots, \widehat{\sigma}_H)$. Recall that $\widehat{\boldsymbol{\beta}}$ is composed of the first $\widehat{d}$ columns of $\widehat{\mathbf{M}}$. Then

$$\|\widehat{\mathbf{B}}_j\|_2 = \|\widehat{\mathbf{M}}_j^\top \widehat{\mathbf{D}}\widehat{\mathbf{N}}^\top\|_2 = (\widehat{\mathbf{M}}_j^\top \widehat{\mathbf{D}}^2 \widehat{\mathbf{M}}_j)^{1/2} \ge \widehat{\sigma}_{\widehat{d}}\|\widehat{\boldsymbol{\beta}}_j\|_2.$$

By definition, $\widehat{d} = \#\{i \mid \widehat{\sigma}_i \ge \delta\}$. Then, $\widehat{\sigma}_{\widehat{d}} \ge \delta$. Thus, we have

$$\delta^{-1}\|\widehat{\mathbf{B}}_j\|_2 \ge \|\widehat{\boldsymbol{\beta}}_j\|_2. \tag{G.39}$$

Since $\widehat{\mathcal{A}} = \{j : \|\widehat{\boldsymbol{\beta}}_j\|_2 > 0\}$, (G.39) indicates that $\widehat{\mathcal{A}} \subseteq \widehat{\mathcal{B}} \subseteq \mathcal{A}$. Moreover, according to Corollary 2, $\mathcal{A} \subseteq \widehat{\mathcal{A}}$. Hence, we have that $\widehat{\mathcal{A}} = \mathcal{A}$ with probability at least $1 - C\exp\{-C'(s \wedge \log p)\}$. □

H Proof of Lemmas and Corollaries

H.1 Proof of Lemma 1

First, we show that the definition of $\mathcal{A}$ does not rely on the choice of basis matrix of central subspace $\mathcal{S}_{Y|\mathbf{X}}$. Let $\boldsymbol{\beta}$ and $\widetilde{\boldsymbol{\beta}}$ be two basis matrices of the central subspace $\mathcal{S}_{Y|\mathbf{X}}$, then $\mathcal{S}_{Y|\mathbf{X}} = \mathcal{S}_{\boldsymbol{\beta}} = \mathcal{S}_{\widetilde{\boldsymbol{\beta}}}$. There exist some non-singular matrix $\mathbf{R} \in \mathbb{R}^{d\times d}$ such that $\widetilde{\boldsymbol{\beta}} = \boldsymbol{\beta}\mathbf{R}$. For each j, $\|\widetilde{\boldsymbol{\beta}}_j\|_2 = \|\boldsymbol{\beta}_j^\top \mathbf{R}\|_2$, which gives us

$$\widetilde{\boldsymbol{\beta}}_j^\top \widetilde{\boldsymbol{\beta}}_j = \boldsymbol{\beta}_j^\top \mathbf{R}\mathbf{R}^\top \boldsymbol{\beta}_j.$$

Since $\mathbf{R}\mathbf{R}^\top$ is positive definite, then $\|\widetilde{\boldsymbol{\beta}}_j\|_2 > 0$ if and only if $\|\boldsymbol{\beta}_j\|_2 > 0$. Hence, the definition of $\mathcal{A}$ is independent of the choice of basis matrix of central subspace.

Let $s = |\mathcal{V}|$ and $\boldsymbol{\alpha}^\star \in \mathbb{R}^{p\times s}$ be a basis matrix of $\mathcal{S}^V_{Y|\mathbf{X}}$, where $\boldsymbol{\alpha}^\star_{\mathcal{V}} = \mathbf{I}_s$ and $\boldsymbol{\alpha}^\star_{\mathcal{V}^c} = \mathbf{0}$. Then $\mathcal{S}^V_{Y|\mathbf{X}} = \mathcal{S}_{\boldsymbol{\alpha}^\star}$. For a central subspace basis $\boldsymbol{\beta}^\star$, we split it into two sub-matrices $\boldsymbol{\beta}^\star_1 \in \mathbb{R}^{s\times d}$ and $\boldsymbol{\beta}^\star_2 \in \mathbb{R}^{(p-s)\times d}$ based on $\mathcal{V}$, where $\boldsymbol{\beta}^\star_1$ is composed of the rows of $\boldsymbol{\beta}^\star$ on indices $\mathcal{V}$. According to the definition of the central variable selection subspace, $\mathcal{S}_{\boldsymbol{\beta}^\star} \subseteq \mathcal{S}_{\boldsymbol{\alpha}^\star}$. Hence, $\mathcal{S}_{\boldsymbol{\beta}^\star_2} \subseteq \{\mathbf{0}\}$ and $\boldsymbol{\beta}^\star_2 = \mathbf{0}$. By the definition of $\mathcal{A}$, we have $\mathcal{A} \subseteq \mathcal{V}$.

It remains to prove $\mathcal{V} \subseteq \mathcal{A}$. According to the definition of central subspace, we have $Y \mid \mathbf{X} \sim Y \mid \mathbf{P}_{\boldsymbol{\beta}^\star}\mathbf{X}$, or equivalently, $Y \perp\!\!\!\perp \mathbf{X} \mid \mathbf{P}_{\boldsymbol{\beta}^\star}\mathbf{X}$. Since there is a one-to-one correspondence between $(\boldsymbol{\beta}^\star)^\top\mathbf{X}$ and $\mathbf{P}_{\boldsymbol{\beta}^\star}\mathbf{X}$, then

$$Y \perp\!\!\!\perp \mathbf{X} \mid (\boldsymbol{\beta}^\star)^\top \mathbf{X}.$$

Let $s' = |\mathcal{A}|$ and $\widetilde{\boldsymbol{\beta}} \in \mathbb{R}^{s' \times d}$ denote the matrix composed of non-zero rows of $\boldsymbol{\beta}$, and we split $\mathbf{X}$ into two sub-vectors $\mathbf{X}_1 \in \mathbb{R}^{s'}$ and $\mathbf{X}_2 \in \mathbb{R}^{p-s'}$ according to $\mathcal{A}$. Then, we have

$$Y \perp\!\!\!\perp (\mathbf{X}_1^\top, \mathbf{X}_2^\top)^\top \mid \widetilde{\boldsymbol{\beta}}^\top \mathbf{X}_1. \tag{H.1}$$

It follows that

$$Y \perp\!\!\!\perp \mathbf{X}_1 \mid \widetilde{\boldsymbol{\beta}}^\top \mathbf{X}_1. \tag{H.2}$$

Combining (H.1) and (H.2), we have

$$Y \perp\!\!\!\perp (\mathbf{X}_1^\top, \mathbf{X}_2^\top)^\top \mid \mathbf{X}_1,$$

or equivalently,

$$Y \mid \mathbf{X} \sim Y \mid \mathbf{X}_1. \tag{H.3}$$

We construct the matrix $\widetilde{\boldsymbol{\alpha}} \in \mathbb{R}^{p \times s'}$ such that $\widetilde{\boldsymbol{\alpha}}_{\mathcal{A}} = \mathbf{I}_{s'}$ and $\widetilde{\boldsymbol{\alpha}}_{\mathcal{A}^c} = \mathbf{0}$. Then by (H.3),

$$Y \mid \mathbf{X} \sim Y \mid \mathbf{P}_{\widetilde{\boldsymbol{\alpha}}} \mathbf{X}. \tag{H.4}$$

Hence, according to the definition of $\mathcal{S}_{Y|\mathbf{X}}^V = \mathcal{S}_{\boldsymbol{\alpha}}$, we have $\mathcal{S}_{\boldsymbol{\alpha}} \subseteq \mathcal{S}_{\widetilde{\boldsymbol{\alpha}}}$ and $\mathcal{V} \subseteq \mathcal{A}$. This completes the proof. $\square$

H.2 Proof of Lemma 2

The update of $\mathbf{B}^{(t)}$ is a convex optimization problem, which is solved by the group-wise coordinate descent algorithm exhibited in Algorithm B. In the next lemma, we first show that for the iterate sequence generated from Algorithm B, the limit point exists and is the global solution of problem (17).

Lemma H.1. *Let $\check{\mathbf{B}}^{(m)}$ denote the iterate from Algorithm B after m iterations, then the limit point of the sequence $\{\check{\mathbf{B}}^{(m)}\}_{m \geq 0}$ exists and is the global solution of* (17)*.*

Proof of Lemma H.1. Let $f(\mathbf{B})$ denote the object function of problem (17), i.e.,

$$f(\mathbf{B}) = \frac{1}{2}\mathrm{tr}\{\mathbf{B}^\top(\widehat{\boldsymbol{\Sigma}} + \gamma \mathbf{I}_p)\mathbf{B}\} - \mathrm{tr}\{\mathbf{B}^\top(\widehat{\mathbf{U}} - \boldsymbol{\omega}^{(t-1)} + \gamma \mathbf{C}^{(t-1)})\} + \lambda_1 \|\mathbf{B}\|_{2,1}.$$

We show that $\mathcal{B} = \{\mathbf{B} \in \mathbb{R}^{p \times H} : f(\mathbf{B}) \leq f(\check{\mathbf{B}}^{(0)})\}$ is a compact set.

Since $f(\mathbf{B})$ is continuous, then it is a closed function. Hence, $\mathcal{B}$ is a closed set. On the other hand, for any $\mathbf{B} \in \mathbb{R}^{p \times H}$, we have

$$\begin{aligned}
&\frac{1}{2}\mathrm{tr}\{\mathbf{B}^\top(\widehat{\boldsymbol{\Sigma}} + \gamma \mathbf{I}_p)\mathbf{B}\} - \mathrm{tr}\{\mathbf{B}^\top(\widehat{\mathbf{U}} - \boldsymbol{\omega}^{(t-1)} + \gamma \mathbf{C}^{(t-1)})\} \\
\geq& -\frac{1}{2}\mathrm{tr}\{(\widehat{\mathbf{U}} - \boldsymbol{\omega}^{(t-1)} + \gamma \mathbf{C}^{(t-1)})^\top (\widehat{\boldsymbol{\Sigma}} + \gamma \mathbf{I}_p)^{-1} (\widehat{\mathbf{U}} - \boldsymbol{\omega}^{(t-1)} + \gamma \mathbf{C}^{(t-1)})\}.
\end{aligned}$$

Therefore, for any $\mathbf{B} \in \mathcal{B}$, we have

$$
\begin{aligned}
&-\frac{1}{2}\mathrm{tr}\{(\widehat{\mathbf{U}} - \boldsymbol{\omega}^{(t-1)} + \gamma\mathbf{C}^{(t-1)})^\top(\widehat{\boldsymbol{\Sigma}} + \gamma\mathbf{I}_p)^{-1}(\widehat{\mathbf{U}} - \boldsymbol{\omega}^{(t-1)} + \gamma\mathbf{C}^{(t-1)})\} + \lambda_1\|\mathbf{B}\|_{2,1} \\
\leq & f(\mathbf{B}) \leq f(\check{\mathbf{B}}^{(0)}),
\end{aligned}
$$

and

$$
\|\mathbf{B}\|_{2,1} \leq \frac{1}{\lambda_1}f(\check{\mathbf{B}}^{(0)}) + \frac{1}{2\lambda_1}\mathrm{tr}\{(\widehat{\mathbf{U}} - \boldsymbol{\omega}^{(t-1)} + \gamma\mathbf{C}^{(t-1)})^\top(\widehat{\boldsymbol{\Sigma}} + \gamma\mathbf{I}_p)^{-1}(\widehat{\mathbf{U}} - \boldsymbol{\omega}^{(t-1)} + \gamma\mathbf{C}^{(t-1)})\}.
$$

Hence, $\mathcal{B}$ is a bounded set. By HeineBorel theorem, $\mathcal{B}$ is a compact set.

