Statistica Sinica: Supplement

Supplementary Materials to "Distributed Sufficient Dimension Reduction for Heterogeneous Massive Data"

Kelin Xu, Liping Zhu, Jianqing Fan

Fudan University, Renmin University of China, Princeton University

Abstract: In the Supplement, we study three additional issues. Firstly, we propose a distributed algorithm when the central subspaces at the local nodes are distinctive from each other, which takes the advantages of the low-rank structure of the kernel matrices. Secondly, we demonstrate the distributed algorithm may outperform its pooled version in the presence of heterogeneity. A bootstrap procedure is also introduced to make a fair comparison. Finally, we provide technical details, such as proofs of theorems and some useful lemmas, in this Supplement.

S1 When the Local Central Subspaces are Distinctive

Throughout the main context, we assume the central subspaces at all local nodes are identical. To be specific, in model (1.2), we assume a common basis, $\mathbf{B}$, is shared by all m local nodes. In some situations, this is perhaps unrealistic. In this Supplement, we allow the central subspaces at the local nodes to be distinctive from each other. In symbols, we assume that:

$$F_j(Y_{i,j} \mid \mathbf{x}_{i,j}) = F_j(Y_{i,j} \mid \mathbf{B}_j^{\mathrm{T}}\mathbf{x}_{i,j}), \text{ for } \ i = 1, \ldots, n, \ j = 1, \ldots, m. \quad \text{(S1.1)}$$

Here, $\mathbf{B}_j$ is a $p \times d_j$ matrix. By the very purposes of dimension reduction, d_j is much smaller than p. Recall the definition of $\mathbf{\Omega}_j$ in Section 2 of the main context. It follows from Li (1991) and Zhu et al. (2010) that $\mathrm{span}(\mathbf{\Omega}_j) \subseteq \mathrm{span}(\mathbf{B}_j)$ under the linearity condition. An important observation is that $\mathbf{\Omega}_j$ is a low-rank matrix, which has at most d_j nonzero eigenvalues. Taking the advantage of this low-rank structure, we can recover $\mathbf{\Omega}_j$ from its principal eigenvectors and the associate eigenvalues. With slight abuse of notations, we let $\mathbf{B}_j \in \mathbb{R}^{p\times d_j}$ be the d_j principal eigenvectors. In addition, lets $\mathbf{\Lambda}_j \in \mathbb{R}^{d_j\times d_j}$ be a diagonal matrix with its diagonal elements being the nonzero eigenvalues of $\mathbf{\Omega}_j$. It follows immediately that $\mathbf{\Omega}_j = \mathbf{B}_j\mathbf{\Lambda}_j\mathbf{B}_j^{\mathrm{T}}$. Therefore, it suffices for a distributed algorithm to transfer the principal eigenvectors and the nonzero eigenvalues to the central node.

Our proposed distributed algorithm proceeds as follows.

Algorithm 3

1. Estimate $\mathbf{\Omega}_j$ at the jth local node, which yields $\widehat{\mathbf{\Omega}}_j$. We can simply use the dense estimate $\widehat{\mathbf{\Omega}}_j$ in Algorithm 1, if all covariance matrices

are invertible, or the sparse estimate $\widehat{\mathbf{\Omega}}_j$ in Algorithm 2 otherwise.

2. Apply a certain criterion to decide the rank of $\widehat{\mathbf{\Omega}}_j$ at the jth local node, which yields $\widehat{d}_j$. In our subsequent illustration, we simply use the maximum eigenvalue ratio criterion of Luo et al. (2009).

3. Apply singular value decomposition to $\widehat{\mathbf{\Omega}}_j$ to obtain its top $\widehat{d}_j$ eigenvectors $\widehat{\mathbf{B}}_{a3,j} \in \mathbb{R}^{p\times \widehat{d}_j}$ and the associated eigenvalues, which are the diagonal elements of $\widehat{\mathbf{\Lambda}}_{a3,j} \in \mathbb{R}^{\widehat{d}_j \times \widehat{d}_j}$. We approximate $\widehat{\mathbf{\Omega}}_j$ with $\widehat{\mathbf{B}}_{a3,j}\widehat{\mathbf{\Lambda}}_{a3,j}\widehat{\mathbf{B}}^{\mathrm{T}}_{a3,j}$.

4. Recall that $\widehat{\mathbf{\Lambda}}_{a3,j}$ is a diagonal matrix. Pass $\widehat{\mathbf{B}}_{a3,j}$ and the diagonal elements of $\widehat{\mathbf{\Lambda}}_{a3,j}$ to the central node to form

$$\widehat{\mathbf{T}}_{a3} \stackrel{\text{def}}{=} m^{-1}\sum_{j=1}^{m}\widehat{\mathbf{B}}_{a3,j}\widehat{\mathbf{\Lambda}}_{a3,j}\widehat{\mathbf{B}}^{\mathrm{T}}_{a3,j} \tag{S1.2}$$

   The communication cost in this step is

$$(p+1)\sum_{j=1}^{m} d_j.$$

5. Apply singular value decomposition to $\widehat{\mathbf{T}}_{a3}$ to obtain the first d_0 top eigenvectors, which yields $\widehat{\mathbf{B}}_{a3}$. If d_0 is unknown, we can again apply a certain criterion, say, Luo et al. (2009), to decide the rank of $\widehat{\mathbf{T}}_{a3}$.

We demonstrate the finite-sample performance of the above distributed algorithm through simulations.

Example 4: We illustrate the performance of Algorithm 3 in this simulated example. We fix $p = 200$, and draw $\mathbf{x} = (X_1, \ldots, X_p)^{\mathrm{T}} \in \mathbb{R}^p$ from multivariate normal distribution with mean zero and covariance matrix $\boldsymbol{\Sigma} = (\rho^{|k-l|})_{p\times p}$. We set $\rho = 0.5$. At each local node, we generate Y from the following models with equal probability 1/2:

$$
\begin{aligned}
Y &= \sin(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x}) + \varepsilon, \\
Y &= \exp(\boldsymbol{\beta}_2^{\mathrm{T}}\mathbf{x}) + \varepsilon,
\end{aligned}
$$

We generate the error term ε from standard normal distribution. We set $\boldsymbol{\beta}_1 = (1, 1, 0, \ldots, 0)^{\mathrm{T}} \in \mathbb{R}^p$ and $\boldsymbol{\beta}_2 = (0, 0, 1, 1, 0, \ldots, 0)^{\mathrm{T}} \in \mathbb{R}^p$. Let $m = \{2^5, 2^6, 2^7, 2^8\}$ and $N = \{2^9p, 2^{10}p, 2^{11}p, 2^{12}p\}$. All the N observations are scattered uniformly across m nodes each of size n.

We implement Algorithm 3 for both sliced inverse regression and cumulative slicing estimation under the case when all sample covariance matrices are invertible. In other words, we use the dense estimate $\widehat{\boldsymbol{\Omega}}_j$.

