FADI: FAST DISTRIBUTED-FRIENDLY PRINCIPAL COMPONENT ANALYSIS WITH HIGH ACCURACY FOR LARGE-SCALE DATA

By Shuting Shen[1,a], Junwei Lu[1,b] and Xihong Lin[1,c]

[1]*Department of Biostatistics, Harvard University,* [a]*shs145@g.harvard.edu;* [b]*junweilu@hsph.harvard.edu;* [c]*xlin@hsph.harvard.edu*

Principal component analysis (PCA) is one of the most popular methods for dimension reduction. In light of the rapidly increasing large-scale data in federated ecosystems, the traditional PCA method is often not applicable due to privacy protection considerations and large computational burdens. Many algorithms have been proposed to lower the computational cost, but few can handle both high dimensionality and massiveness under the distributed setting. For example, fast PCA algorithms [18] have been proposed to accelerate the computation for high-dimensional data but would usually result in higher statistical errors and are not friendly for distributed computing. Distributed PCA algorithms [15] have been developed to handle federated data with large sample sizes but are not computationally efficient when the dimension is large. In this paper, we propose the FAst DIstributed-friendly (FADI) PCA method that can simultaneously perform parallel computing along the dimension and distributed computing along the sample size. Specifically, we utilize L parallel copies of p-dimensional fast sketches to divide the computing burden along the dimension d and aggregate the results distributively along the split samples. Theoretically, we establish comprehensive results applicable to many statistical models and show that FADI enjoys the same non-asymptotic error rate as the traditional PCA when $Lp \asymp d$. We also derive inferential results that characterize the distributional convergence of FADI with the phase-transition phenomenon: when $Lp \gg d$, we achieve the same asymptotic efficiency as the traditional PCA; on the other hand, if one wants to expedite the inference, a different procedure can be performed under the regime $Lp \ll d$ but will lose asymptotic efficiency. We perform extensive simulation studies to show that FADI substantially outperforms the existing methods and validate through numerical experiments the distributional phase-transition phenomenon of FADI. We apply FADI to the analysis of the 1000 Genomes data to study the population structure.

1. Introduction. As one of the most popular methods for dimension reduction, principal component analysis (PCA) finds applications in a broad spectrum of scientific fields including network studies [2], statistical genetics [32] and finance [28]. Methodologically, parameter estimation in many statistical models is dependent on PCA, such as spectral clustering in graphical models [1], missing data imputation through low-rank matrix completion [21] and clustering with subsequent k-means refinement in Gaussian mixture models [11]. When it comes to real data analysis, however, several shortcomings of the traditional PCA method hinder its application to large-scale datasets. On the one side, the high dimensionality and large sample size of modern big data can render the PCA computation infeasible in practice. For instance, PCA has been proven useful in controlling for ancestry confounding in Genome-Wide Association Studies (GWAS) [30], yet biomedical databases such as the UK Biobank [36] often contain hundreds of thousands to millions of Single Nucleotide Polymorphisms (SNPs) and subjects, which entail more scalable algorithms to handle the intensive computation of PCA. On the other side, large-scale datasets in many applications are stored in federated ecosystems, where data cannot leave individual warehouses due to privacy protection considerations [13]. This calls for federated learning methods [24, 27] that provide efficient and privacy-

Keywords and phrases: Computational efficiency | distributed computing | efficient communication | fast PCA | large-scale inference | federated learning | random matrices | random sketches.

protected strategies for joint analysis across multiple data warehouses without the need to exchange individual-level 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 the full-sample data and apply random projection to speed up PCA calculations [9, 18], while the existing distributed PCA algorithms apply the traditional PCA method to the split/site-specific data and aggregate the results [15, 26].

Specifically, fast PCA algorithms utilize the fact that the column space of a low-rank matrix can be represented by a small set of columns and use random projection to approximate the original high-dimensional matrix [3]. For instance, Halko, Martinsson and Tropp [18] proposed to estimate the K leading eigenvectors of a $d \times d$ matrix ($K \ll d$) using Gaussian random sketches, which decreases the PCA computation time by a factor of $O(d)$ at the cost of increasing the statistical error by a factorial power of d. Chen et al. [9] modified Halko, Martinsson and Tropp [18]'s method by repeating the fast sketching multiple times and showed the consistency of the algorithm using the average of i.i.d. random sketches when the number of sketches goes to infinity. However, they did not study the trade-off between computation complexity and the error rate in finite samples, and hence did not recommend the number of fast sketches that optimizes both the computational efficiency and the statistical accuracy. As the fast PCA methods use the full data, they have two major limitations. First, they are often not scalable to large sample sizes, 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" [15, 25, 26]. The horizontal partition splits the data over the sample size n, whereas the vertical partition splits the data over the dimension d. Horizontal partition is useful when the sample size n is large or the data are federated in multiple sites and cannot be shared. Fan et al. [15] considered the horizontally distributed PCA where they estimated the K leading eigenvectors of the $d \times d$ population covariance matrix by applying traditional PCA to each data split and 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 they used the traditional PCA algorithm for each data partition, the computational complexity is at least of order $O(d^3)$, which will be computationally difficult when d is large. Kargupta et al. [26] considered vertical partition and developed a method that collects local principal components (PCs) and then reconstructs global PCs by linear transformations. However, there is no theoretical guarantee on the error rate compared with the traditional full sample PCA, and the method may fail when variables are correlated.

Apart from the aforementioned PCA applications in parameter estimation, inference also constitutes an important part of PCA studies. For example, when studying the ancestry groups of genomes data under the mixed membership models, while the estimation error rate guarantees the overall misclustering rate for all subjects, we might still be interested in testing whether two individuals of interest share the same ancestry membership profile and assessing the associated statistical uncertainty [16]. The demand for real-world PC inference on high-dimensional data calls for fast and general algorithms that can efficiently perform the inferential procedure on possibly federated data systems. In terms of the inferential analyses, despite the rich literature depicting the asymptotic distribution of traditional PCA estimators under different statistical models [16, 29, 38], distributional characterization for fast and distributed PCA methods are not well-studied. For example, Yang et al. [41] characterized the convergence of fast sketching estimators in probability but gave no inferential results. Halko, Martinsson and Tropp [18] provided error bound for the fast PCA algorithm, but there is no characterization of the asymptotic distribution and hence no evaluation of the testing efficiency. Fan et al. [15] derived the non-asymptotic error rate of the distributed PC estimator but did not provide distributional guarantees, and inference based upon their estimator is not fast enough when the dimension is large. Besides, Fan et al. [15]'s method is restricted to the covariance model for i.i.d. samples.

In summary, the existing PCA methods mainly focus on dividing the computing burden along the sample size n, while parallel computing along the dimension d is harder to implement due to the correlated structure of variables and lacks theoretical guarantees. It remains an open question how to develop fast PCA algorithms that can handle high dimensionality and large sample size simultaneously while achieving the same asymptotic efficiency as the traditional PCA. In view of the gaps in existing literature, we propose in this paper a scalable and computationally efficient FAst DIstributed-friendly (FADI) PCA algorithm applicable to large federated data. More specifically, to obtain the K-leading PCs of a $d \times d$ matrix $\mathbf{M}$ from its estimator $\widehat{\mathbf{M}}$, we take the divide-and-conquer strategy to break down the computation complexity along the dimension: we generate the p-dimensional fast sketch $\mathbf{Y} = \widehat{\mathbf{M}}\boldsymbol{\Omega}$ and perform SVD on $\mathbf{Y}$ instead of $\widehat{\mathbf{M}}$ to expedite the PCA computation, where $\boldsymbol{\Omega} \in \mathbb{R}^{d\times p}$ is a Gaussian test matrix with $K \leq p \ll d$; meanwhile, to adjust for the additional variability induced by random approximation we repeat the fast sketching for L times in parallel and aggregate the SVD results to restore statistical accuracy. In the scenario where the data are distributively stored, the federated structure of our estimator also enables its easy implementation without the need of sharing individual-level data, as opposed to previous fast PCA methods that cannot handle distributed data setting. We will show that the computational complexity of FADI is of smaller magnitudes than existing methods (see Table 3). Moreover, we establish general frameworks that cover multiple statistical models. We list below four statistical models as illustrative applications of FADI.

- Spiked covariance model: let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \in \mathbb{R}^d$ be i.i.d. random vectors with spiked covariance $\boldsymbol{\Sigma} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^\top + \sigma^2\mathbf{I}$, where $\mathbf{M} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^\top$ is the rank-K spiked component of interest. Define $\widehat{\mathbf{M}} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i\boldsymbol{X}_i^\top - \widehat{\sigma}^2\mathbf{I}$ to be the estimator for $\mathbf{M}$, where $\widehat{\sigma}^2$ is a consistent estimator for σ^2. We assume that the data are split along the sample size n and stored on m servers.
- Degree-corrected mixed membership (DCMM) model: let $\mathbf{X}$ be the adjacency matrix for an undirected graph of d nodes, where each node is assigned to one of K communities with probability determined by its membership profile, and the connection probabilities between nodes are determined by their community assignments and node-associated degrees. Then the marginal connection probability matrix $\mathbf{M} = \mathbb{E}(\mathbf{X})$ is a rank-K matrix of interest that can be estimated by $\widehat{\mathbf{M}} = \mathbf{X}$.
- Gaussian mixture models (GMM): let $\boldsymbol{W}_1, \ldots, \boldsymbol{W}_d \in \mathbb{R}^n$ be independent random vectors drawn from K Gaussian distributions with different means and identity covariance matrix. We are interested in clustering through $\mathbf{M} = [\mathbb{E}(\boldsymbol{W}_i)^\top\mathbb{E}(\boldsymbol{W}_j)]$, whose estimator is given by $\widehat{\mathbf{M}} = [\boldsymbol{W}_i^\top\boldsymbol{W}_j] - n\mathbf{I}$. Assume the data are distributively stored on m servers along the dimension n.
- Missing matrix inference: we have a low-rank matrix $\mathbf{M}$ of interest, and we observe $\widehat{\mathbf{M}}$ as a perturbed version of $\mathbf{M}$ with missing entries. We make inference on the eigenspace of $\mathbf{M}$ through $\widehat{\mathbf{M}}$.

We will discuss the above examples in detail in Section 2. For distributed data setting such as the spiked covariance model and the GMM, while using fast sketching to untangle the variable dependency and split the computing burden along d, FADI also utilizes the data structure to perform the divide-and-conquer procedure along n, which we will further specify in Section 3. Table 1 provides complexities of FADI for the four examples and suggested choice of parameters for optimal error rates. We will establish in Section 4.1 a general non-asymptotic error bound applicable to multiple statistical models as well as case-specific error rates for each example, and show that the non-asymptotic error rate of the FADI estimator is of the same order as the traditional PCA as long as the number of fast sketches is sufficiently large. Inferentially, we also provide distributional characterizations of the FADI estimator under different regimes of the fast sketching parameters. We observe a phase-transition phenomenon where the asymptotic covariance matrix takes on two different forms as Lp increases. When $Lp \gg d$, the FADI estimator converges in distribution to a multivariate Gaussian, and the asymptotic relative efficiency (ARE) between FADI and the traditional PCA is 1. On the other hand, to accelerate inference, a different inferential procedure will be performed at the regime $Lp \ll d$ at the cost of losing testing efficiency (see Figure 1).

Contributions. We summarize the major contributions of our paper as follows.

TABLE 1
Computational complexities and parameter choice for different statistical models.

	Complexity	p	L
Spiked covariance model	$O(dnp/m + dKp'Lq)$	$K \vee \log d$	d/p
DCMM model	$O(d^2p + dKp'Lq)$	$\sqrt{d}$	$\sqrt{d}$
Gaussian mixture models	$O(dnp/m + dKp'Lq)$	$K \vee \log d$	d/p
Missing matrix inference	$O(d^2p + dKp'Lq)$	$\sqrt{d}$	$\sqrt{d}$

(a) Example 1 Spiked Covariance Model (b) Example 3: Gaussian Mixture Models

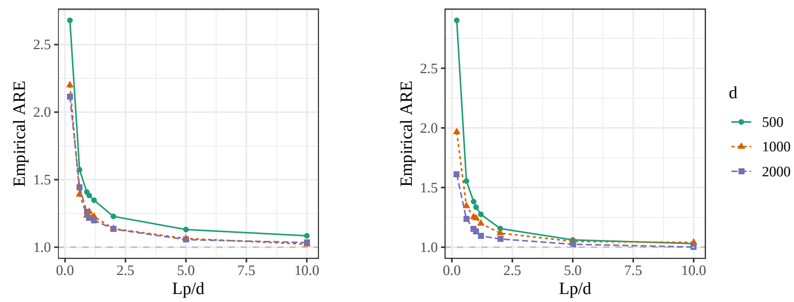


FIGURE 1. *Asymptotic relative efficiency (ARE) between the FADI estimator and the traditional PCA estimator under Example 1 and Example 3, where the ARE is measured by* $\big(\det(\widehat{\mathbf{\Sigma}}_1^{(1)})/\det(\widehat{\mathbf{\Sigma}}_1^{(2)})\big)^{1/K}$ *with* $\widehat{\mathbf{\Sigma}}_1^{(j)}$, $j = 1,2$ *being the empirical covariance matrix for the first row of two estimators [33].*

First, FADI divides complexity along high dimensionality and massiveness with theoretical guarantees: to the best of our knowledge, we propose the first algorithm that can handle parallel computing along d and distributed computing along n simultaneously while achieving the same statistical accuracy as the traditional PCA. Due to the fact that variables are usually dependent, it is challenging to achieve parallel PC computation along the dimension d. By dividing the high-dimensional data into L copies of p-dimensional fast sketches, FADI untangles the variable dependency and overcomes the difficulties of dividing the complexity along d. When the data are computed from large distributed samples, FADI utilizes the data structure and splits the computing burden by aggregating the fast sketches distributively over the split samples. We establish theoretical error bound to show that FADI is as accurate as the traditional PCA so long as $Lp \asymp d$. In comparison, Halko, Martinsson and Tropp [18]'s fast PCA method is not adapted to distributed systems and has elevated error rates; Fan et al. [15]'s distributed PCA is not as efficient when the split data have large d.

Second, we have distributional characterization for inferential analyses: we provide distributional guarantees on the FADI estimator to facilitate inference, which is absent in previous literature on fast and distributed PCA methods. More specifically, we depict the trade-off between computational complexity and testing efficiency by studying FADI's asymptotic distribution under the regimes $Lp \ll d$ and $Lp \gg d$ respectively. We show that the same asymptotic efficiency as the traditional PCA can be achieved at $Lp \gg d$ with a compromise on computational efficiency, while faster inferential procedures can be performed at $Lp \ll d$ with suboptimal testing efficiency. We further validate the distributional phase transition via numerical experiments.

Third, we provide a general framework applicable to multiple statistical models: compared to Fan et al. [15]'s distributed PCA that is restricted to estimating the covariance structure of i.i.d. samples, our method is inclusive of multiple statistical settings. We establish a framework that is comprehensive both methodologically and theoretically, and our theoretical results can be applied to multiple statistical models with mild assumptions.

1.1. *Related Papers.* We have previously reviewed papers on the application of PCA in parameter estimation. Now we move on to related literature for inferential analysis of PCA.

There has been a great amount of literature depicting the asymptotic distribution of traditional PCA estimators under different statistical models. Anderson [4] characterized the asymptotic normality of eigenvectors and eigenvalues for traditional PCA on the sample covariance matrix with fixed dimension. Paul [29] and Wang and Fan [38] extended the analysis to the high-dimensional regime and established distributional results under the spiked covariance model. Similar efforts were made by Johnstone [23] and Baik, Arous and Péché [5], where they studied the limiting distribution of the largest eigenvalue of the sample covariance matrix when both the dimension and the sample size go to infinity. Apart from inference on the sample covariance matrix of i.i.d. data, previous works also made progress in inferential analyses for a variety of statistical models including the DCMM model [16], the missing matrix completion problem [10], and high-dimensional data with heteroskedastic noise and missingness under the spiked covariance model [40]. Specifically, Fan et al. [16] employed statistics based on estimators of the principal eigenspace of the adjacency matrix to perform inference on whether two given nodes share the same membership profile under the DCMM model. Chen et al. [10] constructed entry-wise confidence intervals (CIs) for a low-rank matrix with missing data and Gaussian noise based on debiased convex/nonconvex PC estimators. A similar missing data inference problem was conducted in Yan, Chen and Fan [40], where they adopted a refined spectral method with imputed diagonal for CI construction of the underlying spiked covariance matrix of corrupted samples with missing data.

The aforementioned works were all based upon the traditional PCA approach and considered no distributed data setting, and hence will suffer from low computational efficiency when the data are high-dimensional or distributively stored across different sites. Our paper filled the gap in the literature and provided general inferential results on the fast sketching method with high computational efficiency adapted to high-dimensional federated data.

Paper Organization. The rest of the paper is organized as follows. Section 2 introduces the problem setting and provides an overview of FADI and its intuition. Section 3 discusses FADI's implementation details, as well as the computational complexity of FADI and its modifications when K is unknown. Section 4 presents the theoretical results of the statistical error and asymptotic normality of the FADI estimator. Section 5 shows the numerical evaluation of FADI and comparison with several existing methods. The application of FADI to the 1000 Genomes Data is given in Section 6, followed by discussions.

Notations. We use $\mathbf{1}_d \in \mathbb{R}^d$ to denote the vector of length d with all entries equal to 1, and use $\{\mathbf{e}_i\}_{i=1}^d$ to denote the canonical basis of $\mathbb{R}^d$. For a matrix $\mathbf{A} = [\mathbf{A}_{ij}] \in \mathbb{R}^{m\times n}$, we use $\sigma_i(\mathbf{A})$ (respectively $\lambda_i(\mathbf{A})$) to represent the i-th largest singular value (respectively eigenvalue) of $\mathbf{A}$, and $\sigma_{\max}(\mathbf{A})$ or $\sigma_{\min}(\mathbf{A})$ (respectively $\lambda_{\max}(\mathbf{A})$ or $\lambda_{\min}(\mathbf{A})$) stands for the largest or smallest singular value (respectively eigenvalue) of $\mathbf{A}$. If $\mathbf{A}$ has the singular value decomposition $\mathbf{A} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{V}^\top = \sum_{j=1}^K \sigma_j \mathbf{u}_j \mathbf{v}_j^\top$, then we denote by $\mathbf{A}^\dagger = \mathbf{V}\boldsymbol{\Lambda}^{-1}\mathbf{U}^\top$ the pseudo-inverse of $\mathbf{A}$, $\mathbf{P}_{\mathbf{A}} = \mathbf{A}\mathbf{A}^\dagger$ the projection matrix onto the column space of $\mathbf{A}$, and $\operatorname{sgn}(\mathbf{A}) = \sum_{\sigma_j>0} \mathbf{u}_j\mathbf{v}_j^\top$ the matrix signum. If $\mathbf{A}$ is positive definite with eigen-decomposition $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{U}^\top$, we define $\mathbf{A}^{1/2} = \mathbf{U}\mathbf{D}^{1/2}\mathbf{U}^\top$ and $\mathbf{A}^{-1/2} = \mathbf{U}\mathbf{D}^{-1/2}\mathbf{U}^\top$. 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. For two orthonormal matrices $\mathbf{V}, \mathbf{U} \in \mathbb{R}^{n_1\times n_2}$ with $n_1 > n_2$, we measure the distance between their column spaces by $\rho(\mathbf{U}, \mathbf{V}) := \|\mathbf{U}\mathbf{U}^\top - \mathbf{V}\mathbf{V}^\top\|_{\mathrm{F}}$. For a vector $\mathbf{v} = (v_1, \ldots, v_d)^\top$, we use $\|\mathbf{v}\|_2 = \left(\sum_{i=1}^d |v_i|^2\right)^{1/2}$ to denote the vector ℓ_2-norm, and $\|\mathbf{v}\|_\infty = \max_i |v_i|$ to denote the vector ℓ_∞-norm. For a matrix $\mathbf{A} = [A_{ij}]$, we use $\|\mathbf{A}\|_2$ to denote the matrix spectral norm, $\|\mathbf{A}\|_{\mathrm{F}}$ to denote the Frobenius norm, $\|\mathbf{A}\|_{2,\infty} := \sup_{\|\mathbf{x}\|_2=1} \|\mathbf{A}\mathbf{x}\|_\infty = \max_i \|\mathbf{A}^\top\mathbf{e}_i\|_2$ to denote the 2-to-∞ norm and $\|\mathbf{A}\|_{\max} = \max_{i,j} |A_{ij}|$ to denote the matrix max norm. For two integers $j > i \geq 1$, denote by $[i]$ the set $\{1, 2, \ldots, i\}$, $i : j$ 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 $x_n \leq C y_n$ 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 $x_n \ll y_n$. Let $\mathbb{I}\{\cdot\}$ denote an indicator function, which takes 1 if the statement inside $\{\cdot\}$ is true and 0 otherwise. Through out the paper, we use c and C to represent generic constants and their values might change from place to place.

2. Preliminaries and Problem Setup. We aim to estimate the eigenspace of the rank-K symmetric matrix $\mathbf{M} \in \mathbb{R}^{d \times d}$, whose eigen-decomposition is given by $\mathbf{M} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top$, where $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_K)$, $|\lambda_1| \geq |\lambda_2| \geq \ldots \geq |\lambda_K| > 0$ and $\mathbf{V}$ is the stacking of the K leading eigenvectors.[1] We denote by $\Delta = |\lambda_K|$ the eigengap of $\mathbf{M}$, and assume without loss of generality that $\lambda_1 > 0$. $\widehat{\mathbf{M}}$ is a corrupted version of $\mathbf{M}$ obtained from observed data, with $\mathbf{E} = \widehat{\mathbf{M}} - \mathbf{M}$ representing the error matrix. Our goal is to estimate the column space of $\mathbf{V}$ from the observed matrix $\widehat{\mathbf{M}}$, possibly distributively in certain settings. The following four examples provide concrete statistical setups for the above problem.

EXAMPLE 1 (Spiked Covariance Model [23]). Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \in \mathbb{R}^d$ be i.i.d. sub-Gaussian random vectors with $\mathbb{E}(\boldsymbol{X}_1) = \mathbf{0}$ and $\mathbb{E}(\boldsymbol{X}_1\boldsymbol{X}_1^\top) = \mathbf{\Sigma}$.[2] We assume the following decomposition for the covariance matrix: $\mathbf{\Sigma} = \sigma^2\mathbf{I}_d + \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top$, where $\mathbf{V} \in \mathbb{R}^{d \times K}$ is the stacked K leading eigenvectors and $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_K)$ with $\lambda_1 \geq \ldots \geq \lambda_K > 0$. Assume that the data are split along the sample size n and stored on m different sites. Denote by $\{\boldsymbol{X}_i^{(j)}\}_{i=1}^{n_j}$ the sample split of size n_j on the j-th site, and by $\mathbf{X}^{(j)} = (\boldsymbol{X}_1^{(j)}, \ldots, \boldsymbol{X}_{n_j}^{(j)})^\top$ the corresponding data matrix split ($j = 1, \ldots, m$ and $\sum_{j=1}^m n_j = n$). Denote by $\mathbf{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)^\top$ the full $n \times d$ data matrix. Then $\mathbf{M} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top$, and $\widehat{\mathbf{M}} = \widehat{\mathbf{\Sigma}} - \widehat{\sigma}^2\mathbf{I}_d$, where $\widehat{\mathbf{\Sigma}} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i\boldsymbol{X}_i^\top$ is the sample covariance matrix and $\widehat{\sigma}^2$ is a consistent estimator for σ^2.

EXAMPLE 2 (Degree-Corrected Mixed Membership (DCMM) Model). Let $\mathbf{X} \in \mathbb{R}^{d \times d}$ be a symmetric adjacency matrix for an undirected graph of d nodes, where $\mathbf{X}_{ij} = 1$ if nodes $i, j \in [d]$ are connected and $\mathbf{X}_{ij} = 0$ otherwise. Assume $\mathbf{X}_{ij}$'s are independent for $i \leq j$ and $\mathbb{E}\mathbf{X} = \mathbf{\Theta}\mathbf{\Pi}\mathbf{P}\mathbf{\Pi}^\top\mathbf{\Theta}$, where $\mathbf{\Theta} = \mathrm{diag}(\theta_1, \ldots, \theta_d)$ stands for the degree heterogeneity matrix, $\mathbf{\Pi} = (\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_d)^\top \in \mathbb{R}^{d \times K}$ is the stacked community assignment probability vectors and $\mathbf{P} \in \mathbb{R}^{K \times K}$ is a symmetric rank-K matrix with non-zero constant entries and $\mathbf{P}_{kl} \in (0, 1]$ for $k, l \in [d]$. Then $\mathbf{M} = \mathbb{E}\mathbf{X}$ and $\widehat{\mathbf{M}} = \mathbf{X}$.[3] Since $\mathbf{V}$ and $\mathbf{\Theta}\mathbf{\Pi}$ share the same column space, we can make inference on the community membership profiles through $\mathbf{V}$. [4] In this paper, we assume that there exist constants $C \geq c > 0$ such that $\sigma_K(\mathbf{\Pi}) \geq c\sqrt{d/K}$, $c \leq \lambda_K(\mathbf{P}) \leq \lambda_1(\mathbf{P}) \leq CK$ and $\max_i \theta_i \leq C \min_i \theta_i$, where we define $\theta = \max_i \theta_i^2$ as the rate of signal strength. Please refer to Fan et al. [16] for similar setups.

EXAMPLE 3 (Gaussian Mixture Models (GMM)). Let $\boldsymbol{W}_1, \ldots, \boldsymbol{W}_d \in \mathbb{R}^n$ be independent samples generated from K Gaussian distributions with means $\boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_K^* \in \mathbb{R}^n$. More specifically, each $\boldsymbol{W}_i$ is associated with a predetermined membership parameter $k_i \in [K]$, and $\boldsymbol{W}_i \sim \mathcal{N}(\boldsymbol{\theta}_{k_i}^*, \mathbf{I}_n)$. Denote $\mathbf{X} = (\boldsymbol{W}_1, \ldots, \boldsymbol{W}_d) = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)^\top$, where $\boldsymbol{X}_i$ is the i-th row of $\mathbf{X}$. We consider the distributed setting

[1] When $\mathbf{M}$ is an asymmetric low-rank matrix, we can deploy the "symmetric dilation" trick and take $\mathcal{S}(\mathbf{M}) = \begin{pmatrix} \mathbf{0} & \mathbf{M} \\ \mathbf{M}^\top & \mathbf{0} \end{pmatrix}$ to fit it into the setting.

[2] We assume $\{\boldsymbol{X}_i\}_{i=1}^n$ are i.i.d. for the simplicity of presentation. We will generalize the theoretical results to non-i.i.d. and heterogeneous data in Section 4.1.

[3] In the case where self-loops are absent, $\mathbf{X}$ will be replaced by $\mathbf{X}' = \mathbf{X} - \mathrm{diag}(\mathbf{X})$ and $\mathbf{E}$ will be replaced by $\mathbf{E}' = \mathbf{E} - \mathrm{diag}(\mathbf{X})$. We will abuse the notation and use $\mathbf{E}$ to represent both $\mathbf{E}$ and $\mathbf{E}'$. Our theoretical results hold for both cases.

[4] To address the degree heterogeneity, one can perform the SCORE normalization to cancel out $\mathbf{\Theta}$ [22].

where the data are split along the dimension n and distributively stored on m sites. Denote by $\mathbf{X}^{(j)} = (\boldsymbol{X}_1^{(j)}, \ldots, \boldsymbol{X}_{n_j}^{(j)})^\top$ the data split on the j-th site of size n_j ($j \in [m]$). Without loss of generality, we order $\boldsymbol{W}_i$'s such that $\mathbb{E}(\mathbf{X}) = \boldsymbol{\Theta}^* \mathbf{F}^{*\top}$, where

$$\boldsymbol{\Theta}^* = (\boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_K^*) \in \mathbb{R}^{n \times K}, \quad \mathbf{F}^* = \mathrm{diag}(\mathbf{1}_{d_1}, \ldots, \mathbf{1}_{d_k}) \in \mathbb{R}^{d \times K},$$

with d_k denoting the number of samples drawn from the Gaussian distribution with mean $\boldsymbol{\theta}_k^*$. Then we define $\mathbf{M} = \mathbb{E}[\mathbf{X}^\top \mathbf{X}] - n\mathbf{I}_d = \mathbf{F}^* \boldsymbol{\Theta}^{*\top} \boldsymbol{\Theta}^* \mathbf{F}^{*\top}$ and $\widehat{\mathbf{M}} = \mathbf{X}^\top \mathbf{X} - n\mathbf{I}_d$. Since $\mathbf{V}$ and $\mathbf{F}^*$ share the same column space, we can recover the memberships from $\mathbf{V}$. In this paper, we consider the regime where $n > d$. Besides, we assume that there exists a constant $C > 0$ such that $\max_k d_k \le C \min_k d_k$ and $\sigma_1(\boldsymbol{\Theta}^*) \le C\sigma_K(\boldsymbol{\Theta}^*)$. Please refer to Chen et al. [11] for similar setups.

EXAMPLE 4 (Missing Matrix Inference). Assume that $\mathbf{M} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^\top$ is a symmetric rank-K matrix, and $\mathcal{S} \subseteq [d] \times [d]$ is a subset of indices. We only observe the perturbed entries of $\mathbf{M}$ in the subset $\mathcal{S}$. More specifically, for $i \le j$, we denote $\delta_{ij} = \delta_{ji} = \mathbb{I}\{(i,j) \in \mathcal{S}\}$, and $\delta_{ij} \overset{\text{i.i.d}}{\sim} \mathrm{Bernoulli}(\theta)$. Then for $i, j \in [d]$, the observation for $\mathbf{M}_{ij}$ is $(\mathbf{M}_{ij} + \varepsilon_{ij})\delta_{ij}$, where $\varepsilon_{ij} = \varepsilon_{ji}$ are i.i.d. random variables satisfying $\mathbb{E}(\varepsilon_{ij}) = 0$, $\mathbb{E}(\varepsilon_{ij}^2) = \sigma^2$ and $\sup_{i \le j} |\varepsilon_{ij}| \lesssim \sigma \log d$.[5] Then we define $\widehat{\mathbf{M}}_{ij} = \widehat{\theta}^{-1}(\mathbf{M}_{ij} + \varepsilon_{ij})\delta_{ij}$, where $\widehat{\theta} = 2|\mathcal{S}|/\big(d(d+1)\big)$.[6] Please refer to Chen et al. [10] for more details.

3. Method.

In this section, we will elaborate on the FADI algorithm and its application to different examples. We then discuss the computational complexity of FADI and compare it with the existing methods. We will show how FADI is scalable when the data are large-scale or federated, and discuss how to estimate the rank K.

3.1. *Fast Distributed PCA (FADI): Overview and Intuition.*

We provide in this section an overview of the proposed FAst DIstributed-friendly (FADI) PCA method and its intuition and will present the detailed algorithm in Section 3.2. For a given matrix $\widehat{\mathbf{M}} \in \mathbb{R}^{d \times d}$, the computational cost of the traditional PCA on $\widehat{\mathbf{M}}$ will take $O(d^3)$ flops. In the case where $\widehat{\mathbf{M}}$ is computed from observed data, e.g., the sample covariance matrix $\widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i \boldsymbol{X}_i^\top$, extra computational burden comes from calculating $\widehat{\mathbf{M}}$, e.g., $O(nd^2)$ flops for computing the sample covariance matrix. Hence performing traditional PCA for large-scale data with high dimensions and huge sample sizes can be considerably expensive.

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 [18]. For instance, for a low-rank matrix $\mathbf{M}$ of rank K, its column space can be represented by a low-dimensional fast sketch $\mathbf{M}\boldsymbol{\Omega} \in \mathbb{R}^{d \times p}$ when p is sufficiently large, where $\boldsymbol{\Omega} \in \mathbb{R}^{d \times p}$ is a random Gaussian matrix of dimension p and $K < p \ll d$. In our paper, we perform the fast sketching on the almost low-rank corrupted matrix $\widehat{\mathbf{M}}$, and obtain the fast sketch $\widehat{\mathbf{Y}} = \widehat{\mathbf{M}}\boldsymbol{\Omega} \approx \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^\top \boldsymbol{\Omega}$ that almost maintains the same left singular space as $\mathbf{M} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^\top$. It is hence reasonable to estimate $\mathbf{V}$ by performing SVD on the $d \times p$ matrix $\widehat{\mathbf{Y}}$ that has a much smaller computational cost than directly performing PCA on $\widehat{\mathbf{M}}$. As information might be lost from $\widehat{\mathbf{M}}$ due to fast sketching, it motivates us to repeat the fast sketching multiple times and aggregate the results to reduce the statistical error. We will see in Section 4.1 that when the number of repeated fast sketches is sufficiently large, FADI enjoys the same error rate as the traditional PCA. From this perspective, FADI can be viewed as a "vertically" distributed PCA method as it allocates the computational burden along

[5] We can generalize the results to sub-Gaussian error ε_{ij}'s with variance proxy σ^2 by taking the truncated error $\varepsilon_{ij}^t = \varepsilon_{ij}\mathbb{I}\{|\varepsilon_{ij}| \le 4\sigma\sqrt{\log d}\}$, and by the maximal inequality for sub-Gaussian random variables we know that with probability at least $1 - O(d^{-6})$, $\varepsilon_{ij} = \varepsilon_{ij}^t, \forall i, j \in [d]$, and the theorems can be generalized with minor modifications.

[6] In practice, we can estimate $\mathbf{V}$ by $\widehat{\theta}\widehat{\mathbf{M}} = [(\mathbf{M}_{ij} + \varepsilon_{ij})\delta_{ij}]$ rather than by $\widehat{\mathbf{M}} = [\widehat{\theta}^{-1}(\mathbf{M}_{ij} + \varepsilon_{ij})\delta_{ij}]$, since the two matrices share exactly the same eigenvectors.

the dimension d to several machines using low-dimensional sketches while maintaining high statistical accuracy through the aggregation of local PCs. FADI overcomes the difficulties of vertical splitting caused by the correlation between variables.

To accommodate FADI to the distributed data setting, we need special structures for $\widehat{\mathbf{M}}$. More specifically, assume that the data are stored across m different sites, and we have the decomposition $\widehat{\mathbf{M}} = \sum_{j=1}^{m} \widehat{\mathbf{M}}^{(j)}$, where $\widehat{\mathbf{M}}^{(j)}$ is the component that can be computed locally on the j-th machine ($j \in [m]$). This representation of $\widehat{\mathbf{M}}$ is legitimate in many models. For instance, Examples 1 and 3 in Section 2 satisfy the aforementioned decomposition of $\widehat{\mathbf{M}}$ under the distributed data setting, i.e., $\widehat{\mathbf{M}}^{(j)} = \frac{1}{n}(\mathbf{X}^{(j)\top}\mathbf{X}^{(j)}) - (\widehat{\sigma}^2/m)\mathbf{I}_d$ in Example 1 and $\widehat{\mathbf{M}}^{(j)} = \mathbf{X}^{(j)\top}\mathbf{X}^{(j)} - (n/m)\mathbf{I}_d$ in Example 3. Then instead of applying random sketching directly to $\widehat{\mathbf{M}}$, FADI computes in parallel the local fast sketching for each component $\widehat{\mathbf{M}}^{(j)}$ and aggregates the results across m sites, which will reduce the cost of computing $\widehat{\mathbf{M}}\mathbf{\Omega}$ by a factor of $1/m$.

3.2. *General Algorithmic Framework.* Now we are ready to present the algorithm. Recall for a rank-K matrix $\mathbf{M}$, we observe $\widehat{\mathbf{M}} = \mathbf{M} + \mathbf{E}$, and we aim to estimate the K leading eigenvectors $\mathbf{V}$ of $\mathbf{M}$ from $\widehat{\mathbf{M}}$. Recall that $\widehat{\mathbf{M}}$ has the representation $\widehat{\mathbf{M}} = \sum_{j=1}^{m} \widehat{\mathbf{M}}^{(j)}$, where $m = 1$ for non-distributed data setting and $m > 1$ for distributed data setting.

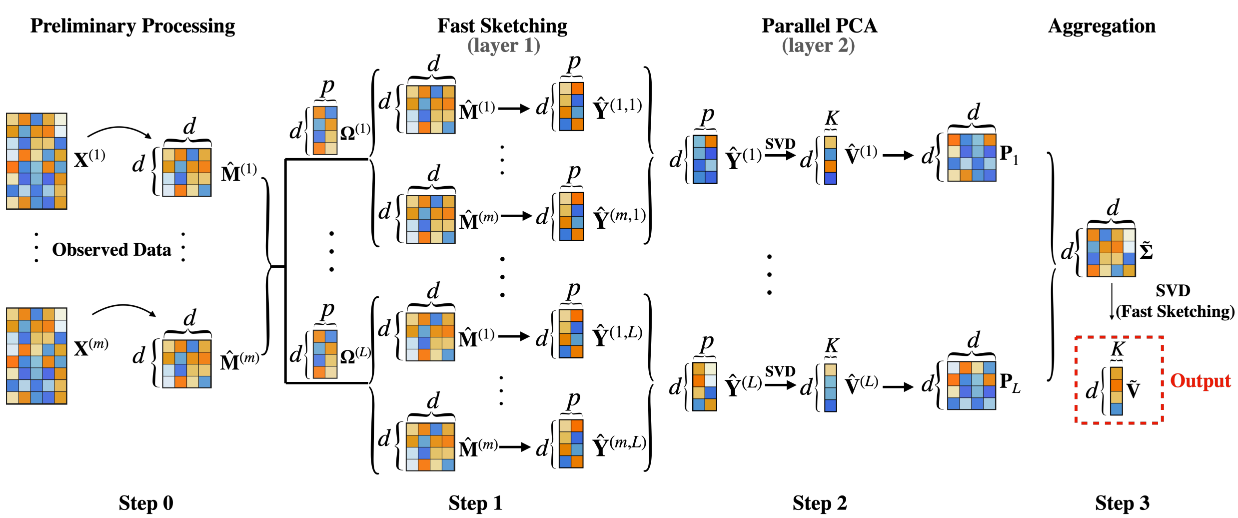


FIGURE 2. *Illustration of FADI. Here $\{\mathbf{X}^{(j)}\}_{j=1}^{m}$ are the raw data stored possibly distributively on m sites, and $\widehat{\mathbf{M}}^{(j)}$ is the j-th component of $\widehat{\mathbf{M}}$ that can be calculated from $\mathbf{X}^{(j)}$. $\widehat{\mathbf{Y}}^{(\ell)} = \sum_{j\in[m]} \widehat{\mathbf{Y}}^{(j,\ell)}$ ($\ell \in [L]$) is the ℓ-th copy of the fast sketch obtained by aggregating the fast sketches calculated distributively for each data split. Note that if $m = 1$ (non-distributed data), we will directly obtain $\widehat{\mathbf{Y}}^{(\ell)} = \widehat{\mathbf{Y}}^{(1,\ell)}$.*

Figure 2 illustrates the fast distributed-friendly PCA (FADI) algorithm:

In Step 0, we perform preliminary processing on the raw data to produce $\{\widehat{\mathbf{M}}^{(j)}\}_{j=1}^{m}$. We will elaborate on the case-specific preprocessing in Section 3.3.