In Algorithm B, $f(\check{\mathbf{B}}^{(m)})$ is decreasing, then $f(\check{\mathbf{B}}^{(m)}) \leq f(\check{\mathbf{B}}^{(0)})$ for $m \geq 0$ and the iterate sequence $\{\check{\mathbf{B}}^{(m)}\}_{m\geq 0} \subset \mathcal{B}$. Since $\mathcal{B}$ is compact, there exists a limit point of $\{\check{\mathbf{B}}^{(m)}\}_{m\geq 0}$ in $\mathcal{B}$. Next, we prove that every limit point is the global minimizer of $f(\mathbf{B})$.

Let $\mathbf{B}_k$ denote the k-th row of $\mathbf{B}$, $k = 1, \ldots, p$, we rewrite $f(\mathbf{B})$ in form of

$$
f(\mathbf{B}) \equiv f(\mathbf{B}_1, \ldots, \mathbf{B}_p) = f_0(\mathbf{B}_1, \ldots, \mathbf{B}_p) + \lambda_1 \sum_{k=1}^{p} f_k(\mathbf{B}_k),
$$

where $f_0 = (1/2)\mathrm{tr}\{\mathbf{B}^\top(\widehat{\boldsymbol{\Sigma}} + \gamma\mathbf{I}_p)\mathbf{B}\} - \mathrm{tr}\{\mathbf{B}^\top(\widehat{\mathbf{U}} - \boldsymbol{\omega}^{(t-1)} + \gamma\mathbf{C}^{(t-1)})\}$, and $f_k(\mathbf{B}_k) = \|\mathbf{B}_k\|_2$, $k = 1, \ldots, p$. Since f_0 is differentiable and convex, and each f_k is convex, conditions (A1), (B1)–(B3), (C2) in Tseng (2001) are satisfied. Therefore, by Lemma 3.1 and Proposition 5.1 in Tseng (2001), any limit point of the iterate sequence $\{\check{\mathbf{B}}^{(m)}\}_{m\geq 0}$ is a global minimizer of $f(\mathbf{B})$. Moreover, $f(\mathbf{B})$ is strictly convex, then the global minimizer of $f(\mathbf{B})$ is unique.

Thus, we have proved that the limit point of the iterate sequence $\{\check{\mathbf{B}}^{(m)}\}_{m\geq 0}$ exists and is the solution of problem (17). □

The proof of Lemma 2 relies on the results in Deng and Yin (2016). Deng and Yin (2016) considered a generic constrained convex optimization problems with separable objective functions in the following form

$$
\begin{aligned}
\min_{\mathbf{x},\mathbf{y}} \quad & g_1(\mathbf{x}) + g_2(\mathbf{y}) \\
\text{s.t.} \quad & \mathbf{M}\mathbf{x} + \mathbf{N}\mathbf{y} = \mathbf{b},
\end{aligned}
\tag{H.5}
$$

where $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{y} \in \mathbb{R}^m$ are unknown variables, $\mathbf{M} \in \mathbb{R}^{p\times n}$ and $\mathbf{N} \in \mathbb{R}^{p\times m}$ are given matrices, $\mathbf{b} \in \mathbb{R}^p$, and $g_1 : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ and $g_2 : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ are closed proper convex functions. The augmented Lagrangian of (H.5) is

$$
L(\mathbf{x}, \mathbf{y}, \boldsymbol{\lambda}) = g_1(\mathbf{x}) + g_2(\mathbf{y}) - \boldsymbol{\lambda}^\top(\mathbf{M}\mathbf{x} + \mathbf{B}\mathbf{y} - \mathbf{b}) + \frac{\beta}{2}\|\mathbf{M}\mathbf{x} + \mathbf{N}\mathbf{y} - \mathbf{b}\|_2^2,
$$

where $\boldsymbol{\lambda} \in \mathbb{R}^p$ is the dual variable and $\beta > 0$ is ADMM parameter. Let $\mathbf{u}^\star = (\mathbf{x}^\star, \mathbf{y}^\star, \boldsymbol{\lambda}^\star)$ be the point satisfying the KKT condition of (H.5), and let $\mathbf{u}^{(t)} = (\mathbf{x}^{(t)}, \mathbf{y}^{(t)}, \boldsymbol{\lambda}^{(t)})$ denote the iterate from ADMM algorithm after t iterations. We use the following proposition:

Proposition H.8. *(Deng and Yin, 2016, ,Theorem 2.2) Assume that the KKT point $\mathbf{u}^\star = (\mathbf{x}^\star, \mathbf{y}^\star, \boldsymbol{\lambda}^\star)$ exists and functions f and g are convex. Further assume that $\{\mathbf{u}^{(t)}\}_{t\geq 0}$ of ADMM algorithm is bounded. We have (i) $\boldsymbol{\lambda}^{(t)} \to \boldsymbol{\lambda}^*$; (ii) $\mathbf{Mx}^{(t)} \to \mathbf{Mx}^\star$; (iii) $\mathbf{Ny}^{(t)} \to \mathbf{Ny}^\star$.*

Proof of Lemma 2. To conform with the notations in (H.5), we rewrite our optimization problem (12) as

$$\begin{aligned} \min_{\mathbf{B},\mathbf{C}\in\mathbb{R}^{p\times H}} \quad & g_1(\mathbf{B}) + g_2(\mathbf{C}), \\ \text{s.t.} \quad & \mathbf{B} - \mathbf{C} = \mathbf{0}, \end{aligned}$$

where $g_1(\mathbf{B}) = (1/2)\mathrm{tr}(\mathbf{B}^\top\widehat{\boldsymbol{\Sigma}}\mathbf{B}) - \mathrm{tr}(\mathbf{B}^\top\widehat{\mathbf{U}}) + \lambda_1\|\mathbf{B}\|_{2,1}$ and $g_2(\mathbf{C}) = \lambda_2\|\mathbf{C}\|_*$. The functions $g_1(\mathbf{B})$ and $g_2(\mathbf{C})$ are closed proper convex functions. Let $\widehat{\boldsymbol{\omega}}$ denote the solution of dual problem of (12). Since (12) is a convex optimization satisfying Slater's condition, then the solution $(\widehat{\boldsymbol{\omega}}, \widehat{\mathbf{B}}, \widehat{\mathbf{C}})$ is the KKT point of (12). Moreover, since we have the constraint $\mathbf{B} - \mathbf{C} = \mathbf{0}$, using the notations in Deng and Yin (2016), we have

$$\mathbf{M} = \mathbf{I},\ \mathbf{N} = -\mathbf{I},\ \mathbf{b} = \mathbf{0}.$$

By Remark 2.2 in Deng and Yin (2016), since $\mathbf{M}$ and $\mathbf{N}$ are full rank, then the iterate $(\boldsymbol{\omega}^{(t)}, \mathbf{B}^{(t)}, \mathbf{C}^{(t)})$ is bounded. Hence, by Proposition H.8, $\mathbf{B}^{(t)} \to \widehat{\mathbf{B}}$ and $\mathbf{C}^{(t)} \to \widehat{\mathbf{C}}$. This completes our proof of Lemma 2. □

H.3 Proof of Lemma 5

To prove Lemma 5, we need Theorem 1 in Tan et al. (2020). The model space considered in Tan et al. (2020) is

$$\begin{aligned} \mathcal{P}'(n, p, s, \rho; \kappa, m) \quad = \quad & \{\mathcal{L}(\mathbf{X}_i, \widetilde{Y}_i), i = 1, \ldots, n : (\mathbf{X}_i, \widetilde{Y}_i)\text{'s are i.i.d. such that} \\ & \mathbf{X}_i \mid (\widetilde{Y}_i = 1) \sim N((1-\alpha)\boldsymbol{\beta}^\star, \mathbf{I}_p - \mathbf{M}),\ \mathbf{X}_i \mid (\widetilde{Y}_i = 2) \sim N(-\alpha\boldsymbol{\beta}^\star, \mathbf{I}_p - \mathbf{M}), \\ & \mathbb{P}(\widetilde{Y}_i = 1) = \alpha,\ \mathbb{P}(\widetilde{Y}_i = 2) = 1 - \alpha; \\ & m^{-1} \leq \varphi_{\min}(\boldsymbol{\Sigma}) \leq \varphi_{\max}(\boldsymbol{\Sigma}) \leq m;\ \boldsymbol{\beta}^\star \text{ has at most } s \text{ nonzero rows;} \\ & \kappa\rho \geq \rho_1 \geq \ldots \geq \rho_{d^\star} \geq \rho > 0 \text{ for a fixed constant } \kappa > 1\}, \end{aligned}$$

where the basis matrix $\boldsymbol{\beta}^\star \in \mathbb{R}^p$ is orthonormal and the matrix $\mathbf{M} = \mathrm{Cov}\{\mathrm{E}(\mathbf{X}_i \mid \widetilde{Y}_i)\}$. It can be shown that $\mathrm{E}(\mathbf{X}_i) = \mathbf{0}$, $\mathbf{M} = \alpha(1-\alpha)\boldsymbol{\beta}^\star(\boldsymbol{\beta}^\star)^\top$ and $\boldsymbol{\Sigma} = \mathbf{I}_p$. Recall that $\widetilde{Y}_i$'s are the discretized responses, and $\rho_1, \ldots, \rho_{d^\star}$ are the generalized eigenvalues in the generalized eigenvalue problem formulation for SIR. Theorem 1 in Tan et al. (2020) is stated in the following proposition.

Proposition H.9. *Assume $C_1 \log(ep/s) \le n\rho^2$ for some sufficiently large constant $C_1 > 0$. Then there exist positive constants C and c_0 such that,*

$$\inf_{\widehat{\boldsymbol{\beta}}} \sup_{\mathbb{P}\in\mathcal{P}'} \mathbb{P}\left\{ \|\mathbf{P}_{\widehat{\boldsymbol{\beta}}} - \mathbf{P}_{\boldsymbol{\beta}^\star}\|_F \ge C\sqrt{\frac{s\log(ep/s)}{n\rho^2}} \wedge c_0 \right\} \ge 0.8,$$

where $\mathcal{P}' = \mathcal{P}'(n, p, s, \rho; \kappa, m)$.

Recall that the model space we considered is

$$\begin{aligned} \mathcal{P}(n,p,s,\rho) \;=\; & \{\mathcal{L}(\mathbf{X}_i, Y_i), i = 1,\ldots,n : (\mathbf{X}_i, Y_i)\text{'s are i.i.d. such that } \mathbf{X}_i \text{ satisfies (A3);} \\ & \boldsymbol{\beta}^\star \text{ has at most } s \text{ nonzero rows; } \rho_{d^\star} = \rho > 0\}. \end{aligned}$$

Since we assume that $d^\star \le H \le A$ for some constant $A > 0$, then $\mathcal{D}(\mathcal{S}_{\mathrm{SIR}}, \mathcal{S}_{\widehat{\boldsymbol{\beta}}}) = \|\mathbf{P}_{\boldsymbol{\beta}^\star} - \mathbf{P}_{\widehat{\boldsymbol{\beta}}}\|_F/\sqrt{2d^\star} \ge \|\mathbf{P}_{\boldsymbol{\beta}^\star} - \mathbf{P}_{\widehat{\boldsymbol{\beta}}}\|_F/\sqrt{2A}$.