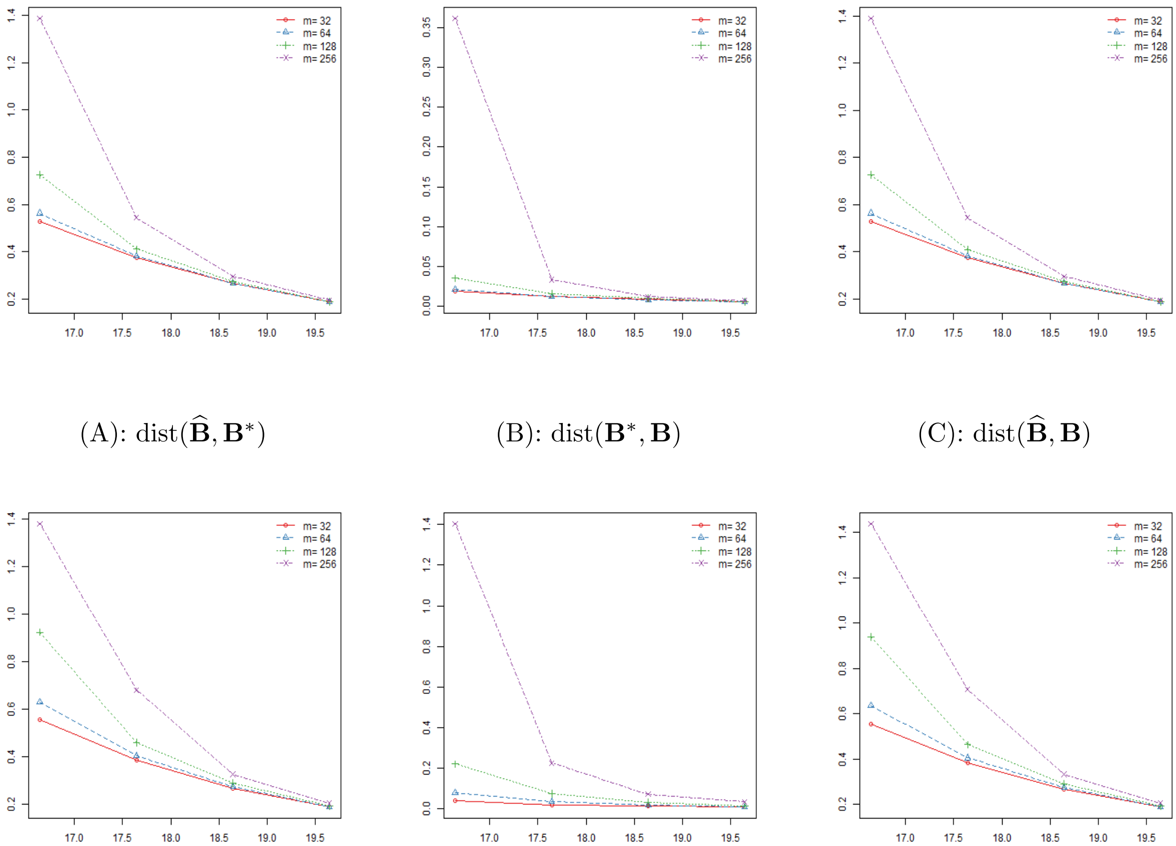
Example 5: We generate the observations in the same way as in Example 6, except for $\rho = 0.8$ in $\boldsymbol{\Sigma} = (\rho^{|k-l|})_{p\times p}$. We implement Algorithm 3 for both sliced inverse regression and cumulative slicing estimation under the case when not all sample covariance matrices are invertible. In other words, we use the sparse estimate $\widehat{\boldsymbol{\Omega}}_j$.

In the above two examples, $\mathbf{B} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) \in \mathbb{R}^{p\times 2}$. We repeat each simulation 1000 times, and report $\mathrm{dist}(\widehat{\mathbf{B}}, \mathbf{B}^*)$, $\mathrm{dist}(\mathbf{B}^*, \mathbf{B})$ and $\mathrm{dist}(\widehat{\mathbf{B}}, \mathbf{B})$ to evaluate the performance of distributed estimates. The simulation results are summarized in Figures 4 and 5 for Examples 4 and 5, respectively.

Figures 4 and 5 deliver similar messages. In both examples, cumulative slicing estimation outperforms sliced inverse regression, and the performance of the latter depends on the number of slices.

S2 American Gut Project Revisited

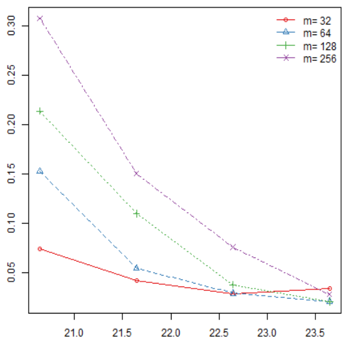
We revisit the American Gut Project in Section 4 of the main context. We explore how to implement a bootstrap procedure in the presence of heterogeneity. We compare two versions of bootstrap methods. There are three steps for both procedures. In the firs step, we bootstrap new observations at each local node. To implement the first version of bootstrap, we perform dimension reduction on the locally bootstrapped observations in the second step, and aggregate the dimension reduction results at all local nodes in the third step. To implement the second version of bootstrap, we pool the locally bootstrapped observations together to form a complete bootstrap sample in the second step, and perform dimension reduction on the complete bootstrap sample in the third step. We replicate the above procedures 100 times, and compare the distances (1.4) of the central sub-



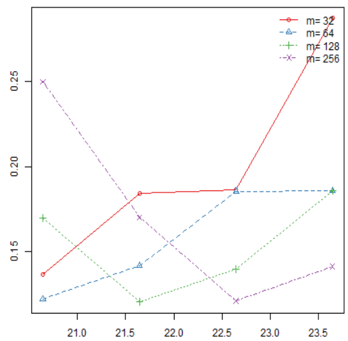
(A): dist($\widehat{\mathbf{B}}, \mathbf{B}^*$) (B): dist($\mathbf{B}^*, \mathbf{B}$) (C): dist($\widehat{\mathbf{B}}, \mathbf{B}$)

(D): dist($\widehat{\mathbf{B}}, \mathbf{B}^*$) (E): dist($\mathbf{B}^*, \mathbf{B}$) (F): dist($\widehat{\mathbf{B}}, \mathbf{B}$)

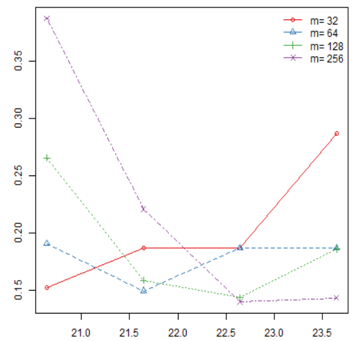
Figure 4: The horizontal axis stands for the log(2)-transformed value of the total sample size N, and the vertical axis stands for dist($\widehat{\mathbf{B}}, \mathbf{B}^*$) in (A) and (D), dist($\mathbf{B}^*, \mathbf{B}$) in (B) and (E), and dist($\widehat{\mathbf{B}}, \mathbf{B}$) in (C) and (F). All the distributed estimates of $\mathbf{B}$ are obtained through Algorithm 3. The distributed estimates of sliced inverse regression are displayed in the subplots (A)-(C) and those of cumulative slicing estimation are displayed in the subplots (D)-(F).