In Step 1, we calculate the Gaussian fast sketches $\widehat{\mathbf{Y}} = \widehat{\mathbf{M}}\mathbf{\Omega} = \sum_{j=1}^{m} \widehat{\mathbf{M}}^{(j)}\mathbf{\Omega}$, where $\mathbf{\Omega}$ is a $d \times p$ standard Gaussian test matrix and $K < p \ll d$. To reduce the statistical error, we repeat the fast sketches L times and aggregate the results from the L copies of $\widehat{\mathbf{Y}}$. Specifically, we generate L i.i.d. Gaussian test matrices $\{\mathbf{\Omega}^{(\ell)}\}_{\ell=1}^{L}$, and for each $\ell \in [L]$, we apply $\mathbf{\Omega}^{(\ell)}$ distributively to $\widehat{\mathbf{M}}^{(j)}$ for each $j \in [m]$ and obtain the ℓ-th fast sketch of $\widehat{\mathbf{M}}^{(j)}$ as $\widehat{\mathbf{Y}}^{(j,\ell)} = \widehat{\mathbf{M}}^{(j)}\mathbf{\Omega}^{(\ell)}$. We send $\widehat{\mathbf{Y}}^{(j,\ell)}$ ($j = 1, \cdots, m$) to the ℓ-th parallel server for aggregation.

In Step 2, on the ℓ-th server, the random sketches $\widehat{\mathbf{Y}}^{(j,\ell)}$ $(j=1,\cdots,m)$ from the m split datasets corresponding to the ℓ-th Gaussian test matrix $\boldsymbol{\Omega}^{(\ell)}$ will be collected and added up to get the ℓ-th fast sketch: $\widehat{\mathbf{Y}}^{(\ell)}=\sum_{j=1}^{m}\widehat{\mathbf{Y}}^{(j,\ell)}$ $(\ell\in[L])$. We next compute in parallel the top K left singular vectors $\widehat{\mathbf{V}}^{(\ell)}$ of $\widehat{\mathbf{Y}}^{(\ell)}$ and send the $\widehat{\mathbf{V}}^{(\ell)}$'s to the central processor for aggregation.

In Step 3, on the central processor, calculate $\widetilde{\boldsymbol{\Sigma}}=\frac{1}{L}\sum_{\ell=1}^{L}\widehat{\mathbf{V}}^{(\ell)}\widehat{\mathbf{V}}^{(\ell)\top}=\frac{1}{L}\sum_{\ell=1}^{L}\mathbf{P}_\ell$, where $\mathbf{P}_\ell=\widehat{\mathbf{V}}^{(\ell)}\widehat{\mathbf{V}}^{(\ell)\top}$ is the projection matrix of $\widehat{\mathbf{V}}^{(\ell)}$. We next calculate the K leading eigenvectors $\widetilde{\mathbf{V}}$ of $\widetilde{\boldsymbol{\Sigma}}$, which will serve as the final estimator of $\mathbf{V}$.

To further improve the computational efficiency, we might conduct another fast sketching in Step 3 to compute $\widetilde{\mathbf{V}}$. More specifically, we apply the power method [18] to $\widetilde{\boldsymbol{\Sigma}}$ by calculating $\widetilde{\mathbf{Y}}=\widetilde{\boldsymbol{\Sigma}}^q\boldsymbol{\Omega}^{\mathrm{F}}=\left(\frac{1}{L}\sum_{l=1}^{L}\widehat{\mathbf{V}}^{(\ell)}\widehat{\mathbf{V}}^{(\ell)\top}\right)^q\boldsymbol{\Omega}^{\mathrm{F}}$, where $\boldsymbol{\Omega}^{\mathrm{F}}\in\mathbb{R}^{d\times p'}$ is a Gaussian test matrix with the dimension p' that can be set different from p for optimal error rate, and $q>1$ is the power. Here, $\widetilde{\mathbf{Y}}$ can be calculated iteratively: $\widetilde{\mathbf{Y}}_{(i)}=\frac{1}{L}\sum_{l=1}^{L}\left(\widehat{\mathbf{V}}^{(\ell)}\widehat{\mathbf{V}}^{(\ell)\top}\widetilde{\mathbf{Y}}_{(i-1)}\right)$ for $i=1,\ldots,q$, where $\widetilde{\mathbf{Y}}_{(0)}=\boldsymbol{\Omega}^{\mathrm{F}}$ and $\widetilde{\mathbf{Y}}=\widetilde{\mathbf{Y}}_{(q)}$. We denote by $\widetilde{\mathbf{V}}^{\mathrm{F}}$ the leading K left singular vectors of $\widetilde{\mathbf{Y}}$. We will show in Section 4 that when q is properly large, the distance between $\widetilde{\mathbf{V}}$ and $\widetilde{\mathbf{V}}^{\mathrm{F}}$ will be negligible.

REMARK 1. We refer to Theorem 4.1 for the choice of p and L. In general, taking $p=2K$ is sufficient. For now, we assume K is known, and the scenarios where K is unknown will be discussed in Section 3.5.

3.3. *Case-Specific Processing of Raw Data.* In this section, we discuss the calculation of $\widehat{\mathbf{M}}$ in Step 0 of FADI specifically for each example.

Example 1: Recall that in Step 0 of FADI, to obtain $\widehat{\mathbf{M}}$, we need a consistent estimator of the residual variance σ^2. Denote by $S=\{i_1,i_2,\ldots,i_{K'}\}\subseteq[d]$ an arbitrary index set of size $K'\geq K+1$. Then we estimate σ^2 by $\widehat{\sigma}^2=\sigma_{\min}(\widehat{\boldsymbol{\Sigma}}_S)$, where $\widehat{\boldsymbol{\Sigma}}_S$ is a $K'\times K'$ principal submatrix of $\widehat{\boldsymbol{\Sigma}}$ computed using only data columns in the set S. Due to the additive structure of the sample covariance matrix, $\widehat{\boldsymbol{\Sigma}}_S$ can be easily computed distributively (see Figure 10 in Supplementary Materials E for reference). Then for $j\in[m]$, we have $\widehat{\mathbf{M}}^{(j)}=\frac{1}{n}(\mathbf{X}^{(j)\top}\mathbf{X}^{(j)})-(\widehat{\sigma}^2/m)\mathbf{I}_d$. Note that since computing $\widehat{\mathbf{M}}^{(j)}\boldsymbol{\Omega}=\frac{1}{n}\mathbf{X}^{(j)\top}(\mathbf{X}^{(j)}\boldsymbol{\Omega})-m^{-1}\widehat{\sigma}^2\boldsymbol{\Omega}$ is much faster than first computing $\widehat{\mathbf{M}}^{(j)}$ then computing $\widehat{\mathbf{M}}^{(j)}\boldsymbol{\Omega}$, we may save the data components of $\widehat{\mathbf{M}}^{(j)}$ for later computation rather than directly calculate $\widehat{\mathbf{M}}^{(j)}$.

Example 2: Since the data are not distributively stored and $\widehat{\mathbf{M}}$ is the adjacency matrix readily obtained from the observed data, $m=1$ and no preliminary processing is needed.

Example 3: Recall that the data $\{\boldsymbol{X}_i\}_{i=1}^{d}\subseteq\mathbb{R}^n$ are vertically distributed across m sites, and $\{\mathbf{X}^{(j)}\}_{j=1}^{m}$ are the corresponding data splits. For the j-th site, we have $\widehat{\mathbf{M}}^{(j)}=\mathbf{X}^{(j)\top}\mathbf{X}^{(j)}-(n/m)\mathbf{I}_d$ and for $\ell\in[L]$ we compute $\widehat{\mathbf{Y}}^{(j,\ell)}$ by $\mathbf{X}^{(j)\top}(\mathbf{X}^{(j)}\boldsymbol{\Omega}^{(\ell)})-(n/m)\boldsymbol{\Omega}^{(\ell)}$.

Example 4: No distributed data structure is assumed and $m=1$. When K is known, we can take $\widehat{\theta\mathbf{M}}=[(\mathbf{M}_{ij}+\varepsilon_{ij})\delta_{ij}]$ instead of $\widehat{\mathbf{M}}$ for the estimation of $\mathbf{V}$, which is the observed corrupted matrix with missing data. When K is unknown or when we need to conduct inference on $\mathbf{V}$ (see Section 4.3), we will calculate $\widehat{\theta}=2|\mathcal{S}|/\{d(d+1)\}$ and take $\widehat{\mathbf{M}}=[\widehat{\theta}^{-1}(\mathbf{M}_{ij}+\varepsilon_{ij})\delta_{ij}]$ for the correct scaling of the eigenvalues .

3.4. *Computational Complexity.* In this section, we provide the computational complexity of FADI for each example given in Section 2. The complexity of each step is listed in Table 2.

For Examples 1 and 3, when the number of data splits can be customized, we recommend taking $m\asymp n/d$ for optimal efficiency. When $p\asymp(K\vee\log d)$, $L\asymp d/p$, $p'\asymp K$ and $q\asymp\log d$, the total computational cost will be $O\big(dn(K\vee\log d)/m+d^2K\log d\big)$.

For Examples 2 and 4, direct SVD on $\widetilde{\boldsymbol{\Sigma}}$ will induce computational cost of order d^3 and we only suggest $\widetilde{\mathbf{V}}^{\mathrm{F}}$ as the eigenspace estimator. If we take $p\asymp\sqrt{d}$, $L\asymp d/p$, $p'\asymp K$ and $q\asymp\log d$, the total computational cost will be $O(d^{2.5})$.

TABLE 2

Computational complexities for Examples 1-4. For Examples 1 and 3, we assume $\max_{j\in[m]} n_j \asymp n/m$ for the simplicity of presentation. In Step 3, the calculation of $\widetilde{\mathbf{V}}$ involves computing $\widetilde{\mathbf{\Sigma}}$ at $O(d^2pL)$ flops and SVD on $\widetilde{\mathbf{\Sigma}}$ at $O(d^3)$ flops, while computing $\widetilde{\mathbf{V}}^F$ involves calculation of $\widetilde{\mathbf{\Sigma}}^q\mathbf{\Omega}^F$ at $O(dKp'Lq)$ flops and SVD on $\widetilde{\mathbf{\Sigma}}^q\mathbf{\Omega}^F$ at $O(dp'^2)$ flops. We recommend computing $\widetilde{\mathbf{V}}^F$ instead of $\widetilde{\mathbf{V}}$ in practice. The total complexity in the last line refers to the total computational cost for $\widetilde{\mathbf{V}}^F$.

	Example 1	Example 2	Example 3	Example 4
Step 0	$\widehat{\mathbf{\Sigma}}_S : O(K^2n/m + K^2m)$ $\widehat{\sigma}^2 : O(K^3)$	N/A	O(1)	$O(d^2)$
Step 1	$\widehat{\mathbf{Y}}^{(j,\ell)} : O(dnp/m)$	$\widehat{\mathbf{Y}}^{(\ell)} : O(d^2p)$	$\widehat{\mathbf{Y}}^{(j,\ell)} : O(dnp/m)$	$\widehat{\mathbf{Y}}^{(\ell)} : O(d^2p)$
Step 2	$\widehat{\mathbf{Y}}^{(\ell)} : O(mdp)$ $\widehat{\mathbf{V}}^{(\ell)} : O(dp^2)$	$\widehat{\mathbf{V}}^{(\ell)} : O(dp^2)$	$\widehat{\mathbf{Y}}^{(\ell)} : O(mdp)$ $\widehat{\mathbf{V}}^{(\ell)} : O(dp^2)$	$\widehat{\mathbf{V}}^{(\ell)} : O(dp^2)$
Step 3	$\widetilde{\mathbf{V}} : O(d^2pL + d^3)$	N/A	$\widetilde{\mathbf{V}} : O(d^2pL + d^3)$	N/A
	$\widetilde{\mathbf{V}}^{\mathrm{F}} : O(dKp'Lq + dp'^2)$			
Total	$O(dnp/m + dKp'Lq)$	$O(d^2p + dKp'Lq)$	$O(dnp/m + dKp'Lq)$	$O(d^2p + dKp'Lq)$

TABLE 3

Error rates and computational complexities for FADI, traditional PCA, fast PCA (one sketching) [18] and distributed PCA [15] for Example 1, where the error rate is evaluated by $\left(\mathbb{E}|\rho(\cdot,\mathbf{V})|^2\right)^{1/2}$. Here $r = \mathrm{tr}(\mathbf{\Sigma})/\|\mathbf{\Sigma}\|_2$ is the effective rank of the covariance matrix and m is the number of sites. For FADI, we take $p \asymp (K \vee \log d)$, $L \asymp d/p$, $p' \asymp K$ and $q \asymp \log d$.

Method	Error Rate	Computational Complexity
FADI	$O(\sqrt{Kr/n})$	$O\left(dn(K \vee \log d)/m + d^2K\log d\right)$
Traditional PCA	$O(\sqrt{Kr/n})$	$O(d^2n + d^3)$
Fast PCA	$O(\sqrt{Kdr/n})$	$O(dnK + d^2K)$
Distributed PCA	$O(\sqrt{Kr/n})$	$O(d^2n/m + d^3)$

Inference on eigenspace will require the calculation of the asymptotic covariance, whose formula and computational costs will be discussed in Sections 4.3 and 4.4.

For a straightforward comparison of FADI with existing works, we provide in Table 3 a list of the theoretical error rates and the computational complexities of FADI against different PCA methods under Example 1 (please refer to Therem 4.1 for the error rates of FADI). We choose Example 1 for illustration because Fan et al. [15]'s distributed PCA method also considers PCA for the covariance matrix of i.i.d. samples. The results show that under the distributed setting FADI has a much lower computational complexity than the other three methods while enjoying the same error rate as the traditional full-sample PCA. In comparison, Fan et al. [15]'s distributed PCA method is slowed down by applying traditional PCA to each data split which is costly for high-dimensional data. The fast PCA algorithm in [18] has suboptimal complexity and theoretical error rate due to their downstream projection that hinders aggregation.

3.5. *Estimation of the Rank K.* FADI requires to input the rank K of the matrix $\mathbf{M}$. In practice, if we are only interested in estimating the leading PCs, 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 as well as facilitate inference on PCs. In fact, the estimation of K can be incorporated into Step 2 and Step 3 of FADI: for the

ℓ-th parallel server ($\ell \in [L]$), after performing the SVD $\widehat{\mathbf{Y}}^{(\ell)} = \widehat{\mathbf{V}}_p^{(\ell)} \widehat{\mathbf{\Lambda}}_p^{(\ell)} \widehat{\mathbf{U}}_p^{(\ell)\top}$, we estimate K by

$$\widehat{K}^{(\ell)} = \min\{k < p : \sigma_{k+1}(\widehat{\mathbf{Y}}^{(\ell)}) - \sigma_p(\widehat{\mathbf{Y}}^{(\ell)}) \le \sqrt{p}\mu_0\},$$

where $\mu_0 > 0$ is a user-specified parameter (we refer to Theorem 4.3 for the choice of μ_0).[7] Then send all the left singular vectors $\widehat{\mathbf{V}}_p^{(\ell)} \in \mathbb{R}^{d\times p}$ and $\widehat{K}^{(\ell)}, \ell \in [L]$ to the central processor. Finally, on the central processor, take $\widehat{K} = \lceil \text{median} \left\{ \widehat{K}^{(1)}, \widehat{K}^{(2)}, \ldots, \widehat{K}^{(L)} \right\} \rceil$ as the estimator for K, and obtain $\widetilde{\mathbf{V}}_{\widehat{K}}$ (respectively $\widetilde{\mathbf{V}}^{\mathrm{F}}_{\widehat{K}}$) by performing PCA (respectively powered fast sketching) on the aggregated average of $\{\widehat{\mathbf{V}}_{\widehat{K}}^{(\ell)}\}_{\ell\in[L]}$ and take the $\widehat{K}$ leading PCs, where $\widehat{\mathbf{V}}_{\widehat{K}}^{(\ell)}$ is the $\widehat{K}$ leading PCs of $\widehat{\mathbf{V}}_p^{(\ell)}$. We will show in Theorem 4.3 that $\widehat{K}$ is a consistent estimator of K.

4. Theory. In this section, we will conduct a theoretical analysis of the FADI estimator. We will establish a theoretical upper bound for the error rate of FADI in Section 4.1, and characterize the asymptotic distribution of the FADI estimator in Section 4.4 and Section 4.3 to facilitate inference.

4.1. *Theoretical Bound on Error Rates.* We need the following condition to guarantee that the error term converges at a proper rate.

ASSUMPTION 1 (Convergence of $\|\mathbf{E}\|_2$). Recall that $\mathbf{E} = \widehat{\mathbf{M}} - \mathbf{M}$ is the error matrix. Assume that $\|\mathbf{E}\|_2$ is sub-exponential, and there exists a rate $r_1(d)$ such that

$$\|\|\mathbf{E}\|_2\|_{\psi_1} = \sup_{q\ge 1} q^{-1} \left(\mathbb{E}\|\mathbf{E}\|_2^q\right)^{1/q} \lesssim r_1(d).$$

REMARK 2. By standard probability theory [37] we know that there exists a constant $c_e > 0$ such that for any $t > 0$ we have $\mathbb{P}(\|\mathbf{E}\|_2 \ge t) \le \exp\left(-c_e t/r_1(d)\right)$ and $\|\mathbf{E}\|_2 = O_P\left(r_1(d)\right)$.

We will conduct a variance-bias decomposition on the error rate $\rho(\widetilde{\mathbf{V}}, \mathbf{V})$, where $\widetilde{\mathbf{V}}$ is the FADI PC estimator defined in Step 3 in Section 3.2. To facilitate this discussion, we introduce the intermediate matrix $\mathbf{\Sigma}' = \mathbb{E}_{\mathbf{\Omega}}\left(\widehat{\mathbf{V}}^{(\ell)}\widehat{\mathbf{V}}^{(\ell)\top}\right)$, where $\widehat{\mathbf{V}}^{(\ell)}$ is the top K left singular vectors of the ℓ-th fast sketch $\widehat{\mathbf{Y}}^{(\ell)}$ defined in Step 2 in Section 3.2, and the expectation is taken with respect to $\mathbf{\Omega}$. Let $\mathbf{V}'$ be the top K eigenvectors of $\mathbf{\Sigma}'$. Note that both $\mathbf{\Sigma}'$ and $\mathbf{V}'$ are random depending on $\widehat{\mathbf{M}}$. For the FADI PC estimator $\widetilde{\mathbf{V}}$, we have the following "variance-bias" decomposition of the error rate:

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

If we condition on all the available data, then the first term characterizes the statistical randomness of $\widetilde{\mathbf{V}}$ due to fast sketching, whereas the second bias term is deterministic and depends on all the information provided by the data. Intuitively, since $\widetilde{\mathbf{\Sigma}} = \frac{1}{L}\sum_{l=1}^{L} \widehat{\mathbf{V}}^{(\ell)}\widehat{\mathbf{V}}^{(\ell)\top}$ would converge to the conditional expectation $\mathbf{\Sigma}'$, $\widetilde{\mathbf{V}}$ would also converge to $\mathbf{V}'$. Hence the first variance term goes to 0 asymptotically. As for the second bias term, let $\widehat{\mathbf{V}}$ be the K leading eigenvectors of $\widehat{\mathbf{M}}$, then we further break the bias term into two components: $\rho(\mathbf{V}', \mathbf{V}) \le \rho(\widehat{\mathbf{V}}, \mathbf{V}) + \rho(\mathbf{V}', \widehat{\mathbf{V}})$. 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 by Lemma B.1 in Supplementary Materials B.1 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 $\widehat{\mathbf{V}}$, which is due to the information we can get from the available data. The following theorem gives the overall error rate of the FADI PC estimator $\widetilde{\mathbf{V}}$. Its proof is given in Supplementary Materials B.2.

[7] Note that the set is non-empty since $p-1$ is contained in the set, and thus $\widehat{K}^{(\ell)}$ is well-defined.

THEOREM 4.1. *Under Assumption 1, if $p \geq \max(2K, K+7)$ and $(\log d)^{-1}\sqrt{p/d}\Delta/r_1(d) \geq C$ for some large enough constant $C > 0$, we have*

$$(1) \qquad \left(\mathbb{E}|\rho(\widetilde{\mathbf{V}}, \mathbf{V})|^2\right)^{1/2} \lesssim \frac{\sqrt{K}}{\Delta} r_1(d) + \sqrt{\frac{Kd}{\Delta^2 pL}} r_1(d).$$

Furthermore, recall $\widetilde{\mathbf{V}}^{\mathrm{F}}$ is 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 1 and the conditions that $p \geq \max(2K, K+8q-1)$ and $(\log d)^{-1}\sqrt{p/d}\Delta/r_1(d) \geq C$, there exists some constant $\eta > 0$ such that

$$(2) \qquad \left(\mathbb{E}|\rho(\widetilde{\mathbf{V}}^{\mathrm{F}}, \mathbf{V})|^2\right)^{1/2} \lesssim \frac{\sqrt{K}}{\Delta} r_1(d) + \sqrt{\frac{Kd}{\Delta^2 pL}} r_1(d) + \sqrt{\frac{Kd}{p'}} \left(\eta q^2 \sqrt{\frac{d}{\Delta^2 p}} r_1(d)\right)^q.$$

REMARK 3. On the RHS of (1), the first term 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 the traditional PCA method. As for (2), the first term and the second term on the RHS are the same as the bias and variance term in (1), while the third term comes from the extra fast sketches. In fact, if we properly choose

$$q = \lceil \left(\log\left(\sqrt{p/d}\Delta/r_1(d)\right)\right)^{-1} \log d \rceil + 1 \leq \log d,$$

the third term in (2) will be negligible. Theorem 4.1 also indicates that p only needs to be of the same order as $K \vee \log d$, which significantly reduces the communication costs from $O(d^2)$ to $O\left(d(K \vee \log d)\right)$ for each server.

Based upon Theorem 4.1, we provide the case-specific error rate for each example given in Section 2 in the following corollary. Please refer to Supplementary Materials B.3 for the proof.

COROLLARY 4.2. *For Examples 1 – 4, we have the following error bounds for each case under corresponding regularity conditions.*

- *Example 1: Define $\kappa_1 = (\lambda_1 + \sigma^2)/\Delta$, then under the conditions that $p' \geq \max(2K, K+7)$, $p \geq \max(2K, K + 8\log d - 1)$, $q = \lceil \log d \rceil$ and $n \geq C(rd/p)\kappa_1^2 \log^4 d$ for some large enough constant $C > 0$, it holds that*

$$(3) \qquad \left(\mathbb{E}|\rho(\widetilde{\mathbf{V}}^{\mathrm{F}}, \mathbf{V})|^2\right)^{1/2} \lesssim \kappa_1 \sqrt{\frac{Kr}{n}} + \kappa_1 \sqrt{\frac{Kdr}{npL}},$$

  where $r = \mathrm{tr}(\mathbf{\Sigma})/\|\mathbf{\Sigma}\|_2$ is the effective rank.
- *Example 2: Suppose $\theta \geq K^2 d^{-1/2+\epsilon}$ for some constant $\epsilon > 0$. If we take $p' \geq \max(2K, K+7)$, $p \gtrsim \sqrt{d}$ and $q = \lceil \log d \rceil$, it holds that*

$$(4) \qquad \left(\mathbb{E}|\rho(\widetilde{\mathbf{V}}^{\mathrm{F}}, \mathbf{V})|^2\right)^{1/2} \lesssim K\sqrt{\frac{K}{d\theta}} + K\sqrt{\frac{K}{pL\theta}}.$$

- *Example 3: Under the conditions that $\Delta_0^2 \geq CK(\log d)^2 \max\left(d(\log d)^2/p,\ \sqrt{n/p}\right)$ for some large enough constant $C > 0$, where $\Delta_0 = \|\mathbf{\Theta}^*\|_2$, if we take $p' \geq \max(2K, K+7)$, $p \geq \max(2K, K + 8\log d - 1)$ and $q = \lceil \log d \rceil$, it holds that*

$$(5) \qquad \left(\mathbb{E}|\rho(\widetilde{\mathbf{V}}^{\mathrm{F}}, \mathbf{V})|^2\right)^{1/2} \lesssim \left(\frac{K}{\Delta_0} + \frac{K}{\Delta_0^2}\sqrt{\frac{Kn}{d}}\right) + \sqrt{\frac{d}{pL}}\left(\frac{K}{\Delta_0} + \frac{K}{\Delta_0^2}\sqrt{\frac{Kn}{d}}\right).$$

- *Example 4: Define $\kappa_2 = |\lambda_1|/\Delta$. Suppose $\theta \ge d^{-1/2+\epsilon}$ for some constant $\epsilon > 0$, $\sigma/\Delta \ll d^{-1}\sqrt{p\theta}$, $\|\mathbf{V}\|_{2,\infty} \le \sqrt{\mu K/d}$ for some $\mu \ge 1$ and $\kappa_2\mu K \ll d^{1/4}$, if we take $p' \ge \max(2K, K+7)$, $p \gtrsim \sqrt{d}$ and $q = \lceil \log d \rceil$, it holds that*

$$\left(\mathbb{E}|\rho(\widetilde{\mathbf{V}}^{\mathrm{F}}, \mathbf{V})|^2\right)^{1/2} \lesssim \sqrt{K}\left(\frac{\kappa_2\mu K}{\sqrt{d\theta}} + \sqrt{\frac{d\sigma^2}{\Delta^2\theta}}\right) + \sqrt{\frac{Kd}{pL}}\left(\frac{\kappa_2\mu K}{\sqrt{d\theta}} + \sqrt{\frac{d\sigma^2}{\Delta^2\theta}}\right). \tag{6}$$

REMARK 4. We can generalize the results of Example 1 to the heterogeneous residual variance model for non-i.i.d. data, under which $\{\boldsymbol{X}_i\}_{i=1}^n \subseteq \mathbb{R}^d$ are centered random vectors with covariance matrices satisfying $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n \mathbb{E}(\boldsymbol{X}_i\boldsymbol{X}_i^\top) = \mathbf{\Sigma} = \mathbf{D} + \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top$, where $\mathbf{D} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2)$ and $\|\mathbf{D}\|_2/\Delta = o(1)$. Then by plugging in $r_1(d) = \|\mathbf{D}\|_2 + \|\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_2\|_{\psi_1}$ we have the error bound under the heterogeneous scenario. While the first term is deterministic depending on the residual variance, the second term depends on the dependence structure of the sample. Many studies have depicted the convergence of the sample covariance matrix for non-i.i.d. data [6, 14].

REMARK 5. Here we compare our theoretical estimation rates derived in Corollary 4.2 with the existing results in literature. For Example 1, when $Lp \gtrsim d$, our error rate in (3) is optimal [15]. Under the distributed data setting, we require the total sample size n to be larger than rd/p, while Fan et al. [15] requires that $n/m > r$, where n/m is the sample size for each data split. Compared with Fan et al. [15]'s method, our method has theoretical guarantees for the error rate regardless of the number of data splits, but our scaling condition $n \gtrsim rd/p$ has an extra factor of d/p in exchange for reduced computation cost. As for Example 2, our estimation rate in (4) matches the inferential results in Fan et al. [16]. Please also refer to Section 4.3 for a detailed comparison with Fan et al. [16]'s method in terms of the limiting distributions. For Example 3, our estimation rate in (5) is the same as in Chen et al. [11]. For Example 4, our error rate in (6) matches the results in Chen et al. [11].

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 3.5 recovers the true K with high probability.

THEOREM 4.3. *Under Assumption 1, define $\eta_0 = 480c_e^{-1}\sqrt{d/(\Delta^2 p)}r_1(d)\log d$, where $c_e > 0$ is the constant defined in Remark 2. When $d \ge 2$, $2K \le p \ll d(\log d)^{-2}$ and $\eta_0 \le (32\log d)^{-2/(p-K+1)}$, if we choose μ_0 such that $\Delta\eta_0/24 \le \mu_0 \le \Delta\sqrt{\eta_0}/12$, then with probability at least $1 - O(d^{-(L\wedge 20)/2})$, $\widehat{K} = K$.*

The proof of Theorem 4.3 is deferred to Supplementary Materials B.4. Note that the $\log d$ in η_0 is for sub-exponential $\|\mathbf{E}\|_2$. In specific examples where $\|\mathbf{E}\|_2$ may converge faster than sub-exponentially, the $\log d$ might be dropped. Theorem 4.3 suggests that so long as the error term converges properly fast, we can easily estimate K in parallel without losing computational efficiency. We provide case-specific choices of the thresholding parameter μ_0 in the following corollary, whose proof can be found in Supplementary Materials B.5.

COROLLARY 4.4. *For Examples 1 to 4, we specify the recommendation of μ_0 under certain regularity conditions.*

- *Example 1: Under the conditions that $2K \le p \ll (\log d)^{-2}d$, $n \gg \kappa_1^2 rd/p(\log d)^4$, $(\lambda_1 + \sigma^2) \ll \left(\sqrt{np}/(d\log d)\right)^{1/4}$ and $\Delta \gg \left(\sigma^{-2}(np)^{-1/2}d\log d\right)^{1/3}$, if we take $\mu_0 = \left(d(np)^{-1/2}\log d\right)^{3/4}/12$, with probability at least $1 - O\left(d^{-(L\wedge 20)/2}\right)$, we have $\widehat{K} = K$.*
- *Example 2: Define $\widehat{\theta} = d^{-2}\sum_{i\le j}\widehat{\mathbf{M}}_{ij}$, then under the condition that $\theta \ge K^2 d^{-1/2+\epsilon}$ for some constant $\epsilon > 0$ and $\sqrt{d} \lesssim p \ll (\log d)^{-2}d$, if we take $\mu_0 = (\widehat{\theta}/p)^{1/2}d\log d/12$, with probability at least $1 - O\left(d^{-(L\wedge 20)/2}\right)$, we have $\widehat{K} = K$.*

- *Example 3: Under the condition that $2K \leq p \ll (\log d)^{-2}d$ and $K(\log d)^3\sqrt{n/p} \ll \Delta_0^2 \ll nK/d(\log d)^2$, if we take $\mu_0 = d(\log d)^2\sqrt{n/p}/12$, with probability at least $1 - O\left(d^{-(L\wedge 20)/2}\right)$, we have $\widehat{K} = K$.*
- *Example 4: Under the condition that $\theta \geq d^{-1/2+\epsilon}$ for some constant $\epsilon > 0$, $\|\mathbf{V}\|_{2,\infty} \leq \sqrt{\mu K/d}$ for some $\mu \geq 1$, $\kappa_2^2\mu^2 K \ll (\log d)^2$, $\sqrt{d} \lesssim p \ll (\log d)^{-2}d$ and $(p\theta)^{-1/4}\sqrt{d\sigma/\Delta}\log d = o(1)$, if we take $\mu_0 = d\widehat{\sigma}_0 \log d(p\widehat{\theta})^{-1/2}/12$, where $\widehat{\sigma}_0 = \left(\sum_{(i,j)\in\mathcal{S}}(\widehat{\theta}\widehat{\mathbf{M}}_{ij})^2/|\mathcal{S}|\right)^{1/2}$, then with probability at least $1 - O\left(d^{-(L\wedge 20)/2}\right)$, we have $\widehat{K} = K$.*

REMARK 6. For Example 3, we impose the upper bound on Δ_0 because in practice the eigengap Δ is unknown, and estimation of Δ requires knowledge of K. Imposing the upper bound on Δ_0 makes the term in μ_0 involving knowledge of Δ vanish and enables the estimation of K from observed data.

4.2. *Inferential Results on Asymptotic Distribution: Intuition and Assumptions.* In Section 4.1, we discuss the theoretical upper bound for the error rate and present the bias-variance decomposition for the FADI estimator $\widetilde{\mathbf{V}}^{\mathrm{F}}$. From (2), we can see that when $Lp \gg d$, the bias term will be the leading term, and the dominating error comes from $\rho(\widehat{\mathbf{V}}, \mathbf{V})$, whereas when $Lp \ll d$, the variance term will be the leading term and the main error derives from $\rho(\widetilde{\mathbf{V}}^{\mathrm{F}}, \widehat{\mathbf{V}})$. This offers insight into conducting inferential analysis on the estimator and implies a possible phase transition in the asymptotic distribution. Before moving on to further discussions, we need the following assumption to ensure that the bias of $\widehat{\mathbf{M}}$ is negligible.

ASSUMPTION 2 (Statistical Rate for Biased Error Term). For the error matrix we have the decomposition $\mathbf{E} = \mathbf{E}_0 + \mathbf{E}_b$, where $\mathbb{E}(\mathbf{E}_0) = \mathbf{0}$ and $\mathbf{E}_b$ is the biased error term satisfying $\lim_{d\to\infty}\mathbb{P}\big(\|\mathbf{E}_b\|_2 \leq r_2(d)\big) = 1$ with $r_2(d) = o\big(r_1(d)\big)$.

In fact, we will later show in Section 4.3 and Section 4.4 that the leading term for the distance between $\widetilde{\mathbf{V}}^{\mathrm{F}}$ and $\mathbf{V}$ takes on two different forms under the two scenarios:

$$\begin{aligned} \widetilde{\mathbf{V}}^{\mathrm{F}}\mathbf{H} - \mathbf{V} &\approx \mathbf{P}_\perp \mathbf{E}_0 \mathbf{V}\mathbf{\Lambda}^{-1} \quad, &&\text{if } Lp \gg d;\\ \widetilde{\mathbf{V}}^{\mathrm{F}}\mathbf{H} - \mathbf{V} &\approx \mathbf{P}_\perp \mathbf{E}_0 \mathbf{\Omega}\mathbf{B}_{\mathbf{\Omega}} L^{-1}, &&\text{if } Lp \ll d, \end{aligned}$$

where $\mathbf{H}$ is some orthogonal matrix aligning $\widetilde{\mathbf{V}}^{\mathrm{F}}$ with $\mathbf{V}$, $\mathbf{P}_\perp = \mathbf{I} - \mathbf{V}\mathbf{V}^\top$ is the projection matrix onto the column space perpendicular to $\mathbf{V}$, $\mathbf{\Omega} = (\mathbf{\Omega}^{(1)}/\sqrt{p}, \ldots, \mathbf{\Omega}^{(L)}/\sqrt{p}) \in \mathbb{R}^{d\times Lp}$ and $\mathbf{B}_{\mathbf{\Omega}} = (\mathbf{B}^{(1)\top}, \ldots, \mathbf{B}^{(L)\top})^\top$ with $\mathbf{B}^{(\ell)} = (\mathbf{\Lambda}\mathbf{V}^\top\mathbf{\Omega}^{(\ell)}/\sqrt{p})^\dagger \in \mathbb{R}^{p\times K}$ for $\ell = 1, \ldots, L$. To get an intuitive understanding on the form of the leading error term, let's start with the scenario $Lp \gg d$ where $\rho(\widetilde{\mathbf{V}}^{\mathrm{F}}, \mathbf{V}) \approx \rho(\widehat{\mathbf{V}}, \mathbf{V})$ and consider the case where $\{|\lambda_k|\}_{k=1}^K$ are well-separated such that $\mathbf{H} \approx \mathbf{I}_K$. Following basic algebra, we have

$$\begin{aligned} \widetilde{\mathbf{V}}^{\mathrm{F}} - \mathbf{V} &\approx \widehat{\mathbf{V}} - \mathbf{V} \approx \mathbf{P}_\perp(\widehat{\mathbf{V}} - \mathbf{V}) = \mathbf{P}_\perp(\widehat{\mathbf{M}}\widehat{\mathbf{V}}\widehat{\mathbf{\Lambda}}^{-1} - \mathbf{M}\mathbf{V}\mathbf{\Lambda}^{-1})\\ &\approx \mathbf{P}_\perp(\widehat{\mathbf{M}} - \mathrm{M})\mathbf{V}\mathbf{\Lambda}^{-1} = \mathbf{P}_\perp\mathbf{E}_0\mathbf{V}\mathbf{\Lambda}^{-1}, \end{aligned}$$

where the second approximation is due to the fact that $\widehat{\mathbf{V}}$ and $\mathbf{V}$ are fairly close and $\mathbf{P}_{\mathbf{V}}(\widehat{\mathbf{V}} - \mathbf{V})$ will be negligible.