Next, we show that $\mathbf{X}_i$ in parameter space $\mathcal{P}'$ is the sub-Gaussian random vector with bounded parameter. For any $\mathbf{a} \in \mathbb{R}^p$, using the moment generating function of normal random variable, we have

$$\begin{aligned} &\mathrm{E}\{\exp(\mathbf{a}^\top \mathbf{X}_i)\} = \mathrm{E}\left[\mathrm{E}\{\exp(\mathbf{a}^\top \mathbf{X}_i) \mid \widetilde{Y}_i\}\right] \\ =&\alpha \mathrm{E}\{\exp(\mathbf{a}^\top \mathbf{X}_i) \mid \widetilde{Y}_i = 1\} + (1-\alpha)\mathrm{E}\{\exp(\mathbf{a}^\top \mathbf{X}_i) \mid \widetilde{Y}_i = 2\} \\ =&\alpha \exp\left[(1-\alpha)\mathbf{a}^\top\boldsymbol{\beta}^\star + \{\mathbf{a}^\top(\mathbf{I}_p - \mathbf{M})\mathbf{a}\}/2\right] + (1-\alpha)\exp\left[-\alpha\mathbf{a}^\top\boldsymbol{\beta}^\star + \{\mathbf{a}^\top(\mathbf{I}_p - \mathbf{M})\mathbf{a}\}/2\right] \\ \le& \exp(\|\mathbf{a}\|_2^2\|\mathbf{I}_p - \mathbf{M}\|/2)\left[\alpha\exp\{(1-\alpha)\|\mathbf{a}\|_2\} + (1-\alpha)\exp(\alpha\|\mathbf{a}\|_2)\right]. \end{aligned}$$

Since $\mathbf{M} = \alpha(1-\alpha)\boldsymbol{\beta}^\star(\boldsymbol{\beta}^\star)^\top$, then $\|\mathbf{I}_p - \mathbf{M}\| = 1$. If $\|\mathbf{a}\|_2 \ge 1$, we have

$$\alpha\exp\{(1-\alpha)\|\mathbf{a}\|_2\} + (1-\alpha)\exp(\alpha\|\mathbf{a}\|_2) \le \exp(\|\mathbf{a}\|_2^2).$$

And if $\|\mathbf{a}\|_2 < 1$, then

$$\alpha\exp\{(1-\alpha)\|\mathbf{a}\|_2\} + (1-\alpha)\exp(\alpha\|\mathbf{a}\|_2) \le \exp(1).$$

Overall, we obtain

$$\mathrm{E}\{\exp(\mathbf{a}^\top\mathbf{X}_i)\} \le \exp(1)\exp(\|\mathbf{a}\|_2^2)\exp(\|\mathbf{a}\|_2^2/2) \le \exp(1)\exp(3\|\mathbf{a}\|_2^2/2).$$

Hence, $\mathbf{X}_i$ in model space $\mathcal{P}'$ is a sub-Gaussian random vector with bounded parameter.

Therefore, the model space $\mathcal{P}'$ is a subset of $\mathcal{P}$, and the proof is complete. □

H.4 Proof of Lemma D.1

Recall that we are considering the linear regression model

$$Y = \boldsymbol{\beta}^\top \mathbf{X} + \epsilon. \tag{H.6}$$

According to the definition of the CS, in model (H.6), we have $\mathcal{S}_{Y|\mathbf{X}} = \text{span}(\boldsymbol{\beta})$. Since $\mathbf{X}$ follows the normal distribution and the structural dimension of $\mathcal{S}_{Y|\mathbf{X}}$ is one, the linearity and coverage conditions are satisfied, which gives us $\mathcal{S}_{\text{SIR}} = \mathcal{S}_{Y|\mathbf{X}}$. In addition, since $\text{Cov}(\mathbf{X}) = \mathbf{I}$, we have $\mathcal{S}_{\text{SIR}} = \text{span}[\text{Cov}\{\text{E}(\mathbf{X} \mid Y)\}]$. Therefore, $\text{span}[\text{Cov}\{\text{E}(\mathbf{X} \mid Y)\}] = \text{span}(\boldsymbol{\beta})$, and $\boldsymbol{\beta}$ is the eigenvector of $\text{Cov}\{\text{E}(\mathbf{X} \mid Y)\}$ corresponding to its non-zero eigenvalue. Let α denote the non-zero eigenvalue of $\text{Cov}\{\text{E}(\mathbf{X} \mid Y)\}$, then we have

$$\alpha = \text{Cov}\{\text{E}(\boldsymbol{\beta}^\top \mathbf{X} \mid Y)\} = \tau^2/(\tau^2 + \omega^2).$$

Note that α is also the signal-to-noise ratio for model (H.6), i.e., $\alpha = \text{Cov}(\boldsymbol{\beta}^\top \mathbf{X})/\text{Var}(Y)$. It is also known that $\text{span}[\text{Cov}\{\text{E}(\mathbf{X} \mid \widetilde{Y})\}] \subseteq \mathcal{S}_{Y|\mathbf{X}} = \text{span}(\boldsymbol{\beta})$. Since the structural dimension of $\mathcal{S}_{Y|\mathbf{X}}$ is one, then $\text{span}[\text{Cov}\{\text{E}(\mathbf{X} \mid \widetilde{Y})\}] = \text{span}(\boldsymbol{\beta})$, and $\boldsymbol{\beta}$ is also the eigenvector of $\text{Cov}\{\text{E}(\mathbf{X} \mid \widetilde{Y})\}$ corresponding to its non-zero eigenvalue. Let $\widetilde{\alpha}$ denote the non-zero eigenvalue of $\text{Cov}\{\text{E}(\mathbf{X} \mid \widetilde{Y})\}$, then we have

$$\widetilde{\alpha} = \text{Cov}\{\text{E}(\boldsymbol{\beta}^\top \mathbf{X} \mid \widetilde{Y})\}.$$

Now we connect $\widetilde{\alpha}$ with α. Recall that $\pi_h = \mathbb{P}(\widetilde{Y} = h)$ for $h = 1, \ldots, H$, then we have

$$\widetilde{\alpha} = \text{Cov}\{\text{E}(\boldsymbol{\beta}^\top \mathbf{X} \mid \widetilde{Y})\} = \text{E}\{\text{E}^2(\boldsymbol{\beta}^\top \mathbf{X} \mid \widetilde{Y})\} = \sum_{h=1}^{H} \pi_h \text{E}^2(\boldsymbol{\beta}^\top \mathbf{X} \mid \widetilde{Y} = h). \tag{H.7}$$

Let $-\infty = a_0 < a_1 < \ldots < a_{H-1} < a_H = +\infty$ be a partition of the range of Y. Without loss of generality, let $\widetilde{Y} = h$ if $Y \in [a_{h-1}, a_h)$, $h = 1, \ldots, H$. Let $T = \boldsymbol{\beta}_0^\top \mathbf{X}$, where $\boldsymbol{\beta}_0 = \boldsymbol{\beta}/\|\boldsymbol{\beta}\|_2$, then $Y = \tau T + \epsilon$ and $T \sim N(0, 1)$. Also let $S = \epsilon/\omega$. We first derive the conditional cumulative distribution function (cdf), $\mathbb{P}(T \leq c \mid \widetilde{Y} = h)$, as follows,

$$\begin{aligned}
\mathbb{P}(T \leq c \mid \widetilde{Y} = h) =& \pi_h^{-1} \mathbb{P}(T \leq c, Y \in [a_{h-1}, a_h)) \\
=& \pi_h^{-1} \mathbb{P}(T \leq c, \tau T + \epsilon \in [a_{h-1}, a_h)) \\
=& \pi_h^{-1} \int_{-\infty}^{c} \int_{(a_{h-1} - \tau t)/\omega}^{(a_h - \tau t)/\omega} f_{T,S}(t, s) \mathrm{d}t \mathrm{d}s \\
=& \pi_h^{-1} \int_{-\infty}^{c} \int_{(a_{h-1} - \tau t)/\omega}^{(a_h - \tau t)/\omega} f_T(t) f_S(s) \mathrm{d}t \mathrm{d}s \\
=& \pi_h^{-1} \int_{-\infty}^{c} f_T(t) \left\{ \Phi\left(\frac{a_h - \tau t}{\omega}\right) - \Phi\left(\frac{a_{h-1} - \tau t}{\omega}\right) \right\} \mathrm{d}t,
\end{aligned}$$

where $f(\cdot)$ denotes the probability density function (pdf) for a random variable, $\Phi(\cdot)$ denotes the cdf of the standard normal distribution, and the fourth equation follows from the independence assumption between $\mathbf{X}$ and ϵ. By taking the derivative of $\mathbb{P}(T \le c \mid \widetilde{Y} = h)$ w.r.t. the variate c, we have the pdf for the conditional distribution $T \mid (\widetilde{Y} = h)$ as follows,

$$f_{T|\widetilde{Y}}(t \mid h) = \pi_h^{-1} f_T(t) \left\{ \Phi\left(\frac{a_h - \tau t}{\omega}\right) - \Phi\left(\frac{a_{h-1} - \tau t}{\omega}\right) \right\}.$$

Therefore,

$$\begin{aligned} \mathrm{E}(T \mid \widetilde{Y} = h) =& \pi_h^{-1} \left\{ \int_{\mathbb{R}} t f_T(t) \Phi\left(\frac{a_h - \tau t}{\omega}\right) \mathrm{d}t - \int_{\mathbb{R}} t f_T(t) \Phi\left(\frac{a_{h-1} - \tau t}{\omega}\right) \mathrm{d}t \right\} \\ =& \pi_h^{-1} (A_h - A_{h-1}). \end{aligned} \tag{H.8}$$

We derive A_h in the following. Since $T = \boldsymbol{\beta}^\top \mathbf{X} \sim N(0,1)$, let $f_N(\cdot)$ denote the pdf of the standard normal distribution, then $f_T(t) = f_N(t)$. And by simple calculation, we have $f_N'(t) = -t f_N(t)$. Then,

$$\left\{ f_N(t) \Phi\left(\frac{a_h - \tau t}{\omega}\right) \right\}' = -t f_N(t) \Phi\left(\frac{a_h - \tau t}{\omega}\right) - \frac{\tau}{\omega} f_N(t) f_N\left(\frac{a_h - \tau t}{\omega}\right).$$

Hence, for $h = 0, \ldots, H$, we have

$$\begin{aligned} A_h =& - f_N(t) \Phi\left(\frac{a_h - \tau t}{\omega}\right) \Bigg|_{-\infty}^{+\infty} - \frac{\tau}{\omega} \int_{\mathbb{R}} f_N(t) f_N\left(\frac{a_h - \tau t}{\omega}\right) \mathrm{d}t \\ =& - \frac{\tau}{\omega} \int_{\mathbb{R}} f_N(t) f_N\left(\frac{a_h - \tau t}{\omega}\right) \mathrm{d}t \\ =& - \sqrt{\frac{\tau^2}{2\pi(\omega^2 + \tau^2)}} \exp\left\{ -\frac{a_h^2}{2(\omega^2 + \tau^2)} \right\}. \end{aligned} \tag{H.9}$$