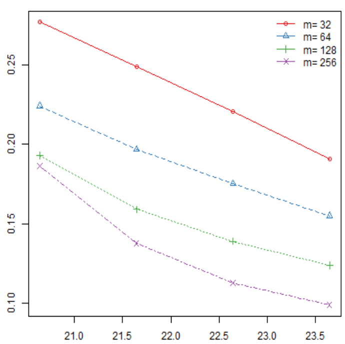
(A): $\text{dist}(\widehat{\mathbf{B}}, \mathbf{B}^*)$



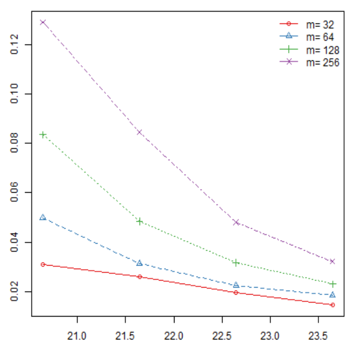
(B): $\text{dist}(\mathbf{B}^*, \mathbf{B})$



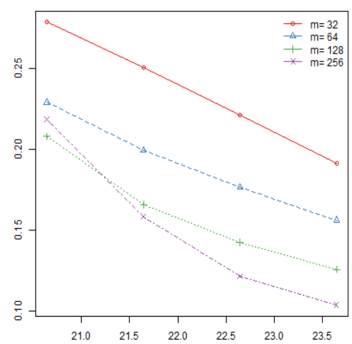
(C): $\text{dist}(\widehat{\mathbf{B}}, \mathbf{B})$



(D): $\text{dist}(\widehat{\mathbf{B}}, \mathbf{B}^*)$



(E): $\text{dist}(\mathbf{B}^*, \mathbf{B})$



(F): $\text{dist}(\widehat{\mathbf{B}}, \mathbf{B})$

Figure 5: The horizontal axis stands for the log(2)-transformed value of the total sample size N, and the vertical axis stands for $\text{dist}(\widehat{\mathbf{B}}, \mathbf{B}^*)$ in (A) and (D), $\text{dist}(\mathbf{B}^*, \mathbf{B})$ in (B) and (E), and $\text{dist}(\widehat{\mathbf{B}}, \mathbf{B})$ in (C) and (F). All the distributed estimates of $\mathbf{B}$ are obtained through Algorithm 2. The distributed estimates of sliced inverse regression are displayed in the subplots (A)-(C) and those of cumulative slicing estimation are displayed in the subplots (D)-(F).

spaces obtained from the bootstrap samples and the original observations. The simulation results are summarized in Figure 6 in this Supplementary Material. It can be clearly seen that, the first version of bootstrap, which in spirit corresponds to the distributed dimension reduction, is much more stable than the second version of bootstrap, which indeed yields a pooled dimension reduction. This indicates that, in the presence of heterogeneity, the distributed dimension reduction methods are perhaps more advantageous than the pooled ones.

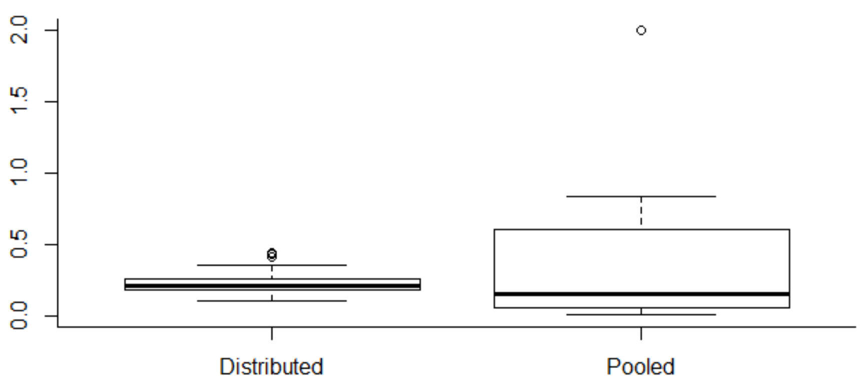


Figure 6: The distances between the estimates obtained by using the bootstrapped samples and those obtained by using the original observations. The left boxplot stands for the distributed estimates and the right one corresponds to the pooled estimates.

S3 Some Useful Lemmas

We first provide some lemmas that pave the road to prove Theorems 1 and 2 . Define the spectral norm $\|\mathbf{A}\| \overset{\text{def}}{=} \lambda_{\max}^{1/2}(\mathbf{A}^{\top}\mathbf{A})$, where $\lambda_{\max}$ stands for the maximum eigenvalue.

Lemma 5. *In addition to Conditions (C1)-(C3), we assume the sample covariance matrices $\widehat{\boldsymbol{\Sigma}}_j^{-1}$s are all invertible. Then there exists an absolute positive constant C such that*

$$\|\|\widehat{\boldsymbol{\Omega}}_{a1,j} - \boldsymbol{\Omega}_j\|\|_{\psi_1} \leq C(p/n)^{1/2}, \text{ for } j = 1, \dots, m.$$

Proof of Lemma 5: Recall that $\widehat{\boldsymbol{\Omega}}_{a1,j} \overset{\text{def}}{=} \widehat{\boldsymbol{\Sigma}}_j^{-1}\widehat{\mathbf{M}}_j\widehat{\boldsymbol{\Sigma}}_j^{-1}$ and $\boldsymbol{\Omega}_j \overset{\text{def}}{=} \boldsymbol{\Sigma}_j^{-1}\mathbf{M}_j\boldsymbol{\Sigma}_j^{-1}$. We further define $Q_1 \overset{\text{def}}{=} \|\widehat{\boldsymbol{\Omega}}_{a1,j} - \widehat{\boldsymbol{\Sigma}}_j^{-1}\widehat{\mathbf{M}}_j\boldsymbol{\Sigma}_j^{-1}\|$, $Q_2 \overset{\text{def}}{=} \|\widehat{\boldsymbol{\Sigma}}_j^{-1}(\widehat{\mathbf{M}}_j - \mathbf{M}_j)\boldsymbol{\Sigma}_j^{-1}\|$, and $Q_3 \overset{\text{def}}{=} \|\widehat{\boldsymbol{\Sigma}}_j^{-1}\mathbf{M}_j\boldsymbol{\Sigma}_j^{-1} - \boldsymbol{\Omega}_j\|$. It follows immediately that $Q_1 = \|\widehat{\boldsymbol{\Sigma}}_j^{-1}\widehat{\mathbf{M}}_j\boldsymbol{\Sigma}_j^{-1}(\boldsymbol{\Sigma}_j - \widehat{\boldsymbol{\Sigma}}_j)\widehat{\boldsymbol{\Sigma}}_j^{-1}\|$ and $Q_3 = \|\boldsymbol{\Sigma}_j^{-1}(\boldsymbol{\Sigma}_j - \widehat{\boldsymbol{\Sigma}}_j)\widehat{\boldsymbol{\Sigma}}_j^{-1}\mathbf{M}_j\boldsymbol{\Sigma}_j^{-1}\|$. By the triangular inequality, $\|\widehat{\boldsymbol{\Omega}}_{a1,j} - \boldsymbol{\Omega}_j\| \leq Q_1 + Q_2 + Q_3$. The Cauchy-Schwartz inequality implies immediately that $\|\mathbf{AB}\| \leq \|\mathbf{A}\|\|\mathbf{B}\|$(Golub and Van Loan, 2013), both Q_1 and Q_3 are dominated by $\|\boldsymbol{\Sigma}_j - \widehat{\boldsymbol{\Sigma}}_j\|$, and Q_2 is controlled by $\|\mathbf{M}_j - \widehat{\mathbf{M}}_j\|$. By Conditions (C2) and (C3), there exists constants C_1 and C_2 independent of j such that

$$\|\widehat{\boldsymbol{\Omega}}_{a1,j} - \boldsymbol{\Omega}_j\| \leq C_1\|\boldsymbol{\Sigma}_j - \widehat{\boldsymbol{\Sigma}}_j\| + C_2\|\mathbf{M}_j - \widehat{\mathbf{M}}_j\|. \tag{A.1}$$

It remains to study the convergence rates of $\|\boldsymbol{\Sigma}_j - \widehat{\boldsymbol{\Sigma}}_j\|$ and $\|\mathbf{M}_j - \widehat{\mathbf{M}}_j\|$.