Now we turn to the scenario $Lp \ll d$, where the error mainly comes from $\widetilde{\mathbf{V}}^{\mathrm{F}} - \widehat{\mathbf{V}}$. For a given $\ell \in [L]$, denote $\mathbf{Y}^{(\ell)} = \mathbf{M}\mathbf{\Omega}^{(\ell)} = \mathbf{V}\mathbf{\Lambda}\widetilde{\mathbf{\Omega}}^{(\ell)}$, where $\widetilde{\mathbf{\Omega}}^{(\ell)} = \mathbf{V}^\top\mathbf{\Omega}^{(\ell)}$ is also a Gaussian test matrix. Intuitively, $p^{-1}\widetilde{\mathbf{\Omega}}^{(\ell)}\widetilde{\mathbf{\Omega}}^{(\ell)\top} \approx \mathbf{I}_K$ when p is much larger than K. Hence $\widetilde{\mathbf{\Omega}}^{(\ell)}$ acts like an orthonormal matrix scaled by $\sqrt{p}$, and the rank-K truncated SVD for $\widehat{\mathbf{Y}}^{(\ell)}/\sqrt{p}$ and $\mathbf{Y}^{(\ell)}/\sqrt{p}$ will approximately be $\widehat{\mathbf{V}}^{(\ell)}\widehat{\mathbf{\Lambda}}(\widetilde{\mathbf{\Omega}}^{(\ell)}/\sqrt{p})$ and $\mathbf{V}\mathbf{\Lambda}(\widetilde{\mathbf{\Omega}}^{(\ell)}/\sqrt{p})$ respectively. Then following similar arguments as when $Lp \gg d$, we have

$$\widehat{\mathbf{V}}^{(\ell)} - \mathbf{V} \approx \mathbf{P}_\perp\left((\widehat{\mathbf{Y}}^{(\ell)}/\sqrt{p})(\widetilde{\mathbf{\Omega}}^{(\ell)}/\sqrt{p})^\top\widehat{\mathbf{\Lambda}}^{-1} - (\mathbf{Y}^{(\ell)}/\sqrt{p})(\widetilde{\mathbf{\Omega}}^{(\ell)}/\sqrt{p})^\top\mathbf{\Lambda}^{-1}\right)$$

$$\approx \mathbf{P}_{\perp}\left(\widehat{\mathbf{Y}}^{(\ell)}/\sqrt{p} - \mathbf{Y}^{(\ell)}/\sqrt{p}\right)(\widetilde{\boldsymbol{\Omega}}^{(\ell)}/\sqrt{p})^{\top}\boldsymbol{\Lambda}^{-1} \approx \mathbf{P}_{\perp}\mathbf{E}_0(\boldsymbol{\Omega}^{(\ell)}/\sqrt{p})\mathbf{B}^{(\ell)},$$

where the last approximation is because when $\widetilde{\boldsymbol{\Omega}}^{(\ell)}/\sqrt{p}$ is almost orthonormal we have $\mathbf{B}^{(\ell)} = (\boldsymbol{\Lambda}\widetilde{\boldsymbol{\Omega}}^{(\ell)}/\sqrt{p})^{\dagger} \approx (\widetilde{\boldsymbol{\Omega}}^{(\ell)}/\sqrt{p})^{\top}\boldsymbol{\Lambda}^{-1}$. Then aggregating the results over $\ell \in [L]$ we have

$$\widetilde{\mathbf{V}}^{\mathrm{F}} - \mathbf{V} \approx \frac{1}{L}\sum_{\ell=1}^{L}\left\{\widehat{\mathbf{V}}^{(\ell)} - \mathbf{V}\right\} \approx \frac{1}{L}\sum_{\ell=1}^{L}\mathbf{P}_{\perp}\mathbf{E}_0(\boldsymbol{\Omega}^{(\ell)}/\sqrt{p})\mathbf{B}^{(\ell)} = \mathbf{P}_{\perp}\mathbf{E}_0\boldsymbol{\Omega}\mathbf{B}_{\boldsymbol{\Omega}}L^{-1}.$$

It is worth noting that

$$\frac{1}{L}\boldsymbol{\Omega}\mathbf{B}_{\boldsymbol{\Omega}} \approx \frac{1}{L}\left(\sum_{\ell=1}^{L}(\boldsymbol{\Omega}/\sqrt{p})(\boldsymbol{\Omega}/\sqrt{p})^{\top}\right)\mathbf{V}\boldsymbol{\Lambda}^{-1} \to \mathbf{V}\boldsymbol{\Lambda}^{-1}, \tag{7}$$

when $Lp \gg d$, which demonstrates the consistency of the leading term across different regimes of Lp. To unify the notations, we denote the leading term for $\widetilde{\mathbf{V}}^{\mathrm{F}}\mathbf{H} - \mathbf{V}$ by

$$\mathbf{V}(\mathbf{E}_0) = \begin{cases}\mathbf{P}_{\perp}\mathbf{E}_0\mathbf{V}\boldsymbol{\Lambda}^{-1}, & \text{if } Lp \gg d;\\ \mathbf{P}_{\perp}\mathbf{E}_0\boldsymbol{\Omega}\mathbf{B}_{\boldsymbol{\Omega}}L^{-1}, & \text{if } Lp \ll d.\end{cases}$$

Before we formally present the theorems, we introduce the following extra regularity conditions necessary for studying the asymptotic features of the eigenspace estimator.

ASSUMPTION 3 (Incoherence Condition). For the eigenspace of the true matrix $\mathbf{M}$, we assume

$$\|\mathbf{V}\|_{2,\infty} \le \sqrt{\mu K/d},$$

where $\mu \ge 1$ may change with d.

ASSUMPTION 4 (Statistical Rates for Eigenspace Convergence). For the unbiased error term $\mathbf{E}_0$ and the traditional PCA estimator $\widehat{\mathbf{V}}$, we have the following statistical rates

$$\lim_{d\to\infty}\mathbb{P}\big(\|\widehat{\mathbf{V}}\,\mathrm{sgn}(\widehat{\mathbf{V}}^{\top}\mathbf{V}) - \mathbf{V}\|_{2,\infty} \le r_3(d)\big) = 1,$$

and

$$\lim_{d\to\infty}\mathbb{P}\big(\|\mathbf{E}_0(\mathbf{I}_d - \widehat{\mathbf{V}}\widehat{\mathbf{V}}^{\top})\mathbf{V}\|_{2,\infty} \le r_4(d)\big) = 1.$$

ASSUMPTION 5 (Central Limit Theorem (CLT)). For the leading term $\mathbf{V}(\mathbf{E}_0)$ and any $i \in [d]$, it holds that

$$\boldsymbol{\Sigma}_i^{-1/2}\mathbf{V}(\mathbf{E}_0)^{\top}\mathbf{e}_i \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{I}_K),$$

where $\boldsymbol{\Sigma}_i = \mathrm{Cov}(\mathbf{V}(\mathbf{E}_0)^{\top}\mathbf{e}_i|\boldsymbol{\Omega})$ when $Lp \ll d$ and $\boldsymbol{\Sigma}_i = \mathrm{Cov}(\mathbf{V}(\mathbf{E}_0)^{\top}\mathbf{e}_i)$ when $Lp \gg d$.

Assumption 3 is the incoherence condition [8] to guarantee that the information of the eigenspace is uniformly spread. In Assumption 4 , $r_3(d)$ bounds the row-wise estimation error for the eigenspace, while $r_4(d)$ characterizes the row-wise convergence rate of the residual error term projected on the spaces spanned by $\widehat{\mathbf{V}}_{\perp}$ and $\mathbf{V}$ consecutively, i.e., $\|\mathbf{E}_0(\mathbf{I}_d - \widehat{\mathbf{V}}\widehat{\mathbf{V}}^{\top})\mathbf{V}\|_{2,\infty} = \|\mathbf{E}_0\mathbf{P}_{\widehat{\mathbf{V}}_{\perp}}\mathbf{P}_{\mathbf{V}}\|_{2,\infty}$. Assumption 5 states that the leading term satisfies the CLT. These assumptions are for the general framework and will be translated into case-specific conditions for concrete examples. With the above assumptions in place, we are ready to present the formal inferential results.

4.3. *Inference When $Lp \gg d$.* Recall that $\widetilde{\mathbf{V}}$ is the K leading eigenvectors of the matrix $\widetilde{\boldsymbol{\Sigma}} = \frac{1}{L}\sum_{\ell=1}^{L}\widehat{\mathbf{V}}^{(\ell)}\widehat{\mathbf{V}}^{(\ell)\top}$, and $\widetilde{\mathbf{V}}^{\mathrm{F}}$ is the K leading left singular vectors of the matrix $\widetilde{\boldsymbol{\Sigma}}^{q}\boldsymbol{\Omega}^{\mathrm{F}}$, where $\boldsymbol{\Omega}^{\mathrm{F}} \in \mathbb{R}^{d\times p'}$ is a Gaussian test matrix. We define $\mathbf{H} = \mathbf{H}_2\mathbf{H}_1\mathbf{H}_0$ to be the alignment matrix between $\widetilde{\mathbf{V}}^{\mathrm{F}}$ and $\mathbf{V}$, where $\mathbf{H}_2 = \mathrm{sgn}(\widetilde{\mathbf{V}}^{\mathrm{F}\top}\widetilde{\mathbf{V}})$, $\mathbf{H}_1 = \mathrm{sgn}(\widetilde{\mathbf{V}}^{\top}\widehat{\mathbf{V}})$ and $\mathbf{H}_0 = \mathrm{sgn}(\widehat{\mathbf{V}}^{\top}\mathbf{V}_K)$. The follow theorem provides the distributional guarantee of FADI when $Lp \gg d$.

THEOREM 4.5. *When $Lp \gg d$, under Assumptions 1 - 5, recall $\boldsymbol{\Sigma}_i = \mathrm{Cov}\left(\mathbf{V}(\mathbf{E}_0)^{\top}\mathbf{e}_i\right)$ for $i \in [d]$, and assume that there exists a statistical rate $\eta_1(d)$ such that*

$$\min_{i\in[d]}\lambda_K\left(\boldsymbol{\Sigma}_i\right) \gtrsim \eta_1(d) \quad \textit{and} \quad \eta_1(d)^{-1/2}r(d) = o(1),$$

where $r(d) := \Delta^{-1}\left(\sqrt{\frac{Kd}{pL}}r_1(d) + r_3(d)r_1(d) + \sqrt{\frac{\mu K}{d\Delta^2}}r_1(d)^2 + r_2(d) + r_4(d)\right)$. If $\Delta^{-1}r_1(d)(\log d)^2\sqrt{d/p} = o(1)$ and we take

$$q \geq 2 + \log(Ld)/\log\log d, \quad p' \geq \max(2K, K+7) \quad \textit{and} \quad p \geq \max(2K, K+8q-1),$$

we have

$$\boldsymbol{\Sigma}_i^{-1/2}(\widetilde{\mathbf{V}}^{\mathrm{F}}\mathbf{H} - \mathbf{V})^{\top}\mathbf{e}_i \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{I}_K), \quad \forall i \in [d]. \tag{8}$$

REMARK 7. Please refer to Supplementary Materials B.9 for the proof of Theorem 4.5. Here $\eta_1(d)$ guarantees that the asymptotic covariance of the leading term is positive definite, and the rate $r(d)$ bounds the remainder term stemming from fast sketching approximation and eigenspace misalignment. When the rate $\eta_1(d)$ is not too small relative to $r(d)$, Theorem 4.10 guarantees the distributional convergence of the FADI estimator. We will see in the concrete examples that the asymptotic covariance of the FADI estimator under the regime $Lp \gg d$ is the same as that of the traditional PCA estimator. In other words, we can increase the number of repeated sketches in exchange for the same testing efficiency as the traditional PCA.

We present the corollaries of Theorem 4.5 for Examples 1 to 4 as follows.

4.3.1. *Spiked Covariance Model.* Recall the set S of size K' defined in Section 3.3 for the estimation of $\widehat{\sigma}^2$. We denote by $\boldsymbol{\Sigma}_S$ the population covariance matrix corresponding to $\widehat{\boldsymbol{\Sigma}}_S$ and define $\widetilde{\sigma}_1 = \|\boldsymbol{\Sigma}_S\|_2$ to be its spectral norm. Denote by $\delta = \lambda_K(\boldsymbol{\Sigma}_S) - \sigma^2$ the eigengap of $\boldsymbol{\Sigma}_S$. We have the following corollary of Theorem 4.10 for Example 1.

COROLLARY 4.6. *Assume that $\{\boldsymbol{X}_i\}_{i=1}^{n}$ are i.i.d. multivariate Gaussian. If we take $K' = K+1$, $p' \geq \max(2K, K+7)$, $q \geq 2 + \log(Ld)/\log\log d$ and $p \geq \max(2K, K+8q-1)$, then when $Lp \gg Kdr\kappa_1^2\lambda_1/\sigma^2$, under Assumption 3 and the conditions that*

$$n \gg \max\left(\kappa_1^4(\log d)^4 r^2\lambda_1/\sigma^2, \left(\kappa_1\lambda_1/\sigma^2\right)^6\right) \quad \textit{and} \quad K \ll \min\left(\left(\widetilde{\sigma}_1/\delta\right)^{-2}\kappa_1 r, \mu^{-2/3}\kappa_1^{-4/3}d^{2/3}\right),$$

we have that (8) holds. Furthermore, we have

$$\widetilde{\boldsymbol{\Sigma}}_i^{-1/2}(\widetilde{\mathbf{V}}^{\mathrm{F}}\mathbf{H} - \mathbf{V})^{\top}\mathbf{e}_i \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{I}_K), \quad \forall i \in [d], \tag{9}$$

where $\widetilde{\boldsymbol{\Sigma}}_i = \frac{\sigma^2}{n}\boldsymbol{\Lambda}^{-1}\mathbf{V}^{\top}\boldsymbol{\Sigma}\mathbf{V}\boldsymbol{\Lambda}^{-1}$ is a simplification of $\boldsymbol{\Sigma}_i$ under Example 1. Besides, if we define $\widetilde{\boldsymbol{\Lambda}} = \widetilde{\mathbf{V}}^{\mathrm{F}\top}\widehat{\mathbf{M}}\widetilde{\mathbf{V}}^{\mathrm{F}}$ and estimate $\widetilde{\boldsymbol{\Sigma}}_i$ by $\widehat{\boldsymbol{\Sigma}}_i = \frac{1}{n}(\widehat{\sigma}^2\widetilde{\boldsymbol{\Lambda}}^{-1} + \widehat{\sigma}^4\widetilde{\boldsymbol{\Lambda}}^{-2})$, then we have

$$\widehat{\boldsymbol{\Sigma}}_i^{-1/2}(\widetilde{\mathbf{V}}^{\mathrm{F}} - \mathbf{V}\mathbf{H}^{\top})^{\top}\mathbf{e}_i \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{I}_K), \quad \forall i \in [d]. \tag{10}$$

REMARK 8. Please refer to Supplementary Materials B.10 for the proof of Corollary 4.6. The computation of $\widetilde{\mathbf{\Lambda}}$ will be performed distributively across the m data splits, and the cost for computing $\widehat{\mathbf{\Sigma}}_i$ is $O(ndK/m)$. We recommend taking $p = \lceil\sqrt{d}\rceil$, $L = \lceil\kappa_1^2 K d^{3/2}\log d\rceil$ and $q = \lceil\log d\rceil \gg 2 + \log(Ld)/\log\log d$ for optimal computational efficiency, where the total computation cost will be $O(K^3 d^{5/2}(\log d)^2)$. Our asymptotic covariance matrix is the same as that of the traditional PCA estimator under the incoherence condition [4, 29, 38]. Specifically, Wang and Fan [38] studied the asymptotic distribution of the traditional PCA estimator by assuming that the spiked eigenvalues are well-separated and diverging to infinity, which is not required by our paper. Our scaling condition is stronger than the estimation results in Corollary 4.2 to cancel out the additional randomness induced by fast sketching and allow for efficient inference.

4.3.2. *Degree-Corrected Mixed Membership Models.*

COROLLARY 4.7. *When $\theta \geq K^2 d^{-1/2+\epsilon}$ for some constant $\epsilon > 0$ and $K = o(d^{1/32})$, if we take $p \gtrsim \sqrt{d}$, $p' \geq \max(2K, K+7)$, $L \gg K^5 d^2/p$ and $q \geq 2 + \log(Ld)/\log\log d$, then (8) holds. Furthermore, we if we denote $\widetilde{\mathbf{\Sigma}}_i = \mathbf{\Lambda}^{-1}\mathbf{V}^\top \operatorname{diag}\big([\mathbf{M}_{ij}(1-\mathbf{M}_{ij})]_{j\in[d]}\big)\mathbf{V}\mathbf{\Lambda}^{-1}$, we have*

$$\widetilde{\mathbf{\Sigma}}_i^{-1/2}(\widetilde{\mathbf{V}}^{\mathrm{F}}\mathbf{H} - \mathbf{V})^\top \mathbf{e}_i \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{I}_K), \quad \forall i \in [d]. \tag{11}$$

Besides, define $\widetilde{\widetilde{\mathbf{M}}} = (\widetilde{\mathbf{V}}^{\mathrm{F}}\widetilde{\mathbf{V}}^{\mathrm{F}\top})\widehat{\mathbf{M}}(\widetilde{\mathbf{V}}^{\mathrm{F}}\widetilde{\mathbf{V}}^{\mathrm{F}\top})$ and $\widetilde{\mathbf{\Lambda}}_K = \widetilde{\mathbf{V}}^{\mathrm{F}\top}\widehat{\mathbf{M}}\widetilde{\mathbf{V}}^{\mathrm{F}}$, then if we estimate $\widetilde{\mathbf{\Sigma}}_i$ by $\widehat{\mathbf{\Sigma}}_i = \widetilde{\mathbf{\Lambda}}_K^{-1}\widetilde{\mathbf{V}}^{\mathrm{F}\top}\operatorname{diag}\big([\widetilde{\widetilde{\mathbf{M}}}_{ij}(1-\widetilde{\widetilde{\mathbf{M}}}_{ij})]_{j\in[d]}\big)\widetilde{\mathbf{V}}^{\mathrm{F}}\widetilde{\mathbf{\Lambda}}_K^{-1}$, we have

$$\widehat{\mathbf{\Sigma}}_i^{-1/2}(\widetilde{\mathbf{V}}^{\mathrm{F}} - \mathbf{V}\mathbf{H}^\top)^\top \mathbf{e}_i \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{I}_K), \quad \forall i \in [d]. \tag{12}$$

REMARK 9. The proof is deferred to Supplementary Materials B.11. The computational cost for $\widehat{\mathbf{\Sigma}}_i$ is $O(d^2K)$. To achieve the optimal computational efficiency, we would take $p = \lceil\sqrt{d}\rceil$ and $L = \lceil K^5 d^{3/2}\log d\rceil$. Hence taking $q = \lceil\log d\rceil$ is sufficient, and the total computational cost will be $O(K^7 d^{5/2}(\log d)^2)$. Inferential analyses on the membership profiles has received attention in previous works [16, 34]. Fan et al. [16] studied the asymptotic normality of the spectral estimator under the DCMM model with complicated assumptions on the eigen-structure (see Conditions 1, 3, 6, 7 in their paper). In comparison, we only impose non-singularity condition on the membership profiles, but have a stronger scaling condition on the signal strength to facilitate the divide-and-conquer process. Our asymptotic covariance is almost the same as Fan et al. [16]'s, suggesting the same level of asymptotic efficiency.

4.3.3. *Gaussian Mixure Models.* We denote by $\mu_\theta = \Delta_0^{-1}\sqrt{n/K}\|\mathbf{\Theta}^*\|_{2,\infty}$ the incoherence parameter for the Gaussian means. Then we have the following corollary of Theorem 4.5 for Example 3.

COROLLARY 4.8. *When $Lp \gg d$, If we take $p' \geq \max(2K, K+7)$, $q \geq 2 + \log(Ld)/\log\log d$ and $p \geq \max(2K, K+8q-1)$, under the conditions that*

$$K = o(d), \quad n \gg d^2, \quad K\sqrt{n}(\log d)^2 \ll \Delta_0^2 \ll \frac{n^{4/3}}{\mu_\theta^2 d} \quad \text{and} \quad L \gg \frac{Kd^2}{p},$$

we have that (8) holds. Furthermore, if we denote $\widetilde{\mathbf{\Sigma}}_i = \mathbf{\Lambda}^{-1}\mathbf{V}^\top\big\{\mathbf{F}^*\mathbf{\Theta}^{*\top}\mathbf{\Theta}^*\mathbf{F}^{*\top} + n\mathbf{I}_d\big\}\mathbf{V}\mathbf{\Lambda}^{-1}$, we have

$$\widetilde{\mathbf{\Sigma}}_i^{-1/2}(\widetilde{\mathbf{V}}^{\mathrm{F}}\mathbf{H} - \mathbf{V})^\top \mathbf{e}_i \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{I}_K), \quad \forall i \in [d]. \tag{13}$$

If we define $\widetilde{\mathbf{\Lambda}} = \widetilde{\mathbf{V}}^{\mathrm{F}\top}\widehat{\mathbf{M}}\widetilde{\mathbf{V}}^{\mathrm{F}}$ and estimate $\widetilde{\mathbf{\Sigma}}_i$ by $\widehat{\mathbf{\Sigma}}_i = \widetilde{\mathbf{\Lambda}}^{-1} + n\widetilde{\mathbf{\Lambda}}^{-2}$, we have

$$\widehat{\mathbf{\Sigma}}_i^{-1/2}(\widetilde{\mathbf{V}}^{\mathrm{F}} - \mathbf{V}\mathbf{H}^\top)^\top \mathbf{e}_i \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{I}_K), \quad \forall i \in [d]. \tag{14}$$

REMARK 10. We impose upper bound on Δ_0 to guarantee that the leading term satisfies the CLT. Similar as Example 1, computation of $\widehat{\mathbf{\Sigma}}_i$ for Example 3 is distributive and takes $O(ndK/m)$. Please refer to Supplementary Materials B.12 for the proof. We recommend taking $p = \lceil\sqrt{d}\rceil$, $L = \lceil Kd^{3/2}\log d\rceil$ and $q = \lceil \log d\rceil$, where the total complexity will be $O(K^3 d^{5/2}(\log d)^2)$. In Corollary 4.8, the scaling condition for n is $n \gg d^2$ compared to $n > d$ in Corollary 4.2, where the extra factor d is to guarantee fast enough convergence rate of the remainder term for inference. It can be verified that the Cramér-Rao lower bound for unbiased estimators of $\mathbf{V}^\top \mathbf{e}_i$ is $\mathbf{\Lambda}^{-1}$, and thus we can also see from (13) that when Δ_0 is large enough, the asymptotic efficiency of $\widetilde{\mathbf{V}}^{\mathrm{F}}$ is 1 under the regime $Lp \gg d$.

4.3.4. *Missing Matrix Inference.*

COROLLARY 4.9. *When $Lp \gg \kappa_2^2 K d^2$ and $\theta \geq d^{-1/2+\epsilon}$ for some constant $\epsilon > 0$, if we take $p' \geq \max(2K, K+7)$, $p \gtrsim \sqrt{d}$ and $q \geq 2 + \log(Ld)/\log\log d$, then under Assumption 3 and the conditions that*

$$\kappa_2^6 K^3 \mu^3 = o(d^{1/2}) \quad \text{and} \quad \sigma/\Delta \ll \sqrt{\theta/d} \cdot \min\left(\left(\kappa_2^2\sqrt{\mu K} + \kappa_2\sqrt{K\log d}\right)^{-1}, \sqrt{p/d}\right),$$

we have that (8) holds. Furthermore, if we denote $\widetilde{\mathbf{\Sigma}}_i = \mathbf{\Lambda}^{-1}\mathbf{V}^\top \operatorname{diag}\left([\mathbf{M}_{ij}^2(1-\theta)/\theta + \sigma^2/\theta]_{j=1}^d\right)\mathbf{V}\mathbf{\Lambda}^{-1}$, we have

$$\widetilde{\mathbf{\Sigma}}_i^{-1/2}(\widetilde{\mathbf{V}}^{\mathrm{F}}\mathbf{H} - \mathbf{V})^\top \mathbf{e}_i \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{I}_K), \quad \forall i \in [d]. \tag{15}$$

Moreover, define $\widetilde{\mathbf{\Lambda}} = \widetilde{\mathbf{V}}^{\mathrm{F}\top}\widehat{\mathbf{M}}\widetilde{\mathbf{V}}^{\mathrm{F}}$ and $\widetilde{\mathbf{M}} = \widetilde{\mathbf{V}}^{\mathrm{F}}\widetilde{\mathbf{\Lambda}}\widetilde{\mathbf{V}}^{\mathrm{F}\top}$. If we estimate σ^2 by $\widehat{\sigma}^2 = \sum_{(i,j)\in\mathcal{S}}(\widehat{\theta}\widehat{\mathbf{M}}_{ij} - \widetilde{\mathbf{M}}_{ij})^2/|\mathcal{S}|$ and $\widetilde{\mathbf{\Sigma}}_i$ by $\widehat{\mathbf{\Sigma}}_i = \widetilde{\mathbf{\Lambda}}^{-1}\widetilde{\mathbf{V}}^{\mathrm{F}\top}\operatorname{diag}\left([\widetilde{\mathbf{M}}_{ij}^2(1-\widehat{\theta})/\widehat{\theta} + \widehat{\sigma}^2/\widehat{\theta}]_{j=1}^d\right)\widetilde{\mathbf{V}}^{\mathrm{F}}\widetilde{\mathbf{\Lambda}}^{-1}$, we have

$$\widehat{\mathbf{\Sigma}}_i^{-1/2}(\widetilde{\mathbf{V}}^{\mathrm{F}} - \mathbf{V}\mathbf{H}^\top)^\top \mathbf{e}_i \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{I}_K), \quad \forall i \in [d]. \tag{16}$$

REMARK 11. Please see Supplementary Materials B.13 for the proof of Corollary 4.9. Computing $\widetilde{\mathbf{\Lambda}}$ and $\widetilde{\mathbf{M}}$ take $O(d^2K)$, while computing $\widehat{\theta}$ and $\widehat{\sigma}^2$ takes $O(d^2)$, and the total computational cost for calculating $\widehat{\mathbf{\Sigma}}_i$ is $O(d^2K)$. We recommend taking $p = \lceil\sqrt{d}\rceil$, $L = \lceil\kappa_2^2 K d^{3/2}\log d\rceil$ and $q = \lceil\log d\rceil$, and the total computational cost will be $O(K^3 d^{5/2}(\log d)^2)$. Chen et al. [10] studied the missing matrix inference problem through penalized optimization, and their testing efficiency is the same as ours.

For Examples 2 - 4, note that by Theorem 4.3 and the law of total probability, the inferential results still hold if we replace $\widetilde{\mathbf{V}}^{\mathrm{F}}$ by $\widetilde{\mathbf{V}}^{\mathrm{F}}_{\widehat{K}}$, where $\widehat{K}$ is the consistent estimator for K defined in Section 3.5.

4.4. *Inference When $Lp \ll d$.* Similar as in the regime of $Lp \gg d$, we first redefine the alignment matrix between $\widetilde{\mathbf{V}}^{\mathrm{F}}$ and $\mathbf{V}$ as $\mathbf{H} = \mathbf{H}_1\mathbf{H}_0$, where $\mathbf{H}_1 = \operatorname{sgn}(\widetilde{\mathbf{V}}^{\mathrm{F}\top}\widetilde{\mathbf{V}})$ and $\mathbf{H}_0 = \operatorname{sgn}(\widetilde{\mathbf{V}}^\top\mathbf{V})$. Then we have the following theorem characterizing the limiting distribution for $\widetilde{\mathbf{V}}^{\mathrm{F}}$.

THEOREM 4.10. *For the case when $Lp \ll d$, under Assumptions 1, 2, 3 and 5, for $i \in [d]$, recall $\mathbf{\Sigma}_i = \operatorname{Cov}(\mathbf{V}(\mathbf{E}_0)^\top \mathbf{e}_i|\mathbf{\Omega})$ and assume that there exists a statistical rate $\eta_2(d)$ such that*

$$\lim_{d\to\infty}\mathbb{P}_{\mathbf{\Omega}}\left(\min_{i\in[d]}\lambda_K\left(\mathbf{\Sigma}_i\right) \geq \eta_2(d)\right) = 1, \qquad \frac{d^2 r_1(d)^4(\log d)^4}{p^2\Delta^4\left(\eta_2(d)\wedge(\log d)^{-1}\right)} = o(1) \quad \text{and} \quad \frac{dr_2(d)^2}{Lp\Delta^2\eta_2(d)} = o(1).$$

Then if we take $K(\log d)^2 \ll p \asymp p' \lesssim d/(\log d)^2$ and $q \geq \log d$ we have

$$\mathbf{\Sigma}_i^{-1/2}(\widetilde{\mathbf{V}}^{\mathrm{F}}\mathbf{H} - \mathbf{V})^\top \mathbf{e}_i \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{I}_K), \quad \forall i \in [d]. \tag{17}$$

REMARK 12. Theorem 4.10 states that with positive definiteness of the asymptotic covariance matrix characterized by $\eta_2(d)$ and proper scaling conditions on the dimension d, the FADI estimator still enjoys asymptotic normality even when the aggregated dimension of fast sketches Lp is much smaller than d. The rate $\eta_2(d)$ is usually at least of order $(d/\lambda_1^2 Lp)\lambda_{\min}(\mathrm{Cov}(\mathbf{E}_0\mathbf{e}_i))$. The proof is deferred to Supplementary Materials B.6.

The following corollaries of Theorem 4.10 provide case-specific distributional guarantee for Examples 1 and 3 under the regime $Lp \ll d$.

4.4.1. *Spiked Covariance Model.*

COROLLARY 4.11. *Assume that $\{\boldsymbol{X}_i\}_{i=1}^n$ are i.i.d. multivariate Gaussian. When $Lp \ll \lambda_1^{-2}\Delta^2 d$, if we take $K' = K+1$, $K(\log d)^2 \ll p \asymp p' \lesssim d/(\log d)^2$ and $q \geq \log d$, under Assumption 3 and the conditions that*

$$n \gg \max\left(\frac{\kappa_1^4\lambda_1^2 dr^2 L}{p\sigma^4}, \frac{\lambda_1^2\widetilde{\sigma}_1^6 K^2}{\Delta^2\delta^4\sigma^4}\right)(\log d)^4 \quad \textit{and} \quad \frac{K\lambda_1^2}{\Delta^2}\sqrt{\frac{\mu}{d}} = o(1),$$

we have that (17) holds. Furthermore, if we define $\widetilde{\mathbf{\Sigma}}_i = \frac{\sigma^2}{nL^2}\mathbf{B}_{\mathbf{\Omega}}^\top\mathbf{\Omega}^\top\mathbf{\Sigma}\mathbf{\Omega}\mathbf{B}_{\mathbf{\Omega}}$, we have

$$\widetilde{\mathbf{\Sigma}}_i^{-1/2}(\widetilde{\mathbf{V}}^{\mathrm{F}}\mathbf{H} - \mathbf{V})^\top\mathbf{e}_i \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{I}_K), \quad \forall i \in [d]. \tag{18}$$

Besides, if we further assume $\sigma^{-2}\lambda_1\kappa_1^4\sqrt{d^2r/(np^2L)} = o(1)$ and estimate $\widetilde{\mathbf{\Sigma}}_i$ by $\widehat{\mathbf{\Sigma}}_i = \frac{\widehat{\sigma}^2}{nL^2}\widehat{\mathbf{B}}_{\mathbf{\Omega}}^\top\mathbf{\Omega}^\top\widehat{\mathbf{\Sigma}}\mathbf{\Omega}\widehat{\mathbf{B}}_{\mathbf{\Omega}}$, where $\widehat{\mathbf{B}}_{\mathbf{\Omega}} = (\widehat{\mathbf{B}}^{(1)\top}, \ldots, \widehat{\mathbf{B}}^{(L)\top})^\top$ with $\widehat{\mathbf{B}}^{(\ell)} = (\widetilde{\mathbf{V}}^{\mathrm{F}\top}\widehat{\mathbf{Y}}^{(\ell)}/\sqrt{p})^\dagger$ for $\ell \in [L]$, we have

$$\widehat{\mathbf{\Sigma}}_i^{-1/2}(\widetilde{\mathbf{V}}^{\mathrm{F}} - \mathbf{V}\mathbf{H}^\top)^\top\mathbf{e}_i \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{I}_K), \quad \forall i \in [d]. \tag{19}$$

REMARK 13. Please refer to Supplementary Materials B.7 for the proof of Corollary 4.11. For the computation of $\widehat{\mathbf{\Sigma}}_i$, apart from $\widehat{\mathbf{V}}^{(\ell)}$, the ℓ-th machine on layer 2 (see Figure 2) will send $\mathbf{\Omega}^{(\ell)}$ and $\widehat{\mathbf{Y}}^{(\ell)}$ to the central processor, and the total communication cost for each server is $O(dp)$. On the central processor, the total computational cost of $\mathbf{B}_{\mathbf{\Omega}}$ will be $O(dpKL)$. Then we will compute $\mathbf{\Omega}^\top\widehat{\mathbf{\Sigma}}\mathbf{\Omega} = \frac{1}{\sqrt{p}}\mathbf{\Omega}^\top(\widehat{\mathbf{Y}}^{(1)}, \ldots, \widehat{\mathbf{Y}}^{(L)}) + \widehat{\sigma}^2\mathbf{\Omega}^\top\mathbf{\Omega}$ with total computational cost of $O\big(d(Lp)^2\big) = o(d^3)$. Compared to Corollary 4.6 under the regime $Lp \gg d$, Corollary 4.11 has stronger scaling conditions on the sample size n to compensate for the extra variability due to less fast sketches. As indicated by (7), the asymptotic covariance matrix of Corollary 4.12 is consistent with Corollary 4.8.

4.4.2. *Gaussian Mixture Models.*

COROLLARY 4.12. *When $Lp \ll d$, if we take $K(\log d)^2 \ll p \asymp p' \lesssim d/(\log d)^2$ and $q \geq \log d$, we have that (17) holds under the conditions that*

$$\sqrt{\frac{K}{d}}\log d = O(1), \quad n \gg \frac{d^3 L}{p}, \quad \textit{and} \quad K(\log d)^2\sqrt{\frac{dnL}{p}} \ll \Delta_0^2 \ll \min\left(n, \frac{n^{4/3}}{\mu_\theta^2 d}\right).$$

Furthermore, if we define $\widetilde{\mathbf{\Sigma}}_i = L^{-2}\mathbf{B}_{\mathbf{\Omega}}^\top\mathbf{\Omega}^\top\Big(\mathbf{F}^*\mathbf{\Theta}^{*\top}\mathbf{\Theta}^*\mathbf{F}^{*\top} + n\mathbf{I}_d\Big)\mathbf{\Omega}\mathbf{B}_{\mathbf{\Omega}}$, then we have

$$\widetilde{\mathbf{\Sigma}}_i^{-1/2}(\widetilde{\mathbf{V}}^{\mathrm{F}}\mathbf{H} - \mathbf{V})^\top\mathbf{e}_i \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{I}_K), \quad \forall i \in [d]. \tag{20}$$

Besides, if we further assume $\Delta_0^2 \ll KLp^2n^2/d^4$ and estimate $\widetilde{\mathbf{\Sigma}}_i$ by $\widehat{\mathbf{\Sigma}}_i = \frac{1}{L^2}\widehat{\mathbf{B}}_{\mathbf{\Omega}}^\top\mathbf{\Omega}^\top\Big(\widehat{\mathbf{M}} + n\mathbf{I}_d\Big)\mathbf{\Omega}\widehat{\mathbf{B}}_{\mathbf{\Omega}}$, where $\widehat{\mathbf{B}}_{\mathbf{\Omega}} = (\widehat{\mathbf{B}}^{(1)\top}, \ldots, \widehat{\mathbf{B}}^{(L)\top})^\top$ with $\widehat{\mathbf{B}}^{(\ell)} = (\widetilde{\mathbf{V}}^{\mathrm{F}\top}\widehat{\mathbf{Y}}^{(\ell)}/\sqrt{p})^\dagger$ for $\ell \in [L]$, we have

$$\widehat{\mathbf{\Sigma}}_i^{-1/2}(\widetilde{\mathbf{V}}^{\mathrm{F}} - \mathbf{V}\mathbf{H}^\top)^\top\mathbf{e}_i \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{I}_K), \quad \forall i \in [d]. \tag{21}$$

REMARK 14. The proof of Corollary 4.12 is deferred to Supplementary Materials B.8. Computation of $\widehat{\boldsymbol{\Sigma}}_i$ is very similar to Example 1 as described in Remark 13, and the total computational cost is $O(d(Lp)^2) = o(d^3)$. The stronger scaling conditions are the trade-off for higher computational efficiency with less fast sketches.

We do not have distributional results for Examples 2 and 4 under the regime $Lp \ll d$. An intuitive explanation would be that the information contained in each entry is independent for Example 2 and Example 4, and when $Lp \ll d$, too much information will be lost from the $d \times d$ graph or matrix. In comparison, we can still recover information from Examples 1 and 3 under the regime $Lp \ll d$ due to the correlation structure of the matrix.

5. Numerical Results. We conduct extensive simulation studies to assess the performance of FADI under each example given in Section 2 and compare it with several existing methods in terms of the error rates and computational costs. We evaluate the distributional convergence of the FADI estimator by computing the empirical coverage probabilities of the FADI confidence interval and comparing it with standard normal distribution by Q-Q plots. We provide in this section the representative results for Example 1 and Example 2. The results for Examples 3 and 4 and some additional simulations for Example 1 are given in Supplementary Materials A.