By combining (H.7)–(H.9), we obtain

$$\begin{aligned} &\mathrm{Cov}\{\mathrm{E}(\boldsymbol{\beta}^\top \mathbf{X} \mid \widetilde{Y})\} = \sum_{h=1}^{H} \pi_h^{-1} (A_h - A_{h-1})^2 \\ =& \sum_{h=1}^{H} \pi_h^{-1} \frac{\tau^2}{2\pi(\omega^2 + \tau^2)} \left[\exp\left\{ -\frac{a_h^2}{2(\omega^2 + \tau^2)} \right\} - \exp\left\{ -\frac{a_{h-1}^2}{2(\omega^2 + \tau^2)} \right\} \right]^2 \\ \ge& \frac{1}{a} \cdot \frac{\tau^2}{2\pi(\omega^2 + \tau^2)} \sum_{h=1}^{H} \left[\exp\left\{ -\frac{a_h^2}{2(\omega^2 + \tau^2)} \right\} - \exp\left\{ -\frac{a_{h-1}^2}{2(\omega^2 + \tau^2)} \right\} \right]^2, \end{aligned}$$

where the inequality comes from the assumption that $a^{-1} \le \pi_h \le a$ for some positive constant a, and for $h = 1, \ldots, H$. Since we assume that ω and τ are bounded from below, then we have

$$
\begin{aligned}
&\sum_{h=1}^{H}\left[\exp\left\{-\frac{a_h^2}{2(\omega^2+\tau^2)}\right\}-\exp\left\{-\frac{a_{h-1}^2}{2(\omega^2+\tau^2)}\right\}\right]^2\\
\geq&\left[\exp\left\{-\frac{a_1^2}{2(\omega^2+\tau^2)}\right\}-\exp\left\{-\frac{a_0^2}{2(\omega^2+\tau^2)}\right\}\right]^2+\left[\exp\left\{-\frac{a_H^2}{2(\omega^2+\tau^2)}\right\}-\exp\left\{-\frac{a_{H-1}^2}{2(\omega^2+\tau^2)}\right\}\right]^2\\
=&\exp^2\left\{-\frac{a_1^2}{2(\omega^2+\tau^2)}\right\}+\exp^2\left\{-\frac{a_{H-1}^2}{2(\omega^2+\tau^2)}\right\}\geq c
\end{aligned}
$$

for some positive constant c. Therefore, we have

$$
\widetilde{\alpha}=\mathrm{Cov}\{\mathrm{E}(\boldsymbol{\beta}^\top\mathbf{X}\mid\widetilde{Y})\}\geq c\frac{\tau^2}{\omega^2+\tau^2}=c\alpha.
$$

In model (H.6), the structural dimension $d^\star=1$ and $\sigma_{\mathrm{SIR},1}=\sigma_{\mathrm{SIR},d^\star}$. Then $\sigma_{\mathrm{SIR},1}/\sigma_{\mathrm{SIR},d^\star}^2=1/\sigma_{\mathrm{SIR},d^\star}$. Recall that $\sigma_{\mathrm{SIR},d^\star}$ is the smallest non-zero singular value of $\mathbf{B}_{\mathrm{SIR}^\star}$, according to Lemma H.3, since $\mathrm{Cov}(\mathbf{X})=\mathbf{I}$, it follows that

$$
\sigma_{\mathrm{SIR},d^\star}^2\geq C\widetilde{\alpha}\geq C'\alpha
$$

for some positive constants C and C'. Hence,

$$
\frac{\sigma_{\mathrm{SIR},1}}{\sigma_{\mathrm{SIR},d^\star}^2}=\frac{1}{\sigma_{\mathrm{SIR},d^\star}}\leq C\frac{1}{\sqrt{\alpha}},
$$

which completes the proof. □

H.5 Proof of Lemma D.2

Since $\mathbb{P}(Y=1)=\mathbb{P}(Y=2)=1/2$, $\mathbf{X}\mid(Y=1)\sim N(\boldsymbol{\mu}/2,\mathbf{I})$ and $\mathbf{X}\mid(Y=2)\sim N(-\boldsymbol{\mu}/2,\mathbf{I})$, we have $\mathrm{E}(\mathbf{X})=0$. And by the law of total covariance, we obtain

$$
\boldsymbol{\Sigma}=\mathrm{Cov}(\mathbf{X})=\mathrm{Cov}\{\mathrm{E}(\mathbf{X}\mid Y)\}+\mathrm{E}\{\mathrm{Cov}(\mathbf{X}\mid Y)\}=\boldsymbol{\mu}\boldsymbol{\mu}^\top/4+\mathbf{I}.
$$

Let $\boldsymbol{\mu}_0=\boldsymbol{\mu}/\|\boldsymbol{\mu}\|_2$ and $\boldsymbol{\mu}_0^\perp\in\mathbb{R}^{p\times(p-1)}$ be the orthogonal complement of $\boldsymbol{\mu}_0$ such that $(\boldsymbol{\mu}_0,\boldsymbol{\mu}_0^\perp)\in\mathbb{R}^{p\times p}$ is an orthogonal matrix. Then,

$$
\boldsymbol{\Sigma}^{-1}=(\boldsymbol{\mu}_0,\boldsymbol{\mu}_0^\perp)\mathrm{diag}\{(1+\|\boldsymbol{\mu}\|_2^2/4)^{-1},1,\ldots,1\}(\boldsymbol{\mu}_0,\boldsymbol{\mu}_0^\perp)^\top.
$$

According to the definition of $\mathbf{B}_{\mathrm{SIR}}^\star$, we have

$$
\mathbf{B}_{\mathrm{SIR}}^\star=\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}/2,-\boldsymbol{\mu}/2).
$$

Then the non-zero singular value of $\mathbf{B}^\star_{\mathrm{SIR}}$ is

$$\sigma_{\mathrm{SIR},d^\star} = \frac{\|\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\|_2}{\sqrt{2}} = \frac{2\sqrt{2}\|\boldsymbol{\mu}\|_2}{4+\|\boldsymbol{\mu}\|_2^2}.$$

Since we assume that $\|\boldsymbol{\mu}\|_2$ is bounded from above, then

$$\sigma_{\mathrm{SIR},d^\star} \geq C\|\boldsymbol{\mu}\|_2$$

for some positive constant C. In this model, the structural dimension $d^\star = 1$ and $\sigma_{\mathrm{SIR},1} = \sigma_{\mathrm{SIR},d^\star}$. Then $\sigma_{\mathrm{SIR},1}/\sigma^2_{\mathrm{SIR},d^\star} = 1/\sigma_{\mathrm{SIR},d^\star}$. Hence,

$$\frac{\sigma_{\mathrm{SIR},1}}{\sigma^2_{\mathrm{SIR},d^\star}} = \frac{1}{\sigma_{\mathrm{SIR},d^\star}} \leq C\frac{1}{\|\boldsymbol{\mu}\|_2}.$$

This completes the proof.

$\square$

H.6 Proof of Corollories 1 & 2

We first prove Corollary 1. Under the same conditions as in Theorem 1, the three results in Theorem 1 hold. As we further assume that $(\sigma_1/\sigma^2_{d^\star})\sqrt{s\log p/n} = o(1)$ and we allow $\sigma_1/\sigma^2_{d^\star}$ to diverge to infinity as n diverge, then $\sqrt{s\log p/n} = o(1)$. Thus, as n diverge to infinity, we have (i) $\|\widehat{\mathbf{B}} - \mathbf{B}^\star\|_F \to 0$, (ii) $\widehat{d} = d^\star$, and (iii) $\mathcal{D}(\mathcal{S}_{\boldsymbol{\beta}^\star}, \mathcal{S}_{\widehat{\boldsymbol{\beta}}}) \to 0$, with probability tending to 1.

Then we prove Corollary 2. By Lemma F.1, under the same conditions as in Theorem 1, we have $\|\widehat{\mathbf{B}}-\mathbf{B}^\star\| \leq \|\widehat{\mathbf{B}}-\mathbf{B}^\star\|_F \leq C_B\sqrt{s\log p/n}$ with probability at least $1-C\exp\{-C'(s\wedge \log p)\}$ for some constants $C_B, C, C' > 0$. And according to Assumption (A1), by selecting $M \geq 3C_B$, we further have $\|\widehat{\mathbf{B}} - \mathbf{B}^\star\| \leq (C_B/M)\sigma_{d^\star} \leq \sigma_1/3$. Moreover, by Theorem 1, we have $\widehat{d} = d^\star$, then the estimated basis $\widehat{\boldsymbol{\beta}}$ is a $p \times d^\star$ matrix. Let $\boldsymbol{\alpha}\mathbf{D}\mathbf{V}^\top$ denote the singular value decomposition of $\mathbf{B}^\star$ where $\boldsymbol{\alpha} \in \mathbb{R}^{p\times d^\star}$ and $\mathbf{V} \in \mathbb{R}^{H\times d^\star} \in \mathbb{R}^{d^\star\times d^\star}$ denote its left and right singular vectors and $\mathbf{D} = \mathrm{diag}(\sigma_1, \ldots, \sigma_{d^\star})$ denotes the diagonal matrix composed of its singular values. Then by Proposition E.2, we have

$$\begin{aligned}
\|\widehat{\boldsymbol{\beta}}\mathbf{O} - \boldsymbol{\alpha}\|_F \leq& \frac{2^{3/2}(2\sigma_1 + \|\widehat{\mathbf{B}} - \mathbf{B}^\star\|)\min\{\sqrt{d^\star}\|\widehat{\mathbf{B}} - \mathbf{B}^\star\|, \|\widehat{\mathbf{B}} - \mathbf{B}^\star\|_{\mathrm{F}}\}}{\sigma^2_{d^\star}} \\
\leq& (14\sqrt{2}/3)(\sigma_1/\sigma^2_{d^\star})\|\widehat{\mathbf{B}} - \mathbf{B}^\star\|_F \leq (14\sqrt{2}C_B/3)(\sigma_1/\sigma^2_{d^\star})\sqrt{s\log p/n}.
\end{aligned}$$

for some orthogonal matrix $\mathbf{O} \in \mathbb{R}^{d^\star\times d^\star}$. And it follows that

$$\|\mathbf{O}^\top\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\alpha}_j\|_2 \leq (14\sqrt{2}C_B/3)(\sigma_1/\sigma^2_{d^\star})\sqrt{s\log p/n}, \quad j = 1, \ldots, p. \tag{H.10}$$

In Assumption (A2), we assume that $\min_{j\in\mathcal{A}}\|\mathbf{B}^\star_j\|_2 \geq N(\sigma_1^2/\sigma_{d^\star}^2)\sqrt{s\log p/n}$ for some constant $N>0$. Since for each j,

$$\|\mathbf{B}^\star_j\|_2 = \|\mathbf{V}\mathbf{D}\boldsymbol{\alpha}_j\|_2 \leq \sigma_1\|\boldsymbol{\alpha}_j\|_2,$$

then we have,

$$\min_{j\in\mathcal{A}}\|\boldsymbol{\alpha}_j\|_2 \geq N(\sigma_1/\sigma_{d^\star}^2)\sqrt{s\log p/n}. \tag{H.11}$$

Combining (H.10) and (H.11), by taking $N > (14\sqrt{2}C_B/3)$, for $j\in\mathcal{A}$, we have

$$\|\widehat{\boldsymbol{\beta}}_j\|_2 = \|\mathbf{O}^\top\widehat{\boldsymbol{\beta}}_j\|_2 \geq \|\boldsymbol{\alpha}_j\|_2 - \|\mathbf{O}^\top\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\alpha}_j\|_2 \geq (N - 14\sqrt{2}C_B/3)(\sigma_1/\sigma_{d^\star}^2)\sqrt{s\log p/n} > 0.$$

Since $\widehat{\mathcal{A}} = \{j : \|\widehat{\boldsymbol{\beta}}_j\|_2 > 0\}$, it follows that $\mathcal{A}\subseteq\widehat{\mathcal{A}}$ with probability at least $1-C\exp\{-C'(s\wedge\log p)\}$.