Define $r(\mathbf{\Sigma}_j) \stackrel{\text{def}}{=} \text{trace}(\mathbf{\Sigma}_j)/\{n\lambda_{\max}(\mathbf{\Sigma}_j)\}$, which is not greater than p/n. Koltchinskii and Lounici (2017) showed that, under Conditions (C1)-(C2), for any $t \geq 1$, there exists a generic constant $C_3 \geq 1$ such that

$$\text{pr}\left(\|\widehat{\mathbf{\Sigma}}_j - \mathbf{\Sigma}_j\| \geq C_3\|\mathbf{\Sigma}_j\| \max\left[\{r(\mathbf{\Sigma}_j)\}^{1/2}, r(\mathbf{\Sigma}_j), (t/n)^{1/2}, t/n\right]\right) \leq \exp(-t).$$

This, together with Lemma 2.2.1 in Van Der Vaart and Wellner (1996), entails that

$$\|\|\widehat{\mathbf{\Sigma}}_j - \mathbf{\Sigma}_j\|\|_{\psi_1} \leq C_4(p/n)^{1/2}. \tag{A.2}$$

Accordingly, C_4 in (A.2) is related to C_3, which has been proved as an absolute constant in Koltchinskii and Lounici (2017). Next we study the convergence rate of $\|\widehat{\mathbf{M}}_j - \mathbf{M}_j\|$. Both $\widehat{\mathbf{M}}_{j,a}$ (5) and $\widehat{\mathbf{M}}_{j,c}$ (6) have the same form. On account of $|\widehat{p}_{h,j} - p_{h,j}| = O_p(n^{-1/2})$, the technical details for processing $\|\widehat{\mathbf{M}}_{j,a} - \mathbf{M}_{j,a}\|$ and $\|\widehat{\mathbf{M}}_{j,c} - \mathbf{M}_{j,c}\|$ are thus very similar. To avoid redundancy, we only provide the details for cumulative slicing estimation in what follows. Define

$$\widetilde{\mathbf{M}}_{j,c} \stackrel{\text{def}}{=} n^{-1}\sum_{i=1}^{n} \mathbf{m}_{j,c}(Y_{i,j})\mathbf{m}_{j,c}(Y_{i,j})^{\mathrm{T}}.$$

By triangular inequality, $\|\widehat{\mathbf{M}}_{j,c} - \mathbf{M}_{j,c}\| \leq \|\widehat{\mathbf{M}}_{j,c} - \widetilde{\mathbf{M}}_{j,c}\| + \|\widetilde{\mathbf{M}}_{j,c} - \mathbf{M}_{j,c}\|$. Following similar arguments for proving (A.2), we can show that

$$\|\|\widetilde{\mathbf{M}}_{j,c} - \mathbf{M}_{j,c}\|\|_{\psi_1} \leq C_5(p/n)^{1/2}, \tag{A.3}$$

where C_5 is an absolute constant. We turn to $\|\widehat{\mathbf{M}}_{j,c} - \widetilde{\mathbf{M}}_{j,c}\|$. By Lemma 5 in Wang et al. (2021), we have

$$\mathrm{pr}(\|\widehat{\mathbf{M}}_{j,c} - \widetilde{\mathbf{M}}_{j,c}\| \geq t^2 + 2c_1^{1/2}t) \leq \exp(2 + \log n + p \log 5 - Cnt^2), \quad \text{(A.4)}$$

where c_1 and C are absolute constants induced by Proposition 5.10 of Vershynin (2010). Following similar arguments for proving (A.2), we can also show that

$$\|\|\widehat{\mathbf{M}}_{j,c} - \widetilde{\mathbf{M}}_{j,c}\|\|_{\psi_1} \leq C_6/n. \quad \text{(A.5)}$$

Similarly, constant C_6 in (A.5) is constructed from constants in (A.4), thus C_6 is independent of j. By definition, $\|\|\widehat{\mathbf{M}}_{j,c} - \mathbf{M}_{j,c}\|\|_{\psi_1} \leq \|\|\widehat{\mathbf{M}}_{j,c} - \widetilde{\mathbf{M}}_{j,c}\|\|_{\psi_1} + \|\|\widetilde{\mathbf{M}}_{j,c} - \mathbf{M}_{j,c}\|\|_{\psi_1}$. This, together with (A.3) and (A.5), yields that $\|\|\widehat{\mathbf{M}}_{j,c} - \mathbf{M}_{j,c}\|\|_{\psi_1} \leq C_7(p/n)^{1/2}$, where C_7 is an absolute positive constant. Thus far we complete the proof for cumulative slicing estimation. With similar arguments we can deal with sliced inverse regression. In other words, $\|\|\widehat{\mathbf{M}}_j - \mathbf{M}_j\|\|_{\psi_1} \leq C_8(p/n)^{1/2}$ and C_8 is independent of j. This, together with (A.1) and (A.2), completes the proof of Lemma 5. □

The following lemma is a direct consequence of Lemma 5.

Lemma 6. *In addition to Conditions (C1)-(C3), we assume the sample covariance matrices* $\widehat{\mathbf{\Sigma}}_j^{-1}$*s are all invertible. Then there exists an absolute* *positive constant C such that*

$$\| dist(\widehat{\mathbf{B}}_{a1,j}, \mathbf{B})\|_{\psi_1} \leq C(d_0 p/n)^{1/2}, \text{ for } j = 1, \ldots, m.$$

Proof of Lemma 6: Denote the column space of $\mathbf{B}$ by span($\mathbf{B}$). Further denote the principal angles between span($\widehat{\mathbf{B}}_j$) and span($\mathbf{B}$) by $\Theta(\widehat{\mathbf{B}}_j, \mathbf{B}) \stackrel{\text{def}}{=} (\theta_{1,j}, \theta_{2,j}, \ldots, \theta_{d_0,j})^\top$. In other words, the singular values of $\widehat{\mathbf{B}}_j^\top \mathbf{B}$ are $\cos(\theta_{1,j})$, $\cos(\theta_{2,j})$, ..., $\cos(\theta_{d_0,j})$. Then

$$\{\text{dist}(\widehat{\mathbf{B}}_{a1,j}, \mathbf{B})\}^2 = 2\sum_{k=1}^{d_0}\{1 - \cos^2(\theta_{k,j})\} = 2\|\sin\{\Theta(\widehat{\mathbf{B}}_j, \mathbf{B})\}\|_F^2. \quad \text{(A.6)}$$

By Conditions (C2)-(C3) and Davis-Kahan-Theorem (Yu et al., 2015, Theorem 2), the right hand side of the above display is bounded by $C_0 d_0^{1/2}\|\widehat{\mathbf{\Omega}}_{a1,j} - \mathbf{\Omega}_j\|$, and C_0 is an absolute constant induced by Conditions (C2) and (C3). Therefore, $\text{dist}(\widehat{\mathbf{B}}_{a1,j}, \mathbf{B}) \leq C_0 d_0^{1/2}\|\widehat{\mathbf{\Omega}}_{a1,j} - \mathbf{\Omega}_j c\|$. The proof is completed by invoking Lemma 5. □

Lemma 7. *In addition to Conditions (C1)-(C3), we assume the sample covariance matrices $\widehat{\mathbf{\Sigma}}_j^{-1}$s are all invertible. Then there exists an absolute positive constant C such that*

$$\|E(\widehat{\mathbf{\Omega}}_{a1,j} - \mathbf{\Omega}_j)\|_F \leq Cp/n, \text{ for } j = 1, \ldots, m.$$

