# Fast Distributed Principal Component Analysis of Large-Scale Federated Data

Shuting Shen, Junwei Lu, and Xihong Lin *

**Abstract**

Principal component analysis (PCA) is one of the most popular methods for dimension reduction. In the light of rapidly increasing large-scale data in federated ecosystems, where data cannot leave individual warehouses, such as banks and healthcare systems, the traditional PCA method is often not applicable due to privacy protection consideration and large computational burden. Fast PCA algorithms have been proposed to lower the computational cost for large-scale data, but they cannot handle federated data. Distributed PCA algorithms have been developed to handle federated data by applying traditional PCA to data at each site and aggregating site-specific PCA results. However, they are not computationally efficient and not scalable when data at each site are large with many samples and variables, such as biobanks. In this paper, we propose the FAst DIstributed (FADI) PCA method that performs PCA analysis of large federated data with high computational efficiency and low statistical error without the need of sharing the data across sites. Specifically, FADI applies fast PCA to site-specific data using multiple random sketches and aggregates the fast PCA results across sites. We perform a non-asymptotic theoretical study to show that under some regularity conditions, FADI enjoys the same error rate as the traditional full sample PCA and a significantly smaller order of computational burden compared to the existing methods. We perform extensive simulation studies to compare the finite sample performance of FADI with the existing algorithms, and show that FADI substantially outperforms the other methods in computational efficiency without sacrificing statistical accuracy. We apply FADI to the analysis of the 1000 Genomes data to study the population structure.

**Keyword:** Computational efficiency; Distributed computing; Efficient communication; Fast PCA; Federated learning; PCA; Random matrices; Random sketches.

*Shuting Shen is PhD student (*shs145@g.harvard.edu*) and Junwei Lu (*junweilu@hsph.harvard.edu*) is Assistant Professor at the Department of Biostatistics at Harvard TH Chan School of Public Health. Xihong Lin is Professor of Biostatistics at Harvard T.H. Chan School of Public Health and Professor of Statistics at Harvard University (*xlin@hsph.harvard.edu*). This work was supported by the National Institutes of Health grants R35-CA197449, U01-HG009088, U01HG012064, U19-CA203654, and P30 ES000002.

# 1 Introduction

As one of the most frequently used methods for dimension reduction, principal component analysis (PCA) finds applications in a broad spectrum of scientific fields including network clustering (Abbe et al., 2020), statistical genetics (Reich et al., 2008) and finance (Pasini, 2017). Taking Genome-Wide Association Studies (GWAS) for instance, ancestry differences can induce confounding in genetic association testing, which makes adjusting for population structure necessary (Visscher et al., 2017). Price et al. (2006) showed that by incorporating the principal components (PCs) of the genotypes across the genome as covariates, population structure can be effectively captured and corrected in genetic association analysis.

Despite its fundamental role in statistics, several shortcomings of the traditional PCA method hinder its application to large-scale data. For example, as a result of rapid reduction in genotyping costs and the launch of large consortia and biobanks, GWAS datasets often contain hundreds of thousands to millions of Single Nucleotide Polymorphisms (SNPs) and subjects. They entail scalable algorithms to handle the intensive computation of PCA. For instance, the UK Biobank (Sudlow et al., 2015) has around $n = 500,000$ subjects with GWAS data and Electronic Health Records (EHRs). After LD pruning, around $d = 160,000$ genetic variants were used for calculating ancestry PCs (Dey et al., 2020). To estimate the leading ancestry PCs using the traditional PCA, the total computational cost will be around $160,000^2 \times 500,000 \approx 10^{16}$ flops. Therefore, for large-scale datasets, when the dimension $d$ and the sample size $n$ are both very large, the traditional PCA will be computationally expensive and even infeasible.

Another challenge is that large-scale datasets in many applications are stored in federated ecosystems, where data cannot leave individual warehouses, such as banks and healthcare systems, due to privacy protection considerations (Dhruva et al., 2020). For example, the Million Veteran Program (MVP) biobank data (Klarin et al., 2018) cannot leave the Veterans

Affairs (VA) system. This calls for federated learning methods (Jordan et al., 2019; Li et al., 2020), which provide efficient and privacy-protected strategies for joint analysis across multiple data warehouses without the need to exchange individual-level data. Traditional PCA methods cannot be applied to federated data, e.g., calculation of overall ancestry PCs of UK biobank and MVP data.

The burgeoning popularity of large-scale data necessitates the development of fast algorithms that can cope with both high dimensionality and massiveness efficiently and distributively. Indeed, efforts have been made in recent years on developing fast PCA algorithms and distributed PCA algorithms. The existing fast PCA algorithms use all the data and apply column selection and random projection to speed up PCA calculations (Halko et al., 2011; Chen et al., 2016), while the existing distributed PCA algorithms apply the traditional PCA method to the split/site-specific data and aggregate the results (Kargupta et al., 2001; Fan et al., 2019).

Specifically, for fast PCA algorithms, column selection and random projection provide an efficient algorithm for calculating PCs when the number of variables $d$ is large by using the fact that the column space of a low-rank matrix can be represented by a small set of columns, and the original covariance matrix of the full data can be approximated by its projection onto the span of the representative columns (Achlioptas, 2003). For instance, Halko et al. (2011) proposed to estimate the $K$ leading eigenvectors of a covariance matrix using Gaussian random sketches, which decrease the PCA computation time by a factor of $O(d)$ at the cost of increasing the statistical error by a polynomial factor of $d$. Chen et al. (2016) modified Halko et al. (2011)'s method by repeating the fast sketching multiple times and showed the consistency of the proposed fast PCA algorithm by integrating i.i.d. random sketches when the number of sketches goes to infinity. However, they did not study the non-asymptotic theoretical results of how the error rate is determined by the number of sketches and the trade-off between computation complexity and the error rate in finite samples. As the fast

PCA methods use the full data, they have two major limitations. First, they are often computationally not scalable to large sample sizes $n$, such as biobank size data. Second, they are not applicable to federated data when data in different sites cannot be shared.

The existing distributed PCA algorithms reduce the PCA computational burden by partitioning the full data "horizontally" or "vertically" (Kargupta et al., 2001; Fan et al., 2019; Kannan et al., 2014). The horizontal partition splits the data over $n$ samples, whereas the vertical partition splits the data over $d$ variables. Horizontal partition is useful when the sample size $n$ is large or the data are federated in multiple sites that cannot be shared. Fan et al. (2019) considered distributed PCA by partitioning the data horizontally and estimating the $K$ leading eigenvectors using traditional PCA for each split dataset followed by aggregating the PCA results across different datasets. They showed when the number of data splits is not too large, the error rate of their algorithm is of the same order as the traditional PCA. Since the traditional PCA algorithm is used for each data partition, the computational complexity of their algorithm is at least of order $O(d^3)$, which will be computationally difficult when $d$ is large, e.g., in biobanks, $d$ corresponds to hundreds of thousands of SNPs. The distributed PCA method is hence not scalable for GWAS analysis in large federated biobanks. Kargupta et al. (2001) considered vertical partition and developed a method that collects local PCs and then reconstructs global PCs by linear transformation. However, there is no theoretical guarantee on the error rate of their algorithm compared with the traditional full sample PCA, and the method may fail when variables are correlated.

In view of the limitations of the existing fast PCA algorithms and distributed PCA algorithms for the analysis of large and federated data, we propose in this paper a scalable and computationally efficient fast distributed PCA method for large federated data when both $n$ and $d$ are large. We call the proposed method FAst DIstributed (FADI) PCA. We consider horizontal partition in this paper to handle federated data. Suppose the data are stored in multiple sites. FADI applies multiple random sketches to calculate PCs of each

partitioned dataset, and then uses a distributed algorithm to aggregate the fast PCA results across sites.

FADI improves over the existing fast PCA algorithms by first partitioning the data and applying fast PCA to the partitioned data instead of the whole data, followed by a more computationally efficient PCA aggregation method that converges much faster. It hence can handle federated data with large $n$ and large $d$. FADI improves over the existing distributed PCA methods by applying random sketches instead of traditional PCA to each partitioned dataset, and hence can handle partitioned data with large $d$. In other words, FADI combines the features of the fast PCA algorithm and the distributed PCA algorithm, and can calculate PCs at high computational efficiency when both the sample size $n$ and the dimension of variables $d$ are large, and is also applicable to large federated data without requiring sharing data between sites.

We focus in this paper on the spiked covariance model (Johnstone, 2001), which assumes the covariance matrix of the full data can be decomposed into a low-rank matrix and a diagonal matrix, where the low-rank matrix captures population structure in GWAS. We consider both the homogeneous and heterogeneous residual variance models. In the FADI method, we first truncate the sample covariance matrix to create an almost low-rank structure. Then for distributed analysis of each partitioned dataset, we conduct Gaussian random sketches on the truncated sample covariance matrix by performing Singular Value Decomposition (SVD) on the low-dimensional sketches to obtain the leading eigenvectors. Finally, the results from multiple random sketches are aggregated across the partitioned datasets to obtain the final PC estimator of the full data.

We show that the computational cost of FADI is an order of magnitude smaller than fast PCA and distributed PCA (Table 1). We study the theoretical properties of FADI by conducting a non-asymptotic variance-bias decomposition of the error rate. We show that the non-asymptotic error rate of the FADI PC estimator is of the same order as the traditional

PCA as long as the number of random sketches is sufficiently large. We perform extensive simulation studies to compare FADI with the existing methods, and apply FADI to the analysis of the 1000 Genomes data.

The rest of the paper is organized as follows. Section 2 introduces the problem setting and briefly discusses the novel features of the FADI algorithm compared with the previous methods. Section 3 proposes FADI, and discusses its implementation details, as well as the computational complexity of FADI and its modifications in a range of scenarios. Section 4 presents the theoretical results of the statistical error of FADI. Section 5 discusses generalizing FADI to the heterogeneous spiked model. Section 6 shows the simulation results by comparing FADI with the existing methods. The application of FADI to the 1000 Genomes data is given in Section 7, followed by discussions.

# 2 Eigenspace Estimation in the Spiked Covariance Model

## 2.1 The Problem Setting

Suppose there are $n$ i.i.d. random vectors $\{\boldsymbol{X}_i\}_{i=1}^n \subseteq \mathbb{R}^d$ with expectation $\mathbb{E}(\boldsymbol{X}_i) = \mathbf{0}$ and covariance matrix $\mathbb{E}(\boldsymbol{X}_i\boldsymbol{X}_i^\top) = \boldsymbol{\Sigma}$. We consider the spiked covariance model (Johnstone, 2001), which assumes the covariance matrix $\boldsymbol{\Sigma}$ has the following spectral decomposition: $\boldsymbol{\Sigma} = \mathbf{V}_K\boldsymbol{\Lambda}_K\mathbf{V}_K^\top + \sigma^2\mathbf{I}_d$, where $\mathbf{V}_K = (\mathbf{v}_1, \ldots, \mathbf{v}_K) \in \mathbb{R}^{d\times K}$ is the stacking of the top $K$ eigenvectors corresponding to a low-rank matrix of rank $K$, and $\boldsymbol{\Lambda}_K = \mathrm{diag}\,(\lambda_1 - \sigma^2, \ldots, \lambda_K - \sigma^2)$ is a diagonal matrix, where $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_K$ are the $K$ spiked eigenvalues, and all the non-spiked eigevalues are identical to $\sigma^2$. Here we start with the homogeneous residual variance model which assumes a constant residual variance $\sigma^2\mathbf{I}_d$ in the $\boldsymbol{\Sigma}$ decomposition. In Section 5, we will extend the results to a heterogeneous residual variance model by replacing $\sigma^2\mathbf{I}_d$ in $\boldsymbol{\Sigma}$ by $\mathrm{diag}\{\sigma_1^2, \ldots, \sigma_d^2\}$.

In the distributed setting, the data are stored on $m$ different servers. For federated

data, $m$ servers refer to $m$ sites, e.g., biobanks, whose individual-level data cannot be shared between sites. Denote by $\{\boldsymbol{X}_i^{(j)}\}_{i=1}^{n_j}$ the sample of size $n_j$ on the $j$-th server, where $\mathbf{X}^{(j)} = (\boldsymbol{X}_1^{(j)}, \ldots, \boldsymbol{X}_{n_j}^{(j)})^\top$ denotes the corresponding data matrix split ($j = 1, \ldots, m$ and $\sum_{j=1}^m n_j = n$). Denote by $\mathbf{X} = (\boldsymbol{X}_1^\top, \ldots, \boldsymbol{X}_n^\top)^\top$ the full $n \times p$ data matrix. We here assume $\{\boldsymbol{X}_i\}_{i=1}^n$ are i.i.d. for simplicity of presentation. In practice, samples from different servers/sites may be heterogeneous. In Section 5, we will show that our method will still enjoy good theoretical properties in these scenarios as long as the sample covariance matrix converges to the population covariance matrix at an appropriate rate.

We aim to estimate the column space of the $K$ leading eigenvectors, i.e., $\mathcal{V}_K = \mathrm{Col}(\mathbf{V}_K)$, where $\mathrm{Col}(\mathbf{V}_K)$ denotes the column space of $\mathbf{V}_K$. For now, we assume $K$ is known. In Section 3.3, we will discuss the scenario where $K$ is unknown and propose a method to estimate $K$. Denote by $\Delta = \lambda_K - \sigma^2$ the eigengap, by $\tau = \lambda_1/\Delta$ the condition number, and by $r = \mathrm{tr}(\mathbf{\Sigma})/\|\mathbf{\Sigma}\|_{\mathrm{op}}$ the effective rank of $\mathbf{\Sigma}$ where $\|\mathbf{\Sigma}\|_{\mathrm{op}} := \sup_{\|\mathbf{v}\|=1} \|\mathbf{\Sigma}\mathbf{v}\|$ is the matrix operator norm (Horn and Johnson, 1990). To ensure that $\mathcal{V}_K$ is identifiable, we need $\Delta > 0$. To estimate $\mathcal{V}_K$, the traditional PCA method conducts spectral decomposition on the sample covariance matrix $\widehat{\mathbf{\Sigma}} = n^{-1}\mathbf{X}^\top\mathbf{X} = n^{-1}\sum_{i=1}^n \boldsymbol{X}_i\boldsymbol{X}_i^\top$ and take the top $K$ eigenvectors $\widehat{\mathbf{V}}_K$ as the estimator of $\mathbf{V}_K$. Since the eigenspace is invariant up to rotation, for the top $K$ eigenvectors $\mathbf{V}_K$ and their estimator $\widehat{\mathbf{V}}_K$, we measure their difference by $\rho(\widehat{\mathbf{V}}_K, \mathbf{V}_K) := \|\widehat{\mathbf{V}}_K\widehat{\mathbf{V}}_K^\top - \mathbf{V}_K\mathbf{V}_K^\top\|_{\mathrm{F}}$, where $\|\cdot\|_{\mathrm{F}}$ is the matrix Frobenius norm (Horn and Johnson, 1990).

In this paper, we consider the estimation of the $K$ leading eigenspace $\mathcal{V}_K$ when both the sample size $n$ and the dimension $d$ are large in large federated data where individual-level data cannot be shared between sites.

## 2.2 Fast Distributed PCA (FADI): Overview and Intuition

We provide in this section an overview of the proposed FAst DIstributed PCA (FADI) method and its intuition, and will present the detailed algorithm in Section 3. The bottleneck of the traditional PCA method lies in two aspects: the computation of the sample covariance matrix $\widehat{\mathbf{\Sigma}}$ has complexity of $O(nd^2)$, and the Singular Value Decomposition (SVD) procedure has complexity of $O(d^3)$. Hence when $d$ and $n$ are large, the computational cost of calculating $\widehat{\mathbf{\Sigma}}$ is quite large.

To reduce the computational cost when $d$ is large, the most straightforward idea is to reduce the dimension $d$ of the data. One popular method for dimension reduction is random sketching (Halko et al., 2011). For instance, for a low-rank matrix $\mathbf{A}$ with rank $K$, instead of directly performing SVD on $\mathbf{A}$, we will conduct SVD on $\mathbf{A\Omega}$, where $\mathbf{\Omega} \in \mathbb{R}^{d\times p}$ is a random Gaussian matrix of dimension $p$ and $K < p \ll d$. Since $\mathbf{A}$ is low-rank, the column space of $\mathbf{A}$ can be represented by $\mathbf{A\Omega}$ when $p$ is sufficiently large. In our paper, the covariance matrix $\mathbf{\Sigma}$ is not low-rank. However, the truncated matrix $\mathbf{\Sigma} - \sigma^2\mathbf{I}_d = \mathbf{V}_K\mathbf{\Lambda}_K\mathbf{V}_K^\top$ has rank $K \ll d$. Therefore, if we can obtain a consistent estimator $\widehat{\sigma}^2$ for $\sigma^2$, then the truncated version of the sample covariance matrix will have an almost low-rank structure

$$\widehat{\mathbf{\Sigma}}^{\text{tr}} := \widehat{\mathbf{\Sigma}} - \widehat{\sigma}^2\mathbf{I}_d \approx \mathbf{V}_K\mathbf{\Lambda}_K\mathbf{V}_K^\top,$$

where the superscript "tr" stands for "truncated". Then instead of directly conducting PCA on $\widehat{\mathbf{\Sigma}}$, we perform PCA on the matrix $\widehat{\mathbf{\Sigma}}^{\text{tr}}\mathbf{\Omega}$. The rationale is as follows: due to the spiked structure of the covariance matrix, the matrix $\widehat{\mathbf{\Sigma}}^{\text{tr}}\mathbf{\Omega} \approx \mathbf{V}_K\mathbf{\Lambda}_K\mathbf{V}_K^\top\mathbf{\Omega} = \mathbf{V}_K\mathbf{\Lambda}_K\widetilde{\mathbf{\Omega}}$, where $\widetilde{\mathbf{\Omega}} = \mathbf{V}_K^\top\mathbf{\Omega} \in \mathbb{R}^{K\times p}$ is also a standard Gaussian matrix. Intuitively, $p^{-1}\widetilde{\mathbf{\Omega}}\widetilde{\mathbf{\Omega}}^\top \approx \mathbf{I}_K$ when $p$ is large, and thus $\widetilde{\mathbf{\Omega}}$ acts like an orthonormal matrix scaled by $\sqrt{p}$. In other words, the matrix $\widehat{\mathbf{\Sigma}}^{\text{tr}}\mathbf{\Omega}$ almost maintains the same structure of the left singular space as $\mathbf{V}_K\mathbf{\Lambda}_K\mathbf{V}_K^\top$. It is hence reasonable to estimate $\mathbf{V}_K$ from the $d \times p$ matrix $\widehat{\mathbf{\Sigma}}^{\text{tr}}\mathbf{\Omega}$ that has a much smaller dimension

than the $d \times d$ full sample covariance matrix $\widehat{\mathbf{\Sigma}}$.

To further reduce the computation cost when $n$ is large or the data are federated, instead of applying random sketches to the truncated full sample covariance $\widehat{\mathbf{\Sigma}}^{\mathrm{tr}}$, FADI computes in parallel the fast sketching of the truncated sample covariance of each partitioned sample and aggregates the results across $m$ sites. Specifically, decompose $\widehat{\mathbf{\Sigma}}^{\mathrm{tr}}\mathbf{\Omega} = n^{-1}\sum_{i=1}^{n}(\boldsymbol{X}_i\boldsymbol{X}_i^{\top}\mathbf{\Omega}) - \widehat{\sigma}^2\mathbf{\Omega}$ along the sample size $n$. For the $j$-th sample split $\{\boldsymbol{X}_i^{(j)}\}_{i=1}^{n_j}$ $(j = 1, \ldots, m)$, compute its fast sketching as $(n/m)^{-1}\sum_{i=1}^{n_j}(\boldsymbol{X}_i^{(j)}\boldsymbol{X}_i^{(j)\top}\mathbf{\Omega}) - \widehat{\sigma}^2\mathbf{\Omega}$. Then the fast sketching of the truncated full sample covariance matrix $\widehat{\mathbf{\Sigma}}^{\mathrm{tr}}\mathbf{\Omega}$ can be calculated by averaging over the fast sketches for the $m$ split samples, and then calculating $\mathbf{V}_K$ by performing PCA of $\widehat{\mathbf{\Sigma}}^{\mathrm{tr}}\mathbf{\Omega}$ .

For each partitioned dataset, as our simple algorithm does not require the down-stream procedure of projecting the covariance matrix of the partitioned data onto the column space of the fast sketches (Halko et al., 2011), it motivates us to repeat the fast sketching multiple times for each partitioned data, and aggregate the results to reduce the elevated statistical error caused by the fast sketching approximation. We will show in Section 4 that when the number of repeated fast sketching is sufficiently large, FADI enjoys the same error rate as the full sample PCA. From this perspective, FADI extends "vertically" distributed PCA in the sense that it allocates the computational burden along the dimension $d$ to several machines utilizing repeated fast sketching, where each fast sketching enjoys a smaller dimension and lower computational cost, while all the repeated fast sketches will be aggregated together to produce higher statistical accuracy.

Table 1 provides a comparison of the theoretical error rates and the computational complexities of different PCA methods. The theoretical rate of FADI will be derived in Section 4, and the computational complexities will be analyzed in Section 3.2. The results show that FADI has a much smaller computational complexity compared with the distributed PCA algorithm and the fast PCA algorithm, while enjoying the same error rate as the full sample PCA. Specifically, the computational complexity of FADI is of $O(nK^2 + d^2K^2\log n)$,

| Method | Error rate | Computational Complexity |
|---|---|---|
| Traditional PCA | $O(\sqrt{Kr/n})$ | $O(d^2n + d^3)$ |
| Fast PCA | $O(\sqrt{Kdr/n})$ | $O(dnK + d^2K)$ |
| Distributed PCA | $O(\sqrt{Kr/n})$ | $O(d^2\sqrt{nr} + d^3)$ |
| FADI | $O(\sqrt{Kr/n})$ | $O(nK^2 + d^2K^2 \log n)$ |

Table 1: A comparison of the error rates and computational complexities between Fast Distributed PCA (FADI), distributed PCA, fast PCA (one sketching), and traditional full sample PCA. Here $r := \mathrm{tr}(\mathbf{\Sigma})/\|\mathbf{\Sigma}\|_{\mathrm{op}}$ is the effective rank of the covariance matrix.

which is of a significantly smaller order when $n$ and $d$ are large compared to $O(dnK + d^2K)$ for fast PCA (Halko et al., 2011), $O(d^2\sqrt{nr} + d^3)$ for distributed PCA (Fan et al., 2019), and $O(d^2n + d^3)$ for traditional PCA. FADI maintains the same level of statistical accuracy as the full sample PCA, while the theoretical error rate of the fast PCA algorithm (Halko et al., 2011) is larger than that of the traditional PCA by a factor of $\sqrt{d}$.