On the other hand,

$$\begin{aligned}\|\widehat{\boldsymbol{\beta}}_{\mathcal{A}^c}\|_F &= \|\widehat{\boldsymbol{\beta}}_{\mathcal{A}^c}\mathbf{O}\|_F \leq \|\widehat{\boldsymbol{\beta}}\mathbf{O} - \boldsymbol{\alpha}\|_F \leq (14\sqrt{2}C_B/3)(\sigma_1/\sigma_{d^\star}^2)\sqrt{s\log p/n}\\ &= C(\sigma_1/\sigma_{d^\star}^2)\sqrt{s\log p/n}\end{aligned}$$

with probability at least $1 - C\exp\{-C'(s\wedge\log p)\}$, which completes the proof. □

H.7 Proof of Corollaries 3–5

We prove the three corollaries by verifying Condition (C1) and (C2) under different method-specific assumptions. Recall Assumptions (A3)–(A6) in Section 4.2:

(A3) The predictors $\mathbf{X}_1,\ldots,\mathbf{X}_n\in\mathbb{R}^p$ are i.i.d. centered sub-Gaussian random vectors with parameter $\kappa^2\leq m/4$ for some constant $m>0$. In other words, $\mathrm{E}(\mathbf{X}_i)=\mathbf{0}$ and, for any $\mathbf{a}\in\mathbb{R}^p$, we have $\mathrm{E}\{\exp(\mathbf{a}^\top\mathbf{X}_i)\}\leq\exp(\kappa^2\|\mathbf{a}\|_2^2/2)$, $i=1,\ldots,n$. We further assume that $\mathrm{Cov}(\mathbf{X}_i)=\boldsymbol{\Sigma}$ and $\varphi_{\min}(\boldsymbol{\Sigma})\geq m^{-1}$.

(A4) There exists some constant $T>0$ such that the leading singular value of $\mathbf{B}^\star$, $\sigma_1\leq T$.

(A5) The variable Y is centered sub-Gaussian with the parameter $\theta^2\leq G$ for some constant $G>0$.

(A6) In the PFC model (10), the fitting functions $\boldsymbol{f}(Y_i)$ are i.i.d. centered sub-Gaussian random vectors with parameter $\xi^2\leq L$ for some constant $L>0$, $i=1,\ldots,n$. Also assume that $\|\boldsymbol{\eta}\|$ is bounded from above and $g^{-1}\leq\varphi_{\min}(\boldsymbol{\Delta})\leq\varphi_{\max}(\boldsymbol{\Delta})\leq g$ for some constant $g>0$.

Under Assumption (A3), there is a straightforward result on the bound of eigenvalues of $\mathbf{\Sigma}$ in Lemma H.2. Moreover, we connect the smallest non-zero singular value of $\mathbf{B}^\star_{\mathrm{SIR}}$ to the smallest non-zero generalized eigenvalue in the generalized eigenvalue problem for SIR in Lemma H.3.

Lemma H.2. *Under Assumption (A3), the smallest and the largest eigenvalues of $\mathbf{\Sigma}$ satisfy that $m^{-1} \le \varphi_{\min}(\mathbf{\Sigma}) \le \varphi_{\max}(\mathbf{\Sigma}) \le m$.*

Proof of Lemma H.2. By Lemma E.6, we have that $\varphi_{\max}(\mathbf{\Sigma}) \le 4\kappa^2 \le m$ for some constant $m > 0$. Since we assume that $\varphi_{\min}(\mathbf{\Sigma}) \ge m^{-1}$, we conclude that $m^{-1} \le \varphi_{\min}(\mathbf{\Sigma}) \le \varphi_{\max}(\mathbf{\Sigma}) \le m$. $\square$

Lemma H.3. *Let $\sigma_{\mathrm{SIR},1}$ and $\sigma_{\mathrm{SIR},d^\star}$ be the largest and the smallest non-zero singular values of $\mathbf{B}^\star_{\mathrm{SIR}} = \mathbf{\Sigma}^{-1}\mathbf{U}_{\mathrm{SIR}}$, and ρ be the smallest non-zero generalized eigenvalue in the following generalized eigenvalue problem:*

$$\mathbf{V}_{\mathrm{SIR}}\boldsymbol{\psi} = \mathbf{\Sigma}\boldsymbol{\psi}\mathbf{D}, \quad \boldsymbol{\psi}^\top\mathbf{\Sigma}\boldsymbol{\psi} = \mathbf{I}_{d^\star},$$

where $\mathbf{V}_{\mathrm{SIR}} = \mathrm{Cov}\{\mathrm{E}(\mathbf{X} \mid \widetilde{Y})\}$, $\mathbf{D} = \mathrm{diag}(\rho_1, \dots, \rho_{d^\star})$ and $\rho_1 \ge \cdots \ge \rho_{d^\star} = \rho > 0$. Under Assumption (A3), we have $\sigma^2_{\mathrm{SIR},d^\star} \ge C\rho$ and $\sigma^2_{\mathrm{SIR},1} \le C'\rho_1$ for some positive constants C and C'.

Proof of Lemma H.3. We rewrite $\mathbf{V}_{\mathrm{SIR}} = \mathrm{Cov}\{\mathrm{E}(\mathbf{X} \mid \widetilde{Y})\} = \mathbf{U}_{\mathrm{SIR}}\mathbf{P}\mathbf{U}_{\mathrm{SIR}}^\top$, where $\mathbf{P} = \mathrm{diag}(\pi_1, \dots, \pi_H)$ and $\pi_h = \mathbb{P}(\widetilde{Y} = h)$ for $h = 1, \dots, H$. Then

$$\mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{V}_{\mathrm{SIR}}\mathbf{\Sigma}^{-\frac{1}{2}} = \mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{U}_{\mathrm{SIR}}\mathbf{P}\mathbf{U}_{\mathrm{SIR}}^\top\mathbf{\Sigma}^{-\frac{1}{2}} = \mathbf{\Sigma}^{\frac{1}{2}}\mathbf{B}^\star_{\mathrm{SIR}}\mathbf{P}(\mathbf{B}^\star_{\mathrm{SIR}})^\top\mathbf{\Sigma}^{\frac{1}{2}},$$

By Proposition E.3, since ρ_i is the i-th eigenvalue of $\mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{V}_{\mathrm{SIR}}\mathbf{\Sigma}^{-\frac{1}{2}}$, it follows that

$$\rho_i = \varphi_i(\mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{V}_{\mathrm{SIR}}\mathbf{\Sigma}^{-\frac{1}{2}}) = \theta_i\varphi_i(\mathbf{B}^\star_{\mathrm{SIR}}\mathbf{P}(\mathbf{B}^\star_{\mathrm{SIR}})^\top).$$

By Lemma H.2, under Assumption (A3), we have $m^{-1} \le \theta_i \le m$. Then, $m^{-1}\rho_i \le \varphi_i(\mathbf{B}^\star_{\mathrm{SIR}}\mathbf{P}(\mathbf{B}^\star_{\mathrm{SIR}})^\top) \le m\rho_i$, $i = 1, \dots, p$.

Since we assumed that $a^{-1} \le \pi_h \le a$ for some constant $a > 0$, $h = 1, \dots, H$, then for any $\mathbf{u} \in \mathbb{R}^p$, we have

$$a^{-1}\mathbf{u}^\top\mathbf{B}^\star_{\mathrm{SIR}}\mathbf{P}(\mathbf{B}^\star_{\mathrm{SIR}})^\top\mathbf{u} \le \mathbf{u}^\top\mathbf{B}^\star_{\mathrm{SIR}}(\mathbf{B}^\star_{\mathrm{SIR}})^\top\mathbf{u} \le a\mathbf{u}^\top\mathbf{B}^\star_{\mathrm{SIR}}\mathbf{P}(\mathbf{B}^\star_{\mathrm{SIR}})^\top\mathbf{u}.$$

By Courant-Fischer Minimax Theorem (Golub and Van Loan, 2013, Theorem 8.1.2), we have

$$\sigma^2_{\mathrm{SIR},d^\star} \ge a^{-1}\varphi_{d^\star}(\mathbf{B}^\star_{\mathrm{SIR}}\mathbf{P}(\mathbf{B}^\star_{\mathrm{SIR}})^\top) \ge a^{-1}m^{-1}\rho,$$

and

$$\sigma_{\text{SIR},1}^2 \le a\varphi_1(\mathbf{B}_{\text{SIR}}^{\star}\mathbf{P}(\mathbf{B}_{\text{SIR}}^{\star})^{\top}) \le am\rho_1,$$

which finishes the proof. $\square$

Under the common Assumptions (A3) and (A4) used by all three SEAS methods, we establish in Lemma H.4 the lower bound for $\mathbb{P}(\|(\widehat{\mathbf{\Sigma}}-\mathbf{\Sigma})\mathbf{b}_h^{\star}\|_\infty \le \epsilon, \forall h)$ in Condition (C1). And in Lemma H.5, we derive Condition (C2).

Lemma H.4. *Under Assumptions (A3) and (A4), we have*

$$\mathbb{P}(\|(\widehat{\mathbf{\Sigma}}-\mathbf{\Sigma})\mathbf{b}_h^{\star}\|_\infty \le \epsilon, \forall h) \ge 1 - Cp\exp(-C'n\epsilon^2), \quad 0<\epsilon\le C',$$

for some constants $C, C' > 0$.

Proof of Lemma H.4. Recall that $\widehat{\mathbf{\Sigma}} = \frac{1}{n}\sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i^{\top}$. Let $\mathbf{b}_{h0}^{\star} = \mathbf{b}_h^{\star}/\|\mathbf{b}_h^{\star}\|_2$, we have

$$\mathbf{e}_i^{\top}(\widehat{\mathbf{\Sigma}}-\mathbf{\Sigma})\mathbf{b}_{h0}^{\star} = \frac{1}{n}\sum_{i=1}^n \mathbf{e}_i^{\top}\mathbf{X}_i\mathbf{X}_i^{\top}\mathbf{b}_{h0}^{\star} - \mathbf{e}_i^{\top}\mathbf{\Sigma}\mathbf{b}_{h0}^{\star},$$

where $\mathbf{e}_i$ is the i-th standard basis vector in $\mathbb{R}^p$. By Assumption (A3), $\mathbf{X}$ is sub-Gaussian random vector with bounded parameter $\kappa^2>0$. We take $0<\epsilon<\kappa^2$, then by Proposition E.6, there exist some constants $C, C'>0$ such that

$$\mathbb{P}(|\mathbf{e}_i^{\top}(\widehat{\mathbf{\Sigma}}-\mathbf{\Sigma})\mathbf{b}_{h0}^{\star}| \ge \epsilon) \le C\exp(-C'n\epsilon^2).$$

Then we have

$$\mathbb{P}(|\mathbf{e}_i^{\top}(\widehat{\mathbf{\Sigma}}-\mathbf{\Sigma})\mathbf{b}_{h}^{\star}| \ge \epsilon\|\mathbf{b}_h^{\star}\|_2) \le C\exp(-C'n\epsilon^2).$$

Since we assumed that σ_1 is bounded from above in Assumption (A4) and $d^{\star} \le H$ for bounded H, then we have $\|\mathbf{b}_h^{\star}\|_2 \le \|\mathbf{B}^{\star}\|_F \le \sqrt{d^{\star}}\|\mathbf{B}^{\star}\|_2 = \sqrt{d^{\star}}\sigma_1 \le C$ for some constant $C>0$. It follows that

$$\mathbb{P}(|\mathbf{e}_i^{\top}(\widehat{\mathbf{\Sigma}}-\mathbf{\Sigma})\mathbf{b}_{h}^{\star}| \ge \epsilon) \le C\exp(-C'n\epsilon^2).$$