Proof of Lemma 7: As defined in Lemma 5,

$$\widehat{\mathbf{\Omega}}_{a1,j} - \mathbf{\Omega}_j = \widehat{\mathbf{\Sigma}}_j^{-1}\widehat{\mathbf{M}}_j\widehat{\mathbf{\Sigma}}_j^{-1} - \mathbf{\Sigma}_j^{-1}\mathbf{M}_j\mathbf{\Sigma}_j^{-1}$$

$$=(\widehat{\mathbf{\Sigma}}_j^{-1} - \mathbf{\Sigma}_j^{-1} + \mathbf{\Sigma}_j^{-1})(\widehat{\mathbf{M}}_j - \mathbf{M}_j + \mathbf{M}_j)(\widehat{\mathbf{\Sigma}}_j^{-1} - \mathbf{\Sigma}_j^{-1} + \mathbf{\Sigma}_j^{-1}) - \mathbf{\Sigma}_j^{-1}\mathbf{M}_j\mathbf{\Sigma}_j^{-1}$$

$$=\widehat{\mathbf{\Sigma}}_j^{-1}(\mathbf{\Sigma}_j - \widehat{\mathbf{\Sigma}}_j)\mathbf{\Sigma}_j^{-1}(\widehat{\mathbf{M}}_j - \mathbf{M}_j)\widehat{\mathbf{\Sigma}}_j^{-1}(\mathbf{\Sigma}_j - \widehat{\mathbf{\Sigma}}_j)\mathbf{\Sigma}_j^{-1} + 2\mathbf{\Sigma}_j^{-1}(\widehat{\mathbf{M}}_j - \mathbf{M}_j)\widehat{\mathbf{\Sigma}}_j^{-1}(\mathbf{\Sigma}_j - \widehat{\mathbf{\Sigma}}_j)\mathbf{\Sigma}_j^{-1}$$

$$+ \widehat{\mathbf{\Sigma}}_j^{-1}(\mathbf{\Sigma}_j - \widehat{\mathbf{\Sigma}}_j)\mathbf{\Sigma}_j^{-1}\mathbf{M}_j\widehat{\mathbf{\Sigma}}_j^{-1}(\mathbf{\Sigma}_j - \widehat{\mathbf{\Sigma}}_j)\mathbf{\Sigma}_j^{-1}$$

Therefore, by Conditions (C2) and (C3), we can find an absolute positive constant C_1, such that $\|E(\widehat{\mathbf{\Omega}}_{a1,j} - \mathbf{\Omega}_j)\|_F \le C_1 \|E(\widehat{\mathbf{\Sigma}}_j - \mathbf{\Sigma}_j)^2\|_F$. The proof is completed by invoking (A.2). □

Next we study the non-asymptotic error bound of the penalized estimates $\widehat{\mathbf{\Omega}}_{a2,j}$s, which do not require all the sample covariance matrices are invertible. Define

$$\|\mathbf{A}\|_\infty \overset{\text{def}}{=} \max_{1\le k\le l\le p} |a_{kl}| \text{ and } D_j \overset{\text{def}}{=} \|\mathbf{\Gamma}^{-1}_{\mathcal{S}_j,\mathcal{S}_j,j}\|_\infty.$$

Lemma 8. *In addition to Conditions (C1)-(C3), we assume there exist generic constants C_1 and C_2 such that $\|\mathbf{\Sigma}_j\|_\infty \ge C_1\{\log(p)/n\}^{1/2}$ and $C_2 s_j D_j \|\mathbf{\Sigma}_j\|_\infty \{\log(p)/n\}^{1/2} < \kappa_j$ for all $j = 1, \ldots, m$. Then there exists an absolute positive constant C, such that,*

$$\|\|\widehat{\mathbf{\Omega}}_{a2,j} - \mathbf{\Omega}_j\|_\infty\|_{\psi_1} \le C\kappa_j^{-1} D_j \{\log(p)/n\}^{1/2}.$$

Proof of Lemma 8: We re-present an equivalent form of (2.7) to stack all p columns into a vector. For notational clarity, we define $\widehat{\mathbf{\Gamma}}_j \overset{\text{def}}{=} \widehat{\mathbf{\Sigma}}_j \otimes \widehat{\mathbf{\Sigma}}_j$. It follows that,

$$\mathrm{vec}(\widehat{\mathbf{\Omega}}_{a2,j}) = \arg\min_{\mathrm{vec}(\mathbf{\Phi}_j)} \left[\mathrm{vec}(\mathbf{\Phi}_j)^{\mathrm{T}}\widehat{\mathbf{\Gamma}}_j\mathrm{vec}(\mathbf{\Phi}_j) - 2\mathrm{vec}(\widehat{\mathbf{M}}_j)^{\mathrm{T}}\mathrm{vec}(\mathbf{\Phi}_j) + \lambda_{n,j}\|\mathrm{vec}(\mathbf{\Phi}_j)\|_1\right].$$

Setting $\widehat{\mathbf{A}} = 2\widehat{\mathbf{\Gamma}}_j$, $\widehat{\mathbf{a}} = 2\mathrm{vec}(\widehat{\mathbf{M}}_j)$, $\mathbf{A} = 2\mathbf{\Gamma}_j$ and $\mathbf{a} = 2\mathrm{vec}(\mathbf{M}_j)$ in Lemma 7 of Wang et al. (2021), we can derive the non-asymptotic error bound of $\mathrm{vec}(\widehat{\mathbf{\Omega}}_{a2,j})$. Towards this goal, we need to verify the conditions required by Lemma 7 of Wang et al. (2021).

By definition, $\|\widehat{\mathbf{\Gamma}}_j - \mathbf{\Gamma}_j\|_\infty = \|\widehat{\mathbf{\Sigma}}_j \otimes (\widehat{\mathbf{\Sigma}}_j - \mathbf{\Sigma}_j) + (\widehat{\mathbf{\Sigma}}_j - \mathbf{\Sigma}_j) \otimes \mathbf{\Sigma}_j\|_\infty$, which is not greater than $(\|\widehat{\mathbf{\Sigma}}_j\|_\infty + \|\mathbf{\Sigma}_j\|_\infty)\|\widehat{\mathbf{\Sigma}}_j - \mathbf{\Sigma}_j\|_\infty$. It follows that

$$\|\mathbf{\Gamma}_{\mathcal{S}_j^c,\mathcal{S}_j,j}\mathbf{\Gamma}^{-1}_{\mathcal{S}_j,\mathcal{S}_j,j}\|_\infty + 2s_j\|\mathbf{\Gamma}^{-1}_{\mathcal{S},\mathcal{S},j}\|_\infty\|\widehat{\mathbf{\Gamma}}_j - \mathbf{\Gamma}_j\|_\infty \le 1-\kappa_j + 2s_jD_j(\|\widehat{\mathbf{\Sigma}}_j\|_\infty + \|\mathbf{\Sigma}_j\|_\infty)\|\widehat{\mathbf{\Sigma}}_j - \mathbf{\Sigma}_j\|_\infty.$$

By Lemma 5 of Wang et al. (2021), we have $\|\|\widehat{\mathbf{\Sigma}}_j - \mathbf{\Sigma}_j\|_\infty\|_{\psi_1} \le C_1\{\log(p)/n\}^{1/2}$, and C_1 is a general constant. Accordingly, $\|\|\widehat{\mathbf{\Sigma}}_j\|_\infty\|_{\psi_1} \le \|\mathbf{\Sigma}_j\|_\infty + C_1\{\log(p)/n\}^{1/2}$. Consequently,

$$\begin{aligned} & \|\|\mathbf{\Gamma}_{\mathcal{S}_j^c,\mathcal{S}_j,j}\mathbf{\Gamma}^{-1}_{\mathcal{S}_j,\mathcal{S}_j,j}\|_\infty + 2s_j\|\mathbf{\Gamma}^{-1}_{\mathcal{S},\mathcal{S},j}\|_\infty\|\widehat{\mathbf{\Gamma}}_j - \mathbf{\Gamma}_j\|_\infty\|_{\psi_1} \\ \le\ & 1-\kappa_j + 2s_jD_j(\|\|\widehat{\mathbf{\Sigma}}_j\|_\infty\|_{\psi_1} + \|\mathbf{\Sigma}_j\|_\infty)\|\|\widehat{\mathbf{\Sigma}}_j - \mathbf{\Sigma}_j\|_\infty\|_{\psi_1}. \end{aligned}$$

The right hand side is smaller than or equal to $1-\kappa_j + C_2s_jD_j\|\mathbf{\Sigma}_j\|_\infty\{\log(p)/n\}^{1/2}$, which, by the assumption we imposed, is strictly smaller than 1. Thus the first set of condition required by Lemma 7 of Wang et al. (2021) is satisfied.