# 3 The Fast Distributed PCA Algorithm (FADI)

In this section, we first provide in Section 3.1 the proposed fast distributed PCA algorithm FADI in detail and its application to different scenarios. We then discuss in Section 3.2 the computational complexity of FADI and compare it with the existing methods. We will show how FADI is scalable when the dimension $d$ and the sample size $n$ are both large and/or the large data are federated. We will discuss in Section 3.3 how to estimate the number of spikes $K$ for $\mathbf{\Sigma}$.

We begin by introducing some notations that will be used later in this section. For two integers $j > i \geq 1$, let $[i]$ denote the set $\{1, 2, \ldots, i\}$, and $i : j$ the set $\{i, i+1, \ldots, j\}$, and $:$ the full index set. For a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$, $\mathbf{A}_{[:,a:b]}$ ($\mathbf{A}_{[a:b,:]}$) denotes the $\{a, a+1, \ldots, b\}$-th columns (rows) of $\mathbf{A}$. For two positive sequences $x_n$ and $y_n$, we say $x_n \lesssim y_n$ or $x_n = O(y_n)$ if $x_n \leq Cy_n$ for $C > 0$ that does not depend on $n$. We say $x_n \asymp y_n$ if $x_n \lesssim y_n$ and $y_n \lesssim x_n$. If $\lim_{n\to\infty} x_n/y_n = 0$ then we say $x_n = o(y_n)$ or $x_n \ll y_n$.

## 3.1 The FADI Algorithm

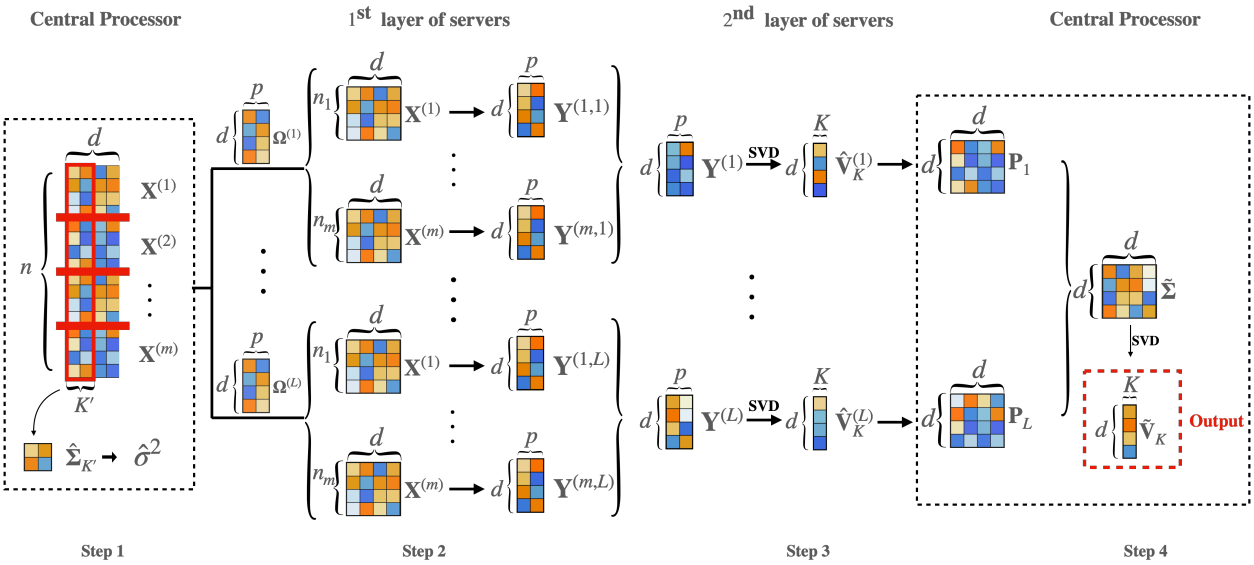



Figure 1: Illustration of FADI. In Step 1, $\sigma^2$ is estimated by calculating the sample covariance matrix $\widehat{\mathbf{\Sigma}}_{K'}$ of the first $K'$ data columns ($K' > K$) and taking its smallest eigenvalue $\widehat{\sigma}^2$. On the first layer, the $\mathbf{\Omega}^{(\ell)}$'s ($\ell = 1, \ldots, L$) represent the $L$ i.i.d. Gaussian test matrices, each assigned to $m$ servers and applied to $\mathbf{X}^{(j)}$ to calculate in parallel the $L$ fast sketches $\mathbf{Y}^{(j,\ell)}$ ($\ell = 1, \ldots, L$) corresponding to $\mathbf{X}^{(j)}$. On the second layer, each of the fast sketches $\mathbf{Y}^{(\ell)}$ ($\ell = 1, \ldots, L$) is calculated by aggregating the $m$ fast sketches $\mathbf{Y}^{(j,\ell)}$ ($j = 1, \ldots, m$) across the $m$ servers, which can be done in parallel. Their leading $K$ left singular vectors $\widehat{\mathbf{V}}_K^{(\ell)}$ are then calculated, and sent to the central processor to produce the corresponding projection matrix $\mathbf{P}_\ell = \widehat{\mathbf{V}}_K^{(\ell)} \widehat{\mathbf{V}}_K^{(\ell)\top}$ for aggregation. The top $K$ eigenvectors $\widetilde{\mathbf{V}}_K$ of the aggregated projection matrix $\widetilde{\mathbf{\Sigma}} = L^{-1} \sum_{\ell=1}^{L} \mathbf{P}_\ell$ within the red dashed line box is the final FADI estimator of the $K$ PCs of the covariance $\mathbf{\Sigma}$.

Figure 1 illustrates the fast distributed PCA (FADI) algorithm:

In Step 1, on the central processor, we first obtain a consistent estimator $\widehat{\sigma}^2$ of $\sigma^2$ by taking the minimum eigenvalue of the sample covariance matrix for the first $K'$ data columns, where $K' \geq K + 1$. Specifically, we compute $\widehat{\mathbf{\Sigma}}_{K'} = n^{-1} \mathbf{X}_{[:,1:K']}^\top \mathbf{X}_{[:,1:K']}$, which can be easily calculated distributively without much computational burden (See Remark 3.1), and then conduct SVD on $\widehat{\mathbf{\Sigma}}_{K'}$. Then take the smallest eigenvalue $\widehat{\sigma}^2 = \lambda_{\min}\big(\widehat{\mathbf{\Sigma}}_{K'}\big)$ as the estimator for $\sigma^2$.

In Step 2, we distributively calculate the Gaussian fast sketches of the truncated sample

covariance matrix $\mathbf{Y} = \widehat{\mathbf{\Sigma}}^{\mathrm{tr}}\mathbf{\Omega}$, where $\widehat{\mathbf{\Sigma}}^{\mathrm{tr}} = \widehat{\mathbf{\Sigma}} - \widehat{\sigma}^2\mathbf{I}_d$ and $\mathbf{\Omega}$ is a $d \times p$ standard Gaussian test matrix with $K < p \ll d$. To reduce the statistical error, we repeat the fast sketches $L$ times and aggregate the results from the $L$ copies of $\mathbf{Y}$. Specifically, suppose that the data are split into $m$ sets. We send the split data $\mathbf{X}^{(j)} \in \mathbb{R}^{n_j \times d}$ $(j = 1, \ldots, m)$ to the $mL$ servers on the first layer. The $(j, \ell)$-th server on the first layer is assigned with $\mathbf{X}^{(j)}$ and the $\ell$-th Gaussian test matrix $\mathbf{\Omega}^{(\ell)} = \{\omega_{ij}^{(\ell)}\} \in \mathbb{R}^{d \times p}$, where $\ell = 1, \ldots, L$ and $\omega_{ij}^{(\ell)} \overset{\text{i.i.d.}}{\sim} N(0, 1)$. Then we calculate the $\ell$-th fast sketch of $\mathbf{X}^{(j)}$ as $\mathbf{Y}^{(j,\ell)} = (n/m)^{-1} \sum_{i=1}^{n_j} \boldsymbol{X}_i^{(j)} \boldsymbol{X}_i^{(j)\top} \mathbf{\Omega}^{(\ell)} - \widehat{\sigma}^2 \mathbf{\Omega}^{(\ell)}$. We send $\mathbf{Y}^{(j,\ell)}$ $(j = 1, \ldots, m)$ to the $\ell$-th server on the second layer.

In Step 3, on the $\ell$-th server of the second layer, the fast sketches $\mathbf{Y}^{(j,\ell)}$ $(j = 1, \ldots, m)$ from the $m$ split datasets corresponding to the $\ell$-th Gaussian test matrix $\mathbf{\Omega}^{(\ell)}$ will be collected and averaged to get the $\ell$-th fast sketches: $\mathbf{Y}^{(\ell)} = m^{-1} \sum_{j=1}^{m} \mathbf{Y}^{(j,\ell)}$ $(\ell = 1, \ldots, L)$. We next compute the top $K$ left singular vectors $\widehat{\mathbf{V}}_K^{(\ell)}$ of $\mathbf{Y}^{(\ell)}$ and send the $\widehat{\mathbf{V}}_K^{(\ell)}$'s to the central processor for aggregation.

In Step 4, on the central processor, calculate $\widetilde{\mathbf{\Sigma}} = L^{-1} \sum_{\ell=1}^{L} \widehat{\mathbf{V}}_K^{(\ell)} \widehat{\mathbf{V}}_K^{(\ell)\top} = L^{-1} \sum_{\ell=1}^{L} \mathbf{P}_\ell$, where $\mathbf{P}_\ell := \widehat{\mathbf{V}}_K^{(\ell)} \widehat{\mathbf{V}}_K^{(\ell)\top}$ is the projection matrix of $\widehat{\mathbf{V}}_K^{(\ell)}$. We next calculate the $K$ leading eigenvectors $\widetilde{\mathbf{V}}_K$ of $\widetilde{\mathbf{\Sigma}}$, which will serve as the final estimator of $\mathbf{V}_K$.

**Remark 3.1.** In Step 1, to obtain $\widehat{\sigma}^2$ we need to calculate $\widehat{\mathbf{\Sigma}}_{K'}$, where $K' > K + 1$. If on the central processor, the whole data matrix is accessible, then $\widehat{\mathbf{\Sigma}}_{K'}$ can be directly computed by operating on the first $K'$ columns of $\mathbf{X}$. If the data are only stored distributively along the sample size $n$, e.g., when data are federated, then $\widehat{\mathbf{\Sigma}}_{K'}$ can be computed by aggregating $\widehat{\mathbf{\Sigma}}_{K'}^{(j)} = \mathbf{X}_{[:,1:K']}^{(j)\top} \mathbf{X}_{[:,1:K']}^{(j)}$ $(j = 1, \ldots, m)$, which is calculated from the first $K'$ columns of the local splits. In Section 3.2, we will see that the computation of $\widehat{\sigma}^2$ will contribute to the total computational complexity by $O(K^2 n)$ flops. Due to the small dimension of $\widehat{\mathbf{\Sigma}}_{K'}$, to further reduce the computational cost, we do not need the entire sample for calculating $\widehat{\mathbf{\Sigma}}_{K'}$. We can just take a subsample of size $n' < n$, as long as $\sqrt{K'/n'} \leq \sqrt{r/n}$. Then the error rate will maintain the same.

**Remark 3.2.** We refer to Theorem 4.3 for the choice of $p$ and $L$, and when the number of data splits $m$ can be specified by the user, we suggest choosing $m \asymp n/d$ to maximize the computational efficiency. In general, taking $p = 2K$ should be sufficient. For now, we assume $K$ is known, and the scenarios where $K$ is unknown will be discussed in Section 3.3. In Section 3.2, we will provide a modified procedure for performing SVD on $\widetilde{\mathbf{\Sigma}}$ using random sketches to further reduce the computational cost in Step 4.

## 3.2 Computational Complexity

In this section, we discuss the computational complexity and the communication cost of FADI. First, for the computational complexity, the computational cost of the traditional PCA is composed of two parts: the computation of the sample covariance matrix and the SVD. As mentioned previously, the computation of the sample covariance matrix takes $O(d^2n)$ flops and usually comprises the dominating part of the computational cost, whereas the SVD step takes $O(d^3)$ flops. Fan et al. (2019)'s distributed PCA method reduces the computational cost to $O(d^2n/m + d^3)$ by dividing the sample into $m$ sets and calculating the PCs of the split data using the traditional PCA method separately on different servers. However, since the traditional PCA method is used for split data, when the number of variables $d$ is large, like in GWAS, its computational cost would be high. Halko et al. (2011)'s fast PCA method improves the computational efficiency of the SVD by applying random sketches to the full data when the number of variables $d$ is large, but their method cannot be applied directly to distributed/federated data due to the extra projection step, and their total computational complexity is of order $O(dnK + d^2K)$, which is not scalable to large $n$.

Now we discuss the computational complexity of FADI. The complexity of each step in FADI is listed as follows.

- **Step 1:** Computation of $\widehat{\mathbf{\Sigma}}_{K'}$: $O(K^2n)$, SVD on $\widehat{\mathbf{\Sigma}}_{K'}$: $O(K^3)$;

- **Step 2:** Computation of $\mathbf{Y}^{(j,\ell)}$: $O(dnp/m)$;
- **Step 3:** Computation of $\mathbf{Y}^{(\ell)}$: $O(mdp)$, SVD on $\mathbf{Y}^{(\ell)}$: $O(dp^2)$;
- **Step 4:** Computation of $\widetilde{\mathbf{\Sigma}}$: $O(d^2pL)$, SVD on $\widetilde{\mathbf{\Sigma}}$: $O(d^3)$.

In summary, since Step 2 and Step 3 are performed in parallel, the total computational complexity for FADI is $O(\max(K^2n, dnp/m, mdp, d^3))$. One may note that when $m$ is large enough, the error rate will be dominated by $O(d^3)$, which is the SVD complexity in Step 4. However, this complexity can be reduced in Step 4 by calculating the singular vectors of $\widetilde{\mathbf{\Sigma}}$ by fast sketching. Specifically, we can replace Step 4 of FADI with the following procedure.

**Step** 4′: On the central processor, generate the Gaussian test matrix $\mathbf{\Omega}^{\mathrm{F}} \in \mathbb{R}^{d\times p'}$, where $p'$ is the dimension of fast sketching that can be set larger than $p$ for smaller error rate, and apply the power method (Halko et al., 2011) by calculating $\widetilde{\mathbf{Y}} = \widetilde{\mathbf{\Sigma}}^q\mathbf{\Omega}^{\mathrm{F}} = \left(L^{-1}\sum_{l=1}^{L}\widehat{\mathbf{V}}_K^{(\ell)}\widehat{\mathbf{V}}_K^{(\ell)\top}\right)^q\mathbf{\Omega}^{\mathrm{F}}$, where $q$ is an integer greater than 1. Perform SVD on $\widetilde{\mathbf{Y}}$ and take the $K$ leading left singular vectors $\widetilde{\mathbf{V}}_K^{\mathrm{F}}$ as the estimator for $\mathbf{V}_K$.

Here, $\widetilde{\mathbf{Y}}$ can be calculated by induction. More specifically, let $\widetilde{\mathbf{Y}}_{(0)} = \mathbf{\Omega}^{\mathrm{F}}$ and for $i = 1,\ldots,q$, we calculate $\widetilde{\mathbf{Y}}_{(i)}$ iteratively: $\widetilde{\mathbf{Y}}_{(i)} = L^{-1}\sum_{l=1}^{L}\left(\widehat{\mathbf{V}}_K^{(\ell)}\widehat{\mathbf{V}}_K^{(\ell)\top}\widetilde{\mathbf{Y}}_{(i-1)}\right)$. Then $\widetilde{\mathbf{Y}} = \widetilde{\mathbf{Y}}_{(q)}$. We can see that this will reduce the total flops in Step 4′ of FADI to $O(dLKp'q + dp'^2)$ flops. For the choice of $q$, refer to Theorem 4.4. If we take $L \asymp Kd/p$, $m \asymp n/d$, $p' \asymp p = 2K$ and $q \asymp \log(n/p')$, then the complexity of Step 4′ is $O(dLKp'q + dp'^2) = O\big(d^2K^2\log(n/K)\big)$, and the total computational cost for FADI is $O\big(d^2K^2\log(n/K) + nK^2\big)$. This suggests that FADI is much faster than the traditional PCA, which is of complexity $O(d^2n + d^3)$; Fan et al. (2019)'s distributed PCA method, which is of complexity $O(d^2\sqrt{nr} + d^3)$; and Halko et al. (2011)'s fast PCA method, which is of complexity $O(dnK + d^2K)$. Furthermore, the error rate of FADI outperforms the fast PCA method by a factor of $\sqrt{d}$ (See Table 1 and Section 4). Note that as we cannot repeat the fast sketches in parallel on the central processor, we take $q$ powers of $\widetilde{\mathbf{\Sigma}}$ to calculate the eigenvectors of $\widetilde{\mathbf{\Sigma}}$ using fast sketches. This approach

maintains the same eigenvectors while making the eigenvalues decrease much faster.

As for the communication cost, for the traditional PCA, if the data are stored distributively on $m$ servers/sites, sharing the entire data or aggregating local covariance matrices will result in communication cost of order $O\big(\min(d^2m, dn)\big)$. The distributed PCA method by Fan et al. (2019) reduces the communication cost to the order $O(mKd)$. For FADI, since the calculation of multiple fast sketches are conducted in parallel, we only need to consider the communication cost for one copy of fast sketches and the communication cost in the aggregation step. It can be seen from Figure 1 that as each server on the first layer needs to send the $d \times p$ matrix $\mathbf{Y}^{(j,\ell)}$ ($j \in [m]$ and $\ell \in [L]$) to the server on the second layer, the communication cost between the first and the second layer will be $O(mpd) = O(mKd)$. For the aggregation step, as each server on the second layer sends the top $K$ eigenvectors $\widehat{\mathbf{V}}_K$ to the central processor, the communication cost is $O(LKd)$. Therefore, the total communication cost for FADI is $O(LKd + mKd)$. When $L$ and $m$ are of a similar order, the communication cost of FADI is of the same order as Fan et al. (2019)'s distributed PCA method. It is worth pointing out that our partition of the sample size $n$ does not induce any statistical error, and thus our error rate is stable with respect to the number of data splits $m$. In comparison, the statistical error in Fan et al. (2019) will be big when $m$ is large.

These discussions suggest that while FADI has a similar communication cost to the distributed PCA method, its computational complexity is of a significantly smaller order than the distributed PCA method. Hence the overall computational burden of FADI is significantly smaller than the distributed PCA.

### 3.3 Estimation of the Number of Spikes $K$

FADI requires to input $K$, the number of spiked eigenvalues of the covariance matrix $\mathbf{\Sigma}$. In practice, the exact value of $K$ is not needed as long as the dimensions of the random Gaussian matrices $\mathbf{\Omega}$ used for fast sketches, $p$ and $p'$, are sufficiently larger than $K$. Yet knowing

the exact value of $K$ will improve the computational efficiency. In fact, the estimation of $K$ can be incorporated into Step 3 and Step 4 of FADI: for the $\ell$-th server on the second layer ( $\ell \in [L]$), after performing the SVD $\mathbf{Y}^{(\ell)} = \widehat{\mathbf{V}}^{(\ell)}\widehat{\mathbf{\Lambda}}^{(\ell)}\widehat{\mathbf{U}}^{(\ell)\top}$, we estimate $K$ by $\widehat{K}^{(\ell)} = \min\{k : \max_{i\geq k}(\sigma_i(\mathbf{Y}^{(\ell)}) - \sigma_p(\mathbf{Y}^{(\ell)})) \leq \sqrt{p}\mu_0\}$ (we set $\widehat{K}^{(\ell)}$ to be $p$ if the set is empty), where $\mu_0$ is a user-specified parameter (we refer to Theorem 4.5 for the choice of $\mu_0$). Then send all the left singular vectors $\widehat{\mathbf{V}}^{(\ell)} \in \mathbb{R}^{d\times p}$ and $\widehat{K}^{(\ell)}, \ell \in [L]$ to the central processor. Finally on the central processor, take $\widehat{K} = \text{median}\left\{\widehat{K}^{(1)}, \widehat{K}^{(2)}, \ldots, \widehat{K}^{(L)}\right\}$ as the estimator for $K$. In Section 6, we will show that $\widehat{K}$ recovers $K$ with high probability under appropriate conditions. Our simulation study also shows that different variations of $\widehat{K}$ in practice have little impact on the error rate of $\widehat{\mathbf{V}}_K$ as long as $\widehat{K} \geq K$.

# 4 Theoretical Bound on Error Rates

In this section, we conduct a theoretical study to derive the error bound of the FADI estimator. We make the following distributional assumptions on the random vectors $\{\boldsymbol{X}_i\}_{i=1}^n$.

**Assumption 4.1.** *$\{\boldsymbol{X}_i\}_{i=1}^n \subseteq \mathbb{R}^d$ are i.i.d. sub-Gaussian random vectors, i.e., for any vector $\mathbf{u} \in \mathbb{R}^d$, there exists some constant $C$ such that* $\sup_{q\geq 1}\left\{(\mathbb{E}|\mathbf{u}^\top \boldsymbol{X}_i|^q)^{1/q}\right\}/\sqrt{q} \leq C\sqrt{\mathbb{E}(\mathbf{u}^\top \boldsymbol{X}_i)^2}$.

Assumption 4.1 guarantees that the sample vectors are light-tailed and the sample covariance matrix will converge in a good manner. We will conduct a variance-bias decomposition on the error rate $\rho(\widetilde{\mathbf{V}}_K, \mathbf{V}_K)$, where $\widetilde{\mathbf{V}}_K$ is the FADI PC estimators defined in Step 4 in Section 3. To facilitate this discussion, we introduce the intermediate matrix $\mathbf{\Sigma}' = \mathbb{E}_{\mathbf{\Omega}}\big(\widehat{\mathbf{V}}_K^{(\ell)}\widehat{\mathbf{V}}_K^{(\ell)\top}\big)$, where $\widehat{\mathbf{V}}_K^{(\ell)}$ is the top K left singular vectors of the $\ell$-th fast sketch $\mathbf{Y}^{(\ell)}$ defined in Step 3 in Section 3, and the expectation is taken with respect to $\mathbf{\Omega}$. Let $\mathbf{V}'_K$ be the top $K$ eigenvectors of $\mathbf{\Sigma}'$. Recall that $\widehat{\mathbf{\Sigma}}^{\text{tr}} = \widehat{\mathbf{\Sigma}} - \widehat{\sigma}^2\mathbf{I}_d$ is the truncated sample covariance matrix, and note that both $\mathbf{\Sigma}'$ and $\mathbf{V}'_K$ are random depending on $\widehat{\mathbf{\Sigma}}^{\text{tr}}$. For the FADI PC estimator $\widetilde{\mathbf{V}}_K$, we have

the following "variance-bias" decomposition of the error rate:

$$\rho(\widetilde{\mathbf{V}}_K, \mathbf{V}_K) \leq \underbrace{\rho(\widetilde{\mathbf{V}}_K, \mathbf{V}'_K)}_{\text{variance}} + \underbrace{\rho(\mathbf{V}'_K, \mathbf{V}_K)}_{\text{bias}}.$$

If we condition on all the available data, then the first term characterizes the statistical randomness of the FADI PC estimator $\widetilde{\mathbf{V}}_K$ due to fast sketching, whereas the second bias term is deterministic and depends on all the information provided by the data. Intuitively, since $\widetilde{\mathbf{\Sigma}} = L^{-1}\sum_{l=1}^{L} \widehat{\mathbf{V}}_K^{(\ell)}\widehat{\mathbf{V}}_K^{(\ell)\top}$ would converge to the conditional expectation $\mathbf{\Sigma}'$, $\widetilde{\mathbf{V}}_K$ would also converge to $\mathbf{V}'_K$. Hence the first variance term goes to 0 asymptotically.