5.1. *Example 1: Spiked Covariance Model.* We generate $\{\boldsymbol{X}_i\}_{i=1}^n$ i.i.d. from $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^\top + \sigma^2\mathbf{I}_d$. We consider $K = 3$, fix the sample size at $n = 20000$ and set $d = 500, 1000, 2000$ respectively to study the asymptotic properties of the FADI estimator under different settings. To ensure the incoherence condition is satisfied, we set $\mathbf{V}$ to be the left singular vectors of a $d \times K$ i.i.d. Gaussian matrix. We take $\boldsymbol{\Lambda} = \mathrm{diag}(6, 4, 2)$ and $\sigma^2 = 1$. For the estimation of σ^2 in Step 0, we set $K' = 6$. We split the data into $m = 20$ subsamples, set $p = p' = 12$ and $q = 7$ in Step 3 to compute $\widetilde{\mathbf{V}}^{\mathrm{F}}$. We set L at a range of values by taking the ratio $Lp/d \in \{0.2, 0.6, 0.9, 1, 1.2, 2, 5, 10\}$ for each setting and compute the asymptotic covariance via Corollary 4.6 and Corollary 4.11 under different regimes of Lp. We define $\widetilde{\mathbf{v}} = \widehat{\boldsymbol{\Sigma}}_1^{-1/2}(\widetilde{\mathbf{V}}^{\mathrm{F}} - \mathbf{V}\mathbf{H}^\top)^\top \mathbf{e}_1$ where $\widehat{\boldsymbol{\Sigma}}_1$ is the asymptotic covariance for the first row of $\widetilde{\mathbf{V}}^{\mathrm{F}}$ and $\mathbf{H} = \mathrm{sgn}(\widetilde{\mathbf{V}}^{\mathrm{F}\top}\mathbf{V})$, and calculate the empirical coverage probability by empirically evaluating $\mathbb{P}\big(\|\widetilde{\mathbf{v}}\|_2^2 \le \chi_3^2(0.95)\big)$ with $\chi_3^2(0.95)$ being the 0.95 quantile of the Chi-squared distribution with degrees of freedom equal to 3. Results under different settings are shown in Figure 3 (with 300 Monte Carlo simulations). Figure 3 (a) shows that as Lp/d increases, the error rate of FADI estimator converges to that of the traditional PCA estimator. From Figure 3 (b) we can see that when Lp/d is approaching 1 from the left, the computational efficiency drops due to the cost of computing $\widehat{\boldsymbol{\Sigma}}_i$. For Figure 3 (c), convergence towards the nominal 95% level can be observed when Lp/d is much smaller or much larger than 1, while the valley at Lp/d around 1 is consistent with the theoretical conditions on Lp/d in Section 4 and implies a possible phase-transition phenomenon on the distributional convergence of FADI. Note that the empirical coverage is closer to the nominal level 0.95 at $d = 2000$ than at $d \in \{500, 1000\}$, which might be caused by the vanishing of some error terms for approximation of the asymptotic covariance matrix as d grows larger (as can be seen in the proof of Corollary 4.6). The good Gaussian approximation of $\widetilde{\mathbf{v}}_1$ is further validated by Figure 3 (d). Based upon the low computational efficiency and poor empirical coverage at Lp/d around 1, we do not recommend the regime $Lp \asymp d$ for conducting inference.

We also compare FADI with Fan et al. [15]'s distributed PCA method, where we mimic the setting of Fan et al. [15] and take $\boldsymbol{\Sigma} = \mathrm{diag}(\lambda, \lambda/2, \lambda/4, 1, \ldots, 1)$. Results over 100 Monte Carlo simulations are given in Table 4. We can see that FADI outperforms both Fan et al. [15]'s method and the traditional PCA under the distributed setting in terms of computational efficiency while maintaining the same level of error rates.

5.2. *Example 2: Degree-Corrected Mixed Membership Models.* We consider the mixed membership model without degree heterogeneity for the simulation. In other words, we take $\boldsymbol{\Theta} = \sqrt{\theta}\mathbf{I}_d$, and $\mathbf{M} =$

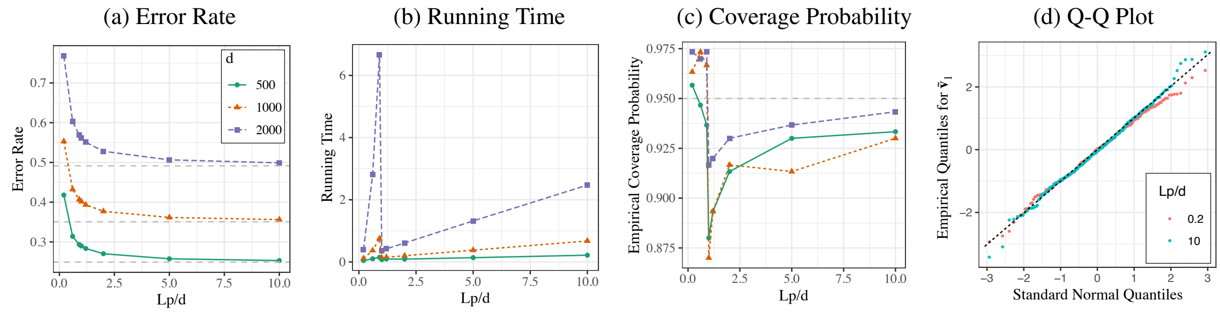


FIGURE 3. *Performance of FADI under different settings for Example 1. (a) Empirical error rate of $\rho(\widetilde{\mathbf{V}}^{\mathrm{F}},\mathbf{V})$, where the grey dashed lines represent the error rate for the traditional PCA estimator $\widehat{\mathbf{V}}$; (b) Running time (in seconds) under different settings (including computation time of $\widehat{\boldsymbol{\Sigma}}_i$). For the traditional PCA, the running time is 4.86 seconds at $d=500$, 20.95 seconds at $d=1000$ and 99.23 seconds at $d=2000$; (c) Empirical coverage probability, where the grey dashed line represents the theoretical rate at 0.95; (d) Q-Q plot for $\widetilde{\mathbf{v}}_1$ at $Lp/d\in\{0.2,10\}$;*

TABLE 4

Comparison of the empirical error rates (of $\rho(\cdot,\mathbf{V})$) and the running times (in seconds) between FADI, traditional full sample PCA and distributed PCA [15] under different settings of d,n and m. 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 σ^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 $\widetilde{\mathbf{V}}^{\mathrm{F}}$ from the fast sketching $\widetilde{\boldsymbol{\Sigma}}^q\boldsymbol{\Omega}^{\mathrm{F}}$.

Parameters				Error rate			Running time (seconds)		
d	n	m	L	FADI	Traditional	Distributed	FADI	Traditional	Distributed
400	30000	15	40	0.068	0.065	0.065	0.07	4.53	0.59
400	60000	30	40	0.048	0.046	0.046	0.05	8.84	0.60
400	100000	50	40	0.037	0.036	0.036	0.05	14.84	0.62
800	100000	50	80	0.052	0.050	0.050	0.10	55.76	3.66
800	5000	50	80	0.230	0.220	0.230	0.05	3.76	2.56
800	25000	50	80	0.106	0.103	0.103	0.07	15.07	2.82
800	50000	50	80	0.073	0.070	0.070	0.07	28.68	3.23
1600	30000	15	160	0.134	0.130	0.130	0.31	80.72	27.02
1600	60000	30	160	0.095	0.092	0.092	0.35	150.75	27.29
1600	100000	50	160	0.074	0.071	0.071	0.34	243.83	27.38

$\theta\boldsymbol{\Pi}\mathbf{P}\boldsymbol{\Pi}^\top$. For two preselected nodes $i,j\in[d]$, we test the hypothesis $\mathrm{H}_0:\boldsymbol{\pi}_i=\boldsymbol{\pi}_j$ vs. $\mathrm{H}_1:\boldsymbol{\pi}_i\neq\boldsymbol{\pi}_j$ by testing whether $\boldsymbol{\pi}_i-\boldsymbol{\pi}_j=0$, which is equivalent to studying whether $\mathbf{V}^\top(\mathbf{e}_i-\mathbf{e}_j)=0$. With modification of Corollary 4.7 involving elementary algebra we can show that

$$\widetilde{\boldsymbol{\Sigma}}_{i,j}^{-1/2}(\widetilde{\mathbf{V}}^{\mathrm{F}}\mathbf{H}-\mathbf{V})^\top(\mathbf{e}_i-\mathbf{e}_j)\xrightarrow{d}\mathcal{N}(\mathbf{0},\mathbf{I}_K), \tag{22}$$

where the asymptotic covariance is defined as $\widetilde{\boldsymbol{\Sigma}}_{i,j}=\widetilde{\boldsymbol{\Sigma}}_i+\widetilde{\boldsymbol{\Sigma}}_j$[8] and can be consistently estimated by plugging in $\widetilde{\mathbf{V}}^{\mathrm{F}}$, $\widetilde{\mathbf{M}}$ and $\widehat{\boldsymbol{\Lambda}}$ defined in Corollary 4.7.

[8]It can be verified that the correlation between two rows of $\widetilde{\mathbf{V}}^{\mathrm{F}}$ is negligible due to the incoherence of $\mathbf{V}$.

In our simulation, we set $\theta = 0.9$, and consider $K = 3$ communities. We set the membership profiles $\mathbf{\Pi}$ and the connection probability matrix $\mathbf{P}$ to be

$$\boldsymbol{\pi}_i = \begin{cases} (1,0,0)^\top & \text{if } 1 \le i \le \lfloor d/6 \rfloor \\ (0,1,0)^\top & \text{if } \lfloor d/6 \rfloor < i \le \lfloor d/3 \rfloor \\ (0,0,1)^\top & \text{if } \lfloor d/3 \rfloor < i \le \lfloor d/2 \rfloor \\ (0.6,0.2,0.2)^\top & \text{if } \lfloor d/2 \rfloor < i \le \lfloor 5d/8 \rfloor \\ (0.2,0.6,0.2)^\top & \text{if } \lfloor 5d/8 \rfloor < i \le \lfloor 3d/4 \rfloor \\ (0.2,0.2,0.6)^\top & \text{if } \lfloor 3d/4 \rfloor < i \le \lfloor 7d/8 \rfloor \\ (1/3,1/3,1/3)^\top & \text{if } \lfloor 7d/8 \rfloor < i \le \lfloor d \rfloor \end{cases}, \quad \mathbf{P} = \begin{pmatrix} 1 & 0.2 & 0.1 \\ 0.2 & 1 & 0.2 \\ 0.1 & 0.2 & 1 \end{pmatrix}.$$

Then we test the performance of FADI under $d \in \{500, 1000, 2000\}$ respectively, and under each setting of d, we take $p = p' = 12$, $q = 7$ and set the number of repeated fast sketches by the ratio $Lp/d \in \{0.2, 0.6, 0.9, 1, 1.2, 2, 5, 10\}$. For each setting, we conduct 300 independent Monte Carlo simulations. We first preselect two nodes, which we denote by i and j, with membership profiles both equal to $(0.6, 0.2, 0.2)^\top$ and calculate the empirical coverage probability of $\mathbb{P}\big(\|\widetilde{\boldsymbol{d}}\|_2^2 \le \chi_3^2(0.95)\big)$, where $\widetilde{\boldsymbol{d}} = \widehat{\boldsymbol{\Sigma}}_{i,j}^{-1/2}\widetilde{\mathbf{V}}^{\mathrm{F}\top}(\mathbf{e}_i - \mathbf{e}_j)$ with $\widehat{\boldsymbol{\Sigma}}_{i,j}$ being the estimator for $\widetilde{\boldsymbol{\Sigma}}_{i,j}$. We also evaluate the power of the test by choosing two nodes with different membership profiles equal to $(0.6, 0.2, 0.2)^\top$ and $(1/3, 1/3, 1/3)^\top$ respectively, which we denote by i and k. We empirically calculate the power $\mathbb{P}\big(\chi_3^2(0.95) \le \|\widetilde{\boldsymbol{d}}'\|_2^2\big)$, where $\widetilde{\boldsymbol{d}}' = \widehat{\boldsymbol{\Sigma}}_{i,k}^{-1/2}\widetilde{\mathbf{V}}^{\mathrm{F}\top}(\mathbf{e}_i - \mathbf{e}_k)$. Under the regime $Lp/d < 1$, we do not have theoretical results for the asymptotic normality and we calculate the asymptotic covariance referring to Theorem 4.10 by

$$\widehat{\boldsymbol{\Sigma}}_{i,j} = L^{-2}\widehat{\mathbf{B}}_{\mathbf{\Omega}}^\top \mathbf{\Omega}\top \operatorname{diag}\left([\widetilde{\mathbf{M}}_{ik}(1-\widetilde{\mathbf{M}}_{ik}) + \widetilde{\mathbf{M}}_{jk}(1-\widetilde{\mathbf{M}}_{jk})]_{k=1}^d\right)\mathbf{\Omega}\widehat{\mathbf{B}}_{\mathbf{\Omega}},$$

where $\widehat{\mathbf{B}}_{\mathbf{\Omega}} = (\widehat{\mathbf{B}}^{(1)\top}, \ldots, \widehat{\mathbf{B}}^{(L)\top})^\top$ with $\widehat{\mathbf{B}}^{(\ell)} = (\widetilde{\mathbf{V}}^{\mathrm{F}\top}\widehat{\mathbf{Y}}^{(\ell)}/\sqrt{p})^\dagger \in \mathbb{R}^{p\times K}$ for $\ell = 1, \ldots, L$. We also apply k-means to $\widetilde{\mathbf{V}}^{\mathrm{F}}$ to differentiate different membership profiles and compare the misclustering rate with the traditional PCA estimator. The results of different settings are shown in Figure 4. We can see from Figure 4 (d) that under the regime $Lp/d < 1$ the empirical coverage probability is zero under all settings, which validates the necessity of $Lp/d \gg 1$ for performance guarantee. Figure 4 (f) demonstrates the asymptotic normality of $\widetilde{\boldsymbol{d}}_1$ at $Lp/d = 10$, whereas $\widetilde{\boldsymbol{d}}_1$ is poorly approximated by the standard Gaussian at $Lp/d = 0.2$.

We also compare with Fan et al. [16]'s method (SIMPLE) on the inference of the membership profiles under DCMM, who conducted inference directly on the traditional PCA estimator $\widehat{\mathbf{V}}$ and adopted a one-step correction to the empirical eigenvalues for calculation of the asymptotic covariance matrix. We compared the inferential performance of FADI at $Lp/d = 10$ with their method (under 100 independent Monte Carlos) and summarize the results in Table 5, where the running time includes both the PCA procedure and the computing time of $\widehat{\boldsymbol{\Sigma}}_{i,j}$. Compared to Fan et al. [16]'s method, our method has similar coverage probability and power but is much more efficient both in conducting the PCA and in computing the covariance matrix.

TABLE 5

Comparison of the empirical coverage probability, power and the running times (in seconds) between FADI and SIMPLE [16] under different settings of d. In all settings, we take $p = p' = 12$, $q = 7$ and set $Lp/d = 10$.

Parameters			Coverage probability		Power		Running time (seconds)	
d	p	L	FADI	SIMPLE	FADI	SIMPLE	FADI	SIMPLE
500	12	417	0.91	0.92	0.87	0.88	0.21	0.73
1000	12	833	0.94	0.94	1.00	1.00	0.69	6.77
2000	12	1667	0.95	0.98	1.00	1.00	2.61	59.42

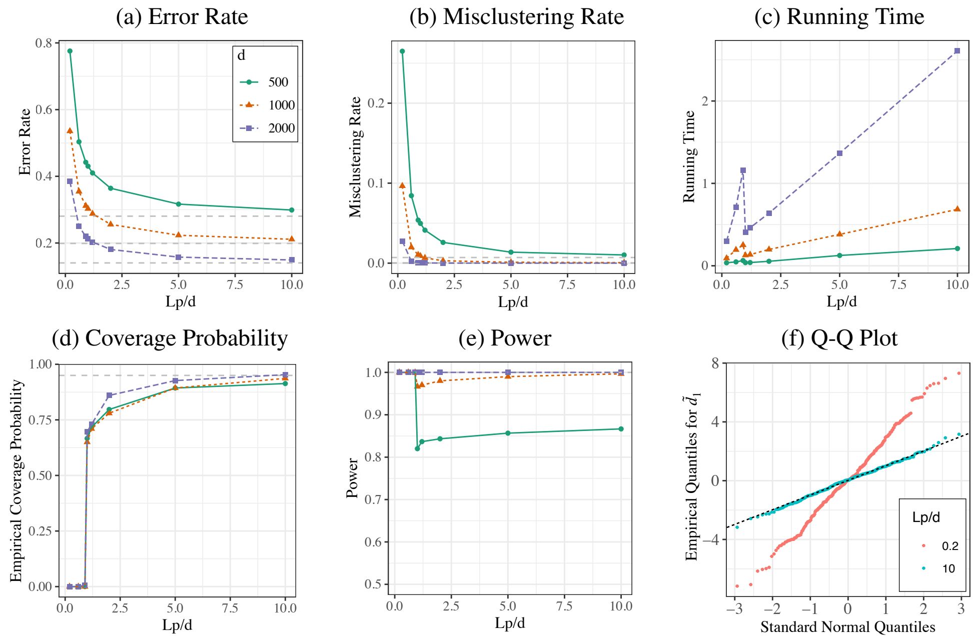
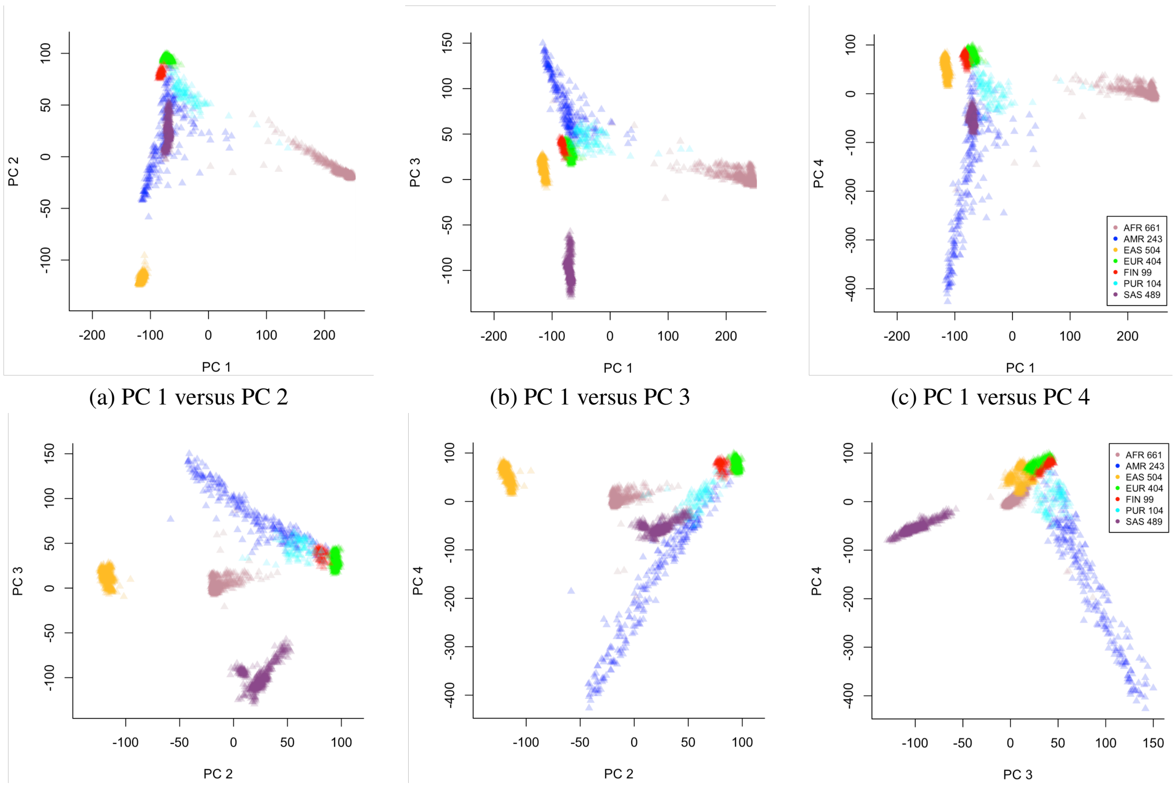


FIGURE 4. *Performance of FADI under different settings for Example 2. (a) Empirical error rate of* $\rho(\widetilde{\mathbf{V}}^{\mathrm{F}}, \mathbf{V})$*; (b) Misclustering rate for* $\widetilde{\mathbf{V}}^{\mathrm{F}}$ *by K-means with grey dashed lines representing the misclustering rate for the traditional PCA estimator* $\widehat{\mathbf{V}}$*; (c) Running time (in seconds) under different settings (including computing* $\widehat{\boldsymbol{\Sigma}}_{i,j}$*). For the traditional PCA, the running time is 0.43 seconds at* $d = 500$*, 3.77 seconds at* $d = 1000$ *and 32.62 seconds at* $d = 2000$*; (d) Empirical coverage probability (1− Type I error); (e) Power of the test; (f) Q-Q plot for* $\widetilde{\boldsymbol{d}}_1$ *at* $Lp/d \in \{0.2, 10\}$*.*

6. Application to the 1000 Genomes Data. In this section, we apply FADI and other existing methods to the 1000 Genomes Data [12]. The 1000 Genomes Project performs whole-genome sequencing of a large number of individuals from diverse populations and aims to establish a comprehensive public catalog of human genetic variants [12]. 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 [31]. There are 2504 subjects in total, and 168,047 independent variants after the LD pruning. As we are interested in the 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 ($d = 2504$) [30]. 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.

6.1. *Estimation of Principal Eigenspace.* For the estimation of the principal components, we assume that the data follows the spiked covariance model specified in Example 1. 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 $\mu_0 = \big(d(np)^{-1/2}\log d\big)^{3/4}/12$. 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 5, where a clear separation can be observed among different super-populations. To estimate the non-spiked eigenvalues/residual variance σ^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 $\mathrm{mean}(\widehat{\sigma}^2) = 0.7804$ with the standard deviation $\mathrm{SD}(\widehat{\sigma}^2) = 0.018$. Thus the estimation is stable with respect to different choices of the samples.



(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 5. *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.*

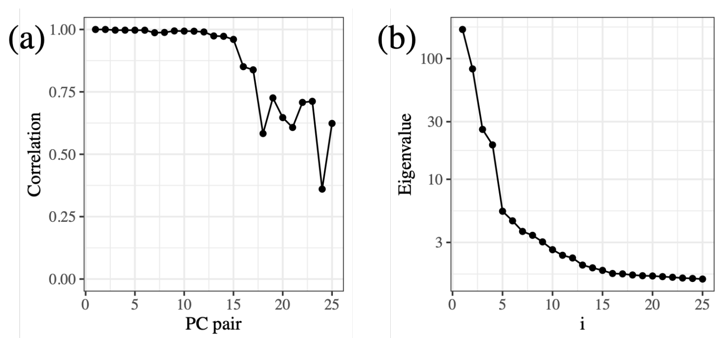


FIGURE 6. *(a) Correlations between the 25 leading PCs calculated by FADI and by full sample PCA on the 1000 Genomes Data; (b) Top 25 eigenvalues for the sample covariance matrix of the 1000 Genomes Data.*

Figure 6 (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 get smaller after the 15th eigenvalue (see Figure 6 (b)).

Figure 11 in the supplementary materials 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 and the distributed PCA method [15] takes 120.2 seconds. These results show that FADI greatly outperforms the existing PCA methods in terms of computational time.

6.2. *Inference on Ancestry Membership Profiles.* We also generate an undirected graph from the 1000 Genomes Data. To increase the randomness for better fitting of the model setting in Example 2, we sample 1000 out of the total 168047 variants for generating the graph. More specifically, we treat each subject as a node, and for each given pair of subjects (i, j), we define a genetic similarity score $s_{ij} = \sum_{k=1}^{1000} \mathbb{I}\{x_{ik} = x_{jk}\}$, where x_{ik} refers to the genotype of the k-th variant for subject i. We denote by $s^{0.95}$ the 0.95 quantile of $\{s_{ij}\}_{i<j}$. Subject i and j are connected if and only if $s_{ij} > s^{0.95}$. Denote by $\mathbf{A}$ the adjacency matrix (allowing no self-loops). We include only four super populations: AFR, EAS, EUR and SAS, with 2058 subjects in total. We are interested in testing whether two given subjects i and j belong to the same super population, i.e.,

$$\mathrm{H}_0 : \mathbf{V}_i = \mathbf{V}_j \quad \text{versus} \quad \mathrm{H}_1 : \mathbf{V}_i \neq \mathbf{V}_j.$$

We perform FADI with $p = 50$, $p' = 50$, $q = 3$ and $L = 1000$. The rank estimator from FADI is $\widehat{K} = 4$ by setting $\mu_0 = (\widehat{\theta}/p)^{1/2} d \log d/12$, where $\widehat{\theta}$ is the average degree estimator defined in Section 3.3. We can see the estimated rank is consistent with the number of super populations. We apply K-means clustering to the FADI estimator $\widetilde{\mathbf{V}}^{\mathrm{F}}_{\widehat{K}}$, and calculate the misclustering rate by treating the self-reported ancestry group as the ground truth. The misclustering rate of FADI is 0.135, with computation time of 5.8 seconds. In comparison, we also perform the traditional PCA and apply K-means to the PC estimator. The misclustering rate for the traditional PCA method is 0.134 with computation time of 26.5 seconds, and the correlation between the top four PCs for the traditional PCA and FADI PCA are 0.997, 0.994, 0.994 and 0.996 respectively. To conduct pairwise inference on the ancestry membership profiles, we preselect 16 subjects, with 4 subjects from each super population. We apply Bonferroni correction to correct for the multiple comparison issue and set the level at $0.05 \times \binom{16}{2}^{-1} = 4.17 \times 10^{-4}$. We estimate the asymptotic covariance matrix by Corollary 4.7 and correct $\widetilde{\mathbf{M}}$ by setting entries larger than 1 to 1 and entries smaller than 0 to 0. The pairwise p-values are summarized in Figure 7. The computational time for computing the covariance matrix is 0.31 seconds. We can see that most of the comparison results are consistent with the true ancestry groups, while the inconsistency could be due to the mixed memberships of certain subjects and the unaccounted sub-population structures.

7. Discussion. In this paper, we develop a FAst DIstributed-friendly PCA algorithm FADI that can deal with high-dimensional PC calculations with low computational cost and high accuracy. The algorithm is applicable to multiple statistical models and is friendly for distributed computing. The main idea is to apply distributed-friendly random sketches so as to reduce the data dimension, and aggregate the results from multiple sketches to improve the statistical accuracy and accommodate federated data. We conduct theoretical analysis as well as simulation studies to demonstrate that FADI enjoys the same non-asymptotic error rate as the traditional full sample PCA while significantly reducing the computational time compared to existing methods. We also establish distributional guarantee for the FADI estimator and perform numerical experiments to validate the potential phase-transition phenomenon in distributional convergence.

Fast PCA algorithms using random sketches usually require the data to have certain "almost low-rank" structures, without which the approximation might not be accurate [18]. It is of future research interest to investigate whether the proposed FADI approach can be extended to non-low-rank settings. In Step 3 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 averages and investigate the best convex combination to reduce the statistical error.

	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*
1	1.0000	0.1290	0.5478	0.0169	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
2	0.1290	1.0000	0.8137	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
3	0.5478	0.8137	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
4	0.0169	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
5	0.0000	0.0000	0.0000	0.0000	1.0000	0.2621	0.1150	0.0129	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
6	0.0000	0.0000	0.0000	0.0000	0.2621	1.0000	0.7227	0.0158	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
7	0.0000	0.0000	0.0000	0.0000	0.1150	0.7227	1.0000	0.0385	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
8	0.0000	0.0000	0.0000	0.0000	0.0129	0.0158	0.0385	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
9	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000	0.2774	0.0067	0.0102	0.0000	0.0000	0.0000	0.0000
10	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.2774	1.0000	0.0130	0.0253	0.0000	0.0000	0.0000	0.0000
11	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0067	0.0130	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000
12	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0102	0.0253	0.0000	1.0000	0.0000	0.0000	0.0000	0.0000
13	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000	0.0855	0.0194	0.0014
14	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0855	1.0000	0.5623	0.0000
15	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0194	0.5623	1.0000	0.0000
16	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0014	0.0000	0.0000	1.0000
Node Degrees	*31*	*11*	*26*	*52*	*376*	*377*	*367*	*391*	*90*	*77*	*123*	*48*	*38*	*20*	*20*	*80*

AFR ***EAS*** ***EUR*** ***SAS***

FIGURE 7. *p-values for pairwise comparison among 16 preselected subjects. For subjects pair* (i,j)*, p-value is defined as* $\mathbb{P}(\chi^2_{\widehat{K}} > \|\widetilde{\boldsymbol{d}}\|_2^2)$*, where* $\chi^2_{\widehat{K}}$ *is Chi-squared distribution with degrees of freedom equal to* $\widehat{K}$*, and* $\widetilde{\boldsymbol{d}} = \widehat{\boldsymbol{\Sigma}}_{i,j}^{-1/2}\widetilde{\mathbf{V}}^{\mathrm{F}}_{\widehat{K}}(\mathbf{e}_i - \mathbf{e}_j)$ *with* $\widehat{\boldsymbol{\Sigma}}_{i,j}$ *being the asymptotic covariance matrix defined in Section 5.2.*

REFERENCES

[1] ABBE, E. (2018). Community detection and stochastic block models: Recent developments. *Journal of Machine Learning Research* **18** 1–86.

[2] ABBE, E., FAN, J., WANG, K. and ZHONG, Y. (2020). Entrywise eigenvector analysis of random matrices with low expected rank. *Annals of Statistics* **48** 1452–1474. https://doi.org/10.1214/19-AOS1854 MR4124330

[3] ACHLIOPTAS, D. (2003). Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *Journal of Computer and System Sciences* **66** 671–687. Special issue on PODS 2001 (Santa Barbara, CA). https://doi.org/10.1016/S0022-0000(03)00025-4 MR2005771

[4] ANDERSON, T. W. (1963). Asymptotic theory for principal component analysis. *Annals of Mathematical Statistics* **34** 122–148. https://doi.org/10.1214/aoms/1177704248 MR145620

[5] BAIK, J., AROUS, G. B. and PÉCHÉ, S. (2005). Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *Annals of Probability* **33** 1643–1697.

[6] BANNA, M., MERLEVÈDE, F. and YOUSSEF, P. (2016). Bernstein-type inequality for a class of dependent random matrices. *Random Matrices: Theory and Applications* **05** 1650006. https://doi.org/10.1142/s2010326316500064

[7] BERNSTEIN, S. (1924). On a modification of Chebyshev's inequality and of the error formula of Laplace. *Annals of Science Institute SAV. Ukraine, Sect. Math* **I**.

[8] CANDÈS, E. J. and TAO, T. (2010). The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory* **56** 2053–2080.

[9] CHEN, T., CHANG, D. D., HUANG, S., CHEN, H., LIN, C. and WANG, W. (2016). Integrating multiple random sketches for singular value decomposition. *arXiv preprint arXiv:1608.08285*.

[10] CHEN, Y., FAN, J., MA, C. and YAN, Y. (2019). Inference and uncertainty quantification for noisy matrix completion. *Proceedings of the National Academy of Sciences* **116** 22931–22937. https://doi.org/10.1073/pnas.1910053116

[11] CHEN, Y., CHI, Y., FAN, J. and MA, C. (2021). Spectral methods for data science: A statistical perspective. *Foundations and Trends® in Machine Learning* **14** 566–806.

[12] 1000 GENOMES PROJECT CONSORTIUM (2015). A global reference for human genetic variation. *Nature* **526** 68.

[13] 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** 60. https://doi.org/10.1038/s41746-020-0265-z

[14] 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** 603–680. https://doi.org/10.1111/rssb.12016 MR3091653

[15] FAN, J., WANG, D., WANG, K. and ZHU, Z. (2019). Distributed estimation of principal eigenspaces. *Annals of Statistics* **47** 3009–3031. https://doi.org/10.1214/18-AOS1713 MR4025733

[16] FAN, J., FAN, Y., HAN, X. and LV, J. (2022). SIMPLE: Statistical inference on membership profiles in large networks. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **84** 630-653. https://doi.org/10.1111/rssb.12505

[17] FRANKLIN, J. N. (2012). *Matrix theory*. Courier Corporation.

[18] HALKO, N., MARTINSSON, P. G. and TROPP, J. A. (2011). Finding structure with randomness: probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review* **53** 217–288. https://doi.org/10.1137/090771806 MR2806637

[19] HOEFFDING, W. (1994). Probability inequalities for sums of bounded random variables. In *The collected works of Wassily Hoeffding* 409–426. Springer.

[20] JENSEN, J. L. W. V. (1906). Sur les fonctions convexes et les inégalités entre les valeurs moyennes. *Acta mathematica* **30** 175–193.

[21] JIANG, B., MA, S., CAUSEY, J., QIAO, L., HARDIN, M. P., BITTS, I., JOHNSON, D., ZHANG, S. and HUANG, X. (2016). SparRec: An effective matrix completion framework of missing data imputation for GWAS. *Scientific Reports* **6** 35534. https://doi.org/10.1038/srep35534

[22] JIN, J. (2015). Fast community detection by score. *Annals of Statistics* **43** 57–89.

[23] JOHNSTONE, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Annals of Statistics* **29** 295–327. https://doi.org/10.1214/aos/1009210544 MR1863961

[24] JORDAN, M. I., LEE, J. D. and YANG, Y. (2019). Communication-efficient distributed statistical inference. *Journal of the American Statistical Association* **114** 668–681.

[25] 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*. *Proceedings of Machine Learning Research* **35** 1040–1057. PMLR, Barcelona, Spain.

[26] KARGUPTA, H., HUANG, W., SIVAKUMAR, K. and JOHNSON, E. (2001). Distributed clustering using collective principal component analysis. *Knowledge and Information Systems* **3** 422–448. https://doi.org/10.1007/PL00011677

[27] LI, T., SAHU, A. K., TALWALKAR, A. and SMITH, V. (2020). Federated learning: challenges, methods, and future directions. *IEEE Signal Processing Magazine* **37** 50–60.

[28] PASINI, G. (2017). Principal component analysis for stock portfolio management. *International Journal of Pure and Applied Mathematics* **115** 153-167.

[29] PAUL, D. (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica* 1617–1642.

[30] PRICE, A. L., PATTERSON, N. J., PLENGE, R. M., WEINBLATT, M. E., SHADICK, N. A. and REICH, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* **38** 904–909. https://doi.org/10.1038/ng1847

[31] 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** 559–575. https://doi.org/10.1086/519795

[32] REICH, D., PRICE, A. L. and PATTERSON, N. (2008). Principal component analysis of genetic data. *Nature Genetics* **40** 491–492. https://doi.org/10.1038/ng0508-491

[33] SERFLING, R. J. (2009). *Approximation theorems of mathematical statistics*. John Wiley & Sons.

[34] SHEN, S. and LU, J. (2020). Combinatorial-Probabilistic Trade-Off: Community Properties Test in the Stochastic Block Models. *arXiv preprint arXiv:2010.15063*.

[35] STEWART, G. W. (1977). On the perturbation of pseudo-inverses, projections and linear least squares problems. *SIAM Review* **19** 634–662. https://doi.org/10.1137/1019104 MR461871

[36] 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** e1001779–e1001779. https://doi.org/10.1371/journal.pmed.1001779

[37] VERSHYNIN, R. (2012). *Introduction to the non-asymptotic analysis of random matrices* In *Compressed Sensing: Theory and Applications* 210–268. Cambridge University Press. https://doi.org/10.1017/CBO9780511794308.006

[38] WANG, W. and FAN, J. (2017). Asymptotics of empirical eigen-structure for high dimensional spiked covariance. *Annals of Statistics* **45** 1342–1374. https://doi.org/10.1214/16-AOS1487

[39] WEDIN, P.-. A. (1972). Perturbation bounds in connection with singular value decomposition. *Nordisk Tidskr. Informations-behandling (BIT)* **12** 99–111. https://doi.org/10.1007/bf01932678 MR309968

[40] YAN, Y., CHEN, Y. and FAN, J. (2021). Inference for Heteroskedastic PCA with Missing Data. *arXiv preprint arXiv:2107.12365*.

[41] YANG, F., LIU, S., DOBRIBAN, E. and WOODRUFF, D. P. (2021). How to reduce dimension with PCA and random projections? *IEEE Transactions on Information Theory* **67** 8154-8189.

[42] YU, Y., WANG, T. and SAMWORTH, R. J. (2015). A useful variant of the Davis-Kahan theorem for statisticians. *Biometrika* **102** 315–323. https://doi.org/10.1093/biomet/asv008 MR3371006

SUPPLEMENTARY MATERIAL

Supplementary material for "FADI: Fast Distributed-Friendly PCA With High Accuracy for Large-Scale Data"

This file contains the supplementary materials to the paper "FADI: Fast Distributed-Friendly PCA With High Accuracy for Large-Scale Data". In Appendix A we provide numerical results for Example 3 and Example 4 along with some additional simulation results for Example 1 under the genetic setting. In Appendix B, we present the proofs for the lemmas, theorems, propositions and corollaries given in Section 4 of the main paper. In Appendix C we give the proofs of some technical lemmas useful for the proofs of the main theorems. In Appendix D, we present the modified version of Wedin's theorem, which is used in several proofs. Appendix E provides the supplementary figures deferred from the main paper.

APPENDIX A: ADDITIONAL SIMULATION RESULTS

In this section we present the simulation results for Example 3 and Example 4, and we provide some additional simulation results for Example 1 to evaluate the performance of FADI under the genetic settings.