Next we study the property of $\Delta \overset{\text{def}}{=} 2\|\mathrm{vec}(\widehat{\mathbf{M}}_j) - \mathrm{vec}(\mathbf{M}_j)\|_\infty + 2\|(\widehat{\mathbf{\Gamma}}_j - \mathbf{\Gamma}_j)\mathrm{vec}(\mathbf{\Omega}_j)\|_\infty$. We only process cumulative slicing estimation here. Let

$\mathbf{e}_k$ be a unit length p-vector with its kth entry being one. By condition (C1) and Proposition 2.5.2 in Vershynin (2018), we can show that $\mathbf{e}_k^{\mathrm{T}}(\mathbf{x}_{i,j} - \overline{\mathbf{x}}_j)$ is sub-Gaussian for all $k, j = 1, \ldots, p$. Using the general Hoeffding's inequality (Vershynin, 2018, Theorem 2.6.3), $\mathbf{e}_k^{\mathrm{T}}\widehat{\mathbf{m}}_{j,c}(y)$ is also sub-Gaussian for all $k, j = 1, \ldots, p$. Therefore, it follows immediately from Lemma 2.7.6 and Bernstein's inequality in Vershynin (2018) that there exists general constants c_1 and c_2, such that $\mathrm{pr}\big\{|\mathbf{e}_k(\widehat{\mathbf{M}}_j - \mathbf{M}_j)\mathbf{e}_l\| \geq t\big\} \leq 2\exp\{-n\min(c_1t^2, c_2t)\}$ for $k, l = 1, \ldots, p$. Setting $t = c_3\{\log(p)/n\}^{1/2}$, we can find an absolute constant C, such that $\|\|\mathrm{vec}(\widehat{\mathbf{M}}_j) - \mathrm{vec}(\mathbf{M}_j)\|_\infty\|_{\psi_1} = \|\|\widehat{\mathbf{M}}_j - \mathbf{M}_j\|_\infty\|_{\psi_1} \leq C\{\log(p)/n\}^{1/2}$. In addition,

$$\begin{aligned}\|(\widehat{\boldsymbol{\Gamma}}_j - \boldsymbol{\Gamma}_j)\mathrm{vec}(\boldsymbol{\Omega}_j)\|_\infty &= \|\widehat{\boldsymbol{\Sigma}}_j\boldsymbol{\Omega}_j\widehat{\boldsymbol{\Sigma}}_j - \boldsymbol{\Sigma}_j\boldsymbol{\Omega}_j\boldsymbol{\Sigma}_j\|_\infty \\ &\leq \|(\widehat{\boldsymbol{\Sigma}}_j - \boldsymbol{\Sigma}_j)\boldsymbol{\Omega}_j(\widehat{\boldsymbol{\Sigma}}_j - \boldsymbol{\Sigma}_j)\|_\infty + 2\|\boldsymbol{\Sigma}_j\boldsymbol{\Omega}_j(\widehat{\boldsymbol{\Sigma}}_j - \boldsymbol{\Sigma}_j)\|_\infty \\ &\leq s_j\|\boldsymbol{\Omega}_j\|_\infty\|\widehat{\boldsymbol{\Sigma}}_j - \boldsymbol{\Sigma}_j\|_\infty^2 + 2\|\boldsymbol{\Omega}_j\boldsymbol{\Sigma}_j\|_\infty\|\widehat{\boldsymbol{\Sigma}}_j - \boldsymbol{\Sigma}_j\|_\infty,\end{aligned}$$

which, by invoking $\|\|\widehat{\boldsymbol{\Sigma}}_j - \boldsymbol{\Sigma}_j\|_\infty\|_{\psi_1} \leq C_1\{\log(p)/n\}^{1/2}$, implies immediately that $\|\|(\widehat{\boldsymbol{\Gamma}}_j - \boldsymbol{\Gamma}_j)\mathrm{vec}(\boldsymbol{\Omega}_j)\|_\infty\|_{\psi_1} \leq C_2\{\log(p)/n\}^{1/2}$, and C_2 is independent of j. It follows that $\|\Delta\|_{\psi_1} \leq C_3\{\log(p)/n\}^{1/2}$. C_3 is also an absoluate constant since it is induced by C and C_2 . Set $\lambda_{n,j} = 3C_3\kappa_j^{-1}\{\log(p)/n\}^{1/2}$. Thus the second set of condition required by Lemma 7 of Wang et al. (2021) is satisfied. Thus we are enabled to complete the proof of Lemma 8 by applying Lemma 7 in Wang et al. (2021) directly. □

Lemma 9. *Assume the conditions of Lemma 8 hold true. Then there exists an absolute positive constant C such that*

$$\|dist(\widehat{\mathbf{B}}_{a2,j}, \mathbf{B})\|_{\psi_1} \leq C\kappa_j^{-1} D_j \{d_0 s_j \log(p)/n\}^{1/2}.$$

Proof of Lemma 9: By Davis-Kahan-Theorem (Yu et al., 2015, Theorem 2), the result in (A.6), and conditions (C2) and (C3), there exist absolute constant C_0 such that

$$\text{dist}(\widehat{\mathbf{B}}_{a2,j}, \mathbf{B}) \leq C_0 (8d_0)^{1/2} \|\widehat{\mathbf{\Omega}}_{a2,j} - \mathbf{\Omega}_j\| \leq C_0 (8s_j d_0)^{1/2} \|\widehat{\mathbf{\Omega}}_{a2,j} - \mathbf{\Omega}_j\|_\infty.$$

This, together with Lemma 8, completes the proof of Lemma 9. □

S4 Proof of Lemma 1

By Jensen's inequality,

$$\|\mathbf{\Omega}^*_{a1} - \mathbf{B}\mathbf{B}^{\mathrm{T}}\|_F \leq m^{-1} \sum_{j=1}^{m} E\|\widehat{\mathbf{B}}_{a1,j}\widehat{\mathbf{B}}^{\mathrm{T}}_{a1,j} - \mathbf{B}\mathbf{B}^{\mathrm{T}}\|_F \leq \max_{1\leq j\leq m} \|\text{dist}(\widehat{\mathbf{B}}_{a1,j}, \mathbf{B})\|_{\psi_1}.$$

For sufficiently large n such that $n \geq 2d_0 p C^2$, Lemma 6 ensures that $\|\text{dist}(\widehat{\mathbf{B}}_{a1,j}, \mathbf{B})\|_{\psi_1} < 1/4$. Because $\mathbf{B}\mathbf{B}^{\mathrm{T}}$ is a projection matrix, $\lambda_{d_0}(\mathbf{B}\mathbf{B}^{\mathrm{T}}) = 1$ and $\lambda_{d_0+1}(\mathbf{B}\mathbf{B}^{\mathrm{T}}) = 0$. This, together with Weyl's inequality, indicates that $\lambda_{d_0}(\mathbf{\Omega}^*) > 3/4$ and $\lambda_{d_0+1}(\mathbf{\Omega}^*) < 1/4$. It follows from Davis-Kahan-Theorem (Yu et al., 2015, Theorem 2) that

$$\text{dist}(\widehat{\mathbf{B}}_{a1}, \mathbf{B}^*_{a1}) \leq 4\|\widehat{\mathbf{T}}_{a1} - \mathbf{\Omega}^*_{a1}\|_F. \tag{A.7}$$