As for the second bias term, let $\widehat{\mathbf{V}}_K$ be the $K$ leading eigenvectors of $\widehat{\mathbf{\Sigma}}^{\text{tr}}$, then we further break the bias term into two components: $\rho(\mathbf{V}'_K, \mathbf{V}_K) \leq \rho(\widehat{\mathbf{V}}_K, \mathbf{V}_K) + \rho(\mathbf{V}'_K, \widehat{\mathbf{V}}_K)$. We can see that the first term is the error rate for the traditional PCA method, whereas the second term is the bias caused by fast sketching. We will show in the following Lemma 4.2 that the second term is 0 with high probability. Therefore, the second term is negligible compared to the first term, and the bias of the FADI estimator is of the same order as the error rate of the traditional PCA. In other words, the bias of the FADI estimator mainly comes from $\widehat{\mathbf{V}}_K$, which is due to the information we can get from the available data.

**Lemma 4.2.** *Let $\widehat{\mathbf{V}}\widehat{\mathbf{\Lambda}}\widehat{\mathbf{V}}^\top$ be the SVD of $\widehat{\mathbf{\Sigma}}^{tr}$ and $\widehat{\mathbf{V}}_K$ be the $K$ leading eigenvectors of $\widehat{\mathbf{\Sigma}}^{tr}$. When $\|\widehat{\mathbf{V}}_K\widehat{\mathbf{V}}_K^\top - \mathbf{\Sigma}'\|_{\text{op}} < 1/2$, we have that $\widehat{\mathbf{V}}^\top\mathbf{\Sigma}'\widehat{\mathbf{V}}$ is diagonal and $\|\widehat{\mathbf{V}}_K\widehat{\mathbf{V}}_K^\top - \mathbf{V}'_K\mathbf{V}'^\top_K\|_{\text{op}} = 0$.*

Lemma 4.2 shows that as long as $\mathbf{\Sigma}'$ and $\mathbf{V}_K\mathbf{V}_K^\top$ are not too far apart, $\mathbf{V}'_K$ and $\widehat{\mathbf{V}}_K$ will share the same column space. In fact, Lemma B.3 in Supplementary Materials B shows that the probability that $\mathbf{\Sigma}'$ and $\widehat{\mathbf{V}}_K\widehat{\mathbf{V}}_K^\top$ are not sufficiently close converges to 0. The proof of Lemma 4.2 is deferred to Supplementary Materials A. Recall $\tau = \lambda_1/\Delta$ is the condition number and $r = \text{tr}(\mathbf{\Sigma})/\|\mathbf{\Sigma}\|_{\text{op}}$ is the effective rank. The following theorem gives the overall error rate of the FADI PC estimator $\widetilde{\mathbf{V}}_K$. Its proof is given in Supplementary Materials B.

**Theorem 4.3.** *When $p \geq \max(2K, K+7)$ and $n \geq C(dr/p)\tau^2 \log^2(n/r)$ for some large enough constant $C$, if Assumption 4.1 holds, for the FADI PC estimator $\widetilde{\mathbf{V}}_K$, we have*

$$\left(\mathbb{E}\|\widetilde{\mathbf{V}}_K\widetilde{\mathbf{V}}_K^\top - \mathbf{V}_K\mathbf{V}_K^\top\|_{\mathrm{F}}^2\right)^{1/2} \lesssim \tau\sqrt{\frac{Kr}{n}} + \tau\sqrt{\frac{Kdr}{npL}}. \tag{1}$$

**Remark 4.1.** The first term in (1) is the bias term, while the second term is the variance term. We can see that when the number of sketches $L$ reaches the order $d/p$, the variance term will be of the same order as the bias term, which is the same as the error rate of directly conducting the SVD on the full sample covariance matrix $\widehat{\mathbf{\Sigma}}$. Comparing with Halko et al. (2011)'s fast PCA method, which has the sub-optimal error rate $O(\tau\sqrt{(dKr)/n})$, FADI achieves the optimal rate with a smaller computation complexity for large $n$. Theorem 4.3 also indicates that $p$ only needs to be of the same order as $K$, which significantly reduces the communication costs from $O(d^2)$ to $O(dK)$ for each server.

As mentioned in Section 3.2, when $d$ is too large, we recommend users conduct another fast sketching in the final step. The following theorem characterizes the error rate of the estimator obtained by fast sketching in Step $4'$ of FADI.

**Theorem 4.4.** *In Step $4'$ of FADI as described Section 3.2, recall that $\widetilde{\mathbf{V}}_K^{\mathrm{F}}$ is the estimator obtained by taking the $K$ leading left singular vectors of $\widetilde{\mathbf{\Sigma}}^q\mathbf{\Omega}^{\mathrm{F}}$ for some power $q \geq 1$, where $\mathbf{\Omega}^{\mathrm{F}} \in \mathbb{R}^{d\times p'}$ is a random Gaussian matrix and $p' \geq \max(2K, K+7)$, then under Assumption 4.1 and the condition that $p \geq 8q + K - 1$ and $n \geq C(dr/p)\tau^2\log^2(n/r)$, there exists some constant $\eta$ such that*

$$\left(\mathbb{E}\|\widetilde{\mathbf{V}}_K^{\mathrm{F}}\widetilde{\mathbf{V}}_K^{\mathrm{F}\top} - \mathbf{V}_K\mathbf{V}_K^\top\|_{\mathrm{F}}^2\right)^{1/2} \lesssim \tau\sqrt{\frac{Kr}{n}} + \tau K\sqrt{\frac{dr}{npL}} + \sqrt{\frac{Kd}{p'}}\left(\eta q^2\tau\sqrt{\frac{dr}{np}}\right)^q. \tag{2}$$

The first term is the same as the bias term in Theorem 4.3, while the extra factor $\sqrt{K}$ in the second term and the third term in (2) come from the extra fast sketches. For the

proof of Theorem 4.4, see Supplementary Materials C. In practice, we recommend to take $q \geq \log(n/p')/\log p$ and $L \geq dK/p$ such that the error rate is of the same order as the full sample PCA. We can see that $q$ only needs to be logarithmic to $n$. A larger $q$ will lead to a smaller statistical error yet a greater computational cost, which reflects the trade-off between computational efficiency and statistical accuracy. When the number of spikes $K$ is unknown and estimated by FADI, the following theorem shows that under appropriate conditions, our estimator $\widehat{K}$ presented in Section 3.3 recovers the true $K$ with high probability.

**Theorem 4.5.** *Let* $\eta_0 = (d/\sqrt{np})\log d$. *When* $d \geq 2$, *under Assumption 4.1 and the conditions that* $p \geq 2K$, $\sqrt{p/d}\log d = O(1)$, $\lambda_1\eta_0^{1/4} = o(1), \Delta \geq \eta_0^{1/4}$ *and* $\log n \geq \log(d^2/p) + 2.5\log\log d + 28/p$, *if we choose* $\mu_0 = \eta_0^{3/4}/12$, *then with probability at least* $1 - d^{-10} - d^{-L/2}$, *we have* $\widehat{K} = K$.

The proof of Theorem 4.5 is deferred to Supplementary Materials D. Theorem 4.5 suggests that when the sample size is sufficiently large, as long as the truncated population covariance matrix is not too ill-conditioned, we can easily estimate $K$ distributively without losing computational efficiency.

# 5 Generalization to the Heterogeneous Residual Variance Model for Non-i.i.d. Data

Sections 2 to 4 focus on the homogeneous spiked covariance model. In this section, we extend FADI to a more general spiked model $\mathbf{\Sigma} = \mathbf{V}_K\mathbf{\Lambda}_K\mathbf{V}_K^\top + \mathbf{D}$, where $\mathbf{D} = \mathrm{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_d^2)$ allows for heterogeneous residual variances. We make a few modifications to generalize FADI developed for the homogeneous spiked covariance matrix case in Section 3 to the heterogeneous case. Specifically, due to the almost low-rank structure of the sample covariance matrix, we can skip Step 1 and directly move to Step 2. The fast sketches will be of the form $\mathbf{Y} = \widehat{\mathbf{\Sigma}}\mathbf{\Omega}$

instead. Then in Step 2, for the $(j,\ell)$-th server on the first layer ($j \in [m], \ell \in [L]$), we calculate $\mathbf{Y}^{(j,\ell)} = (n/m)^{-1} \sum_{i=1}^{n_j} \boldsymbol{X}_i^{(j)} \boldsymbol{X}_i^{(j)\top} \boldsymbol{\Omega}^{(\ell)}$. The rest of the steps remain the same.

Similar to the homogeneous case, we can also further improve the computational efficiency of FADI by generating a $d \times p'$ Gaussian test matrix $\boldsymbol{\Omega}^{\mathrm{F}}$ and taking the $K$ leading eigenvectors of $\widetilde{\mathbf{Y}} = \widetilde{\boldsymbol{\Sigma}}^q \boldsymbol{\Omega}^{\mathrm{F}}$, following the same procedure given in Section 3.2. Then $\widetilde{\mathbf{V}}_K^{\mathrm{F}}$ defined in Section 3.2 is the estimator for $\mathbf{V}_K$.

To study the theoretical property of FADI in this heterogeneous scenario, we will need an extra assumption on the diagonal part such that $\boldsymbol{\Sigma}$ will have an almost low-rank structure. Also, in practice, data might not be i.i.d.. Under appropriate conditions on the convergence of the sample covariance matrix, FADI is applicable to non-i.i.d. data and still enjoys a good error rate. Specifically, we need the following assumption on the sample covariance matrix $\widehat{\boldsymbol{\Sigma}}$.

**Assumption 5.1.** *Suppose* $\{\boldsymbol{X}_i\}_{i=1}^n \subseteq \mathbb{R}^d$ *are random vectors with* $\mathbb{E}(\boldsymbol{X}_i) = \mathbf{0}$*, and the covariance matrices satisfy that* $\lim_{n\to\infty} n^{-1} \sum_{i=1}^n \mathbb{E}(\boldsymbol{X}_i \boldsymbol{X}_i^\top) = \boldsymbol{\Sigma}$*. We assume* $\boldsymbol{\Sigma}$ *has the decomposition:* $\boldsymbol{\Sigma} = \mathbf{V}_K \boldsymbol{\Lambda}_K \mathbf{V}_K^\top + \mathbf{D}$*, where* $\mathbf{D} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2)$*. Let* $\widehat{\boldsymbol{\Sigma}} = n^{-1} \sum_{i=1}^n \boldsymbol{X}_i \boldsymbol{X}_i^\top$ *be the sample covariance matrix. When* $n$ *is sufficiently large, we have*

$$\max\left(\|\mathbf{D}\|_{\mathrm{op}}, \left\|\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{\mathrm{op}}\right\|_{\psi_1}\right) \leq \lambda_1 g(r,n), \tag{3}$$

*where* $g(r,n)$ *is the statistical rate dependent on* $r$ *and* $n$*,* $\|\cdot\|_{\psi_1} = q^{-1} \sup_{q\geq 1} (\mathbb{E}|\cdot|^q)^{1/q}$ *and* $\lambda_1$ *is the largest eigenvalue of* $\boldsymbol{\Sigma}$*.*

**Remark 5.1.** By Definition 5.13 in Vershynin (2012), Assumption 5.1 implies that there exists some constant $c > 0$ such that $\mathbb{P}(\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{\mathrm{op}} \geq s) \leq \exp\left(1 - cs/\lambda_1 g(r,n)\right), \forall s \geq 0$. For the convergence of $\left\|\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{\mathrm{op}}\right\|_{\psi_1}$, it can be shown that $\|\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{\mathrm{op}}\|_{\psi_1} \leq \|\|\widehat{\boldsymbol{\Sigma}} - n^{-1} \sum_{i=1}^n \mathbb{E}(\boldsymbol{X}_i \boldsymbol{X}_i^\top)\|_{\mathrm{op}}\|_{\psi_1} + \|n^{-1} \sum_{i=1}^n \mathbb{E}(\boldsymbol{X}_i \boldsymbol{X}_i^\top) - \boldsymbol{\Sigma}\|_{\mathrm{op}}$, where the first term depends on the convergence of the sample covariance, and the second term depends on the average of the population covariance matrices converges to the limit. While the second term is deterministic,

the first term depends on the dependence structure of the sample. The convergence of the sample covariance matrix for non-i.i.d. samples has been well depicted in many studies (Banna et al., 2016; Fan et al., 2013).

The following theorem provides the error rate of FADI under the heterogeneous scenario.

**Theorem 5.2.** *Suppose* $\{\boldsymbol{X}_i\}_{i=1}^n \subseteq \mathbb{R}^d$ *are random vectors with* $\mathbb{E}(\boldsymbol{X}_i) = \mathbf{0}$. *Then under Assumption 5.1 and the condition that* $\tau g(r,n) \le c\sqrt{p/d}\big(\log\sqrt{d/p}\big)^{-1}$ *for some small enough constant* $c$ *and* $p \ge \max(2K, K+7)$, *for the estimator* $\widetilde{\mathbf{V}}_K$, *we have*

$$\Big(\mathbb{E}\|\widetilde{\mathbf{V}}_K\widetilde{\mathbf{V}}_K^\top - \mathbf{V}_K\mathbf{V}_K^\top\|_{\mathrm{F}}^2\Big)^{1/2} \lesssim \sqrt{K}\tau g(r,n) + \tau\sqrt{\frac{Kd}{Lp}}g(r,n). \tag{4}$$

*Furthermore, if we replace Step 4 by Step 4′ and estimate* $\mathbf{V}_K$ *by* $\widetilde{\mathbf{V}}_K^F$, *then if* $p' \ge \max(2K, K+7)$ *and* $q \le (p-K+1)/8$, *we have*

$$\Big(\mathbb{E}\|\widetilde{\mathbf{V}}_K^F\widetilde{\mathbf{V}}_K^{F\top} - \mathbf{V}_K\mathbf{V}_K^\top\|_{\mathrm{F}}^2\Big)^{1/2} \lesssim \sqrt{K}\tau g(r,n) + K\tau\sqrt{\frac{d}{Lp}}g(r,n) + \sqrt{\frac{Kd}{p'}}\Big(2\eta q^2\tau\sqrt{\frac{d}{p}}g(r,n)\Big)^q. \tag{5}$$

The proof of Theorem 5.2 is given in Supplementary Materials E. Under the heterogeneous scenario, FADI still enjoys a good error rate as long as the sample covariance matrix converges to the population covariance matrix at a reasonable rate, and the residual variance is negligible compared to the spiked component. The condition $\|\mathbf{D}\|_{\mathrm{op}}/\lambda_1 = o(1)$ implied by Assumption 5.1 may look strong at first glance. However, it is often reasonable in practice, especially in multi-ethnic GWAS where the data are heterogeneous. For example, in the 1000 Genomes data, $\lambda_1$ can be as large as 200 whereas $\sigma^2$ is only around 0.7. We can also see from (4) that when $L \asymp d/p$, FADI would still enjoy the same error rate as the full sample PCA on $\widehat{\boldsymbol{\Sigma}}$.

Similar to Section 3.3, when the number of spikes $K$ is unknown, we can estimate $K$ by $\widehat{K}^{(\ell)} = \min\{k : \max_{i\ge k}(\sigma_i(\mathbf{Y}^{(\ell)}) - \sigma_p(\mathbf{Y}^{(\ell)})) \le \sqrt{p}\mu_0\}$ for the $\ell$-th fast sketching with

some prespecified threshold $\mu_0$, and take $\widehat{K} = \text{median}\{\widehat{K}^{(\ell)}\}_{\ell=1}^{L}$ as the estimator of $K$. The following generalized theorem of Theorem 4.5 characterizes the performance of $\widehat{K}$ under the heterogeneous scenario.

**Theorem 5.3.** *Define* $\eta_0 = (\sqrt{d/p})g(r,n)\log d$. *When* $d \geq 2$, *under Assumption 5.1 and the conditions that* $\sqrt{p/d}\log d = O(1)$, $\lambda_1\eta_0^{1/4} = o(1), \Delta \geq \eta_0^{1/4}$ *and* $\log g(r,n) \leq -\log\sqrt{d/p} - \frac{3\log\log d}{p-K+1}$, *if we choose* $\mu_0 = \eta_0^{3/4}/12$, *then with probability at least* $1 - d^{-10} - d^{-L/2}$, $\widehat{K} = K$.

The proof of Theorem 5.3 is deferred to Supplementary Materials F. With Theorem 5.2 and Theorem 5.3, we are able to generalize the results from the homogeneous case to the heterogeneous case.

# 6 Simulation Results

We conduct extensive simulation studies to evaluate the performance of FADI and compare it with several existing methods, including the fast PCA method (Chen et al., 2016), the distributed PCA method (Fan et al., 2019), and the traditional full sample PCA, in terms of the error rates and the computational costs. Note that Chen et al. (2016)'s fast PCA method essentially adopted Halko et al. (2011)'s fast PCA method except that they repeated the fast sketching several times to improve its performance. As we use multiple random sketches in FADI, we hence choose Chen et al. (2016)'s fast PCA method rather than Halko et al. (2011)'s method to make fair comparison.

We generate $\{\boldsymbol{X}_i\}_{i=1}^{n} \overset{\text{i.i.d.}}{\sim} \text{N}(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = \text{diag}\{\lambda, \lambda/2, \lambda/4, 1, \ldots, 1\}$ and $\lambda = 4(\Delta + 1)$. We set the parameters to different values and conduct 100 independent Monte Carlo simulations to evaluate the performance of different methods in different settings. We evaluate the error rate by the measure $\rho(\widehat{\mathbf{V}}_K, \mathbf{V}_K)$, where we abuse the notation and let $\widehat{\mathbf{V}}_K$ represent the estimate of $\mathbf{V}_K$ in each method.

## 6.1 Error Rate and Running Time in Different Settings

Table 2 shows the error rates and the running times of different methods in different configurations of $d, m$ and $n$. For FADI PCA, the number of sketches $L$ is taken as $d/10$ in each setting. We can see the error rate of FADI is very close to those of the distributed PCA (Fan et al., 2019), the fast PCA (Chen et al., 2016), and the full sample PCA, with the error rate ratios around 0.95. Table 2 also demonstrates that FADI is much faster than these existing methods. These results suggest that FADI is computationally much more efficient than the existing PCA methods without sacrificing statistical accuracy.

The long running time of Chen et al. (2016)'s fast PCA method can be attributed to two reasons: first, since their error rate depends on taking the power of the sample covariance matrix, it is highly inconvenient to split samples along the sample size $n$ unless we allow for multiple rounds of communication between the servers. Second and most importantly, the convergence rate of their integration method is slow. It takes more than 100 iterations for the algorithm to converge and the convergence time takes up most of the running time. In comparison, the aggregation method used in FADI is simpler and much faster.

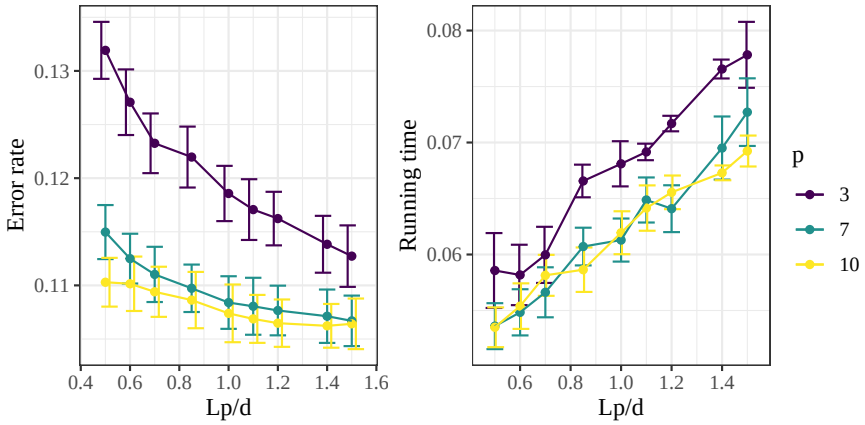



Figure 2: Performance of FADI under different $L$ and $p$ in simulation studies. Under all settings, $K = 3$, $K' = 4$, $p' = p$, $d = 500$, $n = 15000$, $m = 15$, $\Delta = 11.5$ and $q = 7$, where $K$ is the number of spiked eigenvalues, $K'$ is the number of data columns used to estimate $\sigma^2$, $p$ and $p'$ are the dimension of fast sketching in the distributed computing step and the aggregation step, $m$ is the number of splits on $n$ and $\Delta = \lambda_K - \sigma^2$ is the eigengap.

| Error rate | | | | Parameters | | | |
|---|---|---|---|---|---|---|---|
| FADI | Traditional PCA | Distributed PCA | Fast PCA | $d$ | $n$ | $m$ | $L$ |
| 0.068 | 0.065 (0.96) | 0.065 (0.96) | 0.065 (0.96) | 400 | 30000 | 15 | 40 |
| 0.048 | 0.046 (0.96) | 0.046 (0.97) | 0.046 (0.97) | 400 | 60000 | 30 | 40 |
| 0.037 | 0.036 (0.96) | 0.036 (0.97) | 0.036 (0.96) | 400 | 100000 | 50 | 40 |
| 0.052 | 0.050 (0.97) | 0.050 (0.97) | 0.050 (0.97) | 800 | 100000 | 50 | 80 |
| 0.23 | 0.22 (0.96) | 0.23 (0.98) | 0.22 (0.97) | 800 | 5000 | 50 | 80 |
| 0.106 | 0.103 (0.97) | 0.103 (0.97) | 0.101 (0.95) | 800 | 25000 | 50 | 80 |
| 0.073 | 0.070 (0.96) | 0.070 (0.96) | 0.071 (0.98) | 800 | 50000 | 50 | 80 |
| 0.134 | 0.130 (0.96) | 0.130 (0.97) | 0.130 (0.97) | 1600 | 30000 | 15 | 160 |
| 0.095 | 0.092 (0.96) | 0.092 (0.97) | 0.092 (0.96) | 1600 | 60000 | 30 | 160 |
| 0.074 | 0.071 (0.96) | 0.071 (0.97) | 0.071 (0.97) | 1600 | 100000 | 50 | 160 |
| Running time | | | | Parameters | | | |
| FADI | Traditional PCA | Distributed PCA | Fast PCA | $d$ | $n$ | $m$ | $L$ |
| 0.07 | 4.53 (67.9) | 0.59 (8.8) | 3.07 (46.0) | 400 | 30000 | 15 | 40 |
| 0.05 | 8.84 (166.2) | 0.60 (11.2) | 4.47 (84.0) | 400 | 60000 | 30 | 40 |
| 0.05 | 14.84 (285.1) | 0.62 (11.8) | 6.32 (121.5) | 400 | 100000 | 50 | 40 |
| 0.10 | 55.76 (568.7) | 3.66 (37.4) | 16.64 (169.7) | 800 | 100000 | 50 | 80 |
| 0.05 | 3.76 (71.4) | 2.56 (48.6) | 6.36 (120.7) | 800 | 5000 | 50 | 80 |
| 0.07 | 15.07 (220.8) | 2.82 (41.3) | 9.43 (138.1) | 800 | 25000 | 50 | 80 |
| 0.07 | 28.68 (394.7) | 3.23 (44.4) | 11.91 (163.9) | 800 | 50000 | 50 | 80 |
| 0.31 | 80.72 (260.2) | 27.02 (87.1) | 38.87 (125.3) | 1600 | 30000 | 15 | 160 |
| 0.35 | 150.75 (432.7) | 27.29 (78.3) | 45.13 (129.6) | 1600 | 60000 | 30 | 160 |
| 0.34 | 243.83 (721.7) | 27.38 (81.0) | 53.10 (157.2) | 1600 | 100000 | 50 | 160 |

Table 2: Comparison of the error rates and the running times (in seconds) between FADI, full sample traditional PCA, distributed PCA (Fan et al., 2019) and fast PCA (Chen et al., 2016), under different settings of $d, n$ and $m$. Values in the parenthesis represent the error rate ratios or the computational time ratios of each method with respect to FADI. In all settings, $p = p' = 12$, $K = 3$, $K' = 4$, $\Delta = 11.5$ and $q = 7$, where $K'$ is the number of data columns used to estimate $\sigma^2$, $p$ and $p'$ are the dimension of fast sketching in the distributed computation step and the aggregation step, and $q$ is the power parameter for estimating $\widetilde{\mathbf{V}}_K$ from the fast sketching $\widetilde{\mathbf{\Sigma}}^q \mathbf{\Omega}^{\mathrm{F}}$. For the fast PCA method (Chen et al., 2016), the threshold of convergence is taken as $\|D_{\mathrm{F}}(\mathbf{V}_t)\|_{\mathrm{F}} \leq \epsilon = 0.01$, where $D_{\mathrm{F}}(\mathbf{V}_t)$ is the projected gradient evaluated at the $t$-th iteration, and the power of the sample covariance matrix is taken as 3 and there is no splitting over $n$.