A.1. Example 3: Gaussian Mixture Models. Under this setting, we take $K=3$, fix the Gaussian vector dimension at $n=20000$ and set $\Delta_0^2=n^{2/3}$. Then we generate the Gaussian means by $\boldsymbol{\theta}_k^* \overset{\text{i.i.d.}}{\sim} N\left(\mathbf{0}, \frac{\Delta_0^2}{2n}\mathbf{I}_n\right)$, $k\in[K]$. We set the sample size at $d=500,1000,2000$ respectively and generate independent Gaussian samples $\{\boldsymbol{X}_i\}_{i=1}^d\in\mathbb{R}^n$ with Gaussian means $\boldsymbol{\theta}_k^*, k\in[K]$ to study the performance of FADI under different settings. We assign each cluster $k\in[K]$ with d/K Gaussian samples. We divide the data vertically along n into $m=20$ splits, set $p=p'=12$ and $q=7$ for the final powered fast sketching. We take the ratio $Lp/d\in\{0.2,0.6,0.9,1,1.2,2,5,10\}$ for each setting and compute the asymptotic covariance via Corollary 4.8 and Corollary 4.12 under different regimes of Lp. We define $\widetilde{\mathbf{v}}=\widehat{\boldsymbol{\Sigma}}_1^{-1/2}(\widetilde{\mathbf{V}}^{\mathrm{F}}-\mathbf{V}\mathbf{H}^\top)^\top\mathbf{e}_1$ where $\widehat{\boldsymbol{\Sigma}}_1$ is the asymptotic covariance for the first row of $\widetilde{\mathbf{V}}^{\mathrm{F}}$ and $\mathbf{H}=\mathrm{sgn}(\widetilde{\mathbf{V}}^{\mathrm{F}\top}\mathbf{V})$ is the alignment matrix, and calculate the empirical coverage probability by empirically evaluating $\mathbb{P}\big(\|\widetilde{\mathbf{v}}\|_2^2\le\chi_3^2(0.95)\big)$, where $\chi_3^2(0.95)$ is the 0.95 quantile of the Chi-squared distribution with degrees of freedom equal to 3. We perform 300 Monte Carlo simulations and the results under different settings are shown in Figure 8. We can see that the error rate of FADI gets closer to that of the traditional PCA estimator as Lp/d increases while FADI greatly outperforms the traditional PCA in terms of running time under different settings. Note that here d is the sample size, and the decreasing of error rates with increasing d and fixed n (at the same Lp/d ratio) is consistent with Corollary 4.2. Similar to Example 1 in Section 5.1, we can see from Figure 8 (b) the running time is large due to calculation of $\widehat{\boldsymbol{\Sigma}}_i$ at Lp/d approaching 1 from the left, and we do not recommend inference at this regime. Validation of the inferential properties are shown in Figure 8 (c) and (d).

A.2. Example 4: Missing Matrix Inference. For the true matrix $\mathbf{M}$, we consider $K=3$, take $\mathbf{V}$ to be the K left singular vectors of a pregenerated $d\times K$ i.i.d. Gaussian matrix, and take $\boldsymbol{\Lambda}=\mathrm{diag}(6,4,2)$. We set the dimension at $d\in\{500,1000,2000\}$ respectively, and set $\theta=0.4$ and $\sigma=8/d$ for each setting. Then we generate the entry-wise noise by $\varepsilon_{ij}\overset{\text{i.i.d.}}{\sim}\mathcal{N}(0,\sigma^2)$ for $i\le j$, and subsample non-zero entries of $\mathbf{M}$ with probability $\theta=0.4$. Under each setting, we perform FADI at $p=p'=12$, $q=7$ and $Lp/d\in\{0.2,0.6,0.9,1,1.2,2,5,10\}$ for the computation of $\widetilde{\mathbf{V}}^{\mathrm{F}}$. Define $\widetilde{\mathbf{v}}=\widehat{\boldsymbol{\Sigma}}_1^{-1/2}(\widetilde{\mathbf{V}}^{\mathrm{F}}-\mathbf{V}\mathbf{H}^\top)^\top\mathbf{e}_1$ with $\widehat{\boldsymbol{\Sigma}}_1$ being the asymptotic covariance for $\widetilde{\mathbf{V}}^{\mathrm{F}\top}\mathbf{e}_1$ defined in Corollary 4.9 and $\mathbf{H}=\mathrm{sgn}(\widetilde{\mathbf{V}}^{\mathrm{F}\top}\mathbf{V})$, and empirically calculate the coverage probability, i.e., $\mathbb{P}\big(\|\widetilde{\mathbf{v}}\|_2^2\le\chi_3^2(0.95)\big)$. Similar as in Section 5.2, for the regime $Lp<d$, we refer to Theorem 4.10 and calculate $\widehat{\boldsymbol{\Sigma}}_1$ by

$$\widehat{\boldsymbol{\Sigma}}_1=L^{-2}\widehat{\mathbf{B}}_{\boldsymbol{\Omega}}^\top\boldsymbol{\Omega}^\top\,\mathrm{diag}\left([\widetilde{\mathbf{M}}_{1j}^2(1-\widehat{\theta})/\widehat{\theta}+\widehat{\sigma}^2/\widehat{\theta}]_{j=1}^d\right)\boldsymbol{\Omega}\widehat{\mathbf{B}}_{\boldsymbol{\Omega}}.$$

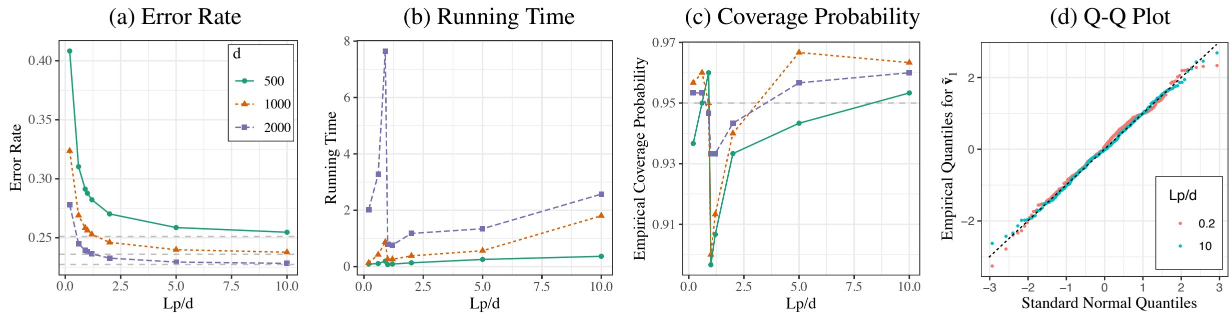


FIGURE 8. *Performance of FADI under different settings for Example 3. (a) Empirical error rate of* $\rho(\widetilde{\mathbf{V}}^{\mathrm{F}}, \mathbf{V})$*, where the grey dashed lines represent the error rate for the traditional PCA estimator* $\widehat{\mathbf{V}}$*; (b) Running time (in seconds) under different settings (including the runtime for computing* $\widehat{\mathbf{\Sigma}}_i$*). For the traditional PCA, the running time is 5.43 seconds at* $d = 500$*, 23.32 seconds at* $d = 1000$ *and 105.58 seconds at* $d = 2000$*; (c) Empirical coverage probability, where the grey dashed line represents the theoretical rate at 0.95; (d) Q-Q plot for* $\widetilde{\mathbf{v}}_1$ *at* $Lp/d \in \{0.2, 10\}$*;*

Results over 300 Monte Carlo simulations are provided in Figure 9. Figure 9 (a) illustrates that the error rate of FADI is almost the same as the traditional PCA as Lp/d gets larger, and Figure 9 (b) shows that the computational efficiency of FADI greatly outperforms the traditional PCA for large dimension d. We can observe from Figure 9 (c) that the confidence interval performs poorly at $Lp/d < 1$ with the coverage probability equal to 1, which is consistent with the theoretical conditions in Corollary 4.9 for distributional convergence. Figure 9 (d) shows the good Gaussian approximation of FADI at $Lp/d = 10$, and the results at $Lp/d = 0.2$ is consistent with Figure 9 (c).

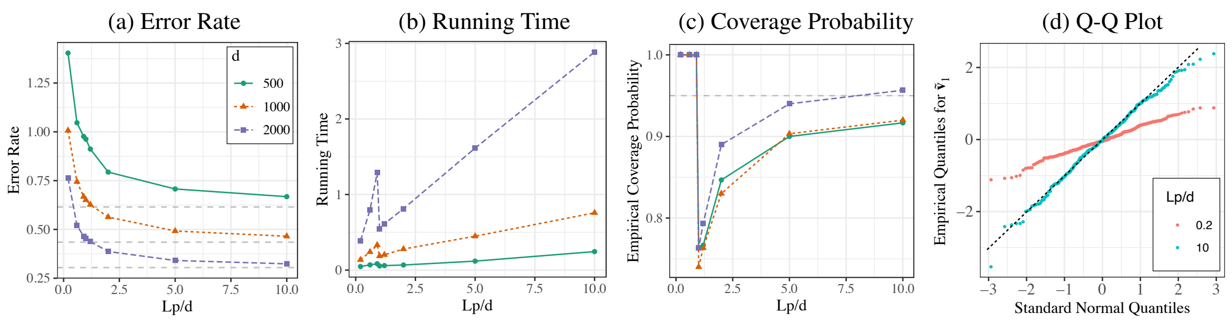


FIGURE 9. *Performance of FADI under different settings for Example 4. (a) Empirical error rate of* $\rho(\widetilde{\mathbf{V}}^{\mathrm{F}}, \mathbf{V})$ *with traditional PCA error rate as the reference; (b) Running time (in seconds) under different settings (including the computational time of* $\widehat{\mathbf{\Sigma}}_i$*). For the traditional PCA, the running time is 0.42 seconds at* $d = 500$*, 3.48 seconds at* $d = 1000$ *and 30.62 seconds at* $d = 2000$*; (c) Empirical coverage probability; (d) Q-Q plot for* $\widetilde{\mathbf{v}}_1$ *at* $Lp/d = 10$*;*

A.3. Additional Results for Example 1 in the Genetic Setting. Section 5.1 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, 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 $\{\boldsymbol{X}_i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$, where

$$\mathbf{\Sigma} = \mathrm{diag}(2.4, 1.2, \underbrace{0.6, \ldots, 0.6}_{K-2}, 0.4 \ldots, 0.4).$$

The dimension is $d = 2504$ and the sample size is $n = 160,000$. Error rates and running times using different algorithms are compared under different number of splits m for the sample size n. For FADI, we take $L = 75$, $p = p' = 40$ and $q = 7$.

Table 6 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. [15]'s distributed PCA increases as m increases. FADI is much faster than the other two 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 this section, we still conduct simulations where the eigengap increases gradually to see how it affects the performance of FADI. Table 7 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 ratio of distributed PCA to FADI gets below 1, but are still above 0.9 when the eigengap is larger than 1. As to the running time, FADI outperforms the other two methods in all the settings. In summary, when the eigengap grows larger, the performance of the three algorithms becomes similar to what we see in Section 5.1.

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

TABLE 6

Comparison of the error rates and running times (in seconds) among FADI, full sample PCA and distributed PCA [15], using different numbers of sample splits m in the genetic setting. Values in the parentheses represent the error rate ratios or the computational time ratios of each method with respect to FADI.

	FADI	Traditional PCA	Distributed PCA	Eigengap
Error Rate	1.28	1.06 (0.82)	1.57 (1.22)	0.4
	0.77	0.65 (0.85)	0.71 (0.92)	0.8
	0.48	0.42 (0.88)	0.43 (0.90)	1.6
	0.31	0.29 (0.92)	0.29 (0.93)	3.2
Running Time	2.76	925.15 (334.7)	115.29 (41.7)	0.4
	2.77	916.52 (331.4)	114.76 (41.5)	0.8
	2.69	922.85 (342.7)	114.75 (42.6)	1.6
	2.77	919.20 (332.2)	115.26 (41.7)	3.2

TABLE 7

Comparison of the error rates and running times (in seconds) among FADI, full sample PCA and distributed PCA [15] for different eigengaps Δ 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 6.

APPENDIX B: PROOF OF MAIN THEORETICAL RESULTS

In this section we provide proofs of the theoretical results in Section 4. For the inferential results, we will present proofs of the theorems under the regime $Lp \ll d$ first, which takes into consideration the extra variability caused by the fast sketching, and then give proofs of the theorems under the regime $Lp \gg d$ where the fast sketching randomness is negligible.

B.1. Unbiasedness of Fast Sketching With Respect to $\widehat{\mathbf{M}}$. We show by the following Lemma B.1 that the fast sketching is unbiased with respect to $\widehat{\mathbf{M}}$ under proper conditions.

LEMMA B.1. *Let $\widehat{\mathbf{V}}_d\widehat{\mathbf{\Lambda}}_d\widehat{\mathbf{V}}_d^\top$ be the eigen-decomposition of $\widehat{\mathbf{M}}$, and let $\widehat{\mathbf{V}} = (\widehat{\mathbf{v}}_1, \ldots, \widehat{\mathbf{v}}_K)$ be the stacked K leading eigenvectors of $\widehat{\mathbf{M}}$ corresponding to the eigenvalues with largest magnitudes. When $\|\mathbf{\Sigma}' - \widehat{\mathbf{V}}\widehat{\mathbf{V}}^\top\|_2 < 1/2$, we have that $\mathrm{Col}(\mathbf{V}') = \mathrm{Col}(\widehat{\mathbf{V}})$, where $\mathrm{Col}(\mathbf{V})$ denotes the column space of $\mathbf{V}$.*

PROOF. We will first show that $\widehat{\mathbf{V}}_d^\top\mathbf{\Sigma}'\widehat{\mathbf{V}}_d$ is diagonal. For any $j \in [d]$, we let $\mathbf{D}_j = \mathbf{I}_d - 2\mathbf{e}_j\mathbf{e}_j^\top$, and $\mathbf{\Omega} \in \mathbb{R}^{d\times p}$ be a random matrix with i.i.d. Gaussian entries, and recall we denote the eigen-decomposition of $\widehat{\mathbf{M}}$ by $\widehat{\mathbf{M}} = \widehat{\mathbf{V}}_d\widehat{\mathbf{\Lambda}}_d\widehat{\mathbf{V}}_d^\top$. Then conditional on $\widehat{\mathbf{M}}$ we have

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

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 $\mathbf{D}_j\widehat{\mathbf{V}}_d^\top\mathbf{\Omega}^{(\ell)} \overset{\mathrm{d}}{=} \widehat{\mathbf{V}}_d^\top\mathbf{\Omega}^{(\ell)}$. Also we know the top K eigenvectors of $\widehat{\mathbf{V}}_d\mathbf{D}_j\widehat{\mathbf{V}}_d^\top\widehat{\mathbf{Y}}^{(\ell)}\widehat{\mathbf{Y}}^{(\ell)\top}\widehat{\mathbf{V}}_d\mathbf{D}_j\widehat{\mathbf{V}}_d^\top$ are $\widehat{\mathbf{V}}_d\mathbf{D}_j\widehat{\mathbf{V}}_d^\top\widehat{\mathbf{V}}^{(\ell)}$, and thus $\widehat{\mathbf{V}}_d\mathbf{D}_j\widehat{\mathbf{V}}_d^\top\widehat{\mathbf{V}}^{(\ell)} \overset{\mathrm{d}}{=} \widehat{\mathbf{V}}^{(\ell)}$. Hence we have

$$\begin{aligned}
&\widehat{\mathbf{V}}_d^\top\mathbb{E}\left(\widehat{\mathbf{V}}^{(\ell)}\widehat{\mathbf{V}}^{(\ell)\top}|\widehat{\mathbf{M}}\right)\widehat{\mathbf{V}}_d = \widehat{\mathbf{V}}_d^\top\widehat{\mathbf{V}}_d\mathbf{D}_j\widehat{\mathbf{V}}_d^\top\mathbb{E}\left(\widehat{\mathbf{V}}^{(\ell)}\widehat{\mathbf{V}}^{(\ell)\top}|\widehat{\mathbf{M}}\right)\widehat{\mathbf{V}}_d\mathbf{D}_j\widehat{\mathbf{V}}_d^\top\widehat{\mathbf{V}}_d \\
&= \mathbf{D}_j\widehat{\mathbf{V}}_d^\top\mathbb{E}\left(\widehat{\mathbf{V}}^{(\ell)}\widehat{\mathbf{V}}^{(\ell)\top}|\widehat{\mathbf{M}}\right)\widehat{\mathbf{V}}_d\mathbf{D}_j = \mathbf{D}_j\widehat{\mathbf{V}}_d^\top\mathbf{\Sigma}'\widehat{\mathbf{V}}_d\mathbf{D}_j.
\end{aligned}$$

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

Now under the condition that $\left\|\mathbf{\Sigma}' - \widehat{\mathbf{V}}\widehat{\mathbf{V}}^\top\right\|_2 < 1/2$, for any $j \in [K]$, we denote by $\widehat{\mathbf{v}}_j$ the j-th column of $\widehat{\mathbf{V}}$, and we have

$$\left\|\mathbf{\Sigma}'\widehat{\mathbf{v}}_j\right\|_2 = \left\|\left(\mathbf{\Sigma}' - \widehat{\mathbf{V}}\widehat{\mathbf{V}}^\top + \widehat{\mathbf{V}}\widehat{\mathbf{V}}^\top\right)\widehat{\mathbf{v}}_j\right\|_2 \geq 1 - \left\|\mathbf{\Sigma}' - \widehat{\mathbf{V}}\widehat{\mathbf{V}}^\top\right\|_2 > 1 - \frac{1}{2} = \frac{1}{2}.$$

In other words, the corresponding eigenvalue of $\widehat{\mathbf{v}}_j$ in $\mathbf{\Sigma}'$ is larger than $1/2$. On the other hand, by Weyl's inequality [17], the rest of the $d-K$ eigenvalues of $\mathbf{\Sigma}'$ should be less than 1/2. Therefore, $\widehat{\mathbf{V}}$ are still the leading K eigenvectors for $\mathbf{\Sigma}'$, and thus $\|\mathbf{V}'\mathbf{V}'^\top - \widehat{\mathbf{V}}\widehat{\mathbf{V}}^\top\|_2 = 0$. □

Recall in Section 4 we discuss that the bias term has the following decomposition $\rho(\mathbf{V}', \mathbf{V}) \leq \rho(\widehat{\mathbf{V}}, \mathbf{V}) + \rho(\mathbf{V}', \widehat{\mathbf{V}})$. Lemma B.1 shows that as long as $\mathbf{\Sigma}'$ and $\mathbf{V}\mathbf{V}^\top$ are not too far apart, $\mathbf{V}'$ and $\widehat{\mathbf{V}}$ will share the same column space. In fact, Lemma B.4 in Section B.2 will show that the probability that $\mathbf{\Sigma}'$ and $\widehat{\mathbf{V}}\widehat{\mathbf{V}}^\top$ are not sufficiently close converges to 0 and $\rho(\mathbf{V}', \mathbf{V}) = \rho(\widehat{\mathbf{V}}, \mathbf{V})$ with high probability. With the help of Lemma B.1, we present the proof of the main error bound results in the following section.

B.2. Proof of Theorem 4.1. Recall the problem setting in Section 2. It is not hard to see that we can write $\mathbf{\Lambda} = \mathbf{P}_0\mathbf{\Lambda}^0$, where $\mathbf{\Lambda}^0 = \mathrm{diag}(|\lambda_1|, \ldots, |\lambda_K|)$ and $\mathbf{P}_0 = \mathrm{diag}\left([\mathrm{sgn}(\lambda_i)]_{i=1}^K\right)$. Then $\mathbf{M} = (\mathbf{V}\mathbf{P}_0)\mathbf{\Lambda}^0\mathbf{V}^\top$ is the SVD of $\mathbf{M}$.

We begin with bounding $\left(\mathbb{E}\|\widetilde{\mathbf{V}}\widetilde{\mathbf{V}}^\top - \mathbf{V}\mathbf{V}^\top\|_{\mathrm{F}}^2\right)^{1/2}$. Before delving into the detailed proof, the following two lemmas provide some important properties of the random Gaussian matrix.

LEMMA B.2. *Let $\mathbf{\Omega} \in \mathbb{R}^{d\times p}$ be a random matrix with i.i.d. Gaussian entries. For a random variable, recall that we define the ψ_1 norm to be $\|\cdot\|_{\psi_1} = \sup_{p\geq 1}(\mathbb{E}|\cdot|^p)^{1/p}/p$. Then we have the following bound on the ψ_1 norm of the matrix $\mathbf{\Omega}/\sqrt{p}$:*

$$\| \|\mathbf{\Omega}/\sqrt{p}\|_2 \|_{\psi_1} \lesssim \sqrt{d/p}. \tag{B.23}$$

LEMMA B.3. *Let $\mathbf{\Omega} \in \mathbb{R}^{K\times p}$ denote a random matrix with i.i.d. Gaussian entries, where $p \geq 2K$. For any integer a such that $1 \leq a \leq (p-K+1)/2$, there exists a constant $C > 0$ such that*

$$\mathbb{E}\left(\{\sigma_{\min}(\mathbf{\Omega}/\sqrt{p})\}^{-a}\right) \leq C^a. \tag{B.24}$$

The following lemma shows that $\|\mathbf{\Sigma}' - \mathbf{V}\mathbf{V}\|_2$ and $\|\mathbf{\Sigma}' - \widehat{\mathbf{V}}\widehat{\mathbf{V}}^\top\|_2$ are bounded by a small constant with high probability.

LEMMA B.4. *If Assumption 1 holds and $p \geq \max(2K, K+3)$, there exists a constant $c_0 > 0$ such that for any $\varepsilon > 0$, we have*

$$\max\left\{\mathbb{P}\Big(\|\mathbf{\Sigma}' - \mathbf{V}\mathbf{V}\|_2 \geq \varepsilon\Big), \mathbb{P}\left(\|\mathbf{\Sigma}' - \widehat{\mathbf{V}}\widehat{\mathbf{V}}^\top\|_2 \geq \varepsilon\right)\right\} \lesssim \exp\left(-c_0\sqrt{\frac{p}{d}}\frac{\Delta\varepsilon}{r_1(d)}\right).$$

The proof of Lemma B.2, Lemma B.3 and Lemma B.4 are deferred to Appendix C. Now we can start with the proof. We first decompose the bias term into two parts,

$$\left(\mathbb{E}|\rho(\widetilde{\mathbf{V}}, \mathbf{V})|^2\right)^{1/2} \leq \underbrace{\left(\mathbb{E}|\rho(\widetilde{\mathbf{V}}, \mathbf{V}')|^2\right)^{1/2}}_{\mathrm{I}} + \underbrace{\left(\mathbb{E}|\rho(\mathbf{V}', \mathbf{V})|^2\right)^{1/2}}_{\mathrm{II}}. \tag{B.25}$$

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

B.2.1. *Control of the Bias Term.* We can see that term II can be further decomposed into two terms

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

We can bound both terms separately. First note that $\|\mathbf{V}'\mathbf{V}'^\top - \widehat{\mathbf{V}}\widehat{\mathbf{V}}^\top\|_{\mathrm{F}} \leq \sqrt{2K}\|\mathbf{V}'\mathbf{V}'^\top - \widehat{\mathbf{V}}\widehat{\mathbf{V}}^\top\|_2 \leq \sqrt{2K}$. Thus we have,

$$\begin{aligned}
\left(\mathbb{E}\|\mathbf{V}'\mathbf{V}'^\top - \widehat{\mathbf{V}}\widehat{\mathbf{V}}^\top\|_{\mathrm{F}}^2\right)^{1/2} &\leq \left(\mathbb{E}\|\mathbf{V}'\mathbf{V}'^\top - \widehat{\mathbf{V}}\widehat{\mathbf{V}}^\top\|_{\mathrm{F}}^2 \mathbb{I}\{\|\mathbf{\Sigma}' - \widehat{\mathbf{V}}\widehat{\mathbf{V}}^\top\|_2 \geq 1/2\}\right)^{1/2} \\
&\quad + \left(\mathbb{E}\|\mathbf{V}'\mathbf{V}'^\top - \widehat{\mathbf{V}}\widehat{\mathbf{V}}^\top\|_{\mathrm{F}}^2 \mathbb{I}\{\|\mathbf{\Sigma}' - \widehat{\mathbf{V}}\widehat{\mathbf{V}}^\top\|_2 < 1/2\}\right)^{1/2} \\
&\lesssim 0 + \sqrt{K}\left(\mathbb{P}\left(\|\mathbf{\Sigma}' - \widehat{\mathbf{V}}\widehat{\mathbf{V}}^\top\|_2 \geq 1/2\right)\right)^{1/2} \lesssim \sqrt{K}\exp\left(-\frac{c_0}{4}\sqrt{\frac{p}{d}}\frac{\Delta}{r_1(d)}\right),
\end{aligned}$$

where the last but one inequality follows from Lemma B.1, and the last inequality is a result of Lemma B.4. As for term II in (B.25), by Davis-Kahan's Theorem [42], we have

$$\begin{aligned}
\left(\mathbb{E}\|\widehat{\mathbf{V}}\widehat{\mathbf{V}}^\top - \mathbf{V}\mathbf{V}^\top\|_{\mathrm{F}}^2\right)^{1/2} &\lesssim \frac{\sqrt{K}}{\Delta}\left(\mathbb{E}\|\widehat{\mathbf{M}} - \mathbf{M}\|_2^2\right)^{1/2} = \frac{\sqrt{K}}{\Delta}\left(\mathbb{E}\|\mathbf{E}\|_2^2\right)^{1/2} \\
&\leq \frac{\sqrt{K}}{\Delta}\| \|\mathbf{E}\|_2 \|_{\psi_1} \lesssim \frac{\sqrt{K}}{\Delta} r_1(d).
\end{aligned}$$

Therefore, the bound for the bias term is

$$\mathrm{II} \lesssim \sqrt{K}\exp\left(-\frac{c_0}{4}\sqrt{\frac{p}{d}}\frac{\Delta}{r_1(d)}\right) + \frac{\sqrt{K}}{\Delta} r_1(d).$$

B.2.2. *Control of the Variance Term.* Now we move on to control the variance term. Suppose that $\left\|\mathbf{\Sigma}'-\mathbf{V}\mathbf{V}^\top\right\|_2<1/4$. Then by Weyl's inequality [17] 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 [42]

$$\begin{aligned}
&\left(\mathbb{E}\left(\|\widetilde{\mathbf{V}}\widetilde{\mathbf{V}}^\top-\mathbf{V}'\mathbf{V}'^\top\|_{\mathrm{F}}^2\mathbb{I}\left\{\left\|\mathbf{\Sigma}'-\mathbf{V}\mathbf{V}^\top\right\|_2<1/4\right\}\right)\right)^{1/2}\\
&\lesssim\left(\mathbb{E}\left(\frac{\|\widetilde{\mathbf{\Sigma}}-\mathbf{\Sigma}'\|_{\mathrm{F}}^2}{(\sigma_K(\mathbf{\Sigma}')-\sigma_{K+1}(\mathbf{\Sigma}'))^2}\mathbb{I}\left\{\left\|\mathbf{\Sigma}'-\mathbf{V}\mathbf{V}^\top\right\|_2<1/4\right\}\right)\right)^{1/2}\\
&\quad\lesssim\left(\mathbb{E}\left(\|\widetilde{\mathbf{\Sigma}}-\mathbf{\Sigma}'\|_{\mathrm{F}}^2\mathbb{I}\left\{\left\|\mathbf{\Sigma}'-\mathbf{V}\mathbf{V}^\top\right\|_2<1/4\right\}\right)\right)^{1/2}\le\underbrace{\left(\mathbb{E}\|\widetilde{\mathbf{\Sigma}}-\mathbf{\Sigma}'\|_{\mathrm{F}}^2\right)^{1/2}}_{\mathrm{III}}.
\end{aligned}$$

We will bound term III later. Also similar as previously, note that $\|\widetilde{\mathbf{V}}\widetilde{\mathbf{V}}^\top-\mathbf{V}'\mathbf{V}'^\top\|_{\mathrm{F}}\le\sqrt{2K}$. Thus by Lemma B.4,

$$\begin{aligned}
&\left(\mathbb{E}\left(\|\widetilde{\mathbf{V}}\widetilde{\mathbf{V}}^\top-\mathbf{V}'\mathbf{V}'^\top\|_{\mathrm{F}}^2\mathbb{I}\left\{\left\|\mathbf{\Sigma}'-\mathbf{V}\mathbf{V}^\top\right\|_2\ge\frac{1}{4}\right\}\right)\right)^{\frac{1}{2}}\lesssim\sqrt{K}\left(\mathbb{P}\left(\left\|\mathbf{\Sigma}'-\mathbf{V}\mathbf{V}^\top\right\|_2\ge\frac{1}{4}\right)\right)^{\frac{1}{2}}\\
&\le\sqrt{K}\exp\left(-\frac{c_0}{8}\sqrt{\frac{p}{d}}\frac{\Delta}{r_1(d)}\right).
\end{aligned}$$

Therefore, we have

$$\left(\mathbb{E}\|\widetilde{\mathbf{V}}\widetilde{\mathbf{V}}^\top-\mathbf{V}'\mathbf{V}'^\top\|_{\mathrm{F}}^2\right)^{1/2}\lesssim\sqrt{K}\exp\left(-\frac{c_0}{8}\sqrt{\frac{p}{d}}\frac{\Delta}{r_1(d)}\right)+\underbrace{\left(\mathbb{E}\|\widetilde{\mathbf{\Sigma}}-\mathbf{\Sigma}'\|_{\mathrm{F}}^2\right)^{1/2}}_{\mathrm{III}}.$$

Now we move on to bound term III.

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

where the last but one equality is due to the independence of estimators from different sketches conditional on $\widehat{\mathbf{M}}$. By Jensen's inequality [20], we have

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

Thus we have

$$\left(\mathbb{E}\|\widetilde{\mathbf{\Sigma}}-\mathbf{\Sigma}'\|_{\mathrm{F}}^2\right)^{1/2}\lesssim\frac{1}{\sqrt{L}}\left(\mathbb{E}\left\|\widehat{\mathbf{V}}^{(\ell)}\widehat{\mathbf{V}}^{(\ell)\top}-\mathbf{V}\mathbf{V}^\top\right\|_{\mathrm{F}}^2\right)^{1/2},\tag{B.26}$$

Before bounding the RHS, let's consider the matrix $\mathbf{Y}^{(\ell)}:=\mathbf{V}\mathbf{P}_0\mathbf{\Lambda}^0\mathbf{V}^\top\mathbf{\Omega}^{(\ell)}$. If $\widetilde{\mathbf{\Omega}}^{(\ell)}:=\mathbf{V}^\top\mathbf{\Omega}^{(\ell)}\in\mathbb{R}^{K\times p}$ does not have 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 $\mathbf{Y}^{(\ell)}$ is of rank K, and thus with probability 1, $\mathbf{V}$ and the top K left singular eigenvectors of $\mathbf{Y}^{(\ell)}/\sqrt{p}$ span the same column space. In other words, if we let $\mathbf{\Gamma}_K^{(\ell)}$ be the left singular vectors of $\mathbf{Y}^{(\ell)}/\sqrt{p}$, then $\mathbf{\Gamma}_K^{(\ell)}\mathbf{\Gamma}_K^{(\ell)\top} = \mathbf{V}\mathbf{V}^\top$.

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

$$\begin{aligned}\sigma_K\left(\mathbf{Y}^{(\ell)}/\sqrt{p}\right) &= \sigma_K\left(\mathbf{V}\mathbf{P}_0\mathbf{\Lambda}^0\widetilde{\mathbf{\Omega}}^{(\ell)}/\sqrt{p}\right) = \sigma_K\left(\mathbf{\Lambda}^0\mathbf{U}_{\widetilde{\mathbf{\Omega}}}\mathbf{D}_{\widetilde{\mathbf{\Omega}}}\right)\\ &= \min_{\|\boldsymbol{x}\|_2=1}\|\mathbf{\Lambda}^0\mathbf{U}_{\widetilde{\mathbf{\Omega}}}\mathbf{D}_{\widetilde{\mathbf{\Omega}}}\boldsymbol{x}\|_2 \overset{(i)}{\geq} \sigma_{\min}\left(\widetilde{\mathbf{\Omega}}^{(\ell)}/\sqrt{p}\right)\min_{\|\boldsymbol{v}_1\|_2=1}\left\|\mathbf{\Lambda}^0\mathbf{U}_{\widetilde{\mathbf{\Omega}}}\boldsymbol{v}_1\right\|_2\\ &\overset{(ii)}{\geq} \sigma_{\min}\left(\widetilde{\mathbf{\Omega}}^{(\ell)}/\sqrt{p}\right)\min_{\|\boldsymbol{v}_2\|_2=1}\left\|\mathbf{\Lambda}^0\boldsymbol{v}_2\right\|_2 \geq \Delta\sigma_{\min}\left(\widetilde{\mathbf{\Omega}}^{(\ell)}/\sqrt{p}\right),\end{aligned}$$

where $\boldsymbol{v}_1 = \mathbf{D}_{\widetilde{\mathbf{\Omega}}}\boldsymbol{x}/\|\mathbf{D}_{\widetilde{\mathbf{\Omega}}}\boldsymbol{x}\|_2$, and $\boldsymbol{v}_2 = \mathbf{U}_{\widetilde{\mathbf{\Omega}}}\boldsymbol{v}_1$. Inequality (i) follows because

$$\|\mathbf{D}_{\widetilde{\mathbf{\Omega}}}\boldsymbol{x}\|_2 \geq \sigma_{\min}\left(\widetilde{\mathbf{\Omega}}^{(\ell)}/\sqrt{p}\right)\|\boldsymbol{x}\|_2 = \sigma_{\min}\left(\widetilde{\mathbf{\Omega}}^{(\ell)}/\sqrt{p}\right),$$

and inequality (ii) is because $\|\boldsymbol{v}_2\|_2 = \|\boldsymbol{v}_1\|_2 = 1$.