By union bound, we have

$$\mathbb{P}(\|(\widehat{\mathbf{\Sigma}}-\mathbf{\Sigma})\mathbf{b}_{h}^{\star}\|_\infty \ge \epsilon) \le Cp\exp(-C'n\epsilon^2).$$

Then, since we assume that H is bounded from above, we have

$$\mathbb{P}(\|(\widehat{\mathbf{\Sigma}}-\mathbf{\Sigma})\mathbf{b}_h^{\star}\|_\infty \le \epsilon, \forall h) \ge 1 - CHp\exp(-C'n\epsilon^2) \ge 1 - Cp\exp(-C'n\epsilon^2), \quad 0<\epsilon\le C',$$

for some positive constants C and C'. $\square$

Lemma H.5. *Assume that Assumption (A3) holds. For any $k < p$ such that $ck\log(ep/k) \le n$, where the constant $c = 2(\log 81 + 1)m^2\{1 \vee (64m^3\kappa^2)^2\}$, we have that*

$$\mathbb{P}\{(2m)^{-1} \le \varphi^{\widehat{\mathbf{\Sigma}}}_{\min}(k) \le \varphi^{\widehat{\mathbf{\Sigma}}}_{\max}(k) \le 2m\} \ge 1 - 2\exp\{-(\log 9)k\}.$$

Proof of Lemma H.5. Under Assumption (A3), $\mathbf{X}_i$ is sub-Gaussian random vector with parameter κ^2, $i = 1, \ldots, n$. And by Lemma H.2, $m^{-1} \le \varphi_{\min}(\mathbf{\Sigma}) \le \varphi_{\max}(\mathbf{\Sigma}) \le m$. For any $\mathcal{A} \subseteq \{1, \ldots, p\}$ such that $|\mathcal{A}| = k$, according to Proposition E.7, for all $t \le 32m^2\kappa^2$, we have

$$\mathbb{P}(\|\widehat{\mathbf{\Sigma}}_{\mathcal{A}\mathcal{A}}\mathbf{\Sigma}^{-1}_{\mathcal{A}\mathcal{A}} - \mathbf{I}_k\| \ge t) \le 2\exp\left\{(\log 9)k - \frac{nt^2}{2(32m^2\kappa^2)^2}\right\}.$$

Since $m^{-1} \le \varphi_{\min}(\mathbf{\Sigma}) \le \varphi_{\max}(\mathbf{\Sigma}) \le m$, then $m^{-1} \le \varphi_{\min}(\mathbf{\Sigma}_{\mathcal{A}\mathcal{A}}) \le \varphi_{\max}(\mathbf{\Sigma}_{\mathcal{A}\mathcal{A}}) \le m$. It follows that, for $t \le 32m^2\kappa^2$,

$$\begin{aligned}\mathbb{P}(\|\widehat{\mathbf{\Sigma}}_{\mathcal{A}\mathcal{A}} - \mathbf{\Sigma}_{\mathcal{A}\mathcal{A}}\| \ge t) &\le \mathbb{P}(\|\widehat{\mathbf{\Sigma}}_{\mathcal{A}\mathcal{A}}\mathbf{\Sigma}^{-1}_{\mathcal{A}\mathcal{A}} - \mathbf{I}_k\| \ge t/m) \\ &\le 2\exp\left\{(\log 9)k - \frac{nt^2}{2(32m^3\kappa^2)^2}\right\}. \end{aligned} \tag{H.12}$$

We take $t = \sqrt{2(\log 81 + 1)(32m^3\kappa^2)^2\{n^{-1}k\log(ep/k)\}}$. Since we assume that $ck\log(ep/k) \le n$, where the constant $c = 2(\log 81 + 1)m^2\{1 \vee (64m^3\kappa^2)^2\}$, then

$$t \le \sqrt{\frac{2(\log 81 + 1)(32m^3\kappa^2)^2}{2(\log 81 + 1)m^2\{1 \vee (64m^3\kappa^2)^2\}}} = (32m^2\kappa^2) \wedge (2m)^{-1}.$$

By (H.12) and the fact that $\binom{p}{k} < (ep/k)^k$, we have

$$\begin{aligned}&\mathbb{P}\left(\sup_{\mathcal{A}\subseteq\{1,\ldots,p\},|\mathcal{A}|=k} \|\widehat{\mathbf{\Sigma}}_{\mathcal{A}\mathcal{A}} - \mathbf{\Sigma}_{\mathcal{A}\mathcal{A}}\| \ge \sqrt{2(\log 81 + 1)(32m^3\kappa^2)^2\{n^{-1}k\log(ep/k)\}}\right) \\ \le& 2(ep/k)^k\exp\{(\log 9)k - (\log 81 + 1)k\log(ep/k)\} \\ =& 2\exp\{k\log(ep/k) + (\log 9)k - (\log 81 + 1)k\log(ep/k)\} \\ \le& 2\exp\{-(\log 9)k\log(ep/k)\} \\ \le& 2\exp\{-(\log 9)k\}.\end{aligned}$$

By Weyl's inequality stated in Proposition E.1, we have

$$\varphi_{\min}(\mathbf{\Sigma}_{\mathcal{A}\mathcal{A}}) - \|\widehat{\mathbf{\Sigma}}_{\mathcal{A}\mathcal{A}} - \mathbf{\Sigma}_{\mathcal{A}\mathcal{A}}\| \le \varphi_{\min}(\widehat{\mathbf{\Sigma}}_{\mathcal{A}\mathcal{A}}) \le \varphi_{\max}(\widehat{\mathbf{\Sigma}}_{\mathcal{A}\mathcal{A}}) \le \varphi_{\max}(\mathbf{\Sigma}_{\mathcal{A}\mathcal{A}}) + \|\widehat{\mathbf{\Sigma}}_{\mathcal{A}\mathcal{A}} - \mathbf{\Sigma}_{\mathcal{A}\mathcal{A}}\|.$$

The definitions of $\varphi^{\widehat{\mathbf{\Sigma}}}_{\max}(k)$ and $\varphi^{\widehat{\mathbf{\Sigma}}}_{\min}(k)$ can be equivalently expressed as follows.

$$\begin{aligned}\varphi^{\widehat{\mathbf{\Sigma}}}_{\max}(k) &= \sup_{\mathcal{A}\subseteq\{1,\ldots,p\},|\mathcal{A}|=k} \varphi_{\max}(\widehat{\mathbf{\Sigma}}_{\mathcal{A}\mathcal{A}}), \\ \varphi^{\widehat{\mathbf{\Sigma}}}_{\min}(k) &= \inf_{\mathcal{A}\subseteq\{1,\ldots,p\},|\mathcal{A}|=k} \varphi_{\min}(\widehat{\mathbf{\Sigma}}_{\mathcal{A}\mathcal{A}}).\end{aligned}$$

Then, since $m^{-1} \le \varphi_{\min}(\mathbf{\Sigma}_{\mathcal{A}\mathcal{A}}) \le \varphi_{\max}(\mathbf{\Sigma}_{\mathcal{A}\mathcal{A}}) \le m$, we have

$$\varphi_{\min}^{\widehat{\mathbf{\Sigma}}}(k) \ge m^{-1} - (2m)^{-1} = (2m)^{-1}, \quad \varphi_{\max}^{\widehat{\mathbf{\Sigma}}}(k) \le m + (2m)^{-1} \le 2m$$

with probability at least $1 - 2\exp\{-(\log 9)k\log(ep/k)\}$, which completes the proof.

□

Part I: The proof of Corollary 3

The concentration result for $\widehat{\mathbf{U}}_{\mathrm{SIR}}$ is presented in the next lemma. Combining Lemmas H.4–H.6, Conditions (C1) and (C2) are satisfied in SIR and Corollary 3 is proved

Lemma H.6. *In SIR method, under Assumption (A3), we have*

$$\mathbf{U}_{\mathrm{SIR}} = \{\mathrm{E}(\mathbf{X}|J^1 = 1), \ldots, \mathrm{E}(\mathbf{X}|J^H = 1)\},\ \widehat{\mathbf{U}}_{\mathrm{SIR}} = (\bar{\mathbf{X}}_1, \ldots, \bar{\mathbf{X}}_H).$$

Then,

$$\mathbb{P}(\|\widehat{\mathbf{U}}_{\mathrm{SIR}} - \mathbf{U}_{\mathrm{SIR}}\|_{\max} \le \epsilon) \ge 1 - Cp\exp(-C'n\epsilon^2), \quad 0 < \epsilon \le C'$$

for some constants $C, C' > 0$.

Proof of Lemma H.6. We have

$$\begin{aligned}
&\mathbb{P}(\|\widehat{\mathbf{U}}_{\mathrm{SIR}} - \mathbf{U}_{\mathrm{SIR}}\|_{\max} > \epsilon/2) \le H\mathbb{P}\{\|\bar{\mathbf{X}}_h - \mathrm{E}(\mathbf{X}|J^h = 1)\|_\infty > \epsilon/2\}\\
=&H\mathbb{P}\{\|n_h^{-1}\sum_{i=1}^n \mathbf{X}_i J_i^h - \pi_h^{-1}\mathrm{E}(\mathbf{X}J^h)\|_\infty > \epsilon/2\}\\
\le&H\left[\mathbb{P}\{\|n_h^{-1}\sum_{i=1}^n \mathbf{X}_i J_i^h - (n/n_h)\mathrm{E}(\mathbf{X}J^h)\|_\infty > \epsilon/4\} + \mathbb{P}\{\|(n/n_h)\mathrm{E}(\mathbf{X}J^h) - \pi_h^{-1}\mathrm{E}(\mathbf{X}J^h)\|_\infty > \epsilon/4\}\right]\\
=:&H(L_1 + L_2),
\end{aligned} \tag{H.13}$$

where we use the fact that $\mathrm{E}(\mathbf{X}J^h) = \mathrm{E}\{\mathrm{E}(\mathbf{X}J^h|J^h)\} = \pi_h \mathrm{E}(\mathbf{X}|J^h = 1)$.