Figure 2 shows the performance of FADI for different values of $L$ and $p$. In general, the error rate decreases as $L$ and $p$ increases, but the running time gets longer as the ratio $Lp/d$ grows. Recall Theorem 4.3 shows that when $L$ reaches $d/p$, the error rate of FADI will be the same as the traditional full sample PCA. Thus to guarantee the error rate as well as stable computational cost, we take $Lp/d = 1.2$ in practice (as we do in the simulation).

## 6.2 Performance of FADI on Estimating the Number of Spikes

As discussed in Section 3, FADI requires inputting the number of spikes $K$. In practice, $K$ is usually unknown and needs to be estimated. We conduct simulation studies to evaluate the performance of FADI when $K$ is estimated using the procedure in Section 3.3. We generate $\{\boldsymbol{X}_i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = \mathrm{diag}\{6, 4, 2, 0.5, \ldots, 0.5\} \in \mathbb{R}^{d\times d}$. Table 3 (a) shows that FADI provides accurate estimates of $K$ in different settings when $K$ is unknown, with the error rate of the estimated PCs almost the same as the case where $K$ is known. Furthermore, Table 3 (b) shows that even when one does not have an accurate estimate of $K$, by choosing the input $\widehat{K}$ to be sufficiently large, the error rate is almost identical to when the input $\widehat{K}$ is exactly $K$. Therefore, in practice, we might only need a crude estimate of the range of $K$ rather than recovering $K$ accurately.

## 6.3 Simulation in the Genetic Setting

Section 6.1 compares FADI with several existing methods under a relatively large eigengap. In practice, the eigengap of the population covariance matrix may not be large. To assess different methods in a more realistic scenario, we imitate the setting of the 1000 Genomes data (Consortium et al., 2015), where we take the number of spikes $K = 20$, $\sigma^2 = 0.4$ and the eigengap to be $\Delta = 0.2$. We generate the data by $\{\boldsymbol{X}_i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = \mathrm{diag}(4.8, 2.4, \underbrace{1.2, \ldots, 1.2}_{K-2}, 1 \ldots, 1)$. The dimension is $d = 2504$ and the sample size is $n =$

| (a) | Error rate ($K$ unknown) | Error rate ($K$ known) | Error rate ratio $K$ unknown / $K$ known | $\mathbb{P}(\widehat{K} \neq K)$ | $d$ | $L$ | $p$ |
|---|---|---|---|---|---|---|---|
| | 0.053 | 0.053 | 1.000 | $< 0.01$ | 150 | 26 | 7 |
| | 0.094 | 0.094 | 1.001 | $< 0.01$ | 500 | 60 | 10 |
| | 0.118 | 0.118 | 1.001 | $< 0.01$ | 800 | 80 | 12 |

| (b) | Error rate under different $\widehat{K}$ | | | | $d$ | $L$ | $p$ |
|---|---|---|---|---|---|---|---|
| | 3 | 5 | 7 | 10 | | | |
| | 0.1179 | 0.1180 | 0.1181 | 0.1180 | 800 | 80 | 12 |

Table 3: (a) The performance of FADI in estimating $K$. In all settings $p' = p$, $n = 100,000$, $m = 50$, $K = 3$, $K' = 5$ and $q = 7$. We take the threshold parameter $\mu_0 = 12^{-1}\left(d(np)^{-1/2}\log d\right)^{3/4}$ in each setting; (b) The error rate of FADI when we choose the input parameter $\widehat{K}$ to be $3, 5, 7, 10$ under the setting $d = 800, K = 3, K' = \widehat{K} + 1, n = 100,000, m = 50, p' = p = 12, L = 80$ and $q = 7$.

$160,000$. Error rates and running times are compared using different algorithms under different number of splits $m$ for the sample size $n$. For FADI, we take $L = 75$, $p = p' = 40$ and $q = 7$. $L$ and $p$ for Chen et al. (2016)'s method are set to be the same as in FADI.

Table 4 shows that the number of sample splits $m$ has little impact on the error rate of FADI as expected, while the error rate of Fan et al. (2019)'s distributed PCA increases as $m$ increases. FADI is much faster than the three existing methods in all the practical settings when the eigengap is small. This suggests that in practical problems where the sample size is large and the eigengap is small, FADI not only enjoys much higher computational efficiency compared to the existing methods, but also gives stable estimation for different sample splits along the sample size $n$. Although the settings of small eigengap are of major interest in the more realistic setting, we still conduct simulations where the eigengap increases gradually to see how it affects the performance of FADI. Table 5 shows that as the eigengap gets larger, the error rate of FADI gets closer to that of the traditional full sample PCA, whereas the error rate ratios of distributed PCA and fast PCA to FADI get below 1, but are still above 0.9 when the eigengap is larger than 1. As to the running time, FADI outperforms the other three existing methods in all the settings. In summary, when the eigengap grows larger, the

performance of the four algorithms become similar to what we see in Section 6.1.

| | FADI | Traditional PCA | Distributed PCA | Fast PCA | $m$ |
|---|---|---|---|---|---|
| Error Rate | 2.296 | 1.811 (0.79) | 2.629 (1.15) | 4.122 (1.80) | 10 |
| | 2.294 | 1.811 (0.79) | 3.412 (1.49) | 4.122 (1.80) | 20 |
| | 2.294 | 1.811 (0.79) | 3.955 (1.72) | 4.122 (1.80) | 40 |
| | 2.294 | 1.811 (0.79) | 4.215 (1.84) | 4.122 (1.80) | 80 |
| Running Time | 5.76 | 983.86 (170.8) | 189.76 (32.9) | 464.24 (80.6) | 10 |
| | 3.82 | 992.09 (259.8) | 144.18 (37.8) | 464.24 (121.6) | 20 |
| | 2.86 | 972.47 (339.5) | 119.29 (41.6) | 464.24 (162.1) | 40 |
| | 2.37 | 968.43 (408.5) | 99.39 (41.9) | 464.24 (195.8) | 80 |

Table 4: Comparison of the error rates and running times (in seconds) among FADI, full sample PCA, distributed PCA (Fan et al., 2019) and fast PCA (Chen et al., 2016), using different numbers of sample splits $m$ in the genetic setting. For the fast PCA method of Chen et al. (2016), the threshold of convergence is taken as $\|\mathbf{V}_t - \mathbf{V}_{t+1}\|_{\mathrm{F}} \leq \epsilon = 0.01$, the power of the sample covariance matrix is taken as 3 and there is no sample splitting over $n$.

| | FADI | Traditional PCA | Distributed PCA | Fast PCA | Eigengap |
|---|---|---|---|---|---|
| Error Rate | 1.28 | 1.06 (0.82) | 1.57 (1.22) | 2.30 (1.79) | 0.4 |
| | 0.77 | 0.65 (0.85) | 0.71 (0.92) | 0.78 (1.01) | 0.8 |
| | 0.48 | 0.42 (0.88) | 0.43 (0.90) | 0.43 (0.90) | 1.6 |
| | 0.31 | 0.29 (0.92) | 0.29 (0.93) | 0.29 (0.92) | 3.2 |
| Running Time | 2.76 | 925.15 (334.7) | 115.29 (41.7) | 456.77 (165.2) | 0.4 |
| | 2.77 | 916.52 (331.4) | 114.76 (41.5) | 460.64 (166.6) | 0.8 |
| | 2.69 | 922.85 (342.7) | 114.75 (42.6) | 463.93 (172.3) | 1.6 |
| | 2.77 | 919.20 (332.2) | 115.26 (41.7) | 466.44 (168.6) | 3.2 |

Table 5: Comparison of the error rates and running times (in seconds) among FADI, full sample PCA, distributed PCA (Fan et al., 2019) and fast PCA (Chen et al., 2016) for different eigengaps $\Delta$ in the genetic setting. The number of sample splits $m$ is 40 for FADI and distributed PCA. The settings of the other parameters are the same as those in Table 4.

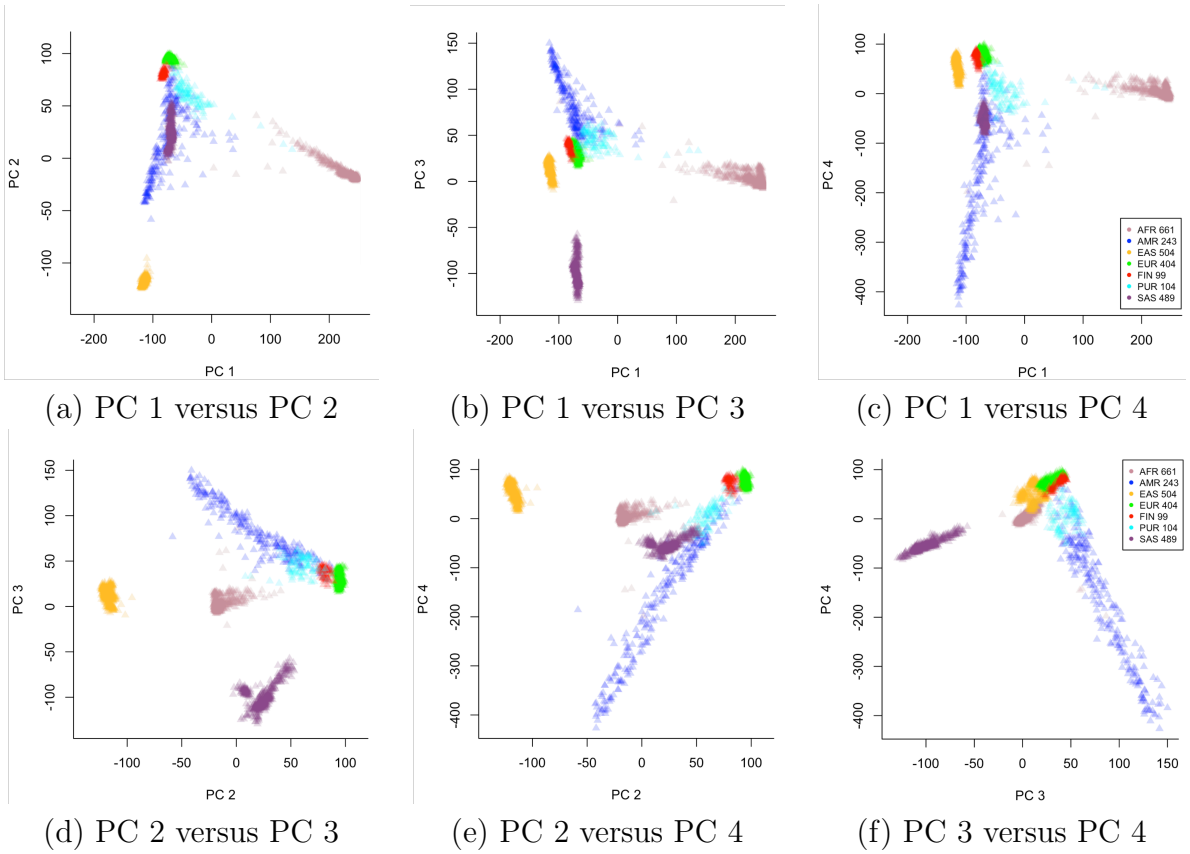
# 7 Application to the 1000 Genomes Data

In this section, we apply FADI and other existing methods to the PCA of the 1000 Genomes data (Consortium et al., 2015). The 1000 Genomes Project performs whole-genome sequencing of a large number of individuals from diverse populations, and aims to establish a

comprehensive public catalogue of human genetic variants (Consortium et al., 2015). We use phase 3 of the 1000 Genomes data and focus on common variants by removing low frequency and rare variants with minor allele frequencies less than 0.05. We perform pair-wise linkage disequilibrium (LD) pruning, with window size of 100, step size of 10 and the $r^2$ threshold at 0.1 (Purcell et al., 2007). There are 2504 subjects in total, and 168,047 independent variants after the LD pruning. As we are interested in estimating ancestry principal components to capture population structure, the sample size $n$ is the number of independent variants after LD pruning ($n = 168,047$), and the dimension $d$ is the number of subjects ($d = 2504$) (Price et al., 2006). The data were collected from 7 super populations: (1) **AFR**: African; (2) **AMR**: Ad Mixed American; (3) **EAS**: East Asian; (4) **EUR**: European; (5) **SAS**: South Asian; (6) **PUR**: Puerto Rican and (7) **FIN**: Finnish; and 26 sub-populations.

We perform FADI with $K' = 27$, $p = 50$, $p' = 100$, $q = 3$, $m = 100$ and $L = 80$. For the estimation of the number of spikes, we take the threshold parameter $\mu_0 = 12^{-1}\left(d(np)^{-1/2}\log d\right)^{3/4}$. The estimated number of spikes from FADI is $\widehat{K} = 26$, which is close to 25, the number of self-reported ethnicity groups minus 1, i.e., $K = 26 - 1$. The results of the 4 leading PCs are shown in Figure 3, where a clear separation can be observed among different super-populations. To estimate the non-spiked eigenvalues/residual variance $\sigma^2$, we sample different 27 individuals out of the 2504 subjects and see how stable the estimate is. We repeat the sampling for 100 times, and the mean of the estimates is $\text{mean}(\widehat{\sigma}^2) = 0.7804$ with the standard deviation $\text{SD}(\widehat{\sigma}^2) = 0.018$. Thus the estimation is stable with respect to different choices of the samples.

Figure 4 (a) demonstrates the correlations between the PCs calculated by FADI and by full sample PCA. We can see that for the 15 leading PCs, the results calculated by FADI are highly correlated to the results calculated by the traditional full sample PCA, whereas the correlations drop afterward. This can be attributed to the fact that the top 15 eigenvalues are well-separated for the sample covariance matrix of the 1000 Genomes data, and the eigengaps



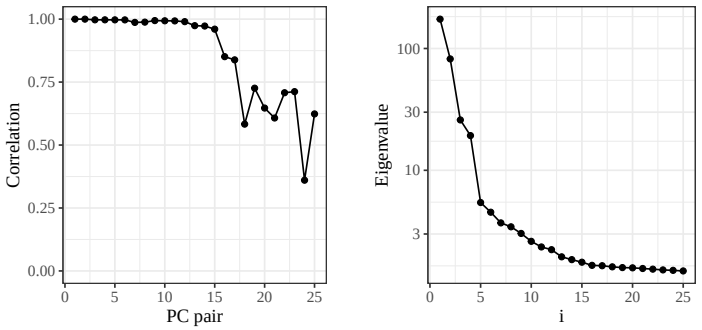

(a) PC 1 versus PC 2 (b) PC 1 versus PC 3 (c) PC 1 versus PC 4

(d) PC 2 versus PC 3 (e) PC 2 versus PC 4 (f) PC 3 versus PC 4

Figure 3: The top 4 principal components of the 1000 Genomes data. For the first two PCs, PC 1 separates African (AFR) super-population from the others, whereas PC 2 separates East Asian (EAS) from the others. As for PC 3 and PC 4, South Asian (SAS) and Ad Mixed American (AMR) are well separated from the rest of the super-populations by PC 3, while PC 4 presents some additional separation.

get smaller after the 15th eigenvalue (see Figure 4 (b)). Figure 5 also shows a good alignment between the PC results calculated by the traditional PCA and FADI.

We compare the computational times of different methods for analyzing the 1000 Genomes data. FADI takes 5.6 seconds at $q = 3$, whereas the traditional PCA method takes 595.4 seconds. When $q = 3$ for each fast sketching, the fast PCA method (Chen et al., 2016) takes 449.2 seconds to complete, and the distributed PCA method (Fan et al., 2019) takes 120.2 seconds. These results show that FADI greatly outperforms the existing PCA methods in terms of computational time.




(a) Correlation for PCs (b) Eigenvalues of $\widehat{\mathbf{\Sigma}}$

Figure 4: (a) Correlations between the 25 leading PCs calculated by FADI and by full sample PCA on the 1000 Genome data; (b) Top 25 eigenvalues for the sample covariance matrix of the 1000 Genomes data.

# 8 Discussions

In this paper, we develop a FAst DIstributed PCA algorithm (FADI) to facilitate scalable PC calculations of large high-dimensional federated data with high computational efficiency and accuracy. FADI is applicable to such data when both the sample size $n$ and the dimension of variables $d$ are large and when the large high-dimensional data are federated and do not allow for sharing individual-level data between sites. The main idea of FADI is to apply random sketches to each split dataset so as to reduce the data dimension and calculate PCs efficiently, and aggregate the results obtained from multiple sketches across split datasets to improve the statistical accuracy and accommodate federated data.

Compared to the traditional PCA that is computationally expensive by directly performing on the full sample covariance matrix and not applicable to federated data, FADI significantly reduces the computational cost from $O(d^2n)$ to $O(nK^2 + d^2K^2 \log n)$, and is applicable to large federated data. Theoretical analysis shows that FADI enjoys the same non-asymptotic error rate as the traditional PCA when the number of repeated sketches $L$ is of order $d/p$. Our simulation results demonstrate that FADI greatly improves the computational efficiency without losing the statistical accuracy compared to the traditional PCA in finite samples.

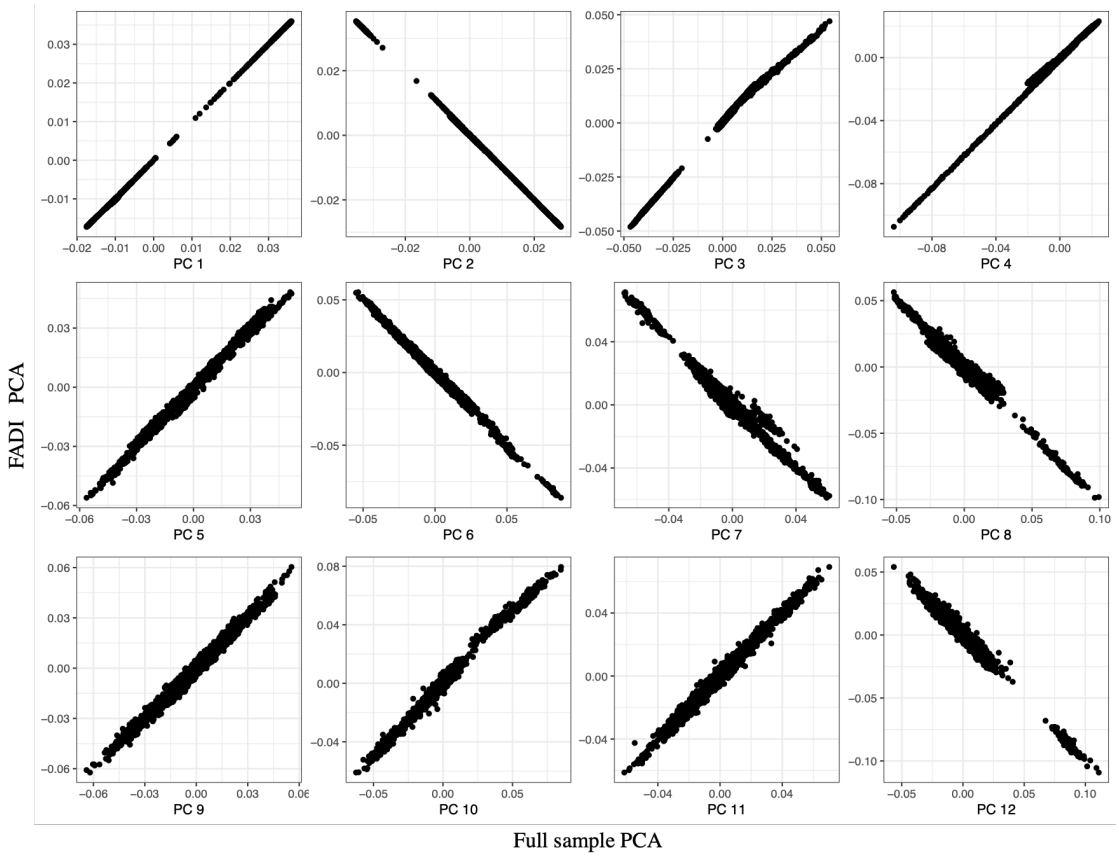



Figure 5: Comparison of the top 12 PCs of the 1000 Genomes data calculated by full sample traditional PCA and by FADI

The existing distributed PCA methods (Liang et al., 2014; Fan et al., 2019) apply traditional PCA to split datasets, but are not scalable when the dimension of variables $d$ is large. The existing fast PCA methods (Halko et al., 2011; Chen et al., 2016) apply random sketches to full data. They allow for large $d$ but are not scalable when the sample size $n$ is large, and are not applicable to federated data. Our proposed fast distributed PCA method FADI overcomes the limitations of the distributed PCA methods and the fast PCA methods by calculating multiple random sketches to split datasets and efficiently aggregating the results across them.