Now by Wedin's Theorem [39] we have the following bound on the RHS of (B.26),

$$\begin{aligned}\frac{1}{\sqrt{L}}\left(\mathbb{E}\left|\rho(\widehat{\mathbf{V}}^{(\ell)},\mathbf{V})\right|^2\right)^{1/2} &\lesssim \frac{\sqrt{K}}{\sqrt{L}}\left(\mathbb{E}\left\|\widehat{\mathbf{Y}}^{(\ell)}/\sqrt{p}-\mathbf{Y}^{(\ell)}/\sqrt{p}\right\|_2^2\Big/\left(\Delta\sigma_{\min}(\widetilde{\mathbf{\Omega}}^{(\ell)}/\sqrt{p})\right)^2\right)^{1/2}\\ &\leq \frac{\sqrt{K}}{\Delta\sqrt{L}}\left(\mathbb{E}\left\|\widehat{\mathbf{Y}}^{(\ell)}/\sqrt{p}-\mathbf{Y}^{(\ell)}/\sqrt{p}\right\|_2^4\right)^{1/4}\left(\mathbb{E}\left(\sigma_{\min}(\widetilde{\mathbf{\Omega}}^{(\ell)}/\sqrt{p})\right)^{-4}\right)^{1/4}\\ &\lesssim \frac{\sqrt{K}}{\Delta\sqrt{L}}\|\|\mathbf{E}\|_2\|_{\psi_1}\cdot\|\|\mathbf{\Omega}^{(\ell)}/\sqrt{p}\|_2\|_{\psi_1} \lesssim \sqrt{\frac{Kd}{\Delta^2 pL}}\|\|\mathbf{E}\|_2\|_{\psi_1} \lesssim \sqrt{\frac{Kd}{\Delta^2 pL}}r_1(d),\end{aligned}$$

where the last but one inequality is due to Lemma B.3. Therefore, we have the final error rate for the estimator $\widetilde{\mathbf{V}}$:

$$\left(\mathbb{E}\|\widetilde{\mathbf{V}}\widetilde{\mathbf{V}}^\top-\mathbf{V}\mathbf{V}\|_{\mathrm{F}}^2\right)^{1/2} \lesssim \underbrace{\sqrt{K}\exp\left(-\frac{c_0}{8}\sqrt{\frac{p}{d}}\frac{\Delta}{r_1(d)}\right)+\frac{\sqrt{K}}{\Delta}r_1(d)}_{\text{bias}}+\underbrace{\sqrt{\frac{Kd}{\Delta^2 pL}}r_1(d)}_{\text{variance}}.$$

Now consider the function $g(x) := \exp(a_0\sqrt{\frac{p}{d}}x)/(\sqrt{d}x^2)$, where $a_0 > 0$ is a fixed constant. We have

$$\frac{d\log g(x)}{dx} = a_0\sqrt{\frac{p}{d}}-\frac{2}{x} > 0, \quad \text{for } x \geq \frac{2}{a_0}\sqrt{\frac{d}{p}}.$$

Thus $g(x)$ is increasing on $x \geq 2\sqrt{d/p}/a_0$, and if we take $x \geq C\sqrt{\frac{d}{p}}\log d$ for some large enough constant $C > 0$, we have that $g(x) \geq 1$. Then by plugging in $x = \Delta/r_1(d)$ and taking $a_0 = c_0/8$, we have that under the condition that $(\log d)^{-1}\sqrt{p/d}\Delta/r_1(d) \geq C$ for some large enough constant $C > 0$, we have

$$\exp\left(-\frac{c_0}{8}\sqrt{\frac{p}{d}}\frac{\Delta}{r_1(d)}\right) \lesssim \frac{1}{\sqrt{d}}\left(\frac{r_1(d)}{\Delta}\right)^2 = o\left(\frac{r_1(d)}{\Delta}\right),$$

and the error rate simplifies to

$$\left(\mathbb{E}\|\widetilde{\mathbf{V}}\widetilde{\mathbf{V}}^\top-\mathbf{V}\mathbf{V}\|_{\mathrm{F}}^2\right)^{1/2} \lesssim \underbrace{\frac{\sqrt{K}}{\Delta}r_1(d)}_{\text{bias}}+\underbrace{\sqrt{\frac{Kd}{\Delta^2 pL}}r_1(d)}_{\text{variance}}.$$

Now we move on to bound $\left(\mathbb{E}\|\widetilde{\mathbf{V}}^{\mathrm{F}}\widetilde{\mathbf{V}}^{\mathrm{F}\top}-\mathbf{V}\mathbf{V}^\top\|_{\mathrm{F}}^2\right)^{1/2}$. Since$\|\cdot\|_2^{2q}$ is convex, by Jensen's inequality [20], under the condition that $p\geq\max(2K,8q+K-1)$ we have that there exists some constant η such that

$$
\begin{aligned}
\mathbb{E}\|\widetilde{\boldsymbol{\Sigma}}-\mathbf{V}\mathbf{V}^\top\|_2^{2q} &\leq \frac{1}{L}\sum_{\ell=1}^{L}\mathbb{E}\|\widehat{\mathbf{V}}^{(\ell)}\widehat{\mathbf{V}}^{(\ell)\top}-\mathbf{V}\mathbf{V}^\top\|_2^{2q}=\mathbb{E}\|\widehat{\mathbf{V}}^{(1)}\widehat{\mathbf{V}}^{(1)\top}-\mathbf{V}\mathbf{V}^\top\|_2^{2q}\\
&\leq \mathbb{E}\biggl(\left\|\widehat{\mathbf{Y}}^{(\ell)}/\sqrt{p}-\mathbf{Y}^{(\ell)}/\sqrt{p}\right\|_2^{2q}\Big/\left(\Delta\sigma_{\min}\left(\widetilde{\boldsymbol{\Omega}}^{(\ell)}/\sqrt{p}\right)\right)^{2q}\biggr)\\
&\leq \frac{1}{\Delta^{2q}}\left(\mathbb{E}\left\|\widehat{\mathbf{Y}}^{(\ell)}/\sqrt{p}-\mathbf{Y}^{(\ell)}/\sqrt{p}\right\|_2^{4q}\right)^{1/2}\left(\mathbb{E}\left(\sigma_{\min}\left(\widetilde{\boldsymbol{\Omega}}^{(\ell)}/\sqrt{p}\right)\right)^{-4q}\right)^{1/2}\\
&\lesssim \left(\eta q^2\sqrt{\frac{d}{\Delta^2 p}}\|\|\mathbf{E}\|_2\|_{\psi_1}\right)^{2q}.
\end{aligned}
$$

Thus by Markov's inequality, we also have

$$
\begin{aligned}
\mathbb{P}\left(\|\widetilde{\boldsymbol{\Sigma}}-\mathbf{V}\mathbf{V}^\top\|_2\geq\frac{1}{2}\right)&=\mathbb{P}\left(\|\widetilde{\boldsymbol{\Sigma}}-\mathbf{V}\mathbf{V}^\top\|_2^{2q}\geq\frac{1}{2^{2q}}\right)\leq 2^{2q}\mathbb{E}\left(\|\widetilde{\boldsymbol{\Sigma}}-\mathbf{V}\mathbf{V}^\top\|_2^{2q}\right)\\
&\lesssim\left(2\eta q^2\sqrt{\frac{d}{\Delta^2 p}}\|\|\mathbf{E}\|_2\|_{\psi_1}\right)^{2q}.
\end{aligned}
$$

Since $\widetilde{\boldsymbol{\Sigma}}$ is the summation of positive semi-definite matrices by construction, $\widetilde{\boldsymbol{\Sigma}}$ is also positive semi-definite. By Weyl's inequality [17], we know that $\sigma_K(\widetilde{\boldsymbol{\Sigma}})\geq 1-\|\widetilde{\boldsymbol{\Sigma}}-\mathbf{V}\mathbf{V}^\top\|_2$ and $\sigma_{K+1}(\widetilde{\boldsymbol{\Sigma}})\leq\|\widetilde{\boldsymbol{\Sigma}}-\mathbf{V}\mathbf{V}^\top\|_2$.

Now if we denote the SVD of $\widetilde{\boldsymbol{\Sigma}}^q$ by $\widetilde{\mathbf{V}}\widetilde{\boldsymbol{\Lambda}}_K^q\widetilde{\mathbf{V}}^\top+\widetilde{\mathbf{V}}_\perp\widetilde{\boldsymbol{\Lambda}}_\perp^q\widetilde{\mathbf{V}}_\perp^\top$, then with probability 1, $\widetilde{\mathbf{V}}\widetilde{\boldsymbol{\Lambda}}_K^q\widetilde{\mathbf{V}}^\top\boldsymbol{\Omega}^{\mathrm{F}}$ and $\widetilde{\mathbf{V}}$ 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 [42], we have

$$
\begin{aligned}
\mathbb{E}\left(\|\widetilde{\mathbf{V}}^{\mathrm{F}}\widetilde{\mathbf{V}}^{\mathrm{F}\top}-\widetilde{\mathbf{V}}\widetilde{\mathbf{V}}^\top\|_{\mathrm{F}}^2\,|\widetilde{\boldsymbol{\Sigma}}\right)&\lesssim\mathbb{E}\left(K\|\widetilde{\boldsymbol{\Sigma}}^q\boldsymbol{\Omega}^{\mathrm{F}}-\widetilde{\mathbf{V}}\widetilde{\boldsymbol{\Lambda}}_K^q\widetilde{\mathbf{V}}^\top\boldsymbol{\Omega}^{\mathrm{F}}\|_2^2/\sigma_{\min}^2(\widetilde{\mathbf{V}}\widetilde{\boldsymbol{\Lambda}}_K^q\widetilde{\mathbf{V}}^\top\boldsymbol{\Omega}^{\mathrm{F}})\,|\widetilde{\boldsymbol{\Sigma}}\right)\\
&\lesssim\left(\frac{\sqrt{K}}{\sigma_K^q(\widetilde{\boldsymbol{\Sigma}})}\|\widetilde{\mathbf{V}}_\perp\widetilde{\boldsymbol{\Lambda}}_\perp^q\widetilde{\mathbf{V}}_\perp^\top\|_2\cdot\|\|\boldsymbol{\Omega}^{\mathrm{F}}/\sqrt{p'}\|_2\|_{\psi_1}\right)^2\\
&\lesssim\frac{Kd}{p'}\frac{\|\widetilde{\boldsymbol{\Sigma}}-\mathbf{V}\mathbf{V}^\top\|_2^{2q}}{\left(1-\|\widetilde{\boldsymbol{\Sigma}}-\mathbf{V}\mathbf{V}^\top\|_2\right)^{2q}}.
\end{aligned}
$$

Therefore we have,

$$
\begin{aligned}
&\left(\mathbb{E}\|\widetilde{\mathbf{V}}^{\mathrm{F}}\widetilde{\mathbf{V}}^{\mathrm{F}\top}-\widetilde{\mathbf{V}}\widetilde{\mathbf{V}}^\top\|_{\mathrm{F}}^2\right)^{1/2}\lesssim\left(\mathbb{E}\|\widetilde{\mathbf{V}}^{\mathrm{F}}\widetilde{\mathbf{V}}^{\mathrm{F}\top}-\widetilde{\mathbf{V}}\widetilde{\mathbf{V}}^\top\|_{\mathrm{F}}^2\mathbb{I}\{\|\widetilde{\boldsymbol{\Sigma}}-\mathbf{V}\mathbf{V}^\top\|_2\leq 1/2\}\right)^{1/2}\\
&\quad+\left(\mathbb{E}\|\widetilde{\mathbf{V}}^{\mathrm{F}}\widetilde{\mathbf{V}}^{\mathrm{F}\top}-\widetilde{\mathbf{V}}\widetilde{\mathbf{V}}^\top\|_{\mathrm{F}}^2\mathbb{I}\{\|\widetilde{\boldsymbol{\Sigma}}-\mathbf{V}\mathbf{V}^\top\|_2>1/2\}\right)^{1/2}\\
&\lesssim 2^q\sqrt{\frac{Kd}{p'}}\left(\mathbb{E}\|\widetilde{\boldsymbol{\Sigma}}-\mathbf{V}\mathbf{V}^\top\|_2^{2q}\right)^{1/2}+\sqrt{K}\left\{\mathbb{P}\left(\|\widetilde{\boldsymbol{\Sigma}}-\mathbf{V}\mathbf{V}^\top\|_2\geq\frac{1}{2}\right)\right\}^{1/2}\\
&\lesssim\sqrt{\frac{Kd}{p'}}\left(2\eta q^2\sqrt{\frac{d}{\Delta^2 p}}\|\|\mathbf{E}\|_2\|_{\psi_1}\right)^q+\sqrt{K}\left(2\eta q^2\sqrt{\frac{d}{\Delta^2 p}}\|\|\mathbf{E}\|_2\|_{\psi_1}\right)^q
\end{aligned}
$$

$$\lesssim \sqrt{\frac{Kd}{p'}} \left(2\eta q^2 \sqrt{\frac{d}{\Delta^2 p}} \| \|\mathbf{E}\|_2 \|_{\psi_1}\right)^q,$$

where the last but one inequality is by Markov's inequality and

$$\mathbb{P}\left(\|\widetilde{\boldsymbol{\Sigma}} - \mathbf{V}\mathbf{V}^\top\|_2 \geq \frac{1}{2}\right) \leq 2^{2q}\mathbb{E}\|\widetilde{\boldsymbol{\Sigma}} - \mathbf{V}\mathbf{V}^\top\|_2^{2q} \lesssim \left(2\eta q^2 \sqrt{\frac{d}{\Delta^2 p}} \| \|\mathbf{E}\|_2 \|_{\psi_1}\right)^{2q}.$$

Thus by previous results and triangle inequality we have

$$\begin{aligned}
\left(\mathbb{E}\left|\rho(\widetilde{\mathbf{V}}^{\mathrm{F}}, \mathbf{V})\right|^2\right)^{1/2} &\lesssim \left(\mathbb{E}\|\widetilde{\mathbf{V}}^{\mathrm{F}}\widetilde{\mathbf{V}}^{\mathrm{F}\top} - \widetilde{\mathbf{V}}\widetilde{\mathbf{V}}^\top\|_{\mathrm{F}}^2\right)^{1/2} + \left(\mathbb{E}\|\widetilde{\mathbf{V}}\widetilde{\mathbf{V}}^\top - \mathbf{V}\mathbf{V}^\top\|_{\mathrm{F}}^2\right)^{1/2} \\
&\lesssim \frac{\sqrt{K}}{\Delta} r_1(d) + \sqrt{\frac{Kd}{\Delta^2 pL}} r_1(d) + \sqrt{\frac{Kd}{p'}} \left(2\eta q^2 \sqrt{\frac{d}{\Delta^2 p}} r_1(d)\right)^q.
\end{aligned}$$

B.3. Proof of Corollary 4.2. The case-specific error rates can be calculated by computing $r_1(d)$ and studying the proper value of q for each example.

• **Example 1:** we know that $\mathbf{E} = \widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma} + (\sigma^2 - \widehat{\sigma}^2)\mathbf{I}$. Now consider the $K' \times K'$ submatrix of $\boldsymbol{\Sigma}$ corresponding to the the index set S, which we denote by $\boldsymbol{\Sigma}_S = \boldsymbol{\Sigma}_{[S,S]}$. We have $\boldsymbol{\Sigma}_S = \sigma^2 \mathbf{I}_{K'} + (\mathbf{V})_{[S,:]}\boldsymbol{\Lambda}(\mathbf{V})_{[S,:]}^\top$, where $(\mathbf{V})_{[S,:]}$ is the submatrix of $\mathbf{V}$ composed of the rows in S. Then since $(\mathbf{V})_{[S,:]}\boldsymbol{\Lambda}(\mathbf{V})_{[S,:]}^\top \succeq \mathbf{0}$ and $\operatorname{rank}\left((\mathbf{V})_{[S,:]}\boldsymbol{\Lambda}(\mathbf{V})_{[S,:]}^\top\right) \leq K$, we know that $\sigma_{\min}(\boldsymbol{\Sigma}_S) = \sigma^2$. By Weyl's inequality [17], we know $|\sigma^2 - \widehat{\sigma}^2| \leq \|\widehat{\boldsymbol{\Sigma}}_S - \boldsymbol{\Sigma}_S\|_2 \leq \|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_2$. Thus we have $\|\mathbf{E}\|_2 \leq \|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_2 + |\sigma^2 - \widehat{\sigma}^2| \leq 2\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_2$. Then by Lemma 3 in Fan et al. [15], we have that there exists some constant $c \geq 1$ such that for any $t \geq 0$, we have

$$\mathbb{P}(\|\mathbf{E}\|_2 \geq t) \leq \mathbb{P}(2\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_2 \geq t) \leq \exp(-\frac{t}{2c(\lambda_1 + \sigma^2)\sqrt{r/n}}),$$

where $r = \operatorname{tr}(\boldsymbol{\Sigma})/\|\boldsymbol{\Sigma}\|_2$ is the effective rank of $\boldsymbol{\Sigma}$. Thus we can see that $\|\mathbf{E}\|_2$ is sub-exponential with

$$\| \|\mathbf{E}\|_2 \|_{\psi_1} \lesssim \| \|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_2 \|_{\psi_1} \lesssim (\lambda_1 + \sigma^2)\sqrt{\frac{r}{n}},$$

and hence we can take $r_1(d) = (\lambda_1 + \sigma^2)\sqrt{\frac{r}{n}}$. When $n \geq C(dr/p)\kappa_1^2(\log d)^4$, by Theorem 4.1 we have

$$\left(\mathbb{E}|\rho(\widetilde{\mathbf{V}}^{\mathrm{F}}, \mathbf{V})|^2\right)^{1/2} \lesssim \kappa_1 \sqrt{\frac{Kr}{n}} + \kappa_1 \sqrt{\frac{Kdr}{npL}} + \sqrt{\frac{Kd}{p'}} \left(\eta q^2 \kappa_1 \sqrt{\frac{dr}{np}}\right)^q,$$

where the third term will be dominated by the first bias term when taking $q = \log d$, and hence (3) holds.

• **Example 2:** Under the problem settings we know that $\mathbf{E} = \widehat{\mathbf{M}} - \mathbf{M} = \mathbf{X} - \mathbb{E}\mathbf{X}$. For the eigenvalues of $\mathbf{M}$, under the given conditions we know that

$$\sigma_K(\mathbf{M}) \gtrsim \theta\sigma_K(\mathbf{P})\sigma_K^2(\boldsymbol{\Pi}) \gtrsim d\theta/K, \quad \sigma_1(\mathbf{M}) \lesssim \theta\sigma_1(\mathbf{P})\sigma_1^2(\boldsymbol{\Pi}) \lesssim Kd\theta\|\boldsymbol{\Pi}\|_{2,\infty}^2 \leq Kd\theta,$$

where the last inequality is because for $i \in [d]$, we have that

$$\|\boldsymbol{\pi}_i\|_2 = \Big(\sum_{k=1}^K \boldsymbol{\pi}_i(k)^2\Big)^{1/2} \leq \Big(\sum_{k=1}^K \boldsymbol{\pi}_i(k)\Big)^{1/2} = 1 \quad \text{and} \quad \|\boldsymbol{\Pi}\|_{2,\infty} \leq 1.$$

Thus we know that $\Delta \gtrsim d\theta/K$.

We then bound the entries of $\mathbf{M}$. We know $\mathbf{M}_{ij} = \theta_i\theta_j \sum_{k=1}^K \sum_{l=1}^K \boldsymbol{\pi}_i(k)\boldsymbol{\pi}_j(l)\mathbf{P}_{kl}$, and thus we have that

$$\mathbf{M}_{ij} \geq \theta_i\theta_j \sum_{k=1}^K \sum_{l=1}^K \boldsymbol{\pi}_i(k)\boldsymbol{\pi}_j(l) \min_{kl}(\mathbf{P}_{kl}) = \theta_i\theta_j \min_{kl}(\mathbf{P}_{kl}) \sum_{k=1}^K \sum_{l=1}^K \boldsymbol{\pi}_i(k)\boldsymbol{\pi}_j(l) = \theta_i\theta_j \min_{kl}(\mathbf{P}_{kl});$$
$$\mathbf{M}_{ij} \leq \theta_i\theta_j \sum_{k=1}^K \sum_{l=1}^K \boldsymbol{\pi}_i(k)\boldsymbol{\pi}_j(l) \max_{kl}(\mathbf{P}_{kl}) = \theta_i\theta_j \max_{kl}(\mathbf{P}_{kl}) \sum_{k=1}^K \sum_{l=1}^K \boldsymbol{\pi}_i(k)\boldsymbol{\pi}_j(l) = \theta_i\theta_j \max_{kl}(\mathbf{P}_{kl}).$$

Thus we can see that $\mathbf{M}_{ij} \asymp \theta$, $\max_{ij} \mathbb{E}(\mathbf{E}_{ij}^2) \lesssim \theta$ and $\max_i \sum_j \mathbb{E}(\mathbf{E}_{ij}^2) \lesssim d\theta$. By Theorem 3.1.4 in [11], we know that there exists some constant $c > 0$ such that for any $t > 0$,

$$\mathbb{P}\{\|\mathbf{E}\|_2 \geq 4\sqrt{d\theta} + t\} \leq d \exp\left(-t^2/c\right).$$

Also, since for $t \geq 5\sqrt{d\theta}$, there exists a constant $c > 0$ such that $\mathbb{P}(\|\mathbf{E}\|_2 \geq t) \leq \exp(-t^2/c)$, we have that $\|\|\mathbf{E}\|_2\|_{\psi_1} \lesssim \sqrt{d\theta}$, and hence we can take $r_1(d) = \sqrt{d\theta}$. Besides, $\sqrt{p/d}\Delta/r_1(d) = \sqrt{p\theta}/K \gtrsim d^{\epsilon/2}$, and hence by Theorem 4.1 we have

$$\left(\mathbb{E}|\rho(\widetilde{\mathbf{V}}^{\mathrm{F}}, \mathbf{V})|^2\right)^{1/2} \lesssim K\sqrt{\frac{K}{d\theta}} + K\sqrt{\frac{K}{pL\theta}} + \sqrt{\frac{Kd}{p'}}\left(\eta q^2 \frac{K}{\sqrt{p\theta}}\right)^q.$$

When

$$q = \log d \gg 1 + 2\epsilon^{-1} > \frac{\log\left(\sqrt{d/p'}\sqrt{d\theta}/K\right)}{\log\left(\sqrt{p/d}\sqrt{d\theta}/K\right)},$$

the third term is negligible and (4) holds.

REMARK 15. It's worth noting that here in Example 2 $\|\mathbf{E}\|_2$ converges faster than sub-Exponential random variables and $\|\mathbf{E}\|_2 \lesssim \sqrt{d\theta}$ with probability at least $1 - d^{-10}$, which we will take into account in later proofs.

REMARK 16. Under the case where no self-loops are present, $\mathbf{E}$ is replaced by $\mathbf{E}' = \mathbf{E} - \mathrm{diag}(\mathbf{X}) = \mathbf{E} - \mathrm{diag}(\mathbf{E}) - \mathrm{diag}(\mathbf{M})$. With similar arguments we can show that

$$\|\|\mathbf{E}'\|_2\|_{\psi_1} \lesssim \|\|\mathbf{E} - \mathrm{diag}(\mathbf{E})\|_2\|_{\psi_1} + \|\mathrm{diag}(\mathbf{M})\|_2 \lesssim \sqrt{d\theta} + \theta \lesssim \sqrt{d\theta},$$
$$\text{and} \quad \|\mathbf{E}'\|_2 \lesssim \|\mathbf{E} - \mathrm{diag}(\mathbf{E})\|_2 + \|\mathrm{diag}(\mathbf{M})\|_2 \lesssim \sqrt{d\theta} + \theta \lesssim \sqrt{d\theta},$$

with probability at least $1 - d^{-10}$, and hence (4) also holds for the no-self-loops case.

• **Example 3:** From the problem setting we know that we can represent $\boldsymbol{W}_i$ as $\boldsymbol{W}_i = \sum_{k=1}^K \mathbb{I}\{k_i = k\}\boldsymbol{\theta}_k^* + \boldsymbol{Z}_i$, where $\boldsymbol{Z}_i \overset{\text{i.i.d}}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$, $i \in [d]$. Denote $\mathbf{Z} = (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_d)$, then it can be seen that $\mathbb{E}(\mathbf{X}^\top\mathbf{X}) = \mathbb{E}(\mathbf{X})^\top\mathbb{E}(\mathbf{X}) + \mathbb{E}(\mathbf{Z}^\top\mathbf{Z}) = \mathbf{F}^*\boldsymbol{\Theta}^{*\top}\boldsymbol{\Theta}^*\mathbf{F}^{*\top} + n\mathbf{I}_d$, and we can write

$$\mathbf{E} = \mathbf{X}^\top\mathbf{X} - \mathbb{E}(\mathbf{X}^\top\mathbf{X}) = \mathbf{F}^*\boldsymbol{\Theta}^{*\top}\mathbf{Z} + \mathbf{Z}^\top\boldsymbol{\Theta}^*\mathbf{F}^{*\top} + \mathbf{Z}^\top\mathbf{Z} - n\mathbf{I}_d,$$

then we know that $\|\mathbf{E}\|_2 \leq 2\|\mathbf{F}^*\boldsymbol{\Theta}^{*\top}\mathbf{Z}\|_2 + n\|\mathbf{Z}^\top\mathbf{Z}/n - \mathbf{I}_d\|_2$. We consider $\|\mathbf{F}^*\boldsymbol{\Theta}^{*\top}\mathbf{Z}\|_2$ first. We know that $\widetilde{\mathbf{Z}} := \boldsymbol{\Theta}^{*\top}\mathbf{Z} = \boldsymbol{\Theta}^{*\top}(\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_d) = (\widetilde{\boldsymbol{Z}}_1, \ldots, \widetilde{\boldsymbol{Z}}_d) \in \mathbb{R}^{K\times d}$, where $\widetilde{\boldsymbol{Z}}_i \overset{\text{i.i.d}}{\sim} \mathcal{N}(\mathbf{0}, \boldsymbol{\Theta}^{*\top}\boldsymbol{\Theta}^*)$. Under the given conditions we know that $\|\boldsymbol{\Theta}^{*\top}\boldsymbol{\Theta}^*\|_2 \leq \Delta_0^2$. Since $(\boldsymbol{\Theta}^{*\top}\boldsymbol{\Theta}^*)^{-1/2}\widetilde{\mathbf{Z}}$ is a $K \times d$ i.i.d. Gaussian matrix, by Lemma B.2, we have that

$$\|\|\widetilde{\mathbf{Z}}\|_2\|_{\psi_1} \lesssim \|(\boldsymbol{\Theta}^{*\top}\boldsymbol{\Theta}^*)^{1/2}\|_2\|\|(\boldsymbol{\Theta}^{*\top}\boldsymbol{\Theta}^*)^{-1/2}\widetilde{\mathbf{Z}}\|_2\|_{\psi_1} \lesssim \Delta_0\sqrt{d}.$$

As for $\|\mathbf{Z}^\top\mathbf{Z}/n-\mathbf{I}_d\|_2$, when $n>d$, by Lemma 3 in Fan et al. [15] we know that $\|\|\mathbf{Z}^\top\mathbf{Z}/n-\mathbf{I}_d\|_2\|_{\psi_1}\lesssim\sqrt{d/n}$, and hence in summary we have

$$\|\|\mathbf{E}\|_2\|_{\psi_1}\lesssim\|\mathbf{F}^*\|_2\|\|\widetilde{\mathbf{Z}}\|_2\|_{\psi_1}+n\|\|\mathbf{Z}^\top\mathbf{Z}/n-\mathbf{I}_d\|_2\|_{\psi_1}\lesssim\Delta_0 d/\sqrt{K}+\sqrt{nd},$$

and we can take $r_1(d)=\Delta_0 d/\sqrt{K}+\sqrt{nd}$. We know that $\Delta=\sigma_{\min}(\mathbf{F}^*\mathbf{\Theta}^{*\top}\mathbf{\Theta}^*\mathbf{F}^{*\top})\gtrsim d\Delta_0^2/K$, and thus under the condition that $\Delta_0^2\geq CK(\log d)^2\left(d(\log d)^2/p\vee\sqrt{n/p}\right)$ for some large enough constant $C>0$, by Theorem 4.1 we have that

$$\left(\mathbb{E}|\rho(\widetilde{\mathbf{V}}^{\mathrm{F}},\mathbf{V})|^2\right)^{1/2}\lesssim\left(\frac{K}{\Delta_0}+\frac{K}{\Delta_0^2}\sqrt{\frac{Kn}{d}}\right)+\sqrt{\frac{d}{pL}}\left(\frac{K}{\Delta_0}+\frac{K}{\Delta_0^2}\sqrt{\frac{Kn}{d}}\right)$$
$$+\sqrt{\frac{Kd}{p'}}\left(\eta q^2\left(\sqrt{\frac{dK}{p\Delta_0^2}}+\frac{K}{\Delta_0^2}\sqrt{\frac{n}{p}}\right)\right)^q.$$

Now for the third term to be dominated by the bias term, we can take

$$q=\log d\geq\frac{\log\left(d/\sqrt{pp'}\right)}{\log\log d}+1\geq\frac{\log\left(d/\sqrt{pp'}\right)}{\log\left(\sqrt{\frac{p}{d}}\frac{\Delta}{r_1(d)}\right)}+1,$$

and hence (5) holds.

REMARK 17. In fact we can derive a slightly sharper tail bound for the convergence rate of $\|\mathbf{E}\|_2$. More specifically, for any $t\geq\Delta_0\sqrt{d}$, by Lemma 3 in Fan et al. [15] there exists some constant $c\geq 1$ such that

$$\begin{aligned}&\mathbb{P}\left(\|\widetilde{\mathbf{Z}}\|_2\geq t\right)=\mathbb{P}\left(\|\widetilde{\mathbf{Z}}\widetilde{\mathbf{Z}}^\top\|_2\geq t^2\right)=\mathbb{P}\left(d\|\widetilde{\mathbf{Z}}\widetilde{\mathbf{Z}}^\top/d-\mathbf{\Theta}^{*\top}\mathbf{\Theta}^*+\mathbf{\Theta}^{*\top}\mathbf{\Theta}^*\|_2\geq t^2\right)\\&\leq\mathbb{P}\left(d\|\widetilde{\mathbf{Z}}\widetilde{\mathbf{Z}}^\top/d-\mathbf{\Theta}^{*\top}\mathbf{\Theta}^*\|_2\geq t^2-d\|\mathbf{\Theta}^{*\top}\mathbf{\Theta}^*\|_2\right)\leq\mathbb{P}\left(\|\widetilde{\mathbf{Z}}\widetilde{\mathbf{Z}}^\top/d-\mathbf{\Theta}^{*\top}\mathbf{\Theta}^*\|_2\geq t^2/d-\Delta_0^2\right)\\&\leq\exp\left(-\frac{t^2/d-\Delta_0^2}{c\Delta_0^2\sqrt{K/d}}\right),\end{aligned}$$

which indicates that $\|\widetilde{\mathbf{Z}}\|_2\lesssim\Delta_0\sqrt{d}$ with probability at least $1-d^{-10}$. Hence under the condition that $\sqrt{K/d}\log d=O(1)$, with probability at least $1-O(d^{-10})$ we have that $\|\mathbf{E}\|_2\lesssim d\Delta_0/\sqrt{K}+\sqrt{dn}\log d$, which will be used as the statistical rate of $\|\mathbf{E}\|_2$ in later proofs.

• **Example 4:** We define $\bar{\mathcal{E}}=[\varepsilon_{ij}]$, then $\widehat{\mathbf{M}}=(1/\widehat{\theta})\mathcal{P}_{\mathcal{S}}(\mathbf{M}+\bar{\mathcal{E}})$, where $\mathcal{P}_{\mathcal{S}}$ is the projection onto the subspace of matrices with non-zero entries only in $\mathcal{S}$. Since $\widehat{\mathbf{M}}$ and $\widehat{\mathbf{M}}':=(\widehat{\theta}/\theta)\widehat{\mathbf{M}}=(1/\theta)\mathcal{P}_{\mathcal{S}}(\mathbf{M}+\bar{\mathcal{E}})$ differ only by a positive factor, $\widehat{\mathbf{M}}$ and $\widehat{\mathbf{M}}'$ share exactly the same sequence of eigenvectors and $\widetilde{\mathbf{V}}^{\mathrm{F}}$ can be viewed as the output by applying FADI to $\widehat{\mathbf{M}}'$. Thus we will establish the results for $\widehat{\mathbf{M}}'$ instead, and abuse the notation by denoting $\mathbf{E}:=\widehat{\mathbf{M}}'-\mathbf{M}$. We first study the order of $\|\mathbf{M}\|_{\max}$. When $\|\mathbf{V}\|_{2,\infty}\leq\sqrt{\mu K/d}$ for some rate $\mu\geq 1$ (that may change with d), for any $i,j\in[d]$, we have that

$$|\mathbf{M}_{ij}|=|\mathbf{e}_i^\top\mathbf{V}\mathbf{\Lambda}(\mathbf{e}_j^\top\mathbf{V})^\top|\leq\|\mathbf{\Lambda}\|_2\|\mathbf{e}_i^\top\mathbf{V}\|_2\|\mathbf{e}_j^\top\mathbf{V}\|_2\leq|\lambda_1|\|\mathbf{V}\|_{2,\infty}^2\leq\frac{|\lambda_1|\mu K}{d}.$$

Thus we have $\|\mathbf{M}\|_{\max}=O(|\lambda_1|\mu K/d)$. Also, we can write $\mathbf{E}=\mathbf{E}_1+\mathbf{E}_2$, where $(\mathbf{E}_1)_{ij}=\mathbf{M}_{ij}(\delta_{ij}-\theta)/\theta$, $(\mathbf{E}_2)_{ij}=\varepsilon_{ij}\delta_{ij}/\theta$, and for $i\leq j$

$$\operatorname{Var}\left((\mathbf{E}_1)_{ij}\right)=\mathbf{M}_{ij}^2(1-\theta)/\theta\leq\|\mathbf{M}\|_{\max}^2/\theta=O\left(\frac{(\lambda_1\mu K)^2}{d^2\theta}\right),\quad\operatorname{Var}\left((\mathbf{E}_2)_{ij}\right)=\sigma^2/\theta.$$

It is not hard to see that $\mathrm{Cov}((\mathbf{E}_1)_{ij}, (\mathbf{E}_2)_{ij}) = 0$. Also, we have that $|(\mathbf{E}_1)_{ij}| \le \|\mathbf{M}\|_{\max}/\theta = O(\frac{|\lambda_1|\mu K}{d\theta})$ and $|(\mathbf{E}_2)_{ij}| \le C\sigma \log d/\theta$ for all $i \le j$. Then we will study $\|\mathbf{E}_1\|_2$ and $\|\mathbf{E}_2\|_2$ separately. We denote $\nu_1 = d\|\mathbf{M}\|_{\max}^2/\theta$ and $\nu_2 = d\sigma^2/\theta$. Under the condition that $\theta \ge d^{-1/2+\epsilon}$ for some constant $\epsilon > 0$, by Theorem 3.1.4 in Chen et al. [11], there exists constant $c > 0$ such that for any $t \ge 4$ we have

$$\begin{aligned}\mathbb{P}\Big(\frac{\|\mathbf{E}_1\|_2}{2\sqrt{\nu_1}} \ge t\Big) &\le \mathbb{P}\big(\|\mathbf{E}_1\|_2/\sqrt{\nu_1} \ge 4 + t\big) = \mathbb{P}\big(\|\mathbf{E}_1\|_2 \ge 4\sqrt{\nu_1} + t\sqrt{\nu_1}\big)\\ &\le d\exp\Big(-\frac{t^2 d\|\mathbf{M}\|_{\max}^2/\theta}{c\|\mathbf{M}\|_{\max}^2/\theta^2}\Big) = \exp(-d\theta t^2/c + \log d)\\ &\le \exp(-\frac{d\theta t^2}{2c}) \le \exp(-t^2).\end{aligned}$$

Very similarly for $\|\mathbf{E}_2\|_2$, there exists $c' > 0$ such that for any $t \ge 4$, we have

$$\begin{aligned}&\mathbb{P}\Big(\frac{\|\mathbf{E}_2\|_2}{2\sqrt{\nu_2}} \ge t\Big) \le \mathbb{P}\big(\|\mathbf{E}_2\|_2 \ge 4\sqrt{\nu_2} + t\sqrt{\nu_2}\big) \le d\exp\Big(-\frac{t^2 d\sigma^2/\theta}{c'\sigma^2(\log d)^2/\theta^2}\Big)\\ &= \exp\Big(-\frac{d\theta t^2}{c'(\log d)^2} + \log d\Big) \le \exp\Big(-\frac{d\theta t^2}{2c'(\log d)^2}\Big) \le \exp(-t^2).\end{aligned}$$

Thus we can see that

$$\big\|\|\mathbf{E}\|_2\big\|_{\psi_1} \le \big\|\|\mathbf{E}_1\|_2\big\|_{\psi_1} + \big\|\|\mathbf{E}_2\|_2\big\|_{\psi_1} \lesssim \sqrt{\nu_1} + \sqrt{\nu_2} \lesssim \frac{|\lambda_1|\mu K}{\sqrt{d\theta}} + \sqrt{\frac{d\sigma^2}{\theta}}.$$

By Theorem 4.1, under the condition that $p = \Omega(\sqrt{d})$, $\sigma/\Delta \ll (\log d)^{-2} d^{-1}\sqrt{p\theta}$ and $\kappa_2\mu K \ll d^{1/4}$, it holds that

$$\begin{aligned}\Big(\mathbb{E}|\rho(\widetilde{\mathbf{V}}^{\mathrm{F}}, \mathbf{V})|_2^2\Big)^{1/2} &\lesssim \sqrt{K}\left(\frac{\kappa_2\mu K}{\sqrt{d\theta}} + \sqrt{\frac{d\sigma^2}{\Delta^2\theta}}\right) + K\sqrt{\frac{d}{pL}}\left(\frac{\kappa_2\mu K}{\sqrt{d\theta}} + \sqrt{\frac{d\sigma^2}{\Delta^2\theta}}\right)\\ &+ \sqrt{\frac{Kd}{p'}}\left(\eta q^2\left(\frac{\kappa_2\mu K}{\sqrt{p\theta}} + \sqrt{\frac{d^2\sigma^2}{p\Delta^2\theta}}\right)\right)^q.\end{aligned}$$

Furthermore, the third term vanishes when $q = \log d$ and (6) holds.

REMARK 18. Here we can also obtain a statistical rate sharper than subexponential rate for $\|\mathbf{E}\|_2$ that would be used in later proofs. Combining the above results for any $t \ge 16\max(\sqrt{\nu_1}, \sqrt{\nu_2})$ we have

$$\begin{aligned}\mathbb{P}\big(\|\mathbf{E}\|_2 \ge t\big) &\le \mathbb{P}\big(\|\mathbf{E}_1\|_2 \ge t/2\big) + \mathbb{P}\big(\|\mathbf{E}_2\|_2 \ge t/2\big) \le \exp(-\frac{d\theta t^2}{32c\nu_1}) + \exp\Big(-\frac{d\theta t^2}{32c'(\log d)^2\nu_2}\Big)\\ &= \exp\Big(-\frac{d^2\theta^2 t^2}{C_1(\lambda_1\mu K)^2}\Big) + \exp\Big(-\frac{\theta^2 t^2}{C_2(\log d)^2\sigma^2}\Big),\end{aligned}$$

where $C_1, C_2 > 0$ are constants. Thus $\|\mathbf{E}\|_2 \lesssim \frac{|\lambda_1|\mu K}{\sqrt{d\theta}} + \sqrt{\frac{d\sigma^2}{\theta}}$ with probability at least $1 - d^{-10}$.

B.4. Proof of Theorem 4.3. We first bound the recovery probability of $\widehat{K}^{(\ell)}$ for each $\ell \in [L]$. Recall that $\widehat{\mathbf{Y}}^{(\ell)}/\sqrt{p} = \mathbf{V}\mathbf{\Lambda}\widetilde{\mathbf{\Omega}}^{(\ell)}/\sqrt{p} + \mathbf{E}\mathbf{\Omega}^{(\ell)}/\sqrt{p}$, where $\widetilde{\mathbf{\Omega}}^{(\ell)} = \mathbf{V}^\top\mathbf{\Omega}^{(\ell)}$.

For the residual term $\mathbf{E}\mathbf{\Omega}^{(\ell)}/\sqrt{p}$, by Lemma 3 in [15], under the condition that $\sqrt{p/d}\log d = o(1)$, with probability at least $1 - d^{-10}$ we have $\|\mathbf{\Omega}^{(\ell)}/\sqrt{p}\|_2 \le 2\sqrt{\frac{d}{p}}$.