We first decompose L_1 into two parts,

$$\begin{aligned}
L_1 =&\mathbb{P}\{\|\sum_{i=1}^n \mathbf{X}_i J_i^h - n\mathrm{E}(\mathbf{X}J^h)\|_\infty > n_h\epsilon/4\}\\
\le&\mathbb{P}\{\|\sum_{i=1}^n \mathbf{X}_i J_i^h - n\mathrm{E}(\mathbf{X}J^h)\|_\infty > n_h\epsilon/4, n_h > \pi_h n/2\} + \mathbb{P}(n_h \le \pi_h n/2)\\
=:&L_{11} + L_{12}.
\end{aligned} \tag{H.14}$$

Since we assume that $a^{-1} \leq \pi_h \leq a$ for some constant $a > 0$, by Hoeffding's inequality stated in Lemma E.5, we have

$$L_{12} = \mathbb{P}(n_h - n\pi_h \leq -\pi_h n/2) \leq \exp(-n\pi_h^2/2) \leq \exp(-Cn). \tag{H.15}$$

for some positive constant C. For L_{11}, we have

$$L_{11} \leq \mathbb{P}\{\|\sum_{i=1}^{n} \mathbf{X}_i J_i^h - n\mathrm{E}(\mathbf{X}J^h)\|_\infty > \pi_h n\epsilon/8\} \leq p\mathbb{P}\{|\sum_{i=1}^{n} X_{ij} J_i^h - n\mathrm{E}(X_j J^h)| > \pi_h n\epsilon/8\}.$$

Since we assume that $\mathbf{X}$ is sub-Gaussian random vector with parameter $\kappa^2 \leq m/4$ in Assumption (A3) for some constant $m > 0$, then by Lemma E.6, $X_j \sim \mathrm{subG}(\kappa^2)$ for each j. According to Lemma E.4, we have

$$\mathrm{E}(|X_{ij}J_i^h|^k) \leq \mathrm{E}(|X_{ij}|^k) \leq (2\kappa^2)^{k/2} k\Gamma(k/2) \leq (m/2)^{k/2} k\Gamma(k/2), \quad \forall k \geq 1.$$

With the fact that $k\Gamma(k/2) \leq k!$ for $k \geq 2$, we have

$$\mathrm{E}(|X_{ij}J_i^h|^k) \leq (\sqrt{m/2})^{k-2} m\frac{k!}{2}, \quad \forall k \geq 2.$$

It follows that

$$\frac{1}{n}\sum_{i=1}^{n} \mathrm{E}(|X_{ij}J_i^h|^k) \leq \frac{k!}{2} v^2 r^{k-2}, \quad \forall k \geq 2,$$

where $v = \sqrt{m}$ and $r = \sqrt{m/2}$. Then according to Lemma E.2, we have

$$\mathbb{P}\{|\sum_{i=1}^{n} X_{ij} J_i^h - n\mathrm{E}(X_j J^h)| > \pi_h n\epsilon/8\} \leq 2\exp\left\{-\frac{n\epsilon^2\pi_h^2}{128(m + \sqrt{m/2}\epsilon)}\right\},$$

By taking small ϵ, we obtain

$$\mathbb{P}\{|\sum_{i=1}^{n} X_{ij} J_i^h - n\mathrm{E}(X_j J^h)| > \pi_h n\epsilon/8\} \leq C\exp(-C' n\epsilon^2).$$

for some positive constants C and C', where we use the assumption that $a^{-1} \leq \pi_h \leq a$ for some constant $a > 0$. Therefore, we have

$$L_{11} \leq Cp\exp(-C' n\epsilon^2). \tag{H.16}$$

Combining (H.14)(H.15)(H.16), we obtain the upper bound for L_1 as

$$L_1 \leq Cp\exp(-C' n\epsilon^2). \tag{H.17}$$

Next, we derive the upper bound for L_2. According to Lemma E.4, we have

$$|\mathrm{E}(X_j J^h)| \leq \mathrm{E}(|X_j|) \leq (2\kappa^2)^{1/2}\Gamma(1/2) \leq \sqrt{m\pi/2}, \quad j = 1, \ldots, p,$$

then $\|\mathrm{E}(\mathbf{X}J^h)\|_\infty \le \sqrt{m\pi/2}$. It follows that,

$$\begin{aligned}
L_2 &\le \mathbb{P}(|n/n_h - \pi_h^{-1}| \cdot \sqrt{m\pi/2} > \epsilon/4) \\
&= \mathbb{P}\{|n/n_h - \pi_h^{-1}| > (4\sqrt{m\pi/2})^{-1}\epsilon\} \\
&= \mathbb{P}\left(n_h < \frac{n}{\pi_h^{-1} + (4\sqrt{m\pi/2})^{-1}\epsilon}\right) + \mathbb{P}\left(n_h > \frac{n}{\pi_h^{-1} - (4\sqrt{m\pi/2})^{-1}\epsilon}\right) \\
&= L_{21} + L_{22}. \qquad \text{(H.18)}
\end{aligned}$$

According to Hoeffding's inequality, since we assume that $a^{-1} \le \pi_h \le a$ for some constant $a > 0$, by taking small ϵ, we have

$$\begin{aligned}
L_{21} &= \mathbb{P}\left(n_h - n\pi_h < -n\left\{\pi_h - \frac{1}{\pi_h^{-1} + (4\sqrt{m\pi/2})^{-1}\epsilon}\right\}\right) \\
&\le \exp\left(-2n\left[\frac{(4\sqrt{m\pi/2})^{-1}\epsilon}{\pi_h^{-1}\{\pi_h^{-1} + (4\sqrt{m\pi/2})^{-1}\epsilon\}}\right]^2\right) \le \exp(-Cn\epsilon^2). \qquad \text{(H.19)}
\end{aligned}$$

for some positive constant C. Similarly, for L_{22},

$$L_{22} = \mathbb{P}\left[n_h - n\pi_h > n\left\{\frac{1}{\pi_h^{-1} - (4\sqrt{m\pi/2})^{-1}\epsilon} - \pi_h\right\}\right] \le \exp(-Cn\epsilon^2). \qquad \text{(H.20)}$$

Thus, combining (H.18)(H.19)(H.20), we obtain the upper bound for L_2 as

$$L_2 \le C\exp(-C'n\epsilon^2). \qquad \text{(H.21)}$$

for some positive constants C and C'.

Combining (H.13)(H.17)(H.21), we obtain the concentration result for $\widehat{\mathbf{U}}_{\mathrm{SIR}}$ that

$$\mathbb{P}(\|\widehat{\mathbf{U}}_{\mathrm{SIR}} - \mathbf{U}_{\mathrm{SIR}}\|_{\max} \le \epsilon) \ge 1 - Cp\exp(-C'n\epsilon^2), \quad 0 < \epsilon \le C'$$

for some positive constants C and C'.

□

Part III: The proof of Corollary 4

In the following Lemma, we present the concentration result for $\widehat{\mathbf{U}}_{\mathrm{Intra}}$. Combining Lemma H.4, Lemma H.5 and Lemma H.7, Conditions (C1) and (C2) are satisfied in intraslice convariance method and Corollary 4 is proved.

Lemma H.7. *In intraslice covariance method, under Assumptions (A3) and (A5), we have*

$$\mathbf{U}_{\mathrm{Intra}} = \{\mathrm{Cov}(\mathbf{X}, YJ^1), \dots, \mathrm{Cov}(\mathbf{X}, YJ^H)\},\ \widehat{\mathbf{U}}_{\mathrm{Intra}} = \{\widehat{\mathrm{Cov}}(\mathbf{X}, YJ^1), \dots, \widehat{\mathrm{Cov}}(\mathbf{X}, YJ^H)\}.$$

Then,

$$\mathbb{P}(\|\widehat{\mathbf{U}}_{\mathrm{Intra}} - \mathbf{U}_{\mathrm{Intra}}\|_{\max} > \epsilon) \leq Cp\exp(-C'n\epsilon^2), \quad 0 < \epsilon \leq C'$$

for some constants $C, C' > 0$.

Proof of Lemma H.7. Since we assume that $\mathbf{X}_i = \mathbf{0}$, $i = 1, \dots, n$, in Assumption (A3), then $\mathrm{Cov}(\mathbf{X}, YJ^h) = \mathrm{E}(\mathbf{X}YJ^h)$ and $\widehat{\mathrm{Cov}}(\mathbf{X}, YJ^h) = n^{-1}\sum_{i=1}^n \mathbf{X}_i Y_i J_i^h$. We have

$$\mathbb{P}(\|\widehat{\mathbf{U}}_{\mathrm{Intra}} - \mathbf{U}_{\mathrm{Intra}}\|_{\max} > \epsilon) \leq \sum_{j=1}^{p}\sum_{h=1}^{H}\mathbb{P}\{|\sum_{i=1}^{n} X_{ij}Y_iJ_i^h - n\mathrm{E}(X_jYJ^h)| > n\epsilon\}.$$

By Cauchy-Schwartz inequality, we have

$$\mathrm{E}(|X_{ij}Y_iJ_i^h|^k) \leq \mathrm{E}(|X_{ij}|^{2k})^{1/2} \cdot \mathrm{E}(|Y_i|^{2k})^{1/2}, \quad i = 1, \dots, n,\ j = 1, \dots, p,\ h = 1, \dots, H.$$

Since we assume that $\mathbf{X}$ is sub-Gaussian random vector with parameter $\kappa^2 \leq m/4$ in Assumption (A3) for some constant $m > 0$, then by Lemma E.6, $X_j \sim \mathrm{subG}(\kappa^2)$ for each j. Also, we assume that $Y \sim \mathrm{subG}(\theta^2)$ in Assumption (A5). According to Lemma E.4, we have

$$\begin{aligned}
\mathrm{E}(|X_{ij}|^{2k})^{1/2} &\leq (2\kappa^2)^{k/2}(2k)^{1/2}\{\Gamma(k)\}^{1/2}, \\
\mathrm{E}(|Y_i|^{2k})^{1/2} &\leq (2\theta^2)^{k/2}(2k)^{1/2}\{\Gamma(k)\}^{1/2}.
\end{aligned}$$

It follows that,

$$\mathrm{E}(|X_{ij}Y_iJ_i^h|^k) \leq \frac{k!}{2}(\sqrt{4mG})^2(\sqrt{mG})^{k-2}, \quad \forall k \geq 1.$$

Then we have

$$\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}(|X_{ij}Y_iJ_i^h|^k) \leq \frac{k!}{2}v^2r^{k-2}, \quad \forall k \geq 1,$$

where $v = \sqrt{4mG}$ and $r = \sqrt{mG}$. According to Lemma E.2, we have

$$\mathbb{P}\{|\sum_{i=1}^{n} X_{ij}Y_iJ_i^h - n\mathrm{E}(X_jYJ^h)| > n\epsilon\} \leq 2\exp\left\{-\frac{n\epsilon^2}{2(4mG + \sqrt{mG}\epsilon)}\right\}.$$

By taking small ϵ, we obtain

$$\mathbb{P}\{|\sum_{i=1}^{n} X_{ij}Y_iJ_i^h - n\mathrm{E}(X_jYJ^h)| > n\epsilon\} \leq C\exp(-C'n\epsilon^2)$$

for some constants $C, C' > 0$. Since we assume that H is bounded from above, then

$$\mathbb{P}(\|\widehat{\mathbf{U}}_{\mathrm{Intra}} - \mathbf{U}_{\mathrm{Intra}}\|_{\max} > \epsilon) \leq CpH\exp(-C'n\epsilon^2) \leq Cp\exp(-C'n\epsilon^2), \quad 0 < \epsilon \leq C'$$

for some constants $C, C' > 0$, which completes the proof.

□

Part IV: The proof of Corollary 5

We first show that under PFC model, Assumption (A3) is implied by Assumption (A6). The result is given in the following lemma.

Lemma H.8. *In PFC model, Assumption (A6) implies Assumption (A3).*

Proof of Lemma H.8. Under Assumption (A6), we have $\mathrm{E}\{\boldsymbol{f}(Y_i)\} = \mathbf{0}$, $i = 1, \ldots, n$. Then $\mathbf{X}_i = \mathbf{0}$ and PFC model is

$$\mathbf{X}_i = \boldsymbol{\Gamma}\boldsymbol{\eta}\boldsymbol{f}(Y_i) + \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, n. \tag{H.22}$$

Since $\boldsymbol{\epsilon}_i \sim N(\mathbf{0}, \boldsymbol{\Delta})$ in PFC model, by Lemma E.7, $\boldsymbol{\epsilon}_i$ is sub-Gaussian random vector with parameter $\|\boldsymbol{\Delta}\|$. We also assume that $\boldsymbol{f}(Y_i)$ is sub-Gaussian random vector with bounded parameter ξ^2 in Assumption (A6). Since $\boldsymbol{\epsilon}_i$ is independent of Y_i, then for any $\mathbf{a} \in \mathbb{R}^p$, we have

$$\begin{aligned}
\mathrm{E}\{\exp(\mathbf{a}^\top \mathbf{X}_i)\} &= \mathrm{E}\{\exp(\mathbf{a}^\top \boldsymbol{\Gamma}\boldsymbol{\eta}\boldsymbol{f}(Y_i))\}\mathrm{E}\{\exp(\mathbf{a}^\top \boldsymbol{\epsilon})\} \\
&\leq \exp(\|\mathbf{a}^\top \boldsymbol{\Gamma}\boldsymbol{\eta}\|_2^2 \xi^2/2) \exp(\|\mathbf{a}\|_2^2 \|\boldsymbol{\Delta}\|/2) \\
&\leq \exp(\|\mathbf{a}\|_2^2 \|\boldsymbol{\eta}\|^2 \xi^2/2 + \|\mathbf{a}\|_2^2 \|\boldsymbol{\Delta}\|/2) \\
&= \exp\{\|\mathbf{a}\|_2^2 (\|\boldsymbol{\eta}\|^2 \xi^2 + \|\boldsymbol{\Delta}\|)/2\}.
\end{aligned}$$

Therefore, $\mathbf{X}_i$ is sub-Gaussian random vector with parameter $\kappa^2 = \|\boldsymbol{\eta}\|^2\xi^2 + \|\boldsymbol{\Delta}\|$. Since we assume that ξ^2, $\|\boldsymbol{\eta}\|$ and $\varphi_{\max}(\boldsymbol{\Delta})$ are bounded from above, the parameter κ^2 is bounded from above.