In addition, by Lemma 4 in Fan et al. (2019), we have

$$\|\|\widehat{\mathbf{T}}_{a1} - \mathbf{\Omega}^*_{a1}\|_F\|_{\psi_1} \le Cm^{-1/2}\|\|\widehat{\mathbf{B}}_{a1,j}\widehat{\mathbf{B}}^{\top}_{a1,j} - E(\widehat{\mathbf{B}}_{a1,j}\widehat{\mathbf{B}}^{\top}_{a1,j})\|_F\|_{\psi_1}. \quad \text{(A.8)}$$

It remains to bound $\|\widehat{\mathbf{B}}_{a1,j}\widehat{\mathbf{B}}^{\top}_{a1,j} - E(\widehat{\mathbf{B}}_{a1,j}\widehat{\mathbf{B}}^{\top}_{a1,j})\|_F$ from above. By Jensen's inequality, $\|E(\widehat{\mathbf{B}}_{a1,j}\widehat{\mathbf{B}}^{\top}_{a1,j}) - \mathbf{B}\mathbf{B}^{\top}\|_F \le$ $E\{\text{dist}(\widehat{\mathbf{B}}_{a1,j}, \mathbf{B})\} \le \|\text{dist}(\widehat{\mathbf{B}}_{a1,j}, \mathbf{B})\|_{\psi_1}$. We apply triangle inequality to obtain that $\|\widehat{\mathbf{B}}_{a1,j}\widehat{\mathbf{B}}^{\top}_{a1,j} - E(\widehat{\mathbf{B}}_{a1,j}\widehat{\mathbf{B}}^{\top}_{a1,j})\|_F \le \text{dist}(\widehat{\mathbf{B}}_{a1,j}, \mathbf{B}) + \|E(\widehat{\mathbf{B}}_{a1,j}\widehat{\mathbf{B}}^{\top}_{a1,j}) - \mathbf{B}\mathbf{B}^{\top}\|_F$, which, by definition, is not greater than $\text{dist}(\widehat{\mathbf{B}}_{a1,j}, \mathbf{B}) + \|\text{dist}(\widehat{\mathbf{B}}_{a1,j}, \mathbf{B})\|_{\psi_1}$. This implies that

$$\|\|\widehat{\mathbf{B}}_{a1,j}\widehat{\mathbf{B}}^{\top}_{a1,j} - E(\widehat{\mathbf{B}}_{a1,j}\widehat{\mathbf{B}}^{\top}_{a1,j})\|_F\|_{\psi_1} \le 2\|\text{dist}(\widehat{\mathbf{B}}_{a1,j}, \mathbf{B})\|_{\psi_1}. \quad \text{(A.9)}$$

Invoking (A.7) - (A.9), we have $\|\text{dist}(\widehat{\mathbf{B}}_{a1}, \mathbf{B}^*_{a1})\|_{\psi_1} \le 8Cm^{-1/2}\|\text{dist}(\widehat{\mathbf{B}}_{a1,j}, \mathbf{B})\|_{\psi_1}$. The proof is completed by invoking Lemma 6. □

S5 Proof of Lemma 2

With similar arguments for proving (A.6), we have $\text{dist}(\mathbf{B}^*_{a1}, \mathbf{B}) = 2^{1/2}\|\sin\{\Theta(\mathbf{B}^*_{a1}, \mathbf{B})\}\|_F$, where $\Theta(\mathbf{B}^*_{a1}, \mathbf{B})$ are the principal angles between $\text{span}(\mathbf{B}^*_{a1})$ and $\text{span}(\mathbf{B})$. Invoking Davis-Kahan-Theorem (Yu et al., 2015, Theorem 2) again, we have

$$\text{dist}(\mathbf{B}^*_{a1}, \mathbf{B}) \le 8^{1/2}\|\mathbf{\Omega}^*_{a1} - \mathbf{B}\mathbf{B}^{\top}\|_F \le 8^{1/2} \max_{1 \le j \le m} \|E(\widehat{\mathbf{B}}_{a1,j}\widehat{\mathbf{B}}^{\top}_{a1,j} - \mathbf{B}\mathbf{B}^{\top})\|_F$$

For notational clarity, we define $\mathbf{E}_{a1,j} \stackrel{\text{def}}{=} \widehat{\mathbf{\Omega}}_{a1,j} - \mathbf{\Omega}_j$. Let $(\widehat{\mathbf{b}}_{1,j}, \widehat{\mathbf{b}}_{2,j}, \ldots, \widehat{\mathbf{b}}_{p,j})$ be the eigenvectors of $\widehat{\mathbf{\Omega}}_{a1,j}$, and $(\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_p)$ be the eigenvectors of $\mathbf{\Omega}$. For any fixed $s \in \{0, 1, \ldots, p - d_0\}$, we define $S \stackrel{\text{def}}{=} \{s+1, \ldots, s+d_0\}$ and

$$\mathbf{G}_k \stackrel{\text{def}}{=} \sum_{l \notin S} \{(\lambda_l(\mathbf{\Omega}_j) - \lambda_{s+k}(\mathbf{\Omega}_j)\}^{-1} \mathbf{b}_l \mathbf{b}_l^{\mathrm{T}}.$$

Let f be a linear map $f: \mathbb{R}^{p\times d_0} \mapsto \mathbb{R}^{p\times d_0}$, $(\mathbf{v}_1, \ldots, \mathbf{v}_{d_0}) \mapsto (-\mathbf{G}_1\mathbf{v}_1, \ldots, -\mathbf{G}_{d_0}\mathbf{v}_{d_0})$. By the linearity of f and the triangular inequality, we have

$$\begin{aligned} & \left\| E\left[\widehat{\mathbf{B}}_{a1,j}\widehat{\mathbf{B}}_{a1,j}^{\mathrm{T}} - \{\mathbf{B}\mathbf{B}^{\mathrm{T}} + f(\mathbf{E}_{a1,j}\mathbf{B})\mathbf{B}^{\mathrm{T}} + \mathbf{B}f(\mathbf{E}_{a1,j}\mathbf{B})^{\mathrm{T}}\}\right]\right\|_F \\ \geq\ & \|E\widehat{\mathbf{B}}_{a1,j}\widehat{\mathbf{B}}_{a1,j}^{\mathrm{T}} - \mathbf{B}\mathbf{B}^{\mathrm{T}}\|_F - \|f\{E(\mathbf{E}_{a1,j})\}\mathbf{B}\mathbf{B}^{\mathrm{T}}\|_F - \|\mathbf{B}f\{E(\mathbf{E}_{a1,j})\}\mathbf{B}^{\mathrm{T}}\|_F. \end{aligned}$$

Define $\epsilon_{a1} \stackrel{\text{def}}{=} \|\mathbf{E}_{a1,j}\| / \min\{\lambda_s(\mathbf{\Omega}_j) - \lambda_{s+1}(\mathbf{\Omega}_j), \lambda_{s+d_0}(\mathbf{\Omega}_j) - \lambda_{s+d_0+1}(\mathbf{\Omega}_j)\}$. By Lemma 2 in Fan et al. (2019) and Jensen's inequality,

$$\left\| E\left[\widehat{\mathbf{B}}_{a1,j}\widehat{\mathbf{B}}_{a1,j}^{\mathrm{T}} - \{\mathbf{B}\mathbf{B}^{\mathrm{T}} + f(\mathbf{E}_{a1,j}\mathbf{B})\mathbf{B}^{\mathrm{T}} + \mathbf{B}f(\mathbf{E}_{a1,j}\mathbf{B})^{\mathrm{T}}\}\right]\right\|_F \leq 24 d_0^{1/2} E(\epsilon_{a1}^2).$$