Our theoretical study shows that the computational complexity of FADI is of an order of magnitude smaller than the existing methods. Our simulation results show that FADI reduces

the running time by almost 100 times for large high dimensional data while maintaining the same level of error rate, compared with the existing distributed PCA method and the existing fast PCA method. Furthermore, in a more realistic genetic setting where the eigengap is small, FADI outperforms distributed PCA (Fan et al., 2019) and fast PCA (Chen et al., 2016) in both the error rate and the running time. Since in practice the number of spikes $K$ is usually unknown, FADI provides a method to estimate $K$ by aggregating the number of spikes in multiple sketches, and can recover $K$ with high probability.

We begin with presenting the method in the homogeneous case with i.i.d. data by assuming the non-spiked eigenvalues to be identical. We then generalize the method to the heterogeneous case with possibly non-i.i.d. data under certain regularity conditions on the norm of the non-spiked component and the convergence rate of the sample covariance matrix. In our theoretical investigation, we require $n \gg d$. Our simulation results indicate that in practice even when $n$ is not much larger than $d$, our method still performs very well. It would be of future research interest to extend the results to the case where $d > n$.

Fast PCA algorithms using random sketches usually require the covariance matrix to have a certain "almost low-rank" structure. Without the spiked structure, approximation by random sketches might not be accurate (Halko et al., 2011). It is of future research interest to investigate whether the proposed FADI approach can be extended to non-spiked models. In Step 4 of FADI, we aggregate local estimators by taking a simple average over the projection matrices. It would be of future research interest to explore the performance of other weighted average and investigate the best convex combination to reduce the statistical error.

# References

Abbe, E., Fan, J., Wang, K., and Zhong, Y. (2020). Entrywise eigenvector analysis of random matrices with low expected rank. *The Annals of Statistics*, 48(3):1452–1474.

Achlioptas, D. (2003). Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *Journal of Computer and System Sciences*, 66(4):671–687. Special issue on PODS 2001 (Santa Barbara, CA).

Banna, M., Merlevède, F., and Youssef, P. (2016). Bernstein-type inequality for a class of dependent random matrices. *Random Matrices: Theory and Applications*, 05(02):1650006.

Chen, T.-L., Chang, D. D., Huang, S.-Y., Chen, H., Lin, C., and Wang, W. (2016). Integrating multiple random sketches for singular value decomposition. *arXiv*.

Consortium, . G. P. et al. (2015). A global reference for human genetic variation. *Nature*, 526(7571):68.

Dey, R., Zhou, W., Kiiskinen, T., et al. (2020). An efficient and accurate frailty model approach for genome-wide survival association analysis controlling for population structure and relatedness in large-scale biobanks. *bioRxiv*.

Dhruva, S. S., Ross, J. S., Akar, J. G., et al. (2020). Aggregating multiple real-world data sources using a patient-centered health-data-sharing platform. *npj Digital Medicine*, 3(1):60.

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

Fan, J., Wang, D., Wang, K., and Zhu, Z. (2019). Distributed estimation of principal eigenspaces. *The Annals of Statistics*, 47(6):3009–3031.

Franklin, J. N. (2012). *Matrix theory*. Courier Corporation.

Halko, N., Martinsson, P. G., and Tropp, J. A. (2011). Finding structure with randomness: probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53(2):217–288.

Horn, R. A. and Johnson, C. R. (1990). Norms for vectors and matrices. *Matrix Analysis*, pages 313–386.

Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *The Annals of Statistics*, 29(2):295–327.

Jordan, M. I., Lee, J. D., and Yang, Y. (2019). Communication-efficient distributed statistical inference. *Journal of the American Statistical Association*, 114(526):668–681. DOI: 10.1080/01621459.2018.1429274.

Kannan, R., Vempala, S., and Woodruff, D. (2014). Principal component analysis and higher correlations for distributed data. In *Proceedings of The 27th Conference on Learning Theory*, volume 35 of *Proceedings of Machine Learning Research*, pages 1040–1057, Barcelona, Spain. PMLR.

Kargupta, H., Huang, W., Sivakumar, K., and Johnson, E. (2001). Distributed clustering using collective principal component analysis. *Knowledge and Information Systems*, 3(4):422–448.

Klarin, D., Damrauer, S. M., Cho, K., et al. (2018). Genetics of blood lipids among ˜300,000 multi-ethnic participants of the million veteran program. *Nature Genetics*, 50(11):1514–1523. PMCID: PMC6521726.

Li, T., Sahu, A. K., Talwalkar, A., and Smith, V. (2020). Federated learning: challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3):50–60. DOI: 10.1109/MSP.2020.2975749.

Liang, Y., Balcan, M.-F. F., Kanchanapally, V., and Woodruff, D. (2014). Improved distributed principal component analysis. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Pasini, G. (2017). Principal component analysis for stock portfolio management. *International Journal of Pure and Applied Mathematics*, 115:153–167.

Price, A. L., Patterson, N. J., Plenge, R. M., et al. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8):904–909.

Purcell, S., Neale, B., Todd-Brown, K., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, 81(3):559–575.

Reich, D., Price, A. L., and Patterson, N. (2008). Principal component analysis of genetic data. *Nature Genetics*, 40(5):491–492.

Sudlow, C., Gallacher, J., Allen, N., et al. (2015). UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLOS Medicine*, 12(3):e1001779–e1001779.

Vershynin, R. (2012). *Introduction to the non-asymptotic analysis of random matrices*, page 210–268. Cambridge University Press.

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., and Yang, J. (2017). 10 years of gwas discovery: biology, function, and translation. *The American Journal of Human Genetics*, 101(1):5–22.

Wedin, P.-A. (1972). Perturbation bounds in connection with singular value decomposition. *Nordisk Tidskrift for Informationsbehandling*, 12:99–111.

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

*Supplementary Materials to*

# Fast Distributed Principal Component Analysis of Large-Scale Federated Data

This file contains the supplementary materials to the paper "Fast Distributed Principal Component Analysis of Large-Scale Federated Data". In Supplementary Materials Section A, we prove Lemma 4.2 that shows that the bias of our proposed fast distributed PCA (FADI) estimator has the same error rate as the traditional PCA. In Supplementary Materials Section B, we give the proof for the main theorem on the error rate in the homogeneous covariance case. In Supplementary Materials Section C, we prove the theorem that controls the statistical error of the modified version of FADI where in the aggregation step we conduct another fast sketching. In Supplementary Materials Section D, we show that FADI can recover the number of spikes $K$ with high probability when $K$ is unknown. Supplementary Materials Section E and Section F provide the proof for the generalized theorems under the heterogeneous residual variance setting. Supplementary Materials Section G gives the proofs for the technical lemmas that are used in the proofs of the main theorems. Finally in Supplementary Materials Section H, we present the modified version of the Wedin's theorem, which is used in several proofs.

Before we begin with the proofs, we introduce a few notations to be used later. For a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$, we use $\sigma_i(\mathbf{A})$ (respectively $\lambda_i(\mathbf{A})$) to represent the $i$-th largest singular value (eigenvalue) of $\mathbf{A}$, and $\sigma_{\min}(\mathbf{A})$ (respectively $\lambda_{\min}(\mathbf{A})$) stands for the smallest singular value (respectively eigenvalue) of $\mathbf{A}$. If $\mathbf{A}$ has the singular value decomposition $\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^\top$, then we denote by $\mathbf{A}^\dagger = \mathbf{U}\mathbf{\Lambda}^{-1}\mathbf{V}^\top$ the pseudo-inverse of $\mathbf{A}$. Let $\mathbb{I}\{\cdot\}$ denote an indicator function, which takes 1 if the statement inside $\{\cdot\}$ is true and 0 otherwise. For two symmetric positive semi-definite matrices $\mathbf{A}$ and $\mathbf{B}$, we say $\mathbf{A} \succeq \mathbf{B}$ if $\mathbf{A} - \mathbf{B}$ is symmetric positive semi-definite.

# A Proof of Lemma 4.2

The proof is similar to the proof of Theorem 2 in Fan et al. (2019), but our randomness comes from the fast sketching along the dimension $d$, while the randomness in Fan et al. (2019)'s distributed PCA algorithm comes from the splitting of the sample size $n$. Besides, we do not require symmetric assumptions on the distribution of $\{\boldsymbol{X}_i\}_{i=1}^n$ as Fan et al. (2019) did in their paper.

Now we start with the proof. For any $j \in [d]$, let $\mathbf{D}_j = \mathbf{I}_d - 2\mathbf{e}_j\mathbf{e}_j^\top$, and $\boldsymbol{\Omega} \in \mathbb{R}^{d\times p}$ be a random matrix with i.i.d. Gaussian entries. Recall that $\widehat{\boldsymbol{\Sigma}}^{\mathrm{tr}} = \widehat{\boldsymbol{\Sigma}} - \widehat{\sigma}^2\mathbf{I}_d = \widehat{\mathbf{V}}\widehat{\boldsymbol{\Lambda}}\widehat{\mathbf{V}}^\top$ is the truncated sample covariance matrix, then conditional on $\widehat{\boldsymbol{\Sigma}}^{\mathrm{tr}}$, we have

$$\begin{aligned}
&\widehat{\mathbf{V}}\mathbf{D}_j\widehat{\mathbf{V}}^\top\mathbf{Y}^{(\ell)}\mathbf{Y}^{(\ell)\top}\widehat{\mathbf{V}}\mathbf{D}_j\widehat{\mathbf{V}}^\top = \widehat{\mathbf{V}}\mathbf{D}_j\widehat{\mathbf{V}}^\top\widehat{\mathbf{V}}\widehat{\boldsymbol{\Lambda}}\widehat{\mathbf{V}}^\top\boldsymbol{\Omega}^{(\ell)}\boldsymbol{\Omega}^{(\ell)\top}\widehat{\mathbf{V}}\widehat{\boldsymbol{\Lambda}}\widehat{\mathbf{V}}^\top\widehat{\mathbf{V}}\mathbf{D}_j\widehat{\mathbf{V}}^\top \\
&= \widehat{\mathbf{V}}\widehat{\boldsymbol{\Lambda}}(\mathbf{D}_j\widehat{\mathbf{V}}^\top\boldsymbol{\Omega}^{(\ell)})(\boldsymbol{\Omega}^{(\ell)\top}\widehat{\mathbf{V}}\mathbf{D}_j)\widehat{\boldsymbol{\Lambda}}\widehat{\mathbf{V}}^\top \overset{d}{=} \widehat{\mathbf{V}}\widehat{\boldsymbol{\Lambda}}\widehat{\mathbf{V}}^\top\boldsymbol{\Omega}^{(\ell)}\boldsymbol{\Omega}^{(\ell)\top}\widehat{\mathbf{V}}\widehat{\boldsymbol{\Lambda}}\widehat{\mathbf{V}}^\top = \mathbf{Y}^{(\ell)}\mathbf{Y}^{(\ell)\top},
\end{aligned}$$

where the second equality is due to the fact that diagonal matrices are commutative, and the last but one equivalence in distribution is due to the fact that $\mathbf{D}_j\widehat{\mathbf{V}}^\top\boldsymbol{\Omega}^{(\ell)} \overset{d}{=} \widehat{\mathbf{V}}^\top\boldsymbol{\Omega}^{(\ell)}$. As we know the top $K$ eigenvectors of $\widehat{\mathbf{V}}\mathbf{D}_j\widehat{\mathbf{V}}^\top\mathbf{Y}^{(\ell)}\mathbf{Y}^{(\ell)\top}\widehat{\mathbf{V}}\mathbf{D}_j\widehat{\mathbf{V}}^\top$ are $\widehat{\mathbf{V}}\mathbf{D}_j\widehat{\mathbf{V}}^\top\widehat{\mathbf{V}}_K^{(\ell)}$, we have $\widehat{\mathbf{V}}\mathbf{D}_j\widehat{\mathbf{V}}^\top\widehat{\mathbf{V}}_K^{(\ell)} \overset{d}{=} \widehat{\mathbf{V}}_K^{(\ell)}$. Hence we have

$$\begin{aligned}
\widehat{\mathbf{V}}^\top\mathbb{E}\left(\widehat{\mathbf{V}}_K^{(\ell)}\widehat{\mathbf{V}}_K^{(\ell)\top}|\widehat{\boldsymbol{\Sigma}}^{\mathrm{tr}}\right)\widehat{\mathbf{V}} &= \widehat{\mathbf{V}}^\top\widehat{\mathbf{V}}\mathbf{D}_j\widehat{\mathbf{V}}^\top\mathbb{E}\left(\widehat{\mathbf{V}}_K^{(\ell)}\widehat{\mathbf{V}}_K^{(\ell)\top}|\widehat{\boldsymbol{\Sigma}}^{\mathrm{tr}}\right)\widehat{\mathbf{V}}\mathbf{D}_j\widehat{\mathbf{V}}^\top\widehat{\mathbf{V}} \\
&= \mathbf{D}_j\widehat{\mathbf{V}}^\top\mathbb{E}\left(\widehat{\mathbf{V}}_K^{(\ell)}\widehat{\mathbf{V}}_K^{(\ell)\top}|\widehat{\boldsymbol{\Sigma}}^{\mathrm{tr}}\right)\widehat{\mathbf{V}}\mathbf{D}_j.
\end{aligned}$$

The above equation holds for any $j \in [d]$, which suggests that $\widehat{\mathbf{V}}^\top\mathbb{E}\left(\widehat{\mathbf{V}}_K^{(\ell)}\widehat{\mathbf{V}}_K^{(\ell)\top}|\widehat{\boldsymbol{\Sigma}}^{\mathrm{tr}}\right)\widehat{\mathbf{V}}$ is diagonal and that $\boldsymbol{\Sigma}'$ and $\widehat{\boldsymbol{\Sigma}}^{\mathrm{tr}}$ share the same set of eigenvectors.

Now under the condition that $\left\|\boldsymbol{\Sigma}' - \widehat{\mathbf{V}}_K\widehat{\mathbf{V}}_K^\top\right\|_{\mathrm{op}} < 1/2$, for any $j \in [K]$, denote by $\widehat{\mathbf{v}}_j$ the

$j$-th column of $\widehat{\mathbf{V}}_K$. We have

$$\|\boldsymbol{\Sigma}'\widehat{\mathbf{v}}_j\| = \left\|\left(\boldsymbol{\Sigma}' - \widehat{\mathbf{V}}_K\widehat{\mathbf{V}}_K^\top + \widehat{\mathbf{V}}_K\widehat{\mathbf{V}}_K^\top\right)\widehat{\mathbf{v}}_j\right\| \geq 1 - \left\|\boldsymbol{\Sigma}' - \widehat{\mathbf{V}}_K\widehat{\mathbf{V}}_K^\top\right\|_{\text{op}} > 1 - \frac{1}{2} = \frac{1}{2}.$$

In other words, the corresponding eigenvalue of $\widehat{\mathbf{v}}_j$ in $\boldsymbol{\Sigma}'$ is larger than $1/2$. On the other hand, by Weyl's inequality (Franklin, 2012), the rest of the $d-K$ eigenvalues of $\boldsymbol{\Sigma}'$ should be less than $1/2$. Therefore, $\widehat{\mathbf{V}}_K$ are still the leading $K$ eigenvectors for $\boldsymbol{\Sigma}'$, and thus $\|\mathbf{V}'_K\mathbf{V}'^\top_K - \widehat{\mathbf{V}}_K\widehat{\mathbf{V}}_K^\top\|_{\text{op}} = 0$.

# B Proof of Theorem 4.3

Before delving into the proof, the following two lemmas provide some important properties of the random Gaussian matrix.

**Lemma B.1.** Let $\boldsymbol{\Omega} \in \mathbb{R}^{d\times p}$ be a random matrix with i.i.d. Gaussian entries, where $p < d$. For a random variable, define the $\psi_1$ norm to be $\|\cdot\|_{\psi_1} = \sup_{q\geq 1}(\mathbb{E}|\cdot|^q)^{1/q}/q$. Then we have the following bound on the $\psi_1$ norm of the matrix $\boldsymbol{\Omega}/\sqrt{p}$:

$$\|\|\boldsymbol{\Omega}/\sqrt{p}\|_{\text{op}}\|_{\psi_1} \leq \sqrt{d/p}. \tag{S.6}$$

**Lemma B.2.** Let $\boldsymbol{\Omega}$ denote a $K$ by $p$ matrix with i.i.d. Gaussian entries, where $p \geq K+1$. For any integer $a$ such that $1 \leq a \leq (p-K+1)/2$, there exists a constant $C > 0$ such that

$$\mathbb{E}\left(\{\sigma_{\min}(\boldsymbol{\Omega}/\sqrt{p})\}^{-a}\right) \leq C^a. \tag{S.7}$$

The following lemma shows that $\|\boldsymbol{\Sigma}' - \mathbf{V}_K\mathbf{V}_K\|_{\text{op}}$ and $\|\boldsymbol{\Sigma}' - \widehat{\mathbf{V}}_K\widehat{\mathbf{V}}_K^\top\|_{\text{op}}$ are bounded by a small constant with high probability.

**Lemma B.3.** If $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \in \mathbb{R}^d$ are i.i.d. sub-Gaussian random vectors with the covariance matrix $\boldsymbol{\Sigma}$, and $n > r$, $p \geq \max(2K, K+7)$, where $r = \mathrm{tr}(\boldsymbol{\Sigma})/\|\boldsymbol{\Sigma}\|_{\mathrm{op}}$, then there exists a constant $C_0 > 0$ such that for any $\varepsilon > 0$, we have

$$\max\left\{\mathbb{P}\Big(\|\boldsymbol{\Sigma}' - \mathbf{V}_K\mathbf{V}_K\|_{\mathrm{op}} \geq \varepsilon\Big), \mathbb{P}\left(\|\boldsymbol{\Sigma}' - \widehat{\mathbf{V}}_K\widehat{\mathbf{V}}_K^\top\|_{\mathrm{op}} \geq \varepsilon\right)\right\} \leq \exp\left(-\frac{C_0\varepsilon}{\tau\sqrt{dr/np}}\right).$$

The proof of Lemma B.1, Lemma B.2 and Lemma B.3 are deferred to Supplementary Materials Section G.

Now we can start with the proof. We first decompose the bias term into two parts,

$$\left(\mathbb{E}\|\widetilde{\mathbf{V}}_K\widetilde{\mathbf{V}}_K^\top - \mathbf{V}_K\mathbf{V}_K^\top\|_{\mathrm{F}}^2\right)^{1/2} \leq \underbrace{\left(\mathbb{E}\|\widetilde{\mathbf{V}}_K\widetilde{\mathbf{V}}_K^\top - \mathbf{V}'_K\mathbf{V}'^\top_K\|_{\mathrm{F}}^2\right)^{1/2}}_{\mathrm{I}} + \underbrace{\left(\mathbb{E}\|\mathbf{V}'_K\mathbf{V}'^\top_K - \mathbf{V}_K\mathbf{V}_K^\top\|_{\mathrm{F}}^2\right)^{1/2}}_{\mathrm{II}}. \tag{S.8}$$

Term I can be regarded as the variance term, whereas term II is the bias term. We will consider the bias term first.

## B.1 Control of the Bias Term

We can see that term II can be further decomposed into two terms

$$\left(\mathbb{E}\|\mathbf{V}'_K\mathbf{V}'^\top_K - \mathbf{V}_K\mathbf{V}_K^\top\|_{\mathrm{F}}^2\right)^{1/2} \leq \left(\mathbb{E}\|\mathbf{V}'_K\mathbf{V}'^\top_K - \widehat{\mathbf{V}}_K\widehat{\mathbf{V}}_K^\top\|_{\mathrm{F}}^2\right)^{1/2} + \left(\mathbb{E}\|\widehat{\mathbf{V}}_K\widehat{\mathbf{V}}_K^\top - \mathbf{V}_K\mathbf{V}_K^\top\|_{\mathrm{F}}^2\right)^{1/2}.$$

We can bound both terms separately. First note that $\|\mathbf{V}'_K\mathbf{V}'^\top_K - \widehat{\mathbf{V}}_K\widehat{\mathbf{V}}_K^\top\|_{\mathrm{F}} \leq \sqrt{2K}\|\mathbf{V}'_K\mathbf{V}'^\top_K - \widehat{\mathbf{V}}_K\widehat{\mathbf{V}}_K^\top\|_{\mathrm{op}} \leq \sqrt{2K}$. Thus we have,

$$\Big(\mathbb{E}\|\mathbf{V}'_K\mathbf{V}'^{\top}_K - \widehat{\mathbf{V}}_K\widehat{\mathbf{V}}_K^{\top}\|_{\mathrm{F}}^2\Big)^{1/2} \le \Big(\mathbb{E}\|\mathbf{V}'_K\mathbf{V}'^{\top}_K - \widehat{\mathbf{V}}_K\widehat{\mathbf{V}}_K^{\top}\|_{\mathrm{F}}^2\mathbb{I}\big\{\|\boldsymbol{\Sigma}' - \widehat{\mathbf{V}}_K\widehat{\mathbf{V}}_K^{\top}\|_{\mathrm{op}} \ge 1/2\big\}\Big)^{1/2}$$
$$+ \Big(\mathbb{E}\|\mathbf{V}'_K\mathbf{V}'^{\top}_K - \widehat{\mathbf{V}}_K\widehat{\mathbf{V}}_K^{\top}\|_{\mathrm{F}}^2\mathbb{I}\big\{\|\boldsymbol{\Sigma}' - \widehat{\mathbf{V}}_K\widehat{\mathbf{V}}_K^{\top}\|_{\mathrm{op}} < 1/2\big\}\Big)^{1/2} \lesssim 0 + \sqrt{K}\left\{\mathbb{P}\Big(\|\boldsymbol{\Sigma}' - \widehat{\mathbf{V}}_K\widehat{\mathbf{V}}_K^{\top}\|_{\mathrm{op}} \ge 1/2\Big)\right\}^{1/2}$$
$$\le \sqrt{K}\exp\left(-\frac{C_0}{4\tau\sqrt{dr/np}}\right),$$

where the last but one inequality follows from Lemma 4.2, and the last inequality is a result of Lemma B.3.