In PFC model (H.22), the covariance of $\mathbf{X}_i$ is expressed as

$$\boldsymbol{\Sigma} = \mathrm{Cov}(\mathbf{X}_i) = \boldsymbol{\Gamma}\boldsymbol{\eta}\mathrm{Cov}(\boldsymbol{f}(Y_i))\boldsymbol{\eta}^\top \boldsymbol{\Gamma}^\top + \boldsymbol{\Delta},$$

then we have

$$\varphi_{\min}(\boldsymbol{\Sigma}) \geq \varphi_{\min}(\boldsymbol{\Gamma}\boldsymbol{\eta}\mathrm{Cov}(\boldsymbol{f}(Y_i))\boldsymbol{\eta}^\top \boldsymbol{\Gamma}^\top) + \varphi_{\min}(\boldsymbol{\Delta}) \geq g^{-1}$$

for some constant $g > 0$, where we use the assumption that $g^{-1} \leq \varphi_{\min}(\boldsymbol{\Delta})$ in Assumption (A6) and the fact that the matrix $\boldsymbol{\Gamma}\boldsymbol{\eta}\mathrm{Cov}(\boldsymbol{f}(Y_i))\boldsymbol{\eta}^\top \boldsymbol{\Gamma}^\top$ is rank-deficient. Thus, Assumption (A3) holds.

□

Next, we present the concentration result for $\widehat{\mathbf{U}}_{\mathrm{PFC}}$. Then Conditions (C1) and (C2) in PFC model are satisfied by combining Lemma H.4, Lemma H.5 and Lemma H.9. And Corollary 5 is proved.

Lemma H.9. *In PFC model, under Assumption (A6), we have*

$$\mathbf{U}_{\text{PFC}} = \text{Cov}\{\mathbf{X}, \boldsymbol{f}(Y)\}, \ \widehat{\mathbf{U}}_{\text{PFC}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i \boldsymbol{f}^\top(Y_i).$$

Then,

$$\mathbb{P}(\|\widehat{\mathbf{U}}_{\text{PFC}} - \mathbf{U}_{\text{PFC}}\|_{\max} > \epsilon/2) \leq Cp\exp\left(-C'n\epsilon^2\right), \quad 0 < \epsilon \leq C'$$

for some constants $C, C' > 0$.

Proof of Lemma H.9. We have

$$\mathbb{P}(\|\widehat{\mathbf{U}}_{\text{PFC}} - \mathbf{U}_{\text{PFC}}\|_{\max} > \epsilon) \leq \sum_{j=1}^{p}\sum_{h=1}^{H} \mathbb{P}\left[\left|\sum_{i=1}^{n} X_{ij} f_h(Y_i) - n\text{E}\{X_j f_h(Y)\}\right| > n\epsilon\right].$$

By Cauchy-Schwartz inequality,

$$\text{E}\{|X_{ij} f_h(Y_i)|^k\} \leq \text{E}(|X_{ij}|^{2k})^{1/2} \cdot \text{E}\{|f_h(Y_i)|^{2k}\}^{1/2}, \quad i = 1, \dots, n, \ j = 1, \dots, p, \ h = 1, \dots, H.$$

By Lemma H.8, Assumption (A3) is implied by Assumption (A6). Since we assume that $\mathbf{X}$ is sub-Gaussian random vector with parameter $\kappa^2 \leq m/4$ in Assumption (A3) for some constant $m > 0$ and $\boldsymbol{f}(Y)$ is sub-Gaussian random vector with parameter $\xi^2 \leq L$ for some constant $L > 0$ in Assumption (A6), then by Lemma E.6, $X_j \sim \text{subG}(\kappa^2)$ and $f_h(Y) \sim \text{subG}(\xi^2)$ for each j and h. According to Lemma E.4, we have

$$\begin{aligned} \text{E}(|X_{ij}|^{2k})^{1/2} &\leq (2\kappa^2)^{k/2}(2k)^{1/2}\{\Gamma(k)\}^{1/2}, \\ \text{E}\{|f_h(Y_i)|^{2k}\}^{1/2} &\leq (2\xi^2)^{k/2}(2k)^{1/2}\{\Gamma(k)\}^{1/2}. \end{aligned}$$

It follows that

$$\text{E}\{|X_{ij} f_h(Y_i)|^k\} \leq \frac{k!}{2}(\sqrt{4mL})^2(\sqrt{mL})^{k-2}, \quad \forall k \geq 1,$$

and

$$\frac{1}{n}\sum_{i=1}^{n} \text{E}\{|X_{ij} f_h(Y_i)|^k\} \leq \frac{k!}{2}v^2 r^{k-2}, \quad \forall k \geq 1,$$

where $v = \sqrt{4mL}$ and $r = \sqrt{mL}$. By Lemma E.2, we have

$$\mathbb{P}\left\{\left|\sum_{i=1}^{n} X_{ij} f_h(Y_i) - n\text{E}(X_j f_h(Y))\right| > n\epsilon\right\} \leq 2\exp\left\{-\frac{n\epsilon^2}{2(4mL + \sqrt{mL}\epsilon)}\right\}.$$

By taking small ϵ, we obtain

$$\mathbb{P}\left\{\left|\sum_{i=1}^{n} X_{ij} f_h(Y_i) - n\text{E}(X_j f_h(Y))\right| > n\epsilon\right\} \leq C\exp(-C'n\epsilon^2)$$

for some constants $C, C' > 0$. Since we assume that H is bounded from above, then

$$\mathbb{P}(\|\widehat{\mathbf{U}}_{\mathrm{PFC}} - \mathbf{U}_{\mathrm{PFC}}\|_{\max} > \epsilon) \leq CpH \exp\left(-C'n\epsilon^2\right) \leq Cp \exp\left(-C'n\epsilon^2\right), \quad 0 < \epsilon \leq C'$$

for some constants $C, C' > 0$, which completes the proof.

□

H.8 Proof of Corollaries C.1–C.3

By Lemma E.7, Assumption (A8) implies that $\mathbf{X}$ is sub-Gaussian random vector with parameter m, and hence implies Assumption (A3). In the following, we prove Corollaries C.1–C.3 by listing the necessary assumptions for each specific SEAS method.

Part I: The proof of Corollary C.1

According to Lemmas H.4–H.6, Conditions (C1) and (C2) are satisfied under Assumptions (A3) and (A4) for SEAS-SIR. Meanwhile, Assumption (A3) is implied directly by Assumption (A8). Then, by Theorem C.1, we obtain the variable selection consistency for SEAS-SIR under Assumptions (A1), (A2), (A4) and (A8)–(A10). □

Part II: The proof of Corollary C.2

According to Lemma H.4, Lemma H.5 and Lemma H.7, Conditions (C1) and (C2) are satisfied under Assumptions (A3)–(A5) for SEAS-Intra. Meanwhile, Assumption (A3) is implied directly by Assumption (A8). Then, by Theorem C.1, we obtain the variable selection consistency for SEAS-Intra under Assumptions (A1), (A2), (A4), (A5) and (A8)–(A10). □

Part III: The proof of Corollary C.3

According to Lemma H.4, Lemma H.5 and Lemma H.9, Conditions (C1) and (C2) are satisfied under Assumptions (A4) and (A6) for SEAS-PFC. Therefore, by Theorem C.1, the variable selection consistency for SEAS-PFC is established under Assumptions (A1), (A2), (A4), (A6) and (A8)–(A10). □

References

Bach, F. R. (2008). Consistency of the group lasso and multiple kernel learning. *Journal of Machine Learning Research*, 9(6).

Birgé, L. and Massart, P. (1998). Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli*, 4(3):329–375.

Cai, T. T., Zhang, A., and Zhou, Y. (2019). Sparse group lasso: Optimal sample complexity, convergence rate, and statistical inference. *arXiv preprint arXiv:1909.09851*.

Chen, X., Cook, R. D., and Zou, C. (2015). Diagnostic studies in sufficient dimension reduction. *Biometrika*, 102(3):545–558.

Cook, R. D. and Zhang, X. (2014). Fused estimators of the central subspace in sufficient dimension reduction. *Journal of the American Statistical Association*, 109(506):815–827.

Cook, R. D. and Zhang, X. (2015). Simultaneous envelopes for multivariate linear regression. *Technometrics*, 57(1):11–25.

Deng, W. and Yin, W. (2016). On the global and linear convergence of the generalized alternating direction method of multipliers. *Journal of Scientific Computing*, 66(3):889–916.

Golub, G. H. and Van Loan, C. F. (2013). *Matrix computations*, volume 3. JHU press.

Hoffman, A. J. and Wielandt, H. W. (1953). The variation of the spectrum of a normal matrix. *Duke Math. J.*, 20(1):37–39.

Horn, R. A. and Johnson, C. R. (2012). *Matrix analysis*. Cambridge university press.

Laurent, B. and Massart, P. (2000). Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pages 1302–1338.

Lin, Q., Li, X., Huang, D., and Liu, J. S. (2021). On the optimality of sliced inverse regression in high dimensions. *The Annals of Statistics*, 49(1):1–20.

Mai, Q., Yang, Y., and Zou, H. (2019). Multiclass sparse discriminant analysis. *Statistica Sinica*, 29:97–111.

Mai, Q., Zou, H., and Yuan, M. (2012). A direct approach to sparse discriminant analysis in ultra-high dimensions. *Biometrika*, 99(1):29–42.

Sæbø, S., Almøy, T., Aarøe, J., and Aastveit, A. H. (2008). St-pls: a multi-directional nearest shrunken centroid type classifier via pls. *Journal of Chemometrics: A Journal of the Chemometrics Society*, 22(1):54–62.

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.

Tan, K., Shi, L., and Yu, Z. (2020). Sparse sir: Optimal rates and adaptive estimation. *The Annals of Statistics*, 48(1):64–85.

Tseng, P. (2001). Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of optimization theory and applications*, 109(3):475–494.

Uematsu, Y., Fan, Y., Chen, K., Lv, J., and Lin, W. (2019). Sofar: large-scale association network learning. *IEEE transactions on information theory*, 65(8):4924–4939.

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press.

Yu, Y., Wang, T., and Samworth, R. J. (2015). A useful variant of the davis–kahan theorem for statisticians. *Biometrika*, 102(2):315–323.

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

Zhou, H. and Li, L. (2014). Regularized matrix regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(2):463–483.

Zhu, L.-X. and Ng, K. W. (1995). Asymptotics of sliced inverse regression. *Statistica Sinica*, pages 727–736.