In addition, $\|f\{E(\mathbf{E}_{a1,j})\}\mathbf{B}\mathbf{B}^{\mathrm{T}}\|_F \leq \|f\{E(\mathbf{E}_{a1,j})\}\|_F \leq C\|E(\mathbf{E}_{a1,j})\|_F$. The first inequality follows from Lemma A1 in Yu et al. (2015), and the second is a direct application of Jensen's inequality. Similarly, $\|\mathbf{B}f\{E(\mathbf{E}_{a1,j})\}\mathbf{B}^{\mathrm{T}}\|_F \leq C\|E(\mathbf{E}_{a1,j})\|_F$. Lemma 7 proves that $\|E(\mathbf{E}_{a1,j})\|_F \leq C_1 p/n$. Besides, Lemma 5 indicates that $\|E(\epsilon_{a1}^2)\|_{\psi_1} \leq C_2 p/n$.

Summarizing the above results, we obtain that there exists an absolute positive constant C such that $\mathrm{dist}(\mathbf{B}_{a1}^*, \mathbf{B}) \leq C d_0^{1/2} p/n$, for $1 \leq j \leq m$. The proof is now completed. □

S6 Proof of Lemma 3

Following similar arguments for proving Lemma 1, we can prove this lemma by using Lemma 9. Details are omitted from the present context. □

S7 Proof of Lemma 4

Invoking Davis-Kahan-Theorem (Yu et al., 2015, Theorem 2) again, we have

$$\mathrm{dist}(\mathbf{B}^*_{a2}, \mathbf{B}) \le 8^{1/2} \|\mathbf{\Omega}^*_{a2} - \mathbf{B}\mathbf{B}^{\mathrm{T}}\|_F \le 8^{1/2} \max_{1\le j\le m} \|E(\widehat{\mathbf{B}}_{a2,j}\widehat{\mathbf{B}}^{\mathrm{T}}_{a2,j} - \mathbf{B}\mathbf{B}^{\mathrm{T}})\|_F$$

In parallel to Lemma 2, we define $\mathbf{E}_{a2,j} \stackrel{\mathrm{def}}{=} \widehat{\mathbf{\Omega}}_{a2,j} - \mathbf{\Omega}_j$ and let $(\widehat{\mathbf{b}}_{1,j}, \widehat{\mathbf{b}}_{2,j}, \ldots, \widehat{\mathbf{b}}_{p,j})$ be the eigenvectors of $\widehat{\mathbf{\Omega}}_{a2,j}$. Recall that $(\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_p)$ are the eigenvectors of $\mathbf{\Omega}$, and f is a linear map defined in Lemma 2. By triangular inequality again, we have

$$\begin{aligned} & \left\| E\left[\widehat{\mathbf{B}}_{a2,j}\widehat{\mathbf{B}}^{\mathrm{T}}_{a2,j} - \{\mathbf{B}\mathbf{B}^{\mathrm{T}} + f(\mathbf{E}_{a2,j}\mathbf{B})\mathbf{B}^{\mathrm{T}} + \mathbf{B}f(\mathbf{E}_{a2,j}\mathbf{B})^{\mathrm{T}}\}\right]\right\|_F \\ \ge\ & \|E\widehat{\mathbf{B}}_{a2,j}\widehat{\mathbf{B}}^{\mathrm{T}}_{a2,j} - \mathbf{B}\mathbf{B}^{\mathrm{T}}\|_F - \|f\{E(\mathbf{E}_{a2,j})\}\mathbf{B}\mathbf{B}^{\mathrm{T}}\|_F - \|\mathbf{B}f\{E(\mathbf{E}_{a2,j})\}\mathbf{B}^{\mathrm{T}}\|_F. \end{aligned}$$

Define $\epsilon_{a2} \stackrel{\mathrm{def}}{=} \|\mathbf{E}_{a2,j}\| / \min\{\lambda_s(\mathbf{\Omega}) - \lambda_{s+1}(\mathbf{\Omega}), \lambda_{s+d_0}(\mathbf{\Omega}) - \lambda_{s+d_0+1}(\mathbf{\Omega})\}$. By Lemma 2 in Fan et al. (2019) and Jensen's inequality,

$$\left\| E\left[\widehat{\mathbf{B}}_{a2,j}\widehat{\mathbf{B}}^{\mathrm{T}}_{a2,j} - \{\mathbf{B}\mathbf{B}^{\mathrm{T}} + f(\mathbf{E}_{a2,j}\mathbf{B})\mathbf{B}^{\mathrm{T}} + \mathbf{B}f(\mathbf{E}_{a2,j}\mathbf{B})^{\mathrm{T}}\}\right]\right\|_F \le 24 d_0^{1/2} E(\epsilon^2_{a2}).$$

In addition, $\|f\{E(\mathbf{E}_{a2,j})\}\mathbf{B}\mathbf{B}^{\top}\|_F \leq \|f\{E(\mathbf{E}_{a2,j})\}\|_F \leq C\|E(\mathbf{E}_{a2,j})\|_F$, Lemma 8 shows that $\|\|\mathbf{E}_{a2,j}\|_\infty\|_{\psi_1} \leq C\kappa_j^{-1}D_j\{\log(p)/n\}^{1/2}$. Therefore, $\|E(\mathbf{E}_{a2,j})\|_F \leq E\|\mathbf{E}_{a2,j}\|_F \leq s_j^{1/2}E\|\mathbf{E}_{a2,j}\|_\infty \leq s_j^{1/2}\|\|\mathbf{E}_{a2,j}\|_\infty\|_{\psi_1}$. Summarizing the above results, we obtain that $\text{dist}(\mathbf{B}_{a2}^*, \mathbf{B}) \leq C\kappa_j^{-1}D_j\{s_j\log(p)/n\}^{1/2}$ for $1 \leq j \leq m$. This completes the proof. □

Bibliography

Fan, J., D. Wang, K. Wang, and Z. Zhu (2019). Distributed estimation of principal eigenspaces. *Annals of Statistics 47*, 3009–3031.

Golub, G. H. and C. F. Van Loan (2013). *Matrix Computations* (4th ed.). The Johns Hopkins University Press.

Koltchinskii, V. and K. Lounici (2017). Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli 23*(1), 110–133.

Li, K.-C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association 86*(414), 316–327.

Luo, R., H. Wang, and C.-L. Tsai (2009). Contour projected dimension reduction. *The Annals of Statistics 37*(6B), 3743–3778.

Van Der Vaart, A. W. and J. A. Wellner (1996). *Weak Convergence and Empirical Processes*. Springer.

Vershynin, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*.

Vershynin, R. (2018). *High-Dimensional Probability: An Introduction with Applications in Data Science*, Volume 47. Cambridge University Press.

Wang, C., B. Jiang, and L. Zhu (2021). Penalized interaction estiamtion for ultrahigh dimensional quadratic regression. *Statistica Sinica 31*, 1549–1570.

Wang, C., Z. Yu, and L. Zhu (2021). On cumulative slicing estimation for high dimensional data. *Statistica Sinica 31*, 223–242.

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

Zhu, L., L. Zhu, and Z. Feng (2010). Dimension reduction in regressions through cumulative slicing estimation. *Journal of the American Statistical Association 105*(492), 1455–1466.