As for term II in (S.8), by Davis-Kahan's Theorem (Yu et al., 2015), we have

$$\Big(\mathbb{E}\|\widehat{\mathbf{V}}_K\widehat{\mathbf{V}}_K^{\top} - \mathbf{V}_K\mathbf{V}_K^{\top}\|_{\mathrm{F}}^2\Big)^{1/2} \lesssim \frac{\sqrt{K}}{\Delta}\Big(\mathbb{E}\|\widehat{\boldsymbol{\Sigma}}^{\mathrm{tr}} - \mathbf{V}_K\boldsymbol{\Lambda}_K\mathbf{V}_K^{\top}\|_{\mathrm{op}}^2\Big)^{1/2} = \frac{\sqrt{K}}{\Delta}\Big(\mathbb{E}\|\widehat{\mathbf{E}}\|_{\mathrm{op}}^2\Big)^{1/2}$$
$$\lesssim \frac{\sqrt{K}}{\Delta}\|\|\widehat{\mathbf{E}}\|_{\mathrm{op}}\|_{\psi_1},$$

where $\widehat{\mathbf{E}} = \widehat{\boldsymbol{\Sigma}}^{\mathrm{tr}} - \mathbf{V}_K\boldsymbol{\Lambda}_K\mathbf{V}_K^{\top} = \widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma} + (\sigma^2 - \widehat{\sigma}^2)\mathbf{I}$. Consider the upper-left $K' \times K'$ submatrix of $\boldsymbol{\Sigma}$, which we denote by $\boldsymbol{\Sigma}_{K'}$. We have $\boldsymbol{\Sigma}_{K'} = \sigma^2\mathbf{I}_{K'} + (\mathbf{V}_K)_{[1:K',:]}\boldsymbol{\Lambda}_K(\mathbf{V}_K)_{[1:K',:]}^{\top}$, where $(\mathbf{V}_K)_{[1:K',:]}$ is the submatrix of $\mathbf{V}_K$ composed of the first $K'$ rows. Then since $(\mathbf{V}_K)_{[1:K',:]}\boldsymbol{\Lambda}_K(\mathbf{V}_K)_{[1:K',:]}^{\top} \succeq \mathbf{0}$ and $\mathrm{rank}\big((\mathbf{V}_K)_{[1:K',:]}\boldsymbol{\Lambda}_K(\mathbf{V}_K)_{[1:K',:]}^{\top}\big) \le K$, we know that $\sigma_{\min}(\boldsymbol{\Sigma}_{K'}) = \sigma^2$. By Weyl's inequality (Franklin, 2012), we know $|\sigma^2 - \widehat{\sigma}^2| \le \|\widehat{\boldsymbol{\Sigma}}_{K'} - \boldsymbol{\Sigma}_{K'}\|_{\mathrm{op}} \le \|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{\mathrm{op}}$. Thus we have $\|\widehat{\mathbf{E}}\|_{\mathrm{op}} \le \|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{\mathrm{op}} + |\sigma^2 - \widehat{\sigma}^2| \lesssim \|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{\mathrm{op}}$. Then by Lemma 3 in Fan et al. (2019), we have $\|\|\widehat{\mathbf{E}}\|_{\mathrm{op}}\|_{\psi_1} \lesssim \|\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{\mathrm{op}}\|_{\psi_1} \lesssim \lambda_1\sqrt{r/n}$. Therefore, the bound for the bias term is

$$\mathrm{II} \lesssim \sqrt{K}\exp\left(-\frac{C_0}{4\tau\sqrt{dr/np}}\right) + \tau\sqrt{\frac{Kr}{n}}.$$

Under the condition that $n \ge C(dr/p)\tau^2\log^2(n/r)$ for some large enough constant $C$, the term $\tau\sqrt{\frac{Kr}{n}}$ is dominant.

## B.2 Control of the Variance Term

Now we move on to control the variance term. Suppose that $\left\|\mathbf{\Sigma}' - \mathbf{V}_K\mathbf{V}_K^T\right\|_{\text{op}} < 1/4$. Then by Weyl's inequality (Franklin, 2012) we have that $\sigma_K(\mathbf{\Sigma}') > 1 - 1/4 = 3/4$ and $\sigma_{K+1}(\mathbf{\Sigma}') < 1/4$. Thus by Davis-Kahan theorem Yu et al. (2015)

$$\begin{aligned}
&\left(\mathbb{E}\left(\|\widetilde{\mathbf{V}}_K\widetilde{\mathbf{V}}_K^\top - \mathbf{V}'_K\mathbf{V}'^\top_K\|_{\text{F}}^2 \mathbb{I}\left\{\left\|\mathbf{\Sigma}' - \mathbf{V}_K\mathbf{V}_K^T\right\|_{\text{op}} < 1/4\right\}\right)\right)^{1/2} \\
&\lesssim \left(\mathbb{E}\left(\frac{\|\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}'\|_{\text{F}}^2}{\left(\sigma_K(\mathbf{\Sigma}') - \sigma_{K+1}(\mathbf{\Sigma}')\right)^2}\mathbb{I}\left\{\left\|\mathbf{\Sigma}' - \mathbf{V}_K\mathbf{V}_K^T\right\|_{\text{op}} < 1/4\right\}\right)\right)^{1/2} \\
&\lesssim \left(\mathbb{E}\left(\|\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}'\|_{\text{F}}^2 \mathbb{I}\left\{\left\|\mathbf{\Sigma}' - \mathbf{V}_K\mathbf{V}_K^T\right\|_{\text{op}} < 1/4\right\}\right)\right)^{1/2} \leq \underbrace{\left(\mathbb{E}\|\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}'\|_{\text{F}}^2\right)^{1/2}}_{\text{III}}.
\end{aligned}$$

We will bound term III later. Also similar as previously, note that $\|\widetilde{\mathbf{V}}_K\widetilde{\mathbf{V}}_K^\top - \mathbf{V}'_K\mathbf{V}'^\top_K\|_{\text{F}} \leq \sqrt{2K}$. Thus by Lemma B.3,

$$\begin{aligned}
&\left(\mathbb{E}\left(\|\widetilde{\mathbf{V}}_K\widetilde{\mathbf{V}}_K^\top - \mathbf{V}'_K\mathbf{V}'^\top_K\|_{\text{F}}^2 \mathbb{I}\left\{\left\|\mathbf{\Sigma}' - \mathbf{V}_K\mathbf{V}_K^T\right\|_{\text{op}} \geq 1/4\right\}\right)\right)^{1/2} \lesssim \sqrt{K}\left\{\mathbb{P}\left(\left\|\mathbf{\Sigma}' - \mathbf{V}_K\mathbf{V}_K^T\right\|_{\text{op}} \geq 1/4\right)\right\}^{1/2} \\
&\leq \sqrt{K}\exp\left(-\frac{C_0}{8\tau\sqrt{dr/np}}\right).
\end{aligned}$$

Therefore, we have

$$\left(\mathbb{E}\|\widetilde{\mathbf{V}}_K\widetilde{\mathbf{V}}_K^\top - \mathbf{V}'_K\mathbf{V}'^\top_K\|_{\text{F}}^2\right)^{1/2} \lesssim \sqrt{K}\exp\left(-\frac{C_0}{8\tau\sqrt{dr/np}}\right) + \underbrace{\left(\mathbb{E}\|\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}'\|_{\text{F}}^2\right)^{1/2}}_{\text{III}}.$$

Now we move on to bound term III.

$$\begin{aligned}
\left(\mathbb{E}\|\widetilde{\mathbf{\Sigma}}-\mathbf{\Sigma}'\|_{\mathrm{F}}^2\right)^{1/2} &= \left(\mathbb{E}\left\|L^{-1}\sum_{j=1}^{L}\widehat{\mathbf{V}}_K^{(\ell)}\widehat{\mathbf{V}}_K^{(\ell)\top}-\mathbb{E}\left(\widehat{\mathbf{V}}_K^{(1)}\widehat{\mathbf{V}}_K^{(1)\top}|\widehat{\mathbf{\Sigma}}^{\mathrm{tr}}\right)\right\|_{\mathrm{F}}^2\right)^{1/2} \\
&= \left(\mathbb{E}\left(\mathbb{E}\left(\left\|L^{-1}\sum_{j=1}^{L}\widehat{\mathbf{V}}_K^{(\ell)}\widehat{\mathbf{V}}_K^{(\ell)\top}-\mathbb{E}\left(\widehat{\mathbf{V}}_K^{(1)}\widehat{\mathbf{V}}_K^{(1)\top}|\widehat{\mathbf{\Sigma}}^{\mathrm{tr}}\right)\right\|_{\mathrm{F}}^2\middle|\widehat{\mathbf{\Sigma}}^{\mathrm{tr}}\right)\right)\right)^{1/2} \\
&= \frac{1}{\sqrt{L}}\left(\mathbb{E}\left\|\widehat{\mathbf{V}}_K^{(\ell)}\widehat{\mathbf{V}}_K^{(\ell)\top}-\mathbb{E}\left(\widehat{\mathbf{V}}_K^{(1)}\widehat{\mathbf{V}}_K^{(1)\top}|\widehat{\mathbf{\Sigma}}^{\mathrm{tr}}\right)\right\|_{\mathrm{F}}^2\right)^{1/2} \\
&\leq \frac{1}{\sqrt{L}}\left(\mathbb{E}\left\|\widehat{\mathbf{V}}_K^{(\ell)}\widehat{\mathbf{V}}_K^{(\ell)\top}-\mathbf{V}_K\mathbf{V}_K^{\top}\right\|_{\mathrm{F}}^2\right)^{1/2}+\frac{1}{\sqrt{L}}\left(\mathbb{E}\left\|\mathbf{V}_K\mathbf{V}_K^{\top}-\mathbf{\Sigma}'\right\|_{\mathrm{F}}^2\right)^{1/2}.
\end{aligned}$$

where the last but one equality is due to the independence of estimators from different sketches conditional on $\widehat{\mathbf{\Sigma}}^{\mathrm{tr}}$ and the properties of Frobenius norm (interchangeability of expectation and trace). By Jensen's inequality, we have

$$\frac{1}{\sqrt{L}}\left(\mathbb{E}\left\|\mathbf{V}_K\mathbf{V}_K^{\top}-\mathbf{\Sigma}'\right\|_{\mathrm{F}}^2\right)^{1/2}\leq\frac{1}{\sqrt{L}}\left(\mathbb{E}\left\|\widehat{\mathbf{V}}_K^{(\ell)}\widehat{\mathbf{V}}_K^{(\ell)\top}-\mathbf{V}_K\mathbf{V}_K^{\top}\right\|_{\mathrm{F}}^2\right)^{1/2}.$$

Thus we have

$$\left(\mathbb{E}\|\widetilde{\mathbf{\Sigma}}-\mathbf{\Sigma}'\|_{\mathrm{F}}^2\right)^{1/2}\lesssim\frac{1}{\sqrt{L}}\left(\mathbb{E}\left\|\widehat{\mathbf{V}}_K^{(\ell)}\widehat{\mathbf{V}}_K^{(\ell)\top}-\mathbf{V}_K\mathbf{V}_K^{\top}\right\|_{\mathrm{F}}^2\right)^{1/2}, \tag{S.9}$$

Before bounding the RHS, consider the matrix $\check{\mathbf{Y}}^{(\ell)}:=\mathbf{V}_K\mathbf{\Lambda}_K\mathbf{V}_K^{\top}\mathbf{\Omega}^{(\ell)}$. If $\widetilde{\mathbf{\Omega}}^{(\ell)}:=\mathbf{V}_K^{\top}\mathbf{\Omega}^{(\ell)}\in\mathbb{R}^{K\times p}$ does not have a full row rank, then the entries will be restricted to a linear space with dimension less than $K\times p$. Since $\widetilde{\mathbf{\Omega}}^{(\ell)}$ is a $K\times p$ standard Gaussian matrix, the probability that $\widetilde{\mathbf{\Omega}}^{(\ell)}$ has full row rank is 1. Also $\check{\mathbf{Y}}^{(\ell)}$ is of rank $K$, and thus with probability 1, $\mathbf{V}_K$ and the top $K$ left singular eigenvectors of $\check{\mathbf{Y}}^{(\ell)}/\sqrt{p}$ span the same column space. In other words, if we let $\mathbf{\Gamma}_K$ be the left singular vectors of $\check{\mathbf{Y}}^{(\ell)}/\sqrt{p}$, then $\mathbf{\Gamma}_K\mathbf{\Gamma}_K^{\top}=\mathbf{V}_K\mathbf{V}_K^{\top}$.

Now consider the $K$-th singular value of $\check{\mathbf{Y}}^{(\ell)}/\sqrt{p}$, let $\mathbf{U}_{\widetilde{\mathbf{\Omega}}}\mathbf{D}_{\widetilde{\mathbf{\Omega}}}\mathbf{V}_{\widetilde{\mathbf{\Omega}}}^{\top}$ be the SVD of $\widetilde{\mathbf{\Omega}}^{(\ell)}/\sqrt{p}$.

We have

$$\begin{aligned}\sigma_K\left(\check{\mathbf{Y}}^{(\ell)}/\sqrt{p}\right) &= \sigma_K\left(\mathbf{V}_K\mathbf{\Lambda}_K\widetilde{\mathbf{\Omega}}^{(\ell)}/\sqrt{p}\right) = \sigma_K\left(\mathbf{\Lambda}_K\mathbf{U}_{\widetilde{\mathbf{\Omega}}}\mathbf{D}_{\widetilde{\mathbf{\Omega}}}\right)\\ &= \min_{\|\boldsymbol{x}\|=1}\|\mathbf{\Lambda}_K\mathbf{U}_{\widetilde{\mathbf{\Omega}}}\mathbf{D}_{\widetilde{\mathbf{\Omega}}}\boldsymbol{x}\| \overset{(i)}{\geq} \sigma_{\min}\left(\widetilde{\mathbf{\Omega}}^{(\ell)}/\sqrt{p}\right)\min_{\|\boldsymbol{v}_1\|=1}\left\|\mathbf{\Lambda}_K\mathbf{U}_{\widetilde{\mathbf{\Omega}}}\boldsymbol{v}_1\right\|\\ &\overset{(ii)}{\geq} \sigma_{\min}\left(\widetilde{\mathbf{\Omega}}^{(\ell)}/\sqrt{p}\right)\min_{\|\boldsymbol{v}_2\|=1}\|\mathbf{\Lambda}_K\boldsymbol{v}_2\| \geq \Delta\sigma_{\min}\left(\widetilde{\mathbf{\Omega}}^{(\ell)}/\sqrt{p}\right),\end{aligned}$$

where $\boldsymbol{v}_1 = \mathbf{D}_{\widetilde{\mathbf{\Omega}}}\boldsymbol{x}/\|\mathbf{D}_{\widetilde{\mathbf{\Omega}}}\boldsymbol{x}\|$, and $\boldsymbol{v}_2 = \mathbf{U}_{\widetilde{\mathbf{\Omega}}}\boldsymbol{v}_1$. Inequality (i) follows because $\|\mathbf{D}_{\widetilde{\mathbf{\Omega}}}\boldsymbol{x}\| \geq \sigma_{\min}\left(\widetilde{\mathbf{\Omega}}^{(\ell)}/\sqrt{p}\right)\|\boldsymbol{x}\| = \sigma_{\min}\left(\widetilde{\mathbf{\Omega}}^{(\ell)}/\sqrt{p}\right)$, and inequality (ii) is because $\|\boldsymbol{v}_2\| = \|\boldsymbol{v}_1\| = 1$.

Now by Wedin's Theorem (Wedin, 1972), under the condition that $p \geq \max(2K, K+7)$ we have the following bound on the RHS of (S.9),

$$\begin{aligned}&\frac{1}{\sqrt{L}}\left(\mathbb{E}\left\|\widehat{\mathbf{V}}_K^{(\ell)}\widehat{\mathbf{V}}_K^{(\ell)\top} - \mathbf{V}_K\mathbf{V}_K^\top\right\|_{\mathrm{F}}^2\right)^{1/2} \lesssim \frac{\sqrt{K}}{\sqrt{L}}\left(\mathbb{E}\left\|\mathbf{Y}^{(\ell)}/\sqrt{p} - \check{\mathbf{Y}}^{(\ell)}/\sqrt{p}\right\|_{\mathrm{op}}^2 / \left(\Delta\sigma_{\min}\big(\widetilde{\mathbf{\Omega}}^{(\ell)}/\sqrt{p}\big)\right)^2\right)^{1/2}\\ &\leq \frac{\sqrt{K}}{\Delta\sqrt{L}}\left(\mathbb{E}\left\|\mathbf{Y}^{(\ell)}/\sqrt{p} - \check{\mathbf{Y}}^{(\ell)}/\sqrt{p}\right\|_{\mathrm{op}}^4\right)^{1/4}\left(\mathbb{E}\left(\sigma_{\min}\big(\widetilde{\mathbf{\Omega}}^{(\ell)}/\sqrt{p}\big)\right)^{-4}\right)^{1/4}\\ &\lesssim \frac{\sqrt{K}}{\Delta\sqrt{L}}\|\|\widehat{\mathbf{E}}\|_{\mathrm{op}}\|_{\psi_1}\cdot\|\|\mathbf{\Omega}^{(\ell)}/\sqrt{p}\|_{\mathrm{op}}\|_{\psi_1} \lesssim \tau\sqrt{\frac{Kdr}{npL}},\end{aligned}$$

where the last but one inequality is due to Lemma B.2 and the last inequality is due to Lemma 3 in Fan et al. (2019). Therefore, we have the final error rate for the estimator $\widetilde{\mathbf{V}}_K$:

$$\left(\mathbb{E}\|\widetilde{\mathbf{V}}_K\widetilde{\mathbf{V}}_K^\top - \mathbf{V}_K\mathbf{V}_K^\top\|_{\mathrm{F}}^2\right)^{1/2} \lesssim \underbrace{\tau\sqrt{\frac{Kr}{n}}}_{\text{bias}} + \underbrace{\tau\sqrt{\frac{Kdr}{npL}}}_{\text{variance}}.$$

# C Proof of Theorem 4.4

Similar to the proof of Theorem 4.3, by the fact that $\|\cdot\|_{\mathrm{op}}^{2q}$ is convex, by Jensen's inequality we have that under the condition that $p \geq 8q + K - 1$ there exists some constant $\eta$ such that

$$\begin{aligned}
\mathbb{E}\|\widetilde{\boldsymbol{\Sigma}} - \mathbf{V}_K\mathbf{V}_K^\top\|_{\mathrm{op}}^{2q} &\leq L^{-1}\sum_{\ell=1}^{L}\mathbb{E}\|\widehat{\mathbf{V}}_K^{(\ell)}\widehat{\mathbf{V}}_K^{(\ell)\top} - \mathbf{V}_K\mathbf{V}_K^\top\|_{\mathrm{op}}^{2q} = \mathbb{E}\|\widehat{\mathbf{V}}_K^{(1)}\widehat{\mathbf{V}}_K^{(1)\top} - \mathbf{V}_K\mathbf{V}_K^\top\|_{\mathrm{op}}^{2q} \\
&\leq \mathbb{E}\Bigg(\left\|\mathbf{Y}^{(\ell)}/\sqrt{p} - \check{\mathbf{Y}}^{(\ell)}/\sqrt{p}\right\|_{\mathrm{op}}^{2q} \Big/ \left(\Delta\sigma_{\min}\left(\widetilde{\boldsymbol{\Omega}}^{(\ell)}/\sqrt{p}\right)\right)^{2q}\Bigg) \\
&\leq \frac{1}{\Delta^{2q}}\left(\mathbb{E}\left\|\mathbf{Y}^{(\ell)}/\sqrt{p} - \check{\mathbf{Y}}^{(\ell)}/\sqrt{p}\right\|_{\mathrm{op}}^{4q}\right)^{1/2}\left(\mathbb{E}\left(\sigma_{\min}\left(\widetilde{\boldsymbol{\Omega}}^{(\ell)}/\sqrt{p}\right)\right)^{-4q}\right)^{1/2} \\
&\lesssim \left(\eta q^2\tau\sqrt{\frac{dr}{np}}\right)^{2q}.
\end{aligned}$$

Also by Markov's inequality, we have

$$\begin{aligned}
\mathbb{P}\left(\|\widetilde{\boldsymbol{\Sigma}} - \mathbf{V}_K\mathbf{V}_K^\top\|_{\mathrm{op}} \geq \frac{1}{2}\right) &\leq \mathbb{P}\left(\|\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}'\|_{\mathrm{op}} \geq \frac{1}{4}\right) + \mathbb{P}\left(\|\boldsymbol{\Sigma}' - \mathbf{V}_K\mathbf{V}_K^\top\|_{\mathrm{op}} \geq \frac{1}{4}\right) \\
&\lesssim \exp\left(-\frac{C_0}{4\tau\sqrt{dr/np}}\right) + \mathbb{E}\|\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}'\|_{\mathrm{F}}^2 \\
&\lesssim \exp\left(-\frac{C_0}{4\tau\sqrt{dr/np}}\right) + \left(\tau\sqrt{\frac{Kdr}{npL}}\right)^2,
\end{aligned}$$

where the penultimate inequality follows from Lemma B.3, and the last inequality follows from the proof of Theorem 4.3. Now by Weyl's inequality (Franklin, 2012), we know that $\sigma_K(\widetilde{\boldsymbol{\Sigma}}) \geq 1 - \|\widetilde{\boldsymbol{\Sigma}} - \mathbf{V}_K\mathbf{V}_K^\top\|_{\mathrm{op}}$ and $\sigma_{K+1}(\widetilde{\boldsymbol{\Sigma}}) \leq \|\widetilde{\boldsymbol{\Sigma}} - \mathbf{V}_K\mathbf{V}_K^\top\|_{\mathrm{op}}$.