Denote by $\mathcal{A}_{\mathbf{E}}$ the event $\{\|\mathbf{E}\|_2 \le 10c_e^{-1}r_1(d)\log d\}$, where $c_e > 0$ is the constant defined in Remark 2. Then conditional on $\mathcal{A}_{\mathbf{E}}$, we have that $\|\mathbf{E}\boldsymbol{\Omega}^{(\ell)}/\sqrt{p}\|_2 \le 20c_e^{-1}\sqrt{\frac{d}{p}}r_1(d)\log d$ with probability at least $1-d^{-10}$ for each $\ell \in [L]$. Recall $\eta_0 = 480c_e^{-1}\sqrt{\frac{d}{\Delta^2 p}}r_1(d)\log d$. From Proposition 10.4 in [18], we know that when $p \ge 2K$,

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

Therefore, with probability at least $1-\eta_0^{(p-K+1)/2}$,

$$\sigma_{\min}(\mathbf{V}\boldsymbol{\Lambda}\widetilde{\boldsymbol{\Omega}}^{(\ell)}/\sqrt{p}) \ge \Delta\sigma_{\min}(\widetilde{\boldsymbol{\Omega}}^{(\ell)}/\sqrt{p}) > \Delta\sqrt{\eta_0}/6 \ge 2\mu_0.$$

By Weyl's inequality [17], we know that conditional on $\mathcal{A}_{\mathbf{E}}$, with probability at least $1-d^{-10}$, $\sigma_{K+1}(\widehat{\mathbf{Y}}^{(\ell)}/\sqrt{p}) \le \|\mathbf{E}\boldsymbol{\Omega}^{(\ell)}/\sqrt{p}\|_2 \le 20c_e^{-1}\sqrt{\frac{d}{p}}r_1(d) = \Delta\eta_0/24 \le \mu_0$ for large enough d, which indicates that $\sigma_{k+1}(\widehat{\mathbf{Y}}^{(\ell)}) - \sigma_p(\widehat{\mathbf{Y}}^{(\ell)}) < \sqrt{p}\mu_0$ for any $k \ge K$. For $k \le K-1$, under the same event we have

$$\begin{aligned}\sigma_{k+1}(\widehat{\mathbf{Y}}^{(\ell)}) - \sigma_p(\widehat{\mathbf{Y}}^{(\ell)}) &\ge \sigma_K(\widehat{\mathbf{Y}}^{(\ell)}) - \sigma_p(\widehat{\mathbf{Y}}^{(\ell)}) \ge \sigma_{\min}(\mathbf{V}\boldsymbol{\Lambda}\widetilde{\boldsymbol{\Omega}}^{(\ell)}) - 2\|\mathbf{E}\boldsymbol{\Omega}^{(\ell)}\|_2 \\ &> \sqrt{p}(\Delta\sqrt{\eta_0}/6 - \Delta\eta_0/12) \ge \sqrt{p}(\Delta\sqrt{\eta_0}/6 - \Delta\sqrt{\eta_0}/12) = \Delta\sqrt{p\eta_0}/12 \ge \sqrt{p}\mu_0.\end{aligned}$$

Then we have

$$\begin{aligned}\mathbb{P}\left(\widehat{K}^{(\ell)} = K \mid \mathcal{A}_{\mathbf{E}}\right) &\ge \mathbb{P}\left(\sigma_K(\widehat{\mathbf{Y}}^{(\ell)}) - \sigma_p(\widehat{\mathbf{Y}}^{(\ell)}) > \sqrt{p}\mu_0,\ \sigma_{K+1}(\widehat{\mathbf{Y}}^{(\ell)}) - \sigma_p(\widehat{\mathbf{Y}}^{(\ell)}) \le \sqrt{p}\mu_0 \,\middle|\, \mathcal{A}_{\mathbf{E}}\right) \\ &\ge \mathbb{P}\left(\sigma_{\min}(\mathbf{V}\boldsymbol{\Lambda}\widetilde{\boldsymbol{\Omega}}^{(\ell)}/\sqrt{p}) \ge \Delta\sqrt{\eta_0}/6, \quad \|\mathbf{E}\boldsymbol{\Omega}^{(\ell)}/\sqrt{p}\|_2 \le \Delta\eta_0/24 \,\middle|\, \mathcal{A}_{\mathbf{E}}\right) \\ &\ge 1 - d^{-10} - \eta_0^{\frac{p-K+1}{2}}.\end{aligned}$$

We know that conditional on $\mathbf{E}$, $\mathbb{I}\{\widehat{K}^{(\ell)} \ne K \mid \mathcal{A}_{\mathbf{E}}\}$ are i.i.d. Bernoulli variables with expectation $p_K := \mathbb{P}(\widehat{K}^{(\ell)} \ne K \mid \mathcal{A}_{\mathbf{E}}) \le d^{-10} + \eta_0^{\frac{p-K+1}{2}} \le 1/4$ 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

$$\begin{aligned}\mathbb{P}(\widehat{K} \ne K) &\le \mathbb{P}\left(\sum_{\ell=1}^L \mathbb{I}\left\{\widehat{K}^{(\ell)} \ne K\right\} - p_K L \ge \frac{L}{4}\right) = \mathbb{E}_{\mathbf{E}}\left(\mathbb{P}\left(\sum_{\ell=1}^L \mathbb{I}\left\{\widehat{K}^{(\ell)} \ne K\right\} - p_K L \ge \frac{L}{4} \,\middle|\, \mathbf{E}\right)\right) \\ &\le \mathbb{P}(\mathcal{A}_{\mathbf{E}})\exp\left\{-(L/4)^2/(2Lp_K)\right\} + 1 - \mathbb{P}(\mathcal{A}_{\mathbf{E}}) \\ &\le \exp\left\{-L/\left(32d^{-10} + 32\eta_0^{\frac{p-K+1}{2}}\right)\right\} + O(d^{-10}).\end{aligned}$$

We know that $32d^{-10} \le (\log d)^{-1}$ for $d \ge 2$, and under the condition that $\eta_0 \le (32\log d)^{-\frac{2}{p-K+1}}$ we have $\mathbb{P}(\widehat{K} \ne K) \le \exp(-L\log d/2) + O(d^{-10}) \lesssim d^{-(L\wedge 20)/2}$.

B.5. Proof of Corollary 4.4. • **Example 1**: From the proof of Corollary 4.2 we know that we can take $r_1(d) = (\lambda_1 + \sigma^2)\sqrt{\frac{r}{n}}\log d$. Then by plugging in each term we know that under the condition that $(\lambda_1 + \sigma^2)\left(d(np)^{-1/2}\log d\right)^{1/4} = o(1)$ and $\Delta \gg \left(\sigma^{-2}(np)^{-1/2}d\log d\right)^{1/3}$, we have $\Delta\eta_0/24 \ll \mu_0 \ll \Delta\sqrt{\eta_0}/12$. Besides, under the condition that $\kappa_1\sqrt{dr/(np)}(\log d)^2 = o(1)$, we also have $\eta_0 \le (32\log d)^{-\frac{2}{p-K+1}}$. Thus the conditions for Theorem 4.3 are satisfied and we have $\widehat{K} = K$ with probaility at least $1 - O(d^{-(L\wedge 20)/2})$.
• **Example 2:** We know from the proof of Corollary 4.2 that $\Delta \gtrsim d\theta/K$. Also from Remark 15 we know that $\|\mathbf{E}\|_2 \lesssim \sqrt{d\theta}$ with probability at least $1-d^{-10}$, and thus we have $\eta_0 \asymp \sqrt{d/(\Delta^2 p)}\sqrt{d\theta} \lesssim K/\sqrt{p\theta} \asymp$

$1/\sqrt{d^{\epsilon-1/2}p}$, $\Delta\eta_0 \asymp d\sqrt{\theta/p}$ and $\Delta\sqrt{\eta_0} \gtrsim d\theta^{3/4}p^{-1/4}K^{-1/2}$. Also recall from the proof of Corollary 4.2 that $\mathbb{E}(\widehat{\mathbf{M}}_{ij}) = \mathbf{M}_{ij} \asymp \theta$ for any $i,j \in [d]$, and hence $d^{-2}\sum_{i\le j}\mathbf{M}_{ij} \asymp \theta$. By Hoeffding's inequality [19], we have that

$$\mathbb{P}\left(\frac{2}{d(d-1)}\left|\sum_{i\le j}\widehat{\mathbf{M}}_{ij} - \sum_{i\le j}\mathbf{M}_{ij}\right| \ge \frac{\sqrt{11\log d}}{d}\right) \lesssim \exp\left(-11d(d-1)\log d/d^2\right) \lesssim d^{-10}.$$

Thus we can see with probability at least $1-O(d^{-10})$, $|\widehat{\theta} - d^{-2}\sum_{i\le j}\mathbf{M}_{ij}| \lesssim \frac{\sqrt{\log d}}{d}$ and $\widehat{\theta} \asymp \theta$, and in turn $\Delta\eta_0/24 \ll \mu_0 \ll \Delta\sqrt{\eta_0}/12$. Thus by Theorem 4.3 the claim follows.

• **Example 3:** We know from the proof of Corollary 4.2 and Remark 17 that $\Delta \gtrsim d\Delta_0^2/K$ and $\|\mathbf{E}\|_2 \lesssim d\Delta_0/\sqrt{K} + \sqrt{dn}\log d$ with probability at least $1-d^{-10}$. Thus we have $\eta_0 \asymp \sqrt{d/(\Delta^2 p)}\big(d\Delta_0/\sqrt{K} + \sqrt{dn}\log d\big)$. Under the condition that $\sqrt{K(\log d)^3}\,(n/p)^{1/4} \ll \Delta_0 \ll \sqrt{nK/d}\log d$, we know that $d\Delta_0/\sqrt{K} + \sqrt{dn}\log d \lesssim \sqrt{dn}\log d$, $\Delta\eta_0 \asymp d\sqrt{n/p}\log d$ and $\sqrt{\eta_0}\log d = o(1)$, and thus $\Delta\eta_0/24 \ll \mu_0 \ll \Delta\sqrt{\eta_0}/12$. By Theorem 4.3 the claim follows.

• **Example 4:** By Hoeffding's inequality [19], with probability at least $1-d^{-10}$ we have that $|\widehat{\theta}-\theta|/\widehat{\theta} \le C\sqrt{\log d}/d\theta$. As for $\widehat{\sigma}_0^2$, we have

$$\widehat{\sigma}_0^2 = \frac{1}{|\mathcal{S}|}\sum_{(i,j)\in\mathcal{S}}(\widehat{\theta}\widehat{\mathbf{M}}_{ij})^2 = \frac{1}{|\mathcal{S}|}\left(\sum_{i\le j}\delta_{ij}\mathbf{M}_{ij}^2 + 2\sum_{(i,j)\in\mathcal{S}}\mathbf{M}_{ij}\varepsilon_{ij} + \sum_{(i,j)\in\mathcal{S}}\varepsilon_{ij}^2\right).$$

We consider the latter two terms first. We know that $|\varepsilon_{ij}| \le C\sigma\log d$ for some constant $C>0$ and $|\mathbf{M}_{ij}| \le |\lambda_1|\mu K/d$, for any $i\le j$. Denote by $\widetilde{\sigma} = (|\lambda_1|\mu K/d)\vee\sigma$, then we have

$$\mathrm{Var}(\mathrm{M}_{ij}\varepsilon_{ij}) \le (\frac{|\lambda_1|\mu K}{d})^2\sigma^2 \le \widetilde{\sigma}^4, \quad |\mathrm{M}_{ij}\varepsilon_{ij}| \le \frac{|\lambda_1|\mu K}{d}C\sigma\log d \le C\widetilde{\sigma}^2\log d, \quad \forall i\le j,$$

and

$$\mathrm{Var}(\varepsilon_{ij}^2) \le C^4\sigma^4(\log d)^4 \le C^4\widetilde{\sigma}^4(\log d)^4, \quad |\varepsilon_{ij}^2| \le C^2\sigma^2(\log d)^2 \le C^2\widetilde{\sigma}^2(\log d)^2, \quad i\le j.$$

Thus by Bernstein inequality [7], conditional on $\mathcal{S}$, with probability at least $1-2d^{-10}$ we have that there exists a constant $C'>0$ independent of $\mathcal{S}$ such that

$$\left|\frac{1}{|\mathcal{S}|}\sum_{(i,j)\in\mathcal{S}}\mathbf{M}_{ij}\varepsilon_{ij}\right| \le C'\left(\frac{\widetilde{\sigma}^2\sqrt{\log d}}{\sqrt{|\mathcal{S}|}} + \frac{\widetilde{\sigma}^2(\log d)^2}{|\mathcal{S}|}\right), \tag{B.27}$$

and

$$\left|\frac{1}{|\mathcal{S}|}\sum_{(i,j)\in\mathcal{S}}\varepsilon_{ij}^2 - \sigma^2\right| \le C'\left(\frac{\widetilde{\sigma}^2(\log d)^{5/2}}{\sqrt{|\mathcal{S}|}} + \frac{\widetilde{\sigma}^2(\log d)^3}{|\mathcal{S}|}\right). \tag{B.28}$$

Now we consider the first term. Since δ_{ij}'s are i.i.d. Bernoulli random variables with expectation θ, we have

$$\mathrm{Var}(\mathbf{M}_{ij}^2\delta_{ij}) \le \theta\widetilde{\sigma}^4, \quad |\mathbf{M}_{ij}^2\delta_{ij}| \le \widetilde{\sigma}^2, \quad i\le j.$$

Also, we know that $\sum_{i\le j}\mathbf{M}_{ij}^2 \ge \|\mathbf{M}\|_{\mathrm{F}}^2/2 \ge K\Delta^2/2$ and $\sum_{i\le j}\mathbf{M}_{ij}^2 \le \|\mathbf{M}\|_{\mathrm{F}}^2 \le K\lambda_1^2$, and hence $K\Delta^2\theta/2 \le \mathbb{E}\Big(\sum_{i\le j}\delta_{ij}\mathbf{M}_{ij}^2\Big) \le K\lambda_1^2\theta$. Then by Bernstein inequality [7] with probability at least $1-d^{-10}$, it holds that

$$\left|\Big(\sum_{i\le j}\delta_{ij}\mathbf{M}_{ij}^2\Big) - \mathbb{E}\Big(\sum_{i\le j}\delta_{ij}\mathbf{M}_{ij}^2\Big)\right| \lesssim d^2\left(\frac{\widetilde{\sigma}^2\sqrt{\theta\log d}}{d} + \frac{\widetilde{\sigma}^2\log d}{d^2}\right) = \widetilde{\sigma}^2(d\sqrt{\theta\log d} + \log d). \tag{B.29}$$

Thus combining (B.27), (B.28) and (B.29) with the fact that $|\mathcal{S}| \asymp d^2\theta$ with probability at least $1 - d^{-10}$, under the condition that $\kappa_2^2\mu^2 K \ll (\log d)^2$, with probability at least $1 - O(d^{-10})$ we have

$$\widetilde{\sigma} \ll \left(\frac{\Delta\sqrt{K}\log d}{d} \vee \sigma\right) + o(\widetilde{\sigma}) \lesssim \widehat{\sigma}_0 \log d \lesssim \left(\frac{|\lambda_1|\sqrt{K}\log d}{d} \vee \sigma\right) + o(\widetilde{\sigma}) \lesssim \widetilde{\sigma}\log d.$$

From the proof of Corollary 4.2 and Remark 18, we know that with probability at least $1 - d^{-10}$,

$$\|\widehat{\mathbf{M}} - \mathbf{M}\|_2 \lesssim \left\|\frac{\widehat{\theta}}{\theta}\widehat{\mathbf{M}} - \widehat{\mathbf{M}}\right\|_2 + \left\|\frac{\widehat{\theta}}{\theta}\widehat{\mathbf{M}} - \mathbf{M}\right\|_2 \lesssim \frac{|\lambda_1|\sqrt{\log d}}{d\theta} + \frac{|\lambda_1|\mu K}{\sqrt{d\theta}} + \sqrt{\frac{d\sigma^2}{\theta}} \lesssim \sqrt{\frac{d\widetilde{\sigma}^2}{\theta}},$$

and hence $\eta_0 \asymp d\widetilde{\sigma}(\Delta\sqrt{p\theta})^{-1}$ and $\Delta\eta_0 \asymp d\widetilde{\sigma}/\sqrt{p\theta}$.

Under the condition that $(p\theta)^{-1/4}\sqrt{d\sigma/\Delta}\log d = o(1)$, with probability at least $1 - O(d^{-10})$ we have $\Delta\eta_0/24 \ll \mu_0 \ll \Delta\sqrt{\eta_0}/12$. Thus by Theorem 4.3 the claim follows.

B.6. Proof of Theorem 4.10. We first decompose $\widetilde{\mathbf{V}}^{\mathrm{F}}\mathbf{H} - \mathbf{V} = \widetilde{\mathbf{V}}^{\mathrm{F}}\mathbf{H} - \widetilde{\mathbf{V}}\mathbf{H}_0 + \widetilde{\mathbf{V}}\mathbf{H}_0 - \mathbf{V}$, and we consider the term $\widetilde{\mathbf{V}}\mathbf{H}_0 - \mathbf{V}$ first.

By Lemma 8 in Fan et al. [15], we have that $\|\widetilde{\mathbf{V}}\mathbf{H}_0 - \mathbf{V} - \mathbf{P}_{\perp}(\widetilde{\boldsymbol{\Sigma}} - \mathbf{V}\mathbf{V}^\top)\mathbf{V}\|_2 \lesssim \|\widetilde{\boldsymbol{\Sigma}} - \mathbf{V}\mathbf{V}^\top\|_2\|\mathbf{P}_{\perp}(\widetilde{\boldsymbol{\Sigma}} - \mathbf{V}\mathbf{V}^\top)\mathbf{V}\|_2$. Note that in Lemma 8 of Fan et al. [15], the norm is Frobenius norm rather than operator norm, and the modification from Frobenius norm to operator norm is trivial and hence omitted. We first study the leading term $\mathbf{P}_{\perp}(\widetilde{\boldsymbol{\Sigma}} - \mathbf{V}\mathbf{V}^\top)\mathbf{V} = \frac{1}{L}\sum_{\ell=1}^{L}\mathbf{P}_{\perp}(\widehat{\mathbf{V}}^{(\ell)}\widehat{\mathbf{V}}^{(\ell)\top} - \mathbf{V}\mathbf{V}^\top)\mathbf{V}$.

For a given $\ell \in [L]$, we know that $\widehat{\mathbf{V}}^{(\ell)}$ is the top K left singular vectors of $\widehat{\mathbf{Y}}^{(\ell)} = \widehat{\mathbf{M}}\boldsymbol{\Omega}^{(\ell)}/\sqrt{p} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^\top\boldsymbol{\Omega}^{(\ell)}/\sqrt{p} + \mathbf{E}\boldsymbol{\Omega}^{(\ell)}/\sqrt{p} = \mathbf{Y}^{(\ell)} + \mathcal{E}^{(\ell)}$, where

$$\mathbf{Y}^{(\ell)} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^\top\boldsymbol{\Omega}^{(\ell)}/\sqrt{p} \quad \text{and} \quad \mathcal{E}^{(\ell)} = \mathbf{E}\boldsymbol{\Omega}^{(\ell)}/\sqrt{p}.$$

By the "symmetric dilation" trick, we denote

$$\mathcal{S}(\widehat{\mathbf{Y}}^{(\ell)}) = \begin{pmatrix} \mathbf{0} & \widehat{\mathbf{Y}}^{(\ell)} \\ \widehat{\mathbf{Y}}^{(\ell)\top} & \mathbf{0} \end{pmatrix}, \quad \mathcal{S}(\mathbf{Y}^{(\ell)}) = \begin{pmatrix} \mathbf{0} & \mathbf{Y}^{(\ell)} \\ \mathbf{Y}^{(\ell)\top} & \mathbf{0} \end{pmatrix},$$

$$\text{and} \quad \mathcal{S}(\mathcal{E}^{(\ell)}) = \mathcal{S}(\widehat{\mathbf{Y}}^{(\ell)}) - \mathcal{S}(\mathbf{Y}^{(\ell)}) = \begin{pmatrix} \mathbf{0} & \mathbf{E}\boldsymbol{\Omega}^{(\ell)}/\sqrt{p} \\ \boldsymbol{\Omega}^{(\ell)\top}\mathbf{E}/\sqrt{p} & \mathbf{0} \end{pmatrix}.$$

We let $\boldsymbol{\Gamma}_K^{(\ell)}\boldsymbol{\Lambda}_K^{(\ell)}\mathbf{U}_K^{(\ell)\top}$ be the SVD of $\mathbf{Y}^{(\ell)}$, and we know that with probability 1 we have $\boldsymbol{\Gamma}_K^{(\ell)} = \mathbf{V}\mathbf{O}_{\boldsymbol{\Omega}^{(\ell)}}$, where $\mathbf{O}_{\boldsymbol{\Omega}^{(\ell)}}$ is an orthonormal matrix depending on $\boldsymbol{\Omega}^{(\ell)}$. It is not hard to verify that the eigen-decomposition of $\mathcal{S}(\mathbf{Y}^{(\ell)})$ is:

$$\mathcal{S}(\mathbf{Y}^{(\ell)}) = \frac{1}{\sqrt{2}}\begin{pmatrix} \boldsymbol{\Gamma}_K^{(\ell)} & \boldsymbol{\Gamma}_K^{(\ell)} \\ \mathbf{U}_K^{(\ell)} & -\mathbf{U}_K^{(\ell)} \end{pmatrix} \cdot \begin{pmatrix} \boldsymbol{\Lambda}_K^{(\ell)} & \mathbf{0} \\ \mathbf{0} & -\boldsymbol{\Lambda}_K^{(\ell)} \end{pmatrix} \cdot \frac{1}{\sqrt{2}}\begin{pmatrix} \boldsymbol{\Gamma}_K^{(\ell)} & \boldsymbol{\Gamma}_K^{(\ell)} \\ \mathbf{U}_K^{(\ell)} & -\mathbf{U}_K^{(\ell)} \end{pmatrix}^\top,$$

where $\boldsymbol{\Lambda}_K^{(\ell)} = \mathrm{diag}(\lambda_1^{(\ell)}, \ldots, \lambda_K^{(\ell)})$. First we study the eigengap $\sigma_{\min}(\boldsymbol{\Lambda}_K^{(\ell)}) = \lambda_K^{(\ell)}$. Recall $\widetilde{\boldsymbol{\Omega}}^{(\ell)} = \mathbf{V}^\top\boldsymbol{\Omega}^{(\ell)} \in \mathbb{R}^{K\times p}$, and it can be seen that the entries of $\widetilde{\boldsymbol{\Omega}}^{(\ell)}$ are i.i.d. standard Gaussian. By Lemma 3 in Fan et al. [15], we know that with probability at least $1 - d^{-10}$, we have that $\|\widetilde{\boldsymbol{\Omega}}^{(\ell)}\widetilde{\boldsymbol{\Omega}}^{(\ell)\top}/p - \mathbf{I}_K\|_2 \lesssim \sqrt{\frac{K}{p}}\log d$, and thus $\sigma_{\min}(\widetilde{\boldsymbol{\Omega}}^{(\ell)}/\sqrt{p}) \geq 1 - O(\sqrt{\frac{K}{p}}\log d)$ with probability at least $1 - d^{-10}$. Thus under the condition that $\sqrt{\frac{K}{p}}\log d = o(1)$, under the same high probability event we have that $\sigma_{\min}(\boldsymbol{\Lambda}_K^{(\ell)}) \geq \Delta/2$. Now we let $\widehat{\mathbf{U}}_K^{(\ell)}$ be the top K right singular vectors of $\widehat{\mathbf{Y}}^{(\ell)}$. For $j \in [K]$ we define

$$\mathbf{G}_j^{(\ell)} = \frac{1}{2}\begin{pmatrix} \boldsymbol{\Gamma}_K^{(\ell)} \\ -\mathbf{U}_K^{(\ell)} \end{pmatrix}(-\boldsymbol{\Lambda}_K^{(\ell)} - \lambda_j^{(\ell)}\mathbf{I}_K)^{-1}\begin{pmatrix} \boldsymbol{\Gamma}_K^{(\ell)} \\ -\mathbf{U}_K^{(\ell)} \end{pmatrix}^\top - \frac{1}{\lambda_j^{(\ell)}}\left\{\mathbf{I}_K - \frac{1}{2}\begin{pmatrix} \boldsymbol{\Gamma}_K^{(\ell)} & \boldsymbol{\Gamma}_K^{(\ell)} \\ \mathbf{U}_K^{(\ell)} & -\mathbf{U}_K^{(\ell)} \end{pmatrix}\begin{pmatrix} \boldsymbol{\Gamma}_K^{(\ell)} & \boldsymbol{\Gamma}_K^{(\ell)} \\ \mathbf{U}_K^{(\ell)} & -\mathbf{U}_K^{(\ell)} \end{pmatrix}^\top\right\}.$$

Then we have $\|\mathbf{G}_j^{(\ell)}\|_2 \le 1/\lambda_K^{(\ell)} \le 2/\Delta$ with probability at least $1-d^{-10}$. Correspondingly we define the linear mapping

$$f: \mathbb{R}^{(d+p)\times K} \to \mathbb{R}^{(d+p)\times K}, \quad (\mathbf{w}_1,\cdots,\mathbf{w}_K) \mapsto \left(-\mathbf{G}_1^{(\ell)}\mathbf{w}_1,\cdots,-\mathbf{G}_K^{(\ell)}\mathbf{w}_K\right),$$

and denote $\widetilde{\boldsymbol{\Gamma}}_K^{(\ell)} = \begin{pmatrix}\boldsymbol{\Gamma}_K^{(\ell)}\\ \mathbf{U}_K^{(\ell)}\end{pmatrix}$. By Lemma 8 in Fan et al. [15], under the condition that $\|\mathcal{S}(\mathcal{E}^{(\ell)})\|_2/\Delta = o(1)$ we have

$$\begin{aligned}
&\left\| \begin{pmatrix}\widehat{\mathbf{V}}^{(\ell)}\\ \widehat{\mathbf{U}}_K^{(\ell)}\end{pmatrix}(\widehat{\mathbf{V}}^{(\ell)\top},\widehat{\mathbf{U}}_K^{(\ell)\top}) - \widetilde{\boldsymbol{\Gamma}}_K^{(\ell)}\widetilde{\boldsymbol{\Gamma}}_K^{(\ell)\top} - f\big(\mathcal{S}(\mathcal{E}^{(\ell)})\widetilde{\boldsymbol{\Gamma}}_K^{(\ell)}\big)\widetilde{\boldsymbol{\Gamma}}_K^{(\ell)\top} - \widetilde{\boldsymbol{\Gamma}}_K^{(\ell)} f\big(\mathcal{S}(\mathcal{E}^{(\ell)})\widetilde{\boldsymbol{\Gamma}}_K^{(\ell)}\big)^\top \right\|_2 \\
&\le \left\| \begin{pmatrix}\widehat{\mathbf{V}}^{(\ell)}\widehat{\mathbf{V}}^{(\ell)\top} - \boldsymbol{\Gamma}_K^{(\ell)}\boldsymbol{\Gamma}_K^{(\ell)\top} & \widehat{\mathbf{V}}^{(\ell)}\widehat{\mathbf{U}}_K^{(\ell)\top} - \boldsymbol{\Gamma}_K^{(\ell)}\mathbf{U}_K^{(\ell)\top}\\ \widehat{\mathbf{U}}_K^{(\ell)}\widehat{\mathbf{V}}^{(\ell)\top} - \mathbf{U}_K^{(\ell)}\boldsymbol{\Gamma}_K^{(\ell)\top} & \widehat{\mathbf{U}}_K^{(\ell)}\widehat{\mathbf{U}}_K^{(\ell)\top} - \mathbf{U}_K^{(\ell)}\mathbf{U}_K^{(\ell)\top}\end{pmatrix}\right. \\
&\qquad \left. - f\big(\mathcal{S}(\mathcal{E}^{(\ell)})\widetilde{\boldsymbol{\Gamma}}_K^{(\ell)}\big)\widetilde{\boldsymbol{\Gamma}}_K^{(\ell)\top} - \widetilde{\boldsymbol{\Gamma}}_K^{(\ell)} f\big(\mathcal{S}(\mathcal{E}^{(\ell)})\widetilde{\boldsymbol{\Gamma}}_K^{(\ell)}\big)^\top \right\|_2 \lesssim \|\mathcal{S}(\mathcal{E}^{(\ell)})\|_2^2/\Delta^2,
\end{aligned}$$

By taking the upper left block of the matrix, we have

$$\begin{aligned}
&\big\|\widehat{\mathbf{V}}^{(\ell)}\widehat{\mathbf{V}}^{(\ell)\top} - \boldsymbol{\Gamma}_K^{(\ell)}\boldsymbol{\Gamma}_K^{(\ell)\top} - f\big(\mathcal{S}(\mathcal{E}^{(\ell)})\widetilde{\boldsymbol{\Gamma}}_K^{(\ell)}\big)_{[1:d,:]}\boldsymbol{\Gamma}_K^{(\ell)\top} - \boldsymbol{\Gamma}_K^{(\ell)} f\big(\mathcal{S}(\mathcal{E}^{(\ell)})\widetilde{\boldsymbol{\Gamma}}_K^{(\ell)}\big)_{[1:d,:]}^\top\big\|_2 \\
&= \big\|\widehat{\mathbf{V}}^{(\ell)}\widehat{\mathbf{V}}^{(\ell)\top} - \mathbf{V}\mathbf{V}^\top - f\big(\mathcal{S}(\mathcal{E}^{(\ell)})\widetilde{\boldsymbol{\Gamma}}_K^{(\ell)}\big)_{[1:d,:]}\boldsymbol{\Gamma}_K^{(\ell)\top} - \boldsymbol{\Gamma}_K^{(\ell)} f\big(\mathcal{S}(\mathcal{E}^{(\ell)})\widetilde{\boldsymbol{\Gamma}}_K^{(\ell)}\big)_{[1:d,:]}^\top\big\|_2 \\
&\lesssim \|\mathcal{S}(\mathcal{E}^{(\ell)})\|_2^2/\Delta^2.
\end{aligned}$$

Now for $j \in [K]$, we study $\mathbf{P}_\perp(\mathbf{G}_j^{(\ell)})_{[1:d,:]}$. Since $\boldsymbol{\Gamma}_K^{(\ell)} = \mathbf{V}\mathbf{O}_{\boldsymbol{\Omega}^{(\ell)}}$, we have $\mathbf{P}_\perp\boldsymbol{\Gamma}_K^{(\ell)} = \mathbf{0}$. Therefore we have,

$$\begin{aligned}
&\mathbf{P}_\perp\boldsymbol{\Gamma}_K^{(\ell)}(-\boldsymbol{\Lambda}_K^{(\ell)} - \lambda_j^{(\ell)}\mathbf{I}_K)^{-1}\begin{pmatrix}\boldsymbol{\Gamma}_K^{(\ell)}\\ -\mathbf{U}_K^{(\ell)}\end{pmatrix}^\top = \mathbf{0}, \quad \text{and} \\
&\mathbf{P}_\perp\left\{\mathbf{I}_{d+p} - \frac{1}{2}\begin{pmatrix}\boldsymbol{\Gamma}_K^{(\ell)} & \boldsymbol{\Gamma}_K^{(\ell)}\\ \mathbf{U}_K^{(\ell)} & -\mathbf{U}_K^{(\ell)}\end{pmatrix}\begin{pmatrix}\boldsymbol{\Gamma}_K^{(\ell)} & \boldsymbol{\Gamma}_K^{(\ell)}\\ \mathbf{U}_K^{(\ell)} & -\mathbf{U}_K^{(\ell)}\end{pmatrix}^\top\right\}_{[1:d,:]} = (\mathbf{P}_\perp,\mathbf{0}) - \frac{1}{2}\mathbf{P}_\perp\boldsymbol{\Gamma}_K^{(\ell)}(\mathbf{I}_d,\mathbf{I}_d)\begin{pmatrix}\boldsymbol{\Gamma}_K^{(\ell)} & \boldsymbol{\Gamma}_K^{(\ell)}\\ \mathbf{U}_K^{(\ell)} & -\mathbf{U}_K^{(\ell)}\end{pmatrix}^\top \\
&= (\mathbf{P}_\perp,\mathbf{0}) + \mathbf{0} = (\mathbf{P}_\perp,\mathbf{0}),
\end{aligned}$$

and as a result we have

$$\mathbf{P}_\perp(\mathbf{G}_j)_{[1:d,:]} = \frac{1}{2}\cdot\mathbf{0} - \frac{1}{\lambda_j^{(\ell)}}\{(\mathbf{P}_\perp,\mathbf{0}) - \mathbf{0}\} = -\frac{1}{\lambda_j^{(\ell)}}(\mathbf{P}_\perp,\mathbf{0}).$$

Thus in turn,

$$\begin{aligned}
&\mathbf{P}_\perp\Big(f\big(\mathcal{S}(\mathcal{E}^{(\ell)})\widetilde{\boldsymbol{\Gamma}}_K^{(\ell)}\big)_{[1:d,:]}\boldsymbol{\Gamma}_K^{(\ell)\top} + \boldsymbol{\Gamma}_K^{(\ell)} f\big(\mathcal{S}(\mathcal{E}^{(\ell)})\widetilde{\boldsymbol{\Gamma}}_K^{(\ell)}\big)_{[1:d,:]}^\top\Big) = \mathbf{P}_\perp f\big(\mathcal{S}(\mathcal{E}^{(\ell)})\widetilde{\boldsymbol{\Gamma}}_K^{(\ell)}\big)_{[1:d,:]}\boldsymbol{\Gamma}_K^{(\ell)\top} \\
&= (\mathbf{P}_\perp,\mathbf{0})\begin{pmatrix}\mathbf{0} & \mathbf{E}\boldsymbol{\Omega}^{(\ell)}/\sqrt{p}\\ \boldsymbol{\Omega}^{(\ell)\top}\mathbf{E}/\sqrt{p} & \mathbf{0}\end{pmatrix}\begin{pmatrix}\boldsymbol{\Gamma}_K^{(\ell)}\\ \mathbf{U}_K^{(\ell)}\end{pmatrix}(\boldsymbol{\Lambda}_K^{(\ell)})^{-1}\boldsymbol{\Gamma}_K^{(\ell)\top} \\
&= \mathbf{P}_\perp\mathbf{E}(\boldsymbol{\Omega}^{(\ell)}/\sqrt{p})\mathbf{U}_K^{(\ell)}(\boldsymbol{\Lambda}_K^{(\ell)})^{-1}\boldsymbol{\Gamma}_K^{(\ell)\top} = \mathbf{P}_\perp\mathbf{E}(\boldsymbol{\Omega}^{(\ell)}/\sqrt{p})(\mathbf{Y}^{(\ell)\top})^\dagger.
\end{aligned}$$

For a given $\ell \in [L]$, under the condition that $\sqrt{p/d}\log d = O(1)$, by Lemma 3 in Fan et al. [15] we have that with probability at least $1-d^{-10}$, $\|\boldsymbol{\Omega}^{(\ell)}\|_2 \lesssim \sqrt{d}$. Combined with previous results on the eigengap

$\sigma_{\min}(\mathbf{\Lambda}_K^{(\ell)})$, we have that with probability $1-O(d^{-9})$, for a fixed constant $C>0$

$$\|\mathbf{\Omega}^{(\ell)}\|_2 \le C\sqrt{d}, \quad \sigma_{\min}(\mathbf{\Lambda}_K^{(\ell)}) \ge \Delta/2, \quad \forall \ell \in [L].$$