Now if we denote the SVD of $\widetilde{\boldsymbol{\Sigma}}^q$ by $\widetilde{\mathbf{V}}_K\widetilde{\boldsymbol{\Lambda}}_K^q\widetilde{\mathbf{V}}_K^\top + \widetilde{\mathbf{V}}_\perp\widetilde{\boldsymbol{\Lambda}}_\perp^q\widetilde{\mathbf{V}}_\perp^\top$, then with probability 1, $\widetilde{\mathbf{V}}_K\widetilde{\boldsymbol{\Lambda}}_K^q\widetilde{\mathbf{V}}_K^\top$ and $\widetilde{\mathbf{V}}_K$ share the same column space. By the relationship $\sigma_k(\widetilde{\boldsymbol{\Sigma}}^q) = \sigma_k^q(\widetilde{\boldsymbol{\Sigma}})$ for

$k \in [d]$ and Davis-Kahan's Theorem (Yu et al., 2015), when $p' \geq \max(2K, K+7)$ we have

$$\begin{aligned}
\mathbb{E}\left(\|\widetilde{\mathbf{V}}_K^{\mathrm{F}}\widetilde{\mathbf{V}}_K^{\mathrm{F}\top} - \widetilde{\mathbf{V}}_K\widetilde{\mathbf{V}}_K^{\top}\|_{\mathrm{F}}^2 \,|\widetilde{\boldsymbol{\Sigma}}\right) &\lesssim \mathbb{E}\left(K\|\widetilde{\boldsymbol{\Sigma}}^q\boldsymbol{\Omega}^{\mathrm{F}} - \widetilde{\mathbf{V}}_K\widetilde{\boldsymbol{\Lambda}}_K^q\widetilde{\mathbf{V}}_K^{\top}\boldsymbol{\Omega}^{\mathrm{F}}\|_{\mathrm{op}}^2/\sigma_{\min}^2(\widetilde{\mathbf{V}}_K\widetilde{\boldsymbol{\Lambda}}_K^q\widetilde{\mathbf{V}}_K^{\top}\boldsymbol{\Omega}^{\mathrm{F}})\,|\widetilde{\boldsymbol{\Sigma}}\right) \\
&\lesssim \left(\frac{\sqrt{K}}{\sigma_K^q(\widetilde{\boldsymbol{\Sigma}})}\|\widetilde{\mathbf{V}}_\perp\widetilde{\boldsymbol{\Lambda}}_\perp^q\widetilde{\mathbf{V}}_\perp^{\top}\|_{\mathrm{op}} \cdot \|\|\boldsymbol{\Omega}^{\mathrm{F}}/\sqrt{p'}\|_{\mathrm{op}}\|_{\psi_1}\right)^2 \\
&\lesssim \frac{Kd}{p'}\frac{\|\widetilde{\boldsymbol{\Sigma}} - \mathbf{V}_K\mathbf{V}_K^{\top}\|_{\mathrm{op}}^{2q}}{\left(1 - \|\widetilde{\boldsymbol{\Sigma}} - \mathbf{V}_K\mathbf{V}_K^{\top}\|_{\mathrm{op}}\right)^{2q}}.
\end{aligned}$$

Therefore we have,

$$\begin{aligned}
&\left(\mathbb{E}\|\widetilde{\mathbf{V}}_K^{\mathrm{F}}\widetilde{\mathbf{V}}_K^{\mathrm{F}\top} - \widetilde{\mathbf{V}}_K\widetilde{\mathbf{V}}_K^{\top}\|_{\mathrm{F}}^2\right)^{1/2} \lesssim \left(\mathbb{E}\|\widetilde{\mathbf{V}}_K^{\mathrm{F}}\widetilde{\mathbf{V}}_K^{\mathrm{F}\top} - \widetilde{\mathbf{V}}_K\widetilde{\mathbf{V}}_K^{\top}\|_{\mathrm{F}}^2\mathbb{I}\{\|\widetilde{\boldsymbol{\Sigma}} - \mathbf{V}_K\mathbf{V}_K^{\top}\|_{\mathrm{op}} \leq 1/2\}\right)^{1/2} \\
&\quad + \left(\mathbb{E}\|\widetilde{\mathbf{V}}_K^{\mathrm{F}}\widetilde{\mathbf{V}}_K^{\mathrm{F}\top} - \widetilde{\mathbf{V}}_K\widetilde{\mathbf{V}}_K^{\top}\|_{\mathrm{F}}^2\mathbb{I}\{\|\widetilde{\boldsymbol{\Sigma}} - \mathbf{V}_K\mathbf{V}_K^{\top}\|_{\mathrm{op}} > 1/2\}\right)^{1/2} \\
&\lesssim 2^q\sqrt{\frac{Kd}{p'}}\left(\mathbb{E}\|\widetilde{\boldsymbol{\Sigma}} - \mathbf{V}_K\mathbf{V}_K^{\top}\|_{\mathrm{op}}^{2q}\right)^{1/2} + \sqrt{K}\left\{\mathbb{P}\left(\|\widetilde{\boldsymbol{\Sigma}} - \mathbf{V}_K\mathbf{V}_K^{\top}\|_{\mathrm{op}} \geq \frac{1}{2}\right)\right\}^{1/2} \\
&\lesssim \sqrt{\frac{Kd}{p'}}\left(2\eta q^2\tau\sqrt{\frac{dr}{np}}\right)^q + \sqrt{K}\exp\left(-\frac{C_0}{8\tau\sqrt{dr/np}}\right) + K\tau\sqrt{\frac{dr}{npL}}.
\end{aligned}$$

Under the condition $n \geq C\frac{dr}{p}\tau^2\log^2\frac{n}{r}$ for some sufficiently large constant $C$, we have $\sqrt{K}\exp\left(-\frac{C_0}{8\tau\sqrt{dr/np}}\right) = o(\tau\sqrt{\frac{Kr}{n}})$, and thus by Theorem 4.3 and triangle inequality, we have

$$\begin{aligned}
\left(\mathbb{E}\|\widetilde{\mathbf{V}}_K^{\mathrm{F}}\widetilde{\mathbf{V}}_K^{\mathrm{F}\top} - \mathbf{V}_K\mathbf{V}_K^{\top}\|_{\mathrm{F}}^2\right)^{1/2} &\lesssim \left(\mathbb{E}\|\widetilde{\mathbf{V}}_K^{\mathrm{F}}\widetilde{\mathbf{V}}_K^{\mathrm{F}\top} - \widetilde{\mathbf{V}}_K\widetilde{\mathbf{V}}_K^{\top}\|_{\mathrm{F}}^2\right)^{1/2} + \left(\mathbb{E}\|\widetilde{\mathbf{V}}_K\widetilde{\mathbf{V}}_K^{\top} - \mathbf{V}_K\mathbf{V}_K^{\top}\|_{\mathrm{F}}^2\right)^{1/2} \\
&\lesssim \sqrt{\frac{Kd}{p'}}\left(2\eta q^2\tau\sqrt{\frac{dr}{np}}\right)^q + \tau\sqrt{\frac{Kr}{n}} + \tau K\sqrt{\frac{dr}{npL}}.
\end{aligned}$$

# D Proof of Theorem 4.5

We first focus on a given $\ell \in [L]$. Recall that $\mathbf{Y}^{(\ell)}/\sqrt{p} = \mathbf{V}_K\boldsymbol{\Lambda}_K\widetilde{\boldsymbol{\Omega}}^{(\ell)}/\sqrt{p} + \widehat{\mathbf{E}}\boldsymbol{\Omega}^{(\ell)}/\sqrt{p}$, where $\widetilde{\boldsymbol{\Omega}}^{(\ell)} = \mathbf{V}_K^{\top}\boldsymbol{\Omega}^{(\ell)}$ and $\widehat{\mathbf{E}} = \widehat{\boldsymbol{\Sigma}}^{\mathrm{tr}} - \mathbf{V}_K\boldsymbol{\Lambda}_K\mathbf{V}_K^{\top}$.

For the residual term $\widehat{\mathbf{E}}\mathbf{\Omega}^{(\ell)}/\sqrt{p}$, from Lemma 3 in Fan et al. (2019), under the condition that $\sqrt{p/d}\log d = O(1)$, the following is true each with probability at least $1 - d^{-10}$ respectively,

$$\|\mathbf{\Omega}^{(\ell)}/\sqrt{p}\|_{\text{op}} \lesssim \sqrt{\frac{d}{p}} \quad \text{and} \quad \|\widehat{\mathbf{E}}\|_{\text{op}} \lesssim \lambda_1\sqrt{\frac{r}{n}}\log d.$$

Denote by $\mathcal{A}$ the event $\big\{\|\widehat{\mathbf{E}}\|_{\text{op}} \lesssim \lambda_1\sqrt{\frac{r}{n}}\log d\big\}$. Then conditional on $\mathcal{A}$, we have that $\|\widehat{\mathbf{E}}\mathbf{\Omega}^{(\ell)}/\sqrt{p}\|_{\text{op}} \lesssim \lambda_1\sqrt{\frac{dr}{np}}\log d \le \lambda_1\eta_0$ with probability at least $1-d^{-10}$ for each $\ell \in [L]$. From Proposition 10.4 in Halko et al. (2011), we know that when $p \ge 2K$

$$\mathbb{P}\left(\sigma_{\min}\big(\widetilde{\mathbf{\Omega}}^{(\ell)}/\sqrt{p}\big) \le \frac{1}{6}\sqrt{\eta_0}\right) \le \mathbb{P}\left(\sigma_{\min}\big(\widetilde{\mathbf{\Omega}}^{(\ell)}/\sqrt{p}\big) \le \frac{p-K+1}{\mathrm{e}p}\sqrt{\eta_0}\right) \le \eta_0^{\frac{p-K+1}{2}}.$$

Therefore, with probability at least $1-\eta_0^{(p-K+1)/2}$, under the condition that $\Delta \ge \eta_0^{1/4}$, $\sigma_{\min}\big(\mathbf{V}_K\mathbf{\Lambda}_K\widetilde{\mathbf{\Omega}}^{(\ell)}/\sqrt{p}\big) \ge \Delta\sigma_{\min}\big(\widetilde{\mathbf{\Omega}}^{(\ell)}/\sqrt{p}\big) \ge \Delta\sqrt{\eta_0}/6 \ge \eta_0^{3/4}/6$.

By Weyl's inequality (Franklin, 2012), we know that when $\lambda_1\eta_0^{1/4} = o(1)$, conditional on $\mathcal{A}$, with probability at least $1-d^{-10}$, $\sigma_{K+1}(\mathbf{Y}^{(\ell)}/\sqrt{p}) \le \|\widehat{\mathbf{E}}\mathbf{\Omega}^{(\ell)}/\sqrt{p}\|_{\text{op}} \lesssim \lambda_1\eta_0 \ll \eta_0^{3/4}/12$ for large enough $n$, which indicates that $\max_{i\ge k}(\sigma_i(\mathbf{Y}^{(\ell)}) - \sigma_p(\mathbf{Y}^{(\ell)})) \le \sqrt{p}\mu_0$ for any $k \ge K+1$. Then we have

$$\begin{aligned}
\mathbb{P}\left(\widehat{K}^{(\ell)} = K \,\middle|\, \mathcal{A}\right) &\ge \mathbb{P}\left(\sigma_K(\frac{\mathbf{Y}^{(\ell)}}{\sqrt{p}}) - \sigma_p(\frac{\mathbf{Y}^{(\ell)}}{\sqrt{p}}) \ge \frac{(\eta_0)^{3/4}}{12}, \quad \sigma_{K+1}(\frac{\mathbf{Y}^{(\ell)}}{\sqrt{p}}) - \sigma_p(\frac{\mathbf{Y}^{(\ell)}}{\sqrt{p}}) < \frac{(\eta_0)^{3/4}}{12} \,\middle|\, \mathcal{A}\right) \\
&\ge \mathbb{P}\left(\sigma_{\min}\big(\mathbf{V}_K\mathbf{\Lambda}_K\widetilde{\mathbf{\Omega}}^{(\ell)}/\sqrt{p}\big) \ge \frac{(\eta_0)^{3/4}}{6}, \quad \|\widehat{\mathbf{E}}\mathbf{\Omega}^{(\ell)}/\sqrt{p}\|_{\text{op}} \lesssim \lambda_1\eta_0 \,\middle|\, \mathcal{A}\right) \\
&\ge 1 - d^{-10} - \eta_0^{\frac{p-K+1}{2}}.
\end{aligned}$$

We know that conditional on $\widehat{\mathbf{E}}$, $\mathbb{I}\{\widehat{K}^{(\ell)} \ne K \,|\, \mathcal{A}\}$ are i.i.d. Bernoulli variables with expectation $p_K \le d^{-10} + \eta_0^{\frac{p-K+1}{2}} \le 1/4$ and variance $p_K(1-p_K) \le p_K$. Since the estimators $\{\widehat{K}^{(\ell)}\}_{\ell=1}^L$ are all integers, we know that if $\widehat{K} \ne K$, at least half of $\{\widehat{K}^{(\ell)}\}_{\ell=1}^L$ are not equal to

$K$. Then by Hoeffding's inequality, we have

$$\mathbb{P}(\widehat{K}\neq K)\leq \mathbb{P}\left(\sum_{\ell=1}^{L}\mathbb{I}\left\{\widehat{K}^{(\ell)}\neq K\right\}-p_K L\geq L/4\right)=\mathbb{E}_{\widehat{\mathbf{E}}}\left(\mathbb{P}\left(\sum_{\ell=1}^{L}\mathbb{I}\left\{\widehat{K}^{(\ell)}\neq K\right\}-p_K L\geq L/4\,\middle|\,\widehat{\mathbf{E}}\right)\right)$$
$$\leq \mathbb{P}(\mathcal{A})\exp\left\{-(L/4)^2/(2Lp_K)\right\}+1-\mathbb{P}(\mathcal{A})\leq \exp\left\{-L/\left(32d^{-10}+32\eta_0^{\frac{p-K+1}{2}}\right)\right\}+d^{-10}.$$

We know that $32d^{-10}\leq(\log d)^{-1}$ for $d\geq 2$ and under the condition that $\log n\geq \log(d^2/p)+2.5\log\log d+8\log 32/p$, we also have $32\eta_0^{\frac{p-K+1}{2}}\leq(\log d)^{-1}$. Therefore, $\mathbb{P}(\widehat{K}\neq K)\leq \exp(-L\log d/2)+d^{-10}\lesssim d^{-10}+d^{-L/2}$.

# E Proof of Theorem 5.2

The proof of Theorem 5.2 basically follows the proof of Theorem 4.3 and Theorem 4.4 except for a few modifications. We will discuss the modifications and will omit the identical steps for conciseness. Before we begin with the proof, we will need to redefine a few notations from the previous proof. Similar as in Section 4, we define $\mathbf{\Sigma}'=\mathbb{E}_{\mathbf{\Omega}}(\widehat{\mathbf{V}}_K^{(1)}\widehat{\mathbf{V}}_K^{(1)\top})$, where $\widehat{\mathbf{V}}_K^{(1)}$ is the estimator from server 1 on the second layer. We also define $\widehat{\mathbf{E}}'=\widehat{\mathbf{\Sigma}}-\mathbf{V}_K\mathbf{\Lambda}_K\mathbf{V}_K^\top$. Then we have that $\|\widehat{\mathbf{E}}'\|_{\mathrm{op}}=\|\widehat{\mathbf{\Sigma}}-\mathbf{V}_K\mathbf{\Lambda}_K\mathbf{V}_K^\top\|_{\mathrm{op}}\leq\|\widehat{\mathbf{\Sigma}}-\mathbf{\Sigma}\|_{\mathrm{op}}+\|\mathbf{D}\|_{\mathrm{op}}$. By Assumption 5.1 and the condition that $\|\mathbf{D}\|_{\mathrm{op}}\lesssim\lambda_1 g(r,n)$, we have $\|\|\widehat{\mathbf{E}}'\|_{\mathrm{op}}\|_{\psi_1}\leq\|\|\widehat{\mathbf{\Sigma}}-\mathbf{\Sigma}\|_{\mathrm{op}}\|_{\psi_1}+\|\mathbf{D}\|_{\mathrm{op}}\lesssim\lambda_1 g(r,n)$.

We will need the following modified versions of Lemma 4.2 and Lemma B.3.

**Lemma E.1.** Let $\widehat{\mathbf{V}}\widehat{\mathbf{\Lambda}}\widehat{\mathbf{V}}^\top$ be the SVD of $\widehat{\mathbf{\Sigma}}$ and $\widehat{\mathbf{V}}_K$ be the $K$ leading eigenvectors of $\widehat{\mathbf{\Sigma}}$. When $\|\mathbf{\Sigma}'-\widehat{\mathbf{V}}_K\widehat{\mathbf{V}}_K^\top\|_{\mathrm{op}}<1/2$, we have that $\widehat{\mathbf{V}}^\top\mathbf{\Sigma}'\widehat{\mathbf{V}}$ is diagonal and $\|\mathbf{V}'_K\mathbf{V}'^\top_K-\widehat{\mathbf{V}}_K\widehat{\mathbf{V}}_K^\top\|_{\mathrm{op}}=0$.

The proof of Lemma E.1 is identical to that of Lemma 4.2 except that we replace $\widehat{\mathbf{\Sigma}}^{\mathrm{tr}}$ by $\widehat{\mathbf{\Sigma}}$.

**Lemma E.2.** $\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n\in\mathbb{R}^d$ are random vectors described in Theorem 5.2. Under Assumption 5.1, $n>r$, $p\geq\max(2K,K+7)$ and the condition that $\tau g(r,n)=o(\sqrt{p/d})$,

there exists a constant $C_0 > 0$ such that for any $\varepsilon > 0$, we have

$$\max\left\{\mathbb{P}\Big(\|\boldsymbol{\Sigma}' - \mathbf{V}_K\mathbf{V}_K\|_{\mathrm{op}} \geq \varepsilon\Big), \mathbb{P}\left(\|\boldsymbol{\Sigma}' - \widehat{\mathbf{V}}_K\widehat{\mathbf{V}}_K^\top\|_{\mathrm{op}} \geq \varepsilon\right)\right\} \leq \exp\left(-\frac{C_0\varepsilon\sqrt{p/d}}{\tau g(r,n)}\right).$$

Please refer to Section G.4 for the proof of Lemma E.2.

It suffices to bound $\left(\mathbb{E}\|\widetilde{\mathbf{V}}_K\widetilde{\mathbf{V}}_K^\top - \mathbf{V}_K\mathbf{V}_K\|_{\mathrm{F}}^2\right)^{1/2}$ and $\left(\mathbb{E}\|\widetilde{\mathbf{V}}_K\widetilde{\mathbf{V}}_K^\top - \widetilde{\mathbf{V}}_K^{\mathrm{F}}\widetilde{\mathbf{V}}_K^{\mathrm{F}\top}\|_{\mathrm{F}}^2\right)^{1/2}$ to prove (4) and (5). We will bound $\left(\mathbb{E}\|\widetilde{\mathbf{V}}_K\widetilde{\mathbf{V}}_K^\top - \mathbf{V}_K\mathbf{V}_K\|_{\mathrm{F}}^2\right)^{1/2}$ first. We can break the error term into the variance term and the bias term, and bound them separately.

$$\left(\mathbb{E}\|\widetilde{\mathbf{V}}_K\widetilde{\mathbf{V}}_K^\top - \mathbf{V}_K\mathbf{V}_K^\top\|_{\mathrm{F}}^2\right)^{1/2} \leq \underbrace{\left(\mathbb{E}\|\widetilde{\mathbf{V}}_K\widetilde{\mathbf{V}}_K^\top - \mathbf{V}'_K\mathbf{V}'^\top_K\|_{\mathrm{F}}^2\right)^{1/2}}_{\mathrm{I}} + \underbrace{\left(\mathbb{E}\|\mathbf{V}'_K\mathbf{V}'^\top_K - \mathbf{V}_K\mathbf{V}_K^\top\|_{\mathrm{F}}^2\right)^{1/2}}_{\mathrm{II}}. \tag{S.10}$$

For the bias term, following similar steps to those in the proof of Theorem 4.3, under the condition that $\tau g(r,n) \leq c\sqrt{p/d}\big(\log\sqrt{d/p}\big)^{-1}$ for some small enough constant $c$, we have

$$\begin{aligned}\left(\mathbb{E}\|\mathbf{V}'_K\mathbf{V}'^\top_K - \mathbf{V}_K\mathbf{V}_K^\top\|_{\mathrm{F}}^2\right)^{1/2} &\lesssim \sqrt{K}\mathbb{P}\Big(\|\boldsymbol{\Sigma}' - \widehat{\mathbf{V}}_K\widehat{\mathbf{V}}_K^\top\|_{\mathrm{op}} \geq 1/2\Big)^{1/2} + \frac{\sqrt{K}}{\Delta}\|\|\widehat{\mathbf{E}}\|_{\mathrm{op}}\|_{\psi_1} \\ &\lesssim \sqrt{K}\exp\left(-\frac{C_0\sqrt{p/d}}{4\tau g(r,n)}\right) + \frac{\sqrt{K}}{\Delta}\lambda_1 g(r,n) \lesssim \sqrt{K}\tau g(r,n).\end{aligned}$$

Similarly, for the variance term, by replacing $\widehat{\boldsymbol{\Sigma}}^{\mathrm{tr}}$ with $\widehat{\boldsymbol{\Sigma}}$ in the proof of Theorem 4.3, we have

$$\begin{aligned}\left(\mathbb{E}\|\widetilde{\mathbf{V}}_K\widetilde{\mathbf{V}}_K^\top - \mathbf{V}'_K\mathbf{V}'^\top_K\|_{\mathrm{F}}^2\right)^{1/2} &\lesssim \sqrt{K}\left\{\mathbb{P}\left(\left\|\boldsymbol{\Sigma}' - \mathbf{V}_K\mathbf{V}_K^T\right\|_{\mathrm{op}} \geq 1/4\right)\right\}^{1/2} + \left(\mathbb{E}\|\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}'\|_{\mathrm{F}}^2\right)^{1/2} \\ &\lesssim \sqrt{K}\exp\left(-\frac{C_0\sqrt{p/d}}{8\tau g(r,n)}\right) + \sqrt{\frac{Kd}{\Delta^2 Lp}}\|\|\widehat{\mathbf{E}}'\|_{\mathrm{op}}\|_{\psi_1} \lesssim \sqrt{K}\exp\left(-\frac{C_0\sqrt{p/d}}{8\tau g(r,n)}\right) + \tau\sqrt{\frac{Kd}{Lp}}g(r,n).\end{aligned}$$

Then combining the bias term and the variance term, we have

$$\Big(\mathbb{E}\|\widetilde{\mathbf{V}}_K\widetilde{\mathbf{V}}_K^\top - \mathbf{V}_K\mathbf{V}_K^\top\|_{\mathrm{F}}^2\Big)^{1/2} \lesssim \sqrt{K}\tau g(r,n) + \tau\sqrt{\frac{Kd}{Lp}}g(r,n).$$

Now we move on to bound $\Big(\mathbb{E}\|\widetilde{\mathbf{V}}_K\widetilde{\mathbf{V}}_K^\top - \widetilde{\mathbf{V}}_K^{\mathrm{F}}\widetilde{\mathbf{V}}_K^{\mathrm{F}\top}\|_{\mathrm{F}}^2\Big)^{1/2}$. Following similar steps as in the proof of Theorem 4.4, if $q \le (p-K+1)/8$, there exists some constant $\eta$ such that

$$\begin{aligned}
&\mathbb{E}\|\widetilde{\mathbf{\Sigma}} - \mathbf{V}_K\mathbf{V}_K^\top\|_{\mathrm{op}}^{2q} \lesssim \Big(\eta q^2\Delta^{-1}\sqrt{\frac{d}{p}}\|\|\widehat{\mathbf{E}}'\|_{\mathrm{op}}\|_{\psi_1}\Big)^{2q} \le \Big(\eta q^2\tau\sqrt{\frac{d}{p}}g(r,n)\Big)^{2q},\\
&\mathbb{P}\Big(\|\widetilde{\mathbf{\Sigma}} - \mathbf{V}_K\mathbf{V}_K^\top\|_{\mathrm{op}} \ge \frac{1}{2}\Big) \lesssim \mathbb{P}\Big(\|\mathbf{\Sigma}' - \mathbf{V}_K\mathbf{V}_K^\top\|_{\mathrm{op}} \ge \frac{1}{4}\Big) + \mathbb{E}\|\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}'\|_{\mathrm{F}}^2\\
&\quad\lesssim \exp\left(-\frac{C_0\sqrt{p/d}}{4\tau g(r,n)}\right) + \left(\tau\sqrt{\frac{Kd}{Lp}}g(r,n)\right)^2.
\end{aligned}$$

Thus following similar steps as in the proof of Theorem 4.4, when $p' \ge \max(2K, K+7)$ we have

$$\begin{aligned}
&\Big(\mathbb{E}\|\widetilde{\mathbf{V}}_K\widetilde{\mathbf{V}}_K^\top - \widetilde{\mathbf{V}}_K^{\mathrm{F}}\widetilde{\mathbf{V}}_K^{\mathrm{F}\top}\|_{\mathrm{F}}^2\Big)^{1/2} \lesssim 2^q\sqrt{\frac{Kd}{p'}}\Big(\mathbb{E}\|\widetilde{\mathbf{\Sigma}} - \mathbf{V}_K\mathbf{V}_K^\top\|_{\mathrm{op}}^{2q}\Big)^{1/2} + \sqrt{K}\Big\{\mathbb{P}\Big(\|\widetilde{\mathbf{\Sigma}} - \mathbf{V}_K\mathbf{V}_K^\top\|_{\mathrm{op}} \ge \frac{1}{2}\Big)\Big\}^{1/2}\\
&\quad\lesssim \sqrt{\frac{Kd}{p'}}\Big(2\eta q^2\tau\sqrt{\frac{d}{p}}g(r,n)\Big)^q + \sqrt{K}\exp\left(-\frac{C_0\sqrt{p/d}}{8\tau g(r,n)}\right) + K\tau\sqrt{\frac{d}{Lp}}g(r,n).
\end{aligned}$$

Then combining with (4), (5) follows.