Besides, under Assumption 1, we have that $\|\mathbf{E}\|_2 \lesssim r_1(d)\log d$ with probability at least $1-d^{-10}$, and in turn by Wedin's Theorem [39], with high probability for all $\ell \in [L]$ we have that

$$\|\widehat{\mathbf{V}}^{(\ell)}\widehat{\mathbf{V}}^{(\ell)\top} - \mathbf{V}\mathbf{V}^\top\|_2 \lesssim \|\mathcal{E}^{(\ell)}\|_2/\sigma_{\min}(\mathbf{\Lambda}_K^{(\ell)}) \lesssim \|\mathbf{E}\|_2\|\mathbf{\Omega}^{(\ell)}/\sqrt{p}\|_2/\Delta \lesssim \frac{r_1(d)}{\Delta}\log d\sqrt{\frac{d}{p}},$$

and thus $\|\widetilde{\mathbf{\Sigma}} - \mathbf{V}\mathbf{V}^\top\|_2 = O_P\big(r_1(d)\log d\sqrt{d/p}/\Delta\big)$. Besides, we have

$$\begin{aligned}
\mathbf{P}_\perp(\widetilde{\mathbf{\Sigma}} - \mathbf{V}\mathbf{V}^\top)\mathbf{V} &= \frac{1}{L}\sum_{\ell=1}^{L}\mathbf{P}_\perp(\widehat{\mathbf{V}}^{(\ell)}\widehat{\mathbf{V}}^{(\ell)\top} - \mathbf{V}\mathbf{V}^\top)\mathbf{V} \\
&= \frac{1}{L}\sum_{\ell=1}^{L}\mathbf{P}_\perp\Big(f\big(\mathcal{S}(\mathcal{E}^{(\ell)})\widetilde{\mathbf{\Gamma}}_K^{(\ell)}\big)_{[1:d,:]}\mathbf{\Gamma}_K^{(\ell)\top} + \mathbf{\Gamma}_K^{(\ell)} f\big(\mathcal{S}(\mathcal{E}^{(\ell)})\widetilde{\mathbf{\Gamma}}_K^{(\ell)}\big)_{[1:d,:]}^\top\Big)\mathbf{V} + \mathbf{R}_1(\widetilde{\mathbf{\Sigma}}) \\
&= \frac{1}{L}\sum_{\ell=1}^{L}\mathbf{P}_\perp\mathbf{E}(\mathbf{\Omega}^{(\ell)}/\sqrt{p})(\mathbf{Y}^{(\ell)\top})^\dagger\mathbf{V} + R_1(\widetilde{\mathbf{\Sigma}}) = \frac{1}{L}\sum_{\ell=1}^{L}\mathbf{P}_\perp\mathbf{E}(\mathbf{\Omega}^{(\ell)}/\sqrt{p})\mathbf{B}^{(\ell)\top} + \mathbf{R}_1(\widetilde{\mathbf{\Sigma}}) \\
&= \frac{1}{L}\mathbf{P}_\perp\mathbf{E}\mathbf{\Omega}\mathbf{B}_\Omega + \mathbf{R}_1(\widetilde{\mathbf{\Sigma}}),
\end{aligned}$$

where $\mathbf{R}_1(\widetilde{\mathbf{\Sigma}})$ is the residual matrix with $\|\mathbf{R}_1(\widetilde{\mathbf{\Sigma}})\|_2 = O_P(\|\mathcal{S}(\mathcal{E}^{(\ell)})\|_2^2/\Delta^2)$. Now we study the matrix $\mathbf{B}^{(\ell)} = (\mathbf{\Lambda}\mathbf{V}^\top\mathbf{\Omega}^{(\ell)}/\sqrt{p})^\dagger$. From previous results we know that with probability at least $1-d^{-9}$, $1/2 \le \sigma_{\min}(\widetilde{\mathbf{\Omega}}^{(\ell)}/\sqrt{p}) \le \sigma_{\max}(\widetilde{\mathbf{\Omega}}^{(\ell)}/\sqrt{p}) \le 3/2$ for any $\ell \in [L]$, and in turn $\frac{2}{3|\lambda_1|} \le \sigma_{\min}(\mathbf{B}^{(\ell)}) \le \sigma_{\max}(\mathbf{B}^{(\ell)}) \le \frac{2}{\Delta}, \quad \forall \ell \in [L]$. Now for any vector $\mathbf{y} \in \mathbb{R}^K$ such that $\|\mathbf{y}\|_2 = 1$, with probability $1-O(d^{-9})$ we have that

$$\|\mathbf{B}_{\mathbf{\Omega}}\mathbf{y}\|_2 = \|(\mathbf{y}^\top\mathbf{B}^{(1)\top}, \ldots, \mathbf{y}^\top\mathbf{B}^{(L)\top})^\top\|_2 = \Big(\sum_{\ell=1}^{L}\|\mathbf{B}^{(\ell)}\mathbf{y}\|_2^2\Big)^{1/2},$$

$$\|\mathbf{B}_{\mathbf{\Omega}}\|_2 = \max_{\|\mathbf{y}\|_2=1}\|\mathbf{B}_{\mathbf{\Omega}}\mathbf{y}\|_2 = \max_{\|\mathbf{y}\|_2=1}\Big(\sum_{\ell=1}^{L}\|\mathbf{B}^{(\ell)}\mathbf{y}\|_2^2\Big)^{1/2} \le \Big(\sum_{\ell=1}^{L}\|\mathbf{B}^{(\ell)}\|_2^2\Big)^{1/2} \le \frac{2\sqrt{L}}{\Delta},$$

$$\sigma_{\min}(\mathbf{B}_{\mathbf{\Omega}}) = \min_{\|\mathbf{y}\|_2=1}\|\mathbf{B}_{\mathbf{\Omega}}\mathbf{y}\|_2 = \min_{\|\mathbf{y}\|_2=1}\Big(\sum_{\ell=1}^{L}\|\mathbf{B}^{(\ell)}\mathbf{y}\|_2^2\Big)^{1/2} \ge \Big(\sum_{\ell=1}^{L}\sigma_{\min}^2(\mathbf{B}^{(\ell)})\Big)^{1/2} \ge \frac{2\sqrt{L}}{3|\lambda_1|}.$$

Now since we know that the entries of $\sqrt{p}\mathbf{\Omega}$ are i.i.d. standard Gaussian, similar as before, under the condition that $Lp \ll d$, by Lemma 3 in Fan et al. [15] we have with high probability that $\frac{1}{2}\sqrt{\frac{d}{p}} \le \sigma_{\min}(\mathbf{\Omega}) \le \sigma_{\max}(\mathbf{\Omega}) \le \frac{3}{2}\sqrt{\frac{d}{p}}$. Therefore, we have the following upper bound on the norm of the leading term

$$\begin{aligned}
\|\mathbf{P}_\perp(\widetilde{\mathbf{\Sigma}} - \mathbf{V}\mathbf{V}^\top)\mathbf{V}\|_2 &\lesssim \|\frac{1}{L}\mathbf{P}_\perp\mathbf{E}\mathbf{\Omega}\mathbf{B}_\Omega\|_2 + \|\mathbf{R}_1(\widetilde{\mathbf{\Sigma}})\|_2 \le \frac{1}{L}\|\mathbf{E}\|_2\|\mathbf{\Omega}\|_2\|\mathbf{B}_{\mathbf{\Omega}}\|_2 + \|\mathbf{R}_1(\widetilde{\mathbf{\Sigma}})\|_2 \\
&= O_P\Big(\sqrt{\frac{d}{Lp}}\frac{r_1(d)\log d}{\Delta} + r_1(d)^2(\log d)^2\frac{d}{p\Delta^2}\Big).
\end{aligned}$$

Thus we have the following decomposition

$$\widetilde{\mathbf{V}}\mathbf{H}_0 - \mathbf{V} = \mathbf{P}_\perp(\widetilde{\mathbf{\Sigma}} - \mathbf{V}\mathbf{V}^\top)\mathbf{V} + \mathbf{R}_0(\widetilde{\mathbf{\Sigma}})$$

$$= \frac{1}{L}\mathbf{P}_\perp\mathbf{E}\boldsymbol{\Omega}\mathbf{B}_\Omega + \mathbf{R}_1(\widetilde{\boldsymbol{\Sigma}}) + \mathbf{R}_0(\widetilde{\boldsymbol{\Sigma}})$$
$$= \frac{1}{L}\mathbf{P}_\perp\mathbf{E}_0\boldsymbol{\Omega}\mathbf{B}_\Omega + \frac{1}{L}\mathbf{P}_\perp\mathbf{E}_b\boldsymbol{\Omega}\mathbf{B}_\Omega + \mathbf{R}_1(\widetilde{\boldsymbol{\Sigma}}) + \mathbf{R}_0(\widetilde{\boldsymbol{\Sigma}}),$$

where $\mathbf{R}_0(\widetilde{\boldsymbol{\Sigma}})$ is a residual matrix with

$$\begin{aligned}\|\mathbf{R}_0(\widetilde{\boldsymbol{\Sigma}})\|_2 &= O_P(\|\widetilde{\boldsymbol{\Sigma}} - \mathbf{V}\mathbf{V}^\top\|_2\|\mathbf{P}_\perp(\widetilde{\boldsymbol{\Sigma}} - \mathbf{V}\mathbf{V}^\top)\mathbf{V}\|_2)\\ &= O_P\Big(\frac{r_1(d)^2(\log d)^2 d}{\sqrt{L}p\Delta^2}\Big) + o_P\Big(r_1(d)^2(\log d)^2\frac{d}{p\Delta^2}\Big).\end{aligned}$$

Thus

$$\|\mathbf{R}_0(\widetilde{\boldsymbol{\Sigma}}) + \mathbf{R}_1(\widetilde{\boldsymbol{\Sigma}})\|_2 = O_P\Big(r_1(d)^2(\log d)^2\frac{d}{p\Delta^2}\Big).$$

Next we consider the term $\widetilde{\mathbf{V}}^{\mathrm{F}}\mathbf{H} - \widetilde{\mathbf{V}}\mathbf{H}_0$. We denote the SVD of $\widetilde{\boldsymbol{\Sigma}}$ by $\widetilde{\mathbf{V}}\widetilde{\boldsymbol{\Lambda}}_K^q\widetilde{\mathbf{V}}^\top + \widetilde{\mathbf{V}}_\perp\widetilde{\boldsymbol{\Lambda}}_\perp^q\widetilde{\mathbf{V}}_\perp^\top$, and by Weyl's inequality [17], we know that $\|\widetilde{\boldsymbol{\Lambda}}_\perp\| \le \|\widetilde{\boldsymbol{\Sigma}} - \mathbf{V}\mathbf{V}^\top\| = O_P\big(r_1(d)\log d\sqrt{d/p}/\Delta\big)$ and $\sigma_K(\widetilde{\boldsymbol{\Lambda}}_K) \ge 1 - \|\widetilde{\boldsymbol{\Sigma}} - \mathbf{V}\mathbf{V}^\top\| \ge 1 - O_P(r_1(d)\log d\sqrt{d/p}/\Delta)$. Thus under the condition that $r_1(d)\log d\sqrt{d/p}/\Delta = o(1)$, for large enough d with high probability we have

$$\|\widetilde{\boldsymbol{\Lambda}}_\perp^q\| \le (r_1(d)\log d\sqrt{d/p}/\Delta)^q \quad \text{and} \quad \sigma_K(\widetilde{\boldsymbol{\Lambda}}_K^q) \ge (1 - O(r_1(d)\log d\sqrt{d/p}/\Delta))^q \ge (1/2)^q.$$

Similar as before, we know that with probability 1 the left singular vector space of $\widetilde{\mathbf{V}}\widetilde{\boldsymbol{\Lambda}}_K^q\widetilde{\mathbf{V}}^\top\boldsymbol{\Omega}^{\mathrm{F}} = \widetilde{\mathbf{V}}\widetilde{\boldsymbol{\Lambda}}_K^q\widetilde{\boldsymbol{\Omega}}^{\mathrm{F}}$ and the column space of $\widetilde{\mathbf{V}}$ are the same, where $\widetilde{\boldsymbol{\Omega}}^{\mathrm{F}} := \widetilde{\mathbf{V}}^\top\boldsymbol{\Omega}^{\mathrm{F}} \in \mathbb{R}^{K\times p'}$ is still a Gaussian test matrix with i.i.d. entries. By Lemma 3 in Fan et al. [15], we have with probability at least $1 - d^{-10}$, $\sigma_{\min}(\widetilde{\boldsymbol{\Omega}}^{\mathrm{F}}/\sqrt{p'}) \ge 1 - O(\sqrt{\frac{K}{p'}}\log d)$. When $\sqrt{\frac{K}{p'}}\log d = o(1)$, by Wedin's Theorem [39], there exists a constant $\eta > 0$ such that with high probability we have

$$\begin{aligned}\|\widetilde{\mathbf{V}}^{\mathrm{F}}\mathbf{H} - \widetilde{\mathbf{V}}\mathbf{H}_0\|_2 &= \|\widetilde{\mathbf{V}}^{\mathrm{F}}\mathbf{H}_1 - \widetilde{\mathbf{V}}\|_2 \lesssim \|\widetilde{\mathbf{V}}_\perp\widetilde{\boldsymbol{\Lambda}}_\perp^q\widetilde{\mathbf{V}}_\perp^\top\boldsymbol{\Omega}^{\mathrm{F}}/\sqrt{p'}\|_2/\sigma_K(\widetilde{\mathbf{V}}\widetilde{\boldsymbol{\Lambda}}_K^q\widetilde{\boldsymbol{\Omega}}^{\mathrm{F}}/\sqrt{p'})\\ &\le \frac{\|\widetilde{\boldsymbol{\Lambda}}_\perp\|_2^q\|\boldsymbol{\Omega}^{\mathrm{F}}/\sqrt{p'}\|_2}{\sigma_K(\widetilde{\boldsymbol{\Lambda}}_K^q)\sigma_K(\widetilde{\boldsymbol{\Omega}}^{\mathrm{F}}/\sqrt{p'})} \lesssim \left(\frac{2\eta r_1(d)\log d\sqrt{d/p}}{\Delta}\right)^q\sqrt{\frac{d}{p'}}.\end{aligned}$$

Denote $r' := 2\eta r_1(d)\log d\sqrt{d/p}/\Delta = o\big((\log d)^{-1/4}\big)$. Then it can be seen that when

$$q \ge \log d \gg 2 + \frac{\log d}{\log\log d} \ge 2 + \frac{\log\sqrt{d/p'}}{\log(1/r')},$$

we have that $(r')^q\sqrt{d/p'} = o\big((r')^2\big)$ and $\|\widetilde{\mathbf{V}}^{\mathrm{F}}\mathbf{H} - \widetilde{\mathbf{V}}\mathbf{H}_0\|_2 = O_P\Big(r_1(d)^2(\log d)^2\frac{d}{p\Delta^2}\Big)$.

Now for $i \in [d]$, recall that with high probability $\sigma_{\min}(\boldsymbol{\Sigma}_i) = \Omega\big(\eta_2(d)\big)$. Therefore, under the condition that $d^2r_1(d)^4(\log d)^4\big(p^2\Delta^4\eta_2(d)\big)^{-1} = o(1)$ and $dr_2(d)^2\big(Lp\Delta^2\eta_2(d)\big)^{-1} = o(1)$, we have with probability $1 - O(d^{-9})$, $\|\frac{1}{L}\mathbf{P}_\perp\mathbf{E}_b\boldsymbol{\Omega}\mathbf{B}_\Omega\|_2 = O_P\Big(\sqrt{\frac{d}{\Delta^2Lp}}r_2(d)\Big) = o_P\big((\sigma_{\min}(\boldsymbol{\Sigma}_i))^{1/2}\big)$, and $\|\mathbf{R}_0(\widetilde{\boldsymbol{\Sigma}}) + \mathbf{R}_1(\widetilde{\boldsymbol{\Sigma}})\|_2 = o_P\big((\sigma_{\min}(\boldsymbol{\Sigma}_i))^{1/2}\big)$. Then under Assumption 5, we have

$$\begin{aligned}&\boldsymbol{\Sigma}_i^{-1/2}(\widetilde{\mathbf{V}}^{\mathrm{F}}\mathbf{H} - \mathbf{V})^\top\mathbf{e}_i = \boldsymbol{\Sigma}_i^{-1/2}(\widetilde{\mathbf{V}}^{\mathrm{F}}\mathbf{H} - \widetilde{\mathbf{V}}\mathbf{H}_0 + \widetilde{\mathbf{V}}\mathbf{H}_0 - \mathbf{V})^\top\mathbf{e}_i\\ &= \boldsymbol{\Sigma}_i^{-1/2}(\frac{1}{L}\mathbf{B}_\Omega^\top\boldsymbol{\Omega}^\top\mathbf{E}_0\mathbf{P}_\perp\mathbf{e}_i) + \boldsymbol{\Sigma}_i^{-1/2}(\widetilde{\mathbf{V}}^{\mathrm{F}}\mathbf{H} - \widetilde{\mathbf{V}}\mathbf{H}_0 + \mathbf{R}_0(\widetilde{\boldsymbol{\Sigma}}) + \mathbf{R}_1(\widetilde{\boldsymbol{\Sigma}}) + \frac{1}{L}\mathbf{P}_\perp\mathbf{E}_b\boldsymbol{\Omega}\mathbf{B}_\Omega)^\top\mathbf{e}_i\\ &= \boldsymbol{\Sigma}_i^{-1/2}\mathbf{V}(\mathbf{E}_0)^\top\mathbf{e}_i + o_P(1) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{I}_K).\end{aligned}$$

B.7. Proof of Corollary 4.11. To prove Corollary 4.11, it suffices for us to show that Assumptions 1, 2 and 5 are met. From the proof of Corollary 4.2, we know that Assumption 1 is satisfied. We move on to show that Assumption 2 is met. Define $\mathbf{V}_d = (\mathbf{V}, \mathbf{V}^{\perp})$ as the stacking of eigenvectors for the covariance matrix $\mathbf{\Sigma}$. Note that $\mathbf{V}^{\perp}$ is not identifiable under the spiked covariance model and is unique up to orthogonal transformation. Let $\boldsymbol{Z}_i = \mathbf{V}_d^{\top}\boldsymbol{X}_i$, and $\boldsymbol{Z}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{\Lambda}_d)$, where $\mathbf{\Lambda}_d = \mathrm{diag}(\mathbf{\Lambda} + \sigma^2\mathbf{I}_K, \sigma^2\mathbf{I}_{d-K})$. We let $\mathbf{\Gamma}_S = (\mathbf{u}_1, \dots, \mathbf{u}_{K+1})$ be the stacking of eigenvectors for the matrix $\mathbf{\Sigma}_S$, and let $\widetilde{\sigma}_1 \geq \widetilde{\sigma}_2 \geq \dots \geq \widetilde{\sigma}_{K+1}$ be the $K+1$ eigenvalues of $\mathbf{\Sigma}_S$. Correspondingly, let $\widehat{\sigma}_1 \geq \dots \geq \widehat{\sigma}_{K+1} = \widehat{\sigma}^2$ be the eigenvalues of the sample covariance matrix $\widehat{\mathbf{\Sigma}}_S$. Since $\mathbf{\Sigma}_S = (\mathbf{V})_{[S,:]}\mathbf{\Lambda}(\mathbf{V})_{[S,:]}^{\top} + \sigma^2\mathbf{I}_{K+1}$, we know that $\widetilde{\sigma}_{K+1} = \sigma^2$ and $\delta = \widetilde{\sigma}_K - \widetilde{\sigma}_{K+1} \geq \Delta\sigma^2_{\min}\big((\mathbf{V})_{[S,:]}\big)$. We define $\widetilde{\mathbf{c}} = (\mathbf{V}^{\perp}_{[S,:]})^{\top}\mathbf{u}_{K+1}$, and denote $\widetilde{\mathbf{c}}_0 = (\mathbf{0}, \mathbf{I}_{d-K})^{\top}\widetilde{\mathbf{c}} \in \mathbb{R}^d$. Then by the proof of Lemma 6.2 in Wang and Fan [38], we know that

$$\widehat{\sigma}^2 - \sigma^2 = \widetilde{\mathbf{c}}_0^{\top}(\frac{1}{n}\sum_{i=1}^n \boldsymbol{Z}_i\boldsymbol{Z}_i^{\top} - \mathbf{\Lambda}_d)\widetilde{\mathbf{c}}_0 + \frac{1}{n}O_P\left(M_{K+1} - \sigma^2 W_{K+1}\right),$$

where $M_{K+1} = \sum_{k\leq K} f_k^2\left(\widetilde{\sigma}_k + (\widehat{\sigma}_k - \widetilde{\sigma}_k)\right), W_{K+1} = \sum_{k\leq K} f_k^2$ and f_k is the $(K+1)$-th element of the k-th eigenvector of $\mathbf{\Gamma}_S^{\top}\widehat{\mathbf{\Sigma}}_S\mathbf{\Gamma}_S$ multiplied by $\sqrt{n}$ for $k \leq K$. We let $\mathbf{f} = (f_1, \dots, f_K)^{\top}/\sqrt{n}$. By Wedin's Theorem [39] and Lemma 3 in Fan et al. [15], we have that with probability at least $1 - d^{-10}$, $|\widehat{\sigma}_k - \widetilde{\sigma}_k| \leq \|\widehat{\mathbf{\Sigma}}_S - \mathbf{\Sigma}_S\|_2 \lesssim \widetilde{\sigma}_1 \log d\sqrt{\frac{K}{n}}$ for $k \leq K$. If we denote by $\mathbf{F}_S := (\mathbf{I}_K, \mathbf{0})^{\top}$ the stacked top K eigenvectors of $\mathbf{\Gamma}_S^{\top}\mathbf{\Sigma}_S\mathbf{\Gamma}_S$, and by $\widehat{\mathbf{F}}_S$ the stacked top K eigenvectors of $\mathbf{\Gamma}_S^{\top}\widehat{\mathbf{\Sigma}}_S\mathbf{\Gamma}_S$, then we know that $\mathbf{f}$ is the $(K+1)$-th row of $\widehat{\mathbf{F}}_S$. By Davis-Kahan's Theorem [42], we also know that there exists an orthonormal matrix $\mathbf{O}_S \in \mathbb{R}^{K\times K}$ such that $\|\mathbf{f}\|_2 = \|\mathbf{O}_S^{\top}\mathbf{f} - \mathbf{0}\|_2 \leq \|\widehat{\mathbf{F}}_S\mathbf{O}_S - \mathbf{F}_S\|_2 \lesssim \frac{\widetilde{\sigma}_1\log d}{\delta}\sqrt{\frac{K}{n}}$, and thus

$$W_{K+1} = \sum_{k\leq K} f_k^2 = n\|\mathbf{f}\|_2^2 \lesssim \frac{\widetilde{\sigma}_1^2 K(\log d)^2}{\delta^2},$$

$$\text{and} \quad M_{K+1} \leq \widetilde{\sigma}_1\sum_{k\leq K} f_k^2 + (\sum_{k\leq K} f_k^2)\|\widehat{\mathbf{\Sigma}}_S - \mathbf{\Sigma}_S\|_2 \lesssim \frac{\widetilde{\sigma}_1^3 K}{\delta^2}(\log d)^2.$$

Thus we can write $\widehat{\sigma}^2 - \sigma^2 = \widetilde{\mathbf{c}}_0^{\top}(\frac{1}{n}\sum_{i=1}^n \boldsymbol{Z}_i\boldsymbol{Z}_i^{\top} - \mathbf{\Lambda}_d)\widetilde{\mathbf{c}}_0 + O_P\big(\frac{\widetilde{\sigma}_1^3 K}{\delta^2 n}(\log d)^2\big)$.

Now we take $\mathbf{E}_0 = \widehat{\mathbf{\Sigma}} - \mathbf{\Sigma} - (\widetilde{\mathbf{c}}_0^{\top}(\frac{1}{n}\sum_{i=1}^n \boldsymbol{Z}_i\boldsymbol{Z}_i^{\top} - \mathbf{\Lambda}_d)\widetilde{\mathbf{c}}_0)\mathbf{I}_d$, and from previous results we know that with high probability $\|\mathbf{E}_b\|_2 = \|\mathbf{E} - \mathbf{E}_0\|_2 \lesssim \frac{\widetilde{\sigma}_1^3 K}{\delta^2 n}(\log d)^2$, such that we have $r_2(d) \asymp \frac{\widetilde{\sigma}_1^3 K}{\delta^2 n}(\log d)^2 = o(r_1(d))$ and Assumption 2 is satisfied.

Now we move on to study the statistical rate $\eta_2(d)$. For any $i \in [d]$, we first study the covariance of $\mathbf{E}_0\mathbf{P}_{\perp}\mathbf{e}_i$. We denote $\widetilde{\mathbf{E}} = \boldsymbol{Z}_1\boldsymbol{Z}_1^{\top} - \mathbf{\Lambda}_d$, then it's not hard to verify that $\mathrm{Cov}(\widetilde{\mathbf{E}}_{ij}, \widetilde{\mathbf{E}}_{gh}) = \lambda_i(\mathbf{\Sigma})\lambda_j(\mathbf{\Sigma})(\mathbb{I}\{i = g, j = h\} + \mathbb{I}\{i = h, j = g\})$. Since $\mathbf{E}_0\mathbf{P}_{\perp}\mathbf{e}_i$ and $\mathbf{V}_d^{\top}\mathbf{E}_0\mathbf{P}_{\perp}\mathbf{e}_i$ share the same eigenvalues, we can study the covariance of $\mathbf{V}_d^{\top}\mathbf{E}_0\mathbf{P}_{\perp}\mathbf{e}_i$ instead. Then $\mathrm{Cov}(\mathbf{V}_d^{\top}\mathbf{E}_0\mathbf{P}_{\perp}\mathbf{e}_i)$ can be calculated as following

$$\begin{aligned}
&\mathrm{Cov}\left\{\mathbf{V}_d^{\top}(\frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i\boldsymbol{X}_i^{\top} - \mathbf{\Sigma})\mathbf{V}^{\perp}(\mathbf{V}^{\perp})^{\top}\mathbf{e}_i - (\widetilde{\mathbf{c}}_0^{\top}(\frac{1}{n}\sum_{i=1}^n \boldsymbol{Z}_i\boldsymbol{Z}_i^{\top} - \mathbf{\Lambda}_d)\widetilde{\mathbf{c}}_0)\mathbf{V}_d^{\top}\mathbf{P}_{\perp}\mathbf{e}_i\right\} \\
&= \mathrm{Cov}\left\{\mathbf{V}_d^{\top}(\frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i\boldsymbol{X}_i^{\top} - \mathbf{\Sigma})\mathbf{V}_d(\mathbf{0}, \mathbf{I}_{d-K})^{\top}\widetilde{\mathbf{e}} - (\widetilde{\mathbf{c}}_0^{\top}(\frac{1}{n}\sum_{i=1}^n \boldsymbol{Z}_i\boldsymbol{Z}_i^{\top} - \mathbf{\Lambda}_d)\widetilde{\mathbf{c}}_0)\widetilde{\mathbf{e}}_0\right\} \\
&= \mathrm{Cov}\left\{(\frac{1}{n}\sum_{i=1}^n \boldsymbol{Z}_i\boldsymbol{Z}_i^{\top} - \mathbf{\Lambda}_d)\widetilde{\mathbf{e}}_0 - (\widetilde{\mathbf{c}}_0^{\top}(\frac{1}{n}\sum_{i=1}^n \boldsymbol{Z}_i\boldsymbol{Z}_i^{\top} - \mathbf{\Lambda}_d)\widetilde{\mathbf{c}}_0)\widetilde{\mathbf{e}}_0\right\},
\end{aligned}$$

where $\widetilde{\mathbf{e}}=(\mathbf{V}^{\perp})^{\top}\mathbf{e}_i$ and $\widetilde{\mathbf{e}}_0=(\mathbf{0},\mathbf{I}_{d-K})^{\top}\widetilde{\mathbf{e}}$. Then we have

$$
\begin{aligned}
\mathrm{Cov}(\mathbf{V}_d^{\top}\mathbf{E}_0\mathbf{P}_{\perp}\mathbf{e}_i) &= \frac{1}{n}\,\mathrm{Cov}(\widetilde{\mathbf{E}}\widetilde{\mathbf{e}}_0-\widetilde{\mathbf{c}}_0^{\top}\widetilde{\mathbf{E}}\widetilde{\mathbf{c}}_0\widetilde{\mathbf{e}}_0)\\
&= \frac{1}{n}\Big\{\,\mathrm{Cov}(\widetilde{\mathbf{E}}\widetilde{\mathbf{e}}_0)+\mathrm{Var}\left(\widetilde{\mathbf{c}}_0^{\top}\widetilde{\mathbf{E}}\widetilde{\mathbf{c}}_0\right)\widetilde{\mathbf{e}}_0\widetilde{\mathbf{e}}_0^{\top}-\mathrm{Cov}(\widetilde{\mathbf{E}}\widetilde{\mathbf{e}}_0,\widetilde{\mathbf{c}}_0^{\top}\widetilde{\mathbf{E}}\widetilde{\mathbf{c}}_0)\widetilde{\mathbf{e}}_0^{\top}-\widetilde{\mathbf{e}}_0\,\mathrm{Cov}(\widetilde{\mathbf{E}}\widetilde{\mathbf{e}}_0,\widetilde{\mathbf{c}}_0^{\top}\widetilde{\mathbf{E}}\widetilde{\mathbf{c}}_0)^{\top}\Big\}\\
&= \frac{1}{n}\{\|\widetilde{\mathbf{e}}_0\|_2^2\sigma^2\mathbf{\Lambda}_d+3\sigma^4\widetilde{\mathbf{e}}_0\widetilde{\mathbf{e}}_0^{\top}-2\sigma^4\langle\widetilde{\mathbf{c}},\widetilde{\mathbf{e}}\rangle(\widetilde{\mathbf{c}}_0\widetilde{\mathbf{e}}_0^{\top}+\widetilde{\mathbf{e}}_0\widetilde{\mathbf{c}}_0^{\top})\}.
\end{aligned}
$$

Thus it can be seen that the covariance matrix is block-diagonal:

$$
\mathrm{Cov}(\mathbf{V}_d^{\top}\mathbf{E}_0\mathbf{P}_{\perp}\mathbf{e}_i)=\frac{1}{n}\begin{pmatrix}\|\widetilde{\mathbf{e}}_0\|_2^2\sigma^2(\mathbf{\Lambda}+\sigma^2\mathbf{I}_K) & \mathbf{0}\\ \mathbf{0} & \|\widetilde{\mathbf{e}}_0\|_2^2\sigma^4(\mathbf{I}_{d-K}+3\tau_1\tau_1^{\top}-2\rho\widetilde{\mathbf{c}}\tau_1^{\top}-2\rho\tau_1\widetilde{\mathbf{c}}^{\top})\end{pmatrix},
$$

where $\tau_1=\widetilde{\mathbf{e}}/\|\widetilde{\mathbf{e}}\|_2$ and $\rho=\langle\widetilde{\mathbf{c}},\tau_1\rangle$. Then following basic algebra, we can write $\mathrm{Cov}(\mathbf{E}_0\mathbf{P}_{\perp}\mathbf{e}_i)$ as:

$$
\frac{1}{n}\Big\{\sigma^2\|\widetilde{\mathbf{e}}_0\|_2^2\mathbf{\Sigma}+3\sigma^4\mathbf{P}_{\perp}\mathbf{e}_i\mathbf{e}_i^{\top}\mathbf{P}_{\perp}-2\sigma^4\rho\|\widetilde{\mathbf{e}}_0\|_2\big[(\mathbf{P}_{\perp})_{[:,S]}\mathbf{u}_{K+1}\mathbf{e}_i^{\top}\mathbf{P}_{\perp}+\mathbf{P}_{\perp}\mathbf{e}_i(\mathbf{u}_{K+1})^{\top}(\mathbf{P}_{\perp})_{[S,:]}\big]\Big\}.
$$

To study $\eta_2(d)$, we will first define $\mathbf{\Sigma}_i'$ as following

$$
\mathbf{\Sigma}_i'=\frac{1}{nL^2}\mathbf{B}_{\mathbf{\Omega}}^{\top}\mathbf{\Omega}^{\top}\Big\{\sigma^2\mathbf{\Sigma}+3\sigma^4\mathbf{e}_i\mathbf{e}_i^{\top}-2\sigma^4\rho\|\widetilde{\mathbf{e}}_0\|_2\big((\mathbf{I}_d)_{[:,S]}\mathbf{u}_{K+1}\mathbf{e}_i^{\top}+\mathbf{e}_i\mathbf{u}_{K+1}^{\top}(\mathbf{I}_d)_{[S,:]}\big)\Big\}\mathbf{\Omega}\mathbf{B}_{\mathbf{\Omega}}.
$$

We know that $\|\widetilde{\mathbf{e}}_0\|_2^2=\|\mathbf{P}_{\perp}\mathbf{e}_i\|_2^2=1-O(\mu K/d)$, thus we have

$$
\Big\|\sigma^2\|\mathbf{P}_{\perp}\mathbf{e}_i\|_2^2\mathbf{\Sigma}-\sigma^2\mathbf{\Sigma}\Big\|_2\le O\big(\frac{\mu K\sigma^2}{d}(\sigma^2+\lambda_1)\big),
$$

$$
\|3\sigma^4\mathbf{P}_{\perp}\mathbf{e}_i\mathbf{e}_i^{\top}\mathbf{P}_{\perp}-3\sigma^4\mathbf{e}_i\mathbf{e}_i^{\top}\|_2\le 3\sigma^4\|(\mathbf{P}_{\perp}\mathbf{e}_i-\mathbf{e}_i)\mathbf{e}_i^{\top}\mathbf{P}_{\perp}\|_2+3\sigma^4\|\mathbf{e}_i(\mathbf{P}_{\perp}\mathbf{e}_i-\mathbf{e}_i)^{\top}\|_2\lesssim\sigma^4\sqrt{\frac{\mu K}{d}},
$$

$$
\begin{aligned}
\|(\mathbf{P}_{\perp})_{[:,S]}\mathbf{u}_{K+1}\mathbf{e}_i^{\top}\mathbf{P}_{\perp}-(\mathbf{I}_d)_{[:,S]}\mathbf{u}_{K+1}\mathbf{e}_i^{\top}\|_2 &\le \|[(\mathbf{P}_{\perp})_{[:,S]}-(\mathbf{I}_d)_{[:,S]}]\mathbf{u}_{K+1}\mathbf{e}_i^{\top}\mathbf{P}_{\perp}\|_2\\
+\|(\mathbf{I}_d)_{[:,S]}]\mathbf{u}_{K+1}\mathbf{e}_i^{\top}(\mathbf{P}_{\perp}-\mathbf{I}_d)\|_2 &\lesssim K\sqrt{\frac{\mu}{d}}+\sqrt{\frac{\mu K}{d}}\lesssim K\sqrt{\frac{\mu}{d}},
\end{aligned}
$$

$$
\begin{aligned}
&2\sigma^4\rho\|\widetilde{\mathbf{e}}_0\|_2\Big\|\big[(\mathbf{P}_{\perp})_{[:,S]}\mathbf{u}_{K+1}\mathbf{e}_i^{\top}\mathbf{P}_{\perp}+\mathbf{P}_{\perp}\mathbf{e}_i(\mathbf{u}_{K+1})^{\top}(\mathbf{P}_{\perp})_{[S,:]}\big]-\big[(\mathbf{I}_d)_{[:,S]}\mathbf{u}_{K+1}\mathbf{e}_i^{\top}+\mathbf{e}_i(\mathbf{u}_{K+1})^{\top}(\mathbf{I}_d)_{[S,:]}\big]\Big\|_2\\
&\lesssim K\sigma^4\sqrt{\frac{\mu}{d}},
\end{aligned}
$$

and in summary we have $\|\mathbf{\Sigma}_i-\mathbf{\Sigma}_i'\|_2=O_P\Big(\frac{Kd\sigma^4}{n\Delta^2Lp}\sqrt{\frac{\mu}{d}}\Big)=O_P\Big(\frac{K\lambda_1^2}{\Delta^2}\sqrt{\frac{\mu}{d}}\Big)\frac{d\sigma^4}{nLp\lambda_1^2}=o_P\big(\frac{d\sigma^4}{nLp\lambda_1^2}\big)$. Now we study $\|\mathbf{\Sigma}_i'-\widetilde{\mathbf{\Sigma}}_i\|_2$. Since the entries of $\sqrt{p}\mathbf{\Omega}$ are i.i.d. standard Gaussian, by Lemma 3 in Fan et al. [15], we know that with probability $1-O(d^{-9})$, we have

$$
\|\mathbf{\Omega}\|_{2,\infty}\lesssim\sqrt{L},\quad \forall i\in[d],\quad \text{and}\quad \|\mathbf{\Omega}_{[S,:]}\|_2\lesssim\sqrt{L}.
$$

Therefore, under the condition that $\frac{\lambda_1^2Lp}{\Delta^2d}=o(1)$ we have

$$
\begin{aligned}
\|\mathbf{\Sigma}_i'-\widetilde{\mathbf{\Sigma}}_i\|_2 &= \sigma^4\Big\|\frac{1}{nL^2}\mathbf{B}_{\mathbf{\Omega}}^{\top}\mathbf{\Omega}^{\top}\Big(3\mathbf{e}_i\mathbf{e}_i^{\top}-2\rho\big((\mathbf{I}_d)_{[:,S]}\mathbf{u}_{K+1}\mathbf{e}_i^{\top}+\mathbf{e}_i\mathbf{u}_{K+1}^{\top}(\mathbf{I}_d)_{[S,:]}\big)\Big)\mathbf{\Omega}\mathbf{B}_{\mathbf{\Omega}}\Big\|_2\\
&\lesssim \frac{\sigma^4}{nL^2}\|\mathbf{B}_{\mathbf{\Omega}}\|_2^2\|\mathbf{\Omega}\|_{2,\infty}\big(\|\mathbf{\Omega}\|_{2,\infty}+\|\mathbf{\Omega}_{[:,S]}\|_2\big)=O_P\big(\frac{\sigma^4}{n\Delta^2}\big)=o_P\big(\frac{d\sigma^4}{nLp\lambda_1^2}\big).
\end{aligned}
$$

As for $\widetilde{\mathbf{\Sigma}}_i$, by Lemma 3 in Fan et al. [15] with high probability we have that $\sigma_K(\mathbf{\Omega}^{\top}\mathbf{V})\gtrsim\sqrt{L}$ and in turn

$$
\sigma_K(\widetilde{\mathbf{\Sigma}}_i)\gtrsim\frac{\sigma^2}{nL^2}\big(\sigma_K(\mathbf{B}_{\mathbf{\Omega}})\big)^2\Big(\big(\sigma_K(\mathbf{\Omega}^{\top}\mathbf{V})\big)^2\Delta+\big(\sigma_K(\mathbf{\Omega})\big)^2\sigma^2\Big)\gtrsim\frac{d\sigma^4}{nLp\lambda_1^2}+\frac{\sigma^2\Delta}{n\lambda_1^2}.
$$

Therefore, combining the previous results, we have that by Weyl's inequality [17], with high probability

$$\lambda_K\big(\mathbf{\Sigma}_i\big) \ge \lambda_K\big(\widetilde{\mathbf{\Sigma}}_i\big) - \|\mathbf{\Sigma}_i - \mathbf{\Sigma}_i'\|_2 - \|\mathbf{\Sigma}_i' - \widetilde{\mathbf{\Sigma}}_i\|_2 \gtrsim \frac{d\sigma^4}{nLp\lambda_1^2} - o\Big(\frac{d\sigma^4}{nLp\lambda_1^2}\Big) \gtrsim \frac{d\sigma^4}{nLp\lambda_1^2} + \frac{\sigma^2\Delta}{n\lambda_1^2}.$$

Thus we know $\eta_2(d) \asymp d\sigma^4/(nLp\lambda_1^2) + \sigma^2\Delta/(n\lambda_1^2)$.

Recall from the proof of Corollary 4.2 with probability $1 - O(d^{-10})$ we have $\|\mathbf{E}_0\|_2 \lesssim (\lambda_1 + \sigma^2)\log d\sqrt{\frac{r}{n}}$. Also recall that $r_2(d) \asymp \frac{\widetilde{\