# F Proof of Theorem 5.3

The proof is very similar to that of Theorem 4.5. Under Assumption 5.1, we know that with probability at least $1-d^{-10}$ we have $\|\widehat{\mathbf{E}}'\|_{\mathrm{op}} \lesssim \lambda_1 g(r,n)\log d$. Denote by $\mathcal{A}'$ the event $\{\|\widehat{\mathbf{E}}'\|_{\mathrm{op}} \lesssim \lambda_1 g(r,n)\log d\}$. Then conditional on $\mathcal{A}'$, for each $\ell \in [L]$ with probability at least $1-d^{-10}$, we have $\|\widehat{\mathbf{E}}'\mathbf{\Omega}^{(\ell)}/\sqrt{p}\|_{\mathrm{op}} \lesssim \lambda_1\eta_0$, and under the same event by Weyl's inequality

(Franklin, 2012) we know that $\max_{i\geq k}(\sigma_i(\mathbf{Y}^{(\ell)}) - \sigma_p(\mathbf{Y}^{(\ell)})) \leq \sqrt{p}\mu_0$ for any $k \geq K+1$ under the condition that $\lambda_1\eta_0^{1/4} = o(1)$. Then similarly, we have

$$\mathbb{P}\left(\widehat{K}^{(\ell)} = K \,\middle|\, \mathcal{A}'\right) \geq \mathbb{P}\left(\sigma_{\min}\big(\mathbf{V}_K\mathbf{\Lambda}_K\widetilde{\mathbf{\Omega}}^{(\ell)}/\sqrt{p}\big) \geq \frac{\eta_0^{3/4}}{6}, \quad \|\widehat{\mathbf{E}}\mathbf{\Omega}^{(\ell)}/\sqrt{p}\|_{\mathrm{op}} \lesssim \lambda_1\eta_0 \quad \middle|\, \mathcal{A}'\right)$$
$$\geq 1 - d^{-10} - \eta_0^{\frac{p-K+1}{2}}.$$

When $d \geq 2$ and $\log g(r,n) \leq -\log\sqrt{d/p} - \frac{3}{p-K+1}\log\log d$, we have that $d^{-10} \leq (\log d)^{-1}/32$ and $\eta_0^{\frac{p-K+1}{2}} \leq (\log d)^{-1}/32$, and hence following similar steps to those in the proof of Theorem 5.3 we have that $\mathbb{I}\big\{\widehat{K}^{(\ell)} \neq K | \mathcal{A}'\big\}$ are i.i.d. Bernoulli variables conditional on $\widehat{\mathbf{E}}'$ with expectation $p_K \leq (\log d)^{-1}/16 \leq 1/4$. Following Hoeffding's inequality, we have

$$\mathbb{P}(\widehat{K} \neq K) \leq \mathbb{P}(\mathcal{A}')\exp\big\{-(L/4)^2/(2Lp_K)\big\} + 1 - \mathbb{P}(\mathcal{A}') \leq d^{-10} + d^{-L/2}.$$

# G Proof of Technical Lemmas

In this section, we provide proofs of the technical lemmas used in the proofs of the main theorems.

## G.1 Proof of Lemma B.1

It can be seen that $\|\mathbf{\Omega}/\sqrt{p}\|_{\mathrm{op}} = (\|\mathbf{\Omega}\mathbf{\Omega}^\top/p\|_{\mathrm{op}})^{1/2} = \big\{(d/p)\|\mathbf{\Omega}^\top\mathbf{\Omega}/d\|_{\mathrm{op}}\big\}^{1/2}$. By Lemma 3 in Fan et al. (2019), we know that $\|\|\mathbf{\Omega}^\top\mathbf{\Omega}/d - \mathbf{I}_p\|_{\mathrm{op}}\|_{\psi_1} \lesssim \sqrt{p/d}$, and thus $\|\|\mathbf{\Omega}^\top\mathbf{\Omega}/d\|_{\mathrm{op}}\|_{\psi_1} \lesssim 1 + \sqrt{p/d} = O(1)$. Therefore, we have $\|\|\mathbf{\Omega}\mathbf{\Omega}^\top/p\|_{\mathrm{op}}\|_{\psi_1} \lesssim d/p$. By Jensen's inequality, we in turn get $\|\|\mathbf{\Omega}/\sqrt{p}\|_{\mathrm{op}}\|_{\psi_1} \lesssim \sqrt{d/p}$.

## G.2 Proof of Lemma B.2

By Proposition 10.4 in Halko et al. (2011), for any $t \geq 1$, we have

$$\mathbb{P}\left(\left\|\mathbf{\Omega}^{\dagger}\right\|_{\mathrm{op}} \geq \frac{\mathrm{e}\sqrt{p}}{p-K+1} \cdot t\right) \leq t^{-(p-K+1)}. \tag{S.11}$$

Since $p \geq 2K$, there exists a constant $c$ such that $\frac{ep}{p-K+1} \leq c$, and thus

$$\mathbb{P}\left(\sqrt{p}\left\|\mathbf{\Omega}^{\dagger}\right\|_{\mathrm{op}} \geq ct\right) \leq t^{-(p-K+1)}. \tag{S.12}$$

Therefore, we have

$$\begin{aligned}
\mathbb{E}\left((\sigma_{\min}(\mathbf{\Omega}/\sqrt{p}))^{-a}\right) &= \mathbb{E}\left(\left\|\sqrt{p}\mathbf{\Omega}^{\dagger}\right\|_{\mathrm{op}}^{a}\right) = \int_{u\geq 0} \mathbb{P}\left(\left\|\sqrt{p}\mathbf{\Omega}^{\dagger}\right\|_{\mathrm{op}}^{a} \geq u\right) du \\
&= \int_{0\leq u\leq c^a} \mathbb{P}\left(\left\|\sqrt{p}\mathbf{\Omega}^{\dagger}\right\|_{\mathrm{op}}^{a} \geq u\right) du + \int_{u\geq c^a} \mathbb{P}\left(\left\|\sqrt{p}\mathbf{\Omega}^{\dagger}\right\|_{\mathrm{op}}^{a} \geq u\right) du \\
&\leq c^a + \int_{u\geq c^a} \mathbb{P}\left(\left\|\sqrt{p}\mathbf{\Omega}^{\dagger}\right\|_{\mathrm{op}} \geq u^{1/a}\right) du \leq c^a + \int_{u\geq c^a} \left(u^{1/a}/c\right)^{-(p-K+1)} du \\
&= c^a\left(1 + \frac{1}{(p-K+1)/a - 1}\right).
\end{aligned}$$

Since $1 + \frac{1}{(p-K+1)/a-1} \leq 2$, the claim follows.

## G.3 Proof of Lemma B.3

We first consider the probability $\mathbb{P}\left(\|\boldsymbol{\Sigma}' - \mathbf{V}_K\mathbf{V}_K\|_{\text{op}} \geq \varepsilon\right)$. Recall the matrix $\check{\mathbf{Y}}^{(\ell)} = \mathbf{V}_K\boldsymbol{\Lambda}_K\mathbf{V}_K^\top\boldsymbol{\Omega}^{(\ell)}$. Now by Wedin's Theorem (Wedin, 1972), we have

$$
\begin{aligned}
\|\boldsymbol{\Sigma}' - \mathbf{V}_K\mathbf{V}_K\|_{\text{op}} &= \|\mathbb{E}\left(\widehat{\mathbf{V}}_K^{(\ell)}\widehat{\mathbf{V}}_K^{(\ell)\top}|\widehat{\boldsymbol{\Sigma}}^{\text{tr}}\right) - \mathbf{V}_K\mathbf{V}_K\|_{\text{op}} \leq \mathbb{E}\left(\left\|\widehat{\mathbf{V}}_K^{(\ell)}\widehat{\mathbf{V}}_K^{(\ell)\top} - \mathbf{V}_K\mathbf{V}_K\right\|_{\text{op}}\Big|\widehat{\boldsymbol{\Sigma}}^{\text{tr}}\right) \\
&\lesssim \mathbb{E}\left(\|\mathbf{Y}^{(\ell)}/\sqrt{p} - \check{\mathbf{Y}}^{(\ell)}/\sqrt{p}\|_{\text{op}}/\sigma_K\left(\check{\mathbf{Y}}^{(\ell)}/\sqrt{p}\right)|\widehat{\boldsymbol{\Sigma}}^{\text{tr}}\right) \leq \frac{\|\widehat{\mathbf{E}}\|_{\text{op}}}{\Delta}\mathbb{E}\left(\frac{\|\boldsymbol{\Omega}^{(\ell)}/\sqrt{p}\|_{\text{op}}}{\sigma_{\min}\left(\widetilde{\boldsymbol{\Omega}}^{(\ell)}/\sqrt{p}\right)}\Big|\widehat{\boldsymbol{\Sigma}}^{\text{tr}}\right) \\
&= \frac{\|\widehat{\mathbf{E}}\|_{\text{op}}}{\Delta}\mathbb{E}\left(\frac{\|\boldsymbol{\Omega}^{(\ell)}/\sqrt{p}\|_{\text{op}}}{\sigma_{\min}\left(\widetilde{\boldsymbol{\Omega}}^{(\ell)}/\sqrt{p}\right)}\right) \leq \frac{\|\widehat{\mathbf{E}}\|_{\text{op}}}{\Delta}\mathbb{E}\left(\|\boldsymbol{\Omega}^{(\ell)}/\sqrt{p}\|_{\text{op}}^2\right)^{1/2}\mathbb{E}\left(\left\{\sigma_{\min}(\boldsymbol{\Omega}^{(\ell)}/\sqrt{p})\right\}^{-2}\right)^{1/2} \\
&\lesssim \frac{\|\widehat{\mathbf{E}}\|_{\text{op}}}{\Delta}\|\|\boldsymbol{\Omega}^{(\ell)}/\sqrt{p}\|_{\text{op}}\|_{\psi_1} \lesssim \frac{\|\widehat{\mathbf{E}}\|_{\text{op}}}{\Delta}\sqrt{d/p},
\end{aligned}
$$

where the last but one inequality is due to Lemma B.2, and the last inequality is due to Lemma B.1. Therefore, by Lemma 3 in Fan et al. (2019), there exists constants $c_1 > 0$ and $C_1 > 0$ such that

$$
\mathbb{P}\left(\|\boldsymbol{\Sigma}' - \mathbf{V}_K\mathbf{V}_K\|_{\text{op}} \geq \varepsilon\right) \leq \mathbb{P}\left(\frac{\|\widehat{\mathbf{E}}\|_{\text{op}}}{\Delta}\sqrt{d/p} \geq c_1\varepsilon\right) \leq \exp\left(-\frac{C_1\varepsilon}{\tau\sqrt{dr/np}}\right).
$$

Similarly, we consider the probability $\mathbb{P}\left(\|\boldsymbol{\Sigma}' - \widehat{\mathbf{V}}_K\widehat{\mathbf{V}}_K^\top\|_{\text{op}} \geq \varepsilon\right)$. By Lemma 3 in Fan et al. (2019), there exist constants $c_2 > 0$, $C_2 > 0$ and $C_3 > 0$ such that

$$
\begin{aligned}
\mathbb{P}\left(\|\boldsymbol{\Sigma}' - \widehat{\mathbf{V}}_K\widehat{\mathbf{V}}_K^\top\|_{\text{op}} \geq \varepsilon\right) &\leq \mathbb{P}\left(\|\boldsymbol{\Sigma}' - \mathbf{V}_K\mathbf{V}_K\|_{\text{op}} \geq \varepsilon/2\right) + \mathbb{P}\left(\|\widehat{\mathbf{V}}_K\widehat{\mathbf{V}}_K^\top - \mathbf{V}_K\mathbf{V}_K\|_{\text{op}} \geq \varepsilon/2\right) \\
&\leq \exp\left(-\frac{C_1\varepsilon}{2\tau\sqrt{dr/np}}\right) + \mathbb{P}\left(\|\widehat{\mathbf{E}}\|_{\text{op}}/\Delta \geq c_2\varepsilon\right) \leq \exp\left(-\frac{C_1\varepsilon}{2\tau\sqrt{dr/np}}\right) + \exp\left(-\frac{C_2\varepsilon}{\tau\sqrt{r/n}}\right) \\
&\leq \exp\left(-\frac{C_3\varepsilon}{\tau\sqrt{dr/np}}\right).
\end{aligned}
$$

Therefore, the claim follows.

## G.4 Proof of Lemma E.2

Recall that $\mathbf{Y}^{(\ell)} = \widehat{\mathbf{\Sigma}}\mathbf{\Omega}^{(\ell)}$, $\check{\mathbf{Y}}^{(\ell)} = \mathbf{V}_K\mathbf{\Lambda}_K\mathbf{V}_K^\top\mathbf{\Omega}^{(\ell)}$ and $\check{\mathbf{Y}}^{(\ell)} - \mathbf{Y}^{(\ell)} = \widehat{\mathbf{E}}'\mathbf{\Omega}^{(\ell)}$. Following the proof of Lemma B.3, by Wedin's Theorem (Wedin, 1972) and Jensen's inequality, we have

$$\begin{aligned}\|\mathbf{\Sigma}' - \mathbf{V}_K\mathbf{V}_K\|_{\mathrm{op}} &\le \mathbb{E}\left(\left\|\widehat{\mathbf{V}}_K^{(\ell)}\widehat{\mathbf{V}}_K^{(\ell)\top} - \mathbf{V}_K\mathbf{V}_K\right\|_{\mathrm{op}}\middle|\widehat{\mathbf{\Sigma}}\right) \lesssim \mathbb{E}\left(\|\mathbf{Y}^{(\ell)}/\sqrt{p} - \check{\mathbf{Y}}^{(\ell)}/\sqrt{p}\|_{\mathrm{op}}/\sigma_K\left(\check{\mathbf{Y}}^{(\ell)}/\sqrt{p}\right)|\widehat{\mathbf{\Sigma}}\right)\\ &\le \frac{\|\widehat{\mathbf{E}}'\|_{\mathrm{op}}}{\Delta}\mathbb{E}\left(\frac{\|\mathbf{\Omega}^{(\ell)}/\sqrt{p}\|_{\mathrm{op}}}{\sigma_{\min}\left(\widetilde{\mathbf{\Omega}}^{(\ell)}/\sqrt{p}\right)}\right) \lesssim \frac{\|\widehat{\mathbf{E}}'\|_{\mathrm{op}}}{\Delta}\|\|\mathbf{\Omega}^{(\ell)}/\sqrt{p}\|_{\mathrm{op}}\|_{\psi_1} \lesssim \frac{\|\widehat{\mathbf{E}}'\|_{\mathrm{op}}}{\Delta}\sqrt{d/p}.\end{aligned}$$

Then by Assumption 5.1, there exist generic constants $c_0, c_1, c_2 > 0$ such that

$$\begin{aligned}\mathbb{P}\left(\|\mathbf{\Sigma}'-\mathbf{V}_K\mathbf{V}_K\|_{\mathrm{op}} \ge \varepsilon\right) &\le \mathbb{P}\left(\frac{\|\widehat{\mathbf{E}}'\|_{\mathrm{op}}}{\Delta}\sqrt{\frac{d}{p}} \ge c_0\varepsilon\right) \le \mathbb{P}\left(\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_{\mathrm{op}} \ge \frac{c_0\varepsilon\Delta}{2}\sqrt{\frac{p}{d}}\right) + \mathbb{P}\left(\|\mathbf{D}\|_{\mathrm{op}} \ge \frac{c_0\varepsilon\Delta}{2}\sqrt{\frac{p}{d}}\right)\\ &\lesssim \exp\left(-\frac{\varepsilon\Delta\sqrt{p/d}}{c_1\lambda_1 g(r,n)}\right) + \mathbb{P}\left(\lambda_1 g(r,n) \ge c_2\varepsilon\Delta\sqrt{\frac{p}{d}}\right).\end{aligned}$$

Under the condition that $\tau g(r,n) = o(\sqrt{p/d})$, we have that $\mathbb{P}\left(\lambda_1 g(r,n) \ge c_2\varepsilon\Delta\sqrt{\frac{p}{d}}\right) = 0$ for large enough $n$. Thus we have

$$\mathbb{P}\left(\|\mathbf{\Sigma}'-\mathbf{V}_K\mathbf{V}_K\|_{\mathrm{op}} \ge \varepsilon\right) \lesssim \exp\left(-\frac{\varepsilon\Delta\sqrt{p/d}}{c_1\lambda_1 g(r,n)}\right).$$

Similarly for the probability $\mathbb{P}\left(\|\mathbf{\Sigma}' - \widehat{\mathbf{V}}_K\widehat{\mathbf{V}}_K^\top\|_{\mathrm{op}} \ge \varepsilon\right)$. By Davis-Kahan's Theorem (Yu et al., 2015), there exists constants $C_0 > 0$ such that

$$\begin{aligned}\mathbb{P}\left(\|\mathbf{\Sigma}' - \widehat{\mathbf{V}}_K\widehat{\mathbf{V}}_K^\top\|_{\mathrm{op}} \ge \varepsilon\right) &\le \mathbb{P}\left(\|\mathbf{\Sigma}' - \mathbf{V}_K\mathbf{V}_K\|_{\mathrm{op}} \ge \varepsilon/2\right) + \mathbb{P}\left(\|\widehat{\mathbf{V}}_K\widehat{\mathbf{V}}_K^\top - \mathbf{V}_K\mathbf{V}_K\|_{\mathrm{op}} \ge \varepsilon/2\right)\\ &\lesssim \exp\left(-\frac{\varepsilon\Delta\sqrt{p/d}}{2c_1\lambda_1 g(r,n)}\right) + \mathbb{P}\left(\|\widehat{\mathbf{E}}'\|_{\mathrm{op}}/\Delta \ge C_0\varepsilon\right) \lesssim \exp\left(-\frac{\varepsilon\Delta\sqrt{p/d}}{2c_1\lambda_1 g(r,n)}\right).\end{aligned}$$

Thus the claim follows.

# H Proof of the Modified Wedin's Theorem

**Lemma H.1** (Modified Wedin's Theorem)**.** Let $\mathbf{M}^\star$ and $\mathbf{M} = \mathbf{M}^\star + \mathbf{E}$ be two matrices in $\mathbb{R}^{n_1 \times n_2}$ (without loss of generality, we assume $n_1 \leq n_2$ ), whose SVDs are given respectively by

$$\mathbf{M}^\star = \sum_{i=1}^{n_1} \sigma_i^\star \mathbf{u}_i^\star \mathbf{v}_i^{\star\top} = \begin{bmatrix} \mathbf{U}^\star & \mathbf{U}_\perp^\star \end{bmatrix} \begin{bmatrix} \mathbf{\Sigma}^\star & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma}_\perp^\star & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{V}^{\star\top} \\ \mathbf{V}_\perp^{\star\star} \end{bmatrix}$$

$$\mathbf{M} = \sum_{i=1}^{n_1} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top = \begin{bmatrix} \mathbf{U} & \mathbf{U}_\perp \end{bmatrix} \begin{bmatrix} \mathbf{\Sigma} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma}_\perp & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{V}^\top \\ \mathbf{V}_\perp^\top \end{bmatrix}$$

Here, $\sigma_1 \geq \cdots \geq \sigma_{n_1}$ (respectively $\sigma_1^\star \geq \cdots \geq \sigma_{n_1}^\star$) stand for the singular values of $\mathbf{M}$ (respectively $\mathbf{M}^\star$) arranged in descending order, $\mathbf{u}_i$ (respectively $\mathbf{u}_i^\star$) denotes the left singular vector associated with the singular value $\sigma_i$ (respectively $\sigma_i^\star$), and $\mathbf{v}_i$ (respectively $\mathbf{v}_i^\star$) represents the right singular vector associated with $\sigma_i$ (respectively $\sigma_i^\star$). $\mathbf{U}$ and $\mathbf{U}^*$ stand for the top $r$ eigenvectors of $\mathbf{M}$ and $\mathbf{M}^*$ respectively. Then,

$$\max\left\{\|\mathbf{U}\mathbf{U}^\top - \mathbf{U}^\star\mathbf{U}^{\star\top}\|_{\mathrm{op}}, \|\mathbf{V}\mathbf{V}^\top - \mathbf{V}^\star\mathbf{V}^{\star\top}\|_{\mathrm{op}}\right\} \lesssim \frac{2\|\mathbf{E}\|_{\mathrm{op}}}{\sigma_r^\star - \sigma_{r+1}^\star}, \tag{S.13}$$

and

$$\max\left\{\|\mathbf{U}\mathbf{U}^\top - \mathbf{U}^\star\mathbf{U}^{\star\top}\|_{\mathrm{F}}, \|\mathbf{V}\mathbf{V}^\top - \mathbf{V}^\star\mathbf{V}^{\star\top}\|_{\mathrm{F}}\right\} \lesssim \frac{2\sqrt{r}\|\mathbf{E}\|_{\mathrm{op}}}{\sigma_r^\star - \sigma_{r+1}^\star}. \tag{S.14}$$

*Proof.* By Wedin's Theorem (Wedin, 1972), if $\|\mathbf{E}\|_{\mathrm{op}} < (1 - 1/\sqrt{2})\left(\sigma_r^\star - \sigma_{r+1}^\star\right)$, (S.13) and (S.14) are true. When $\|\mathbf{E}\|_{\mathrm{op}} \geq (1 - 1/\sqrt{2})\left(\sigma_r^\star - \sigma_{r+1}^\star\right)$, the RHS of (S.13) are larger than or equal to $2 - \sqrt{2}$, whereas the LHS are bounded by 1. Thus (S.13) follows trivially, and so is (S.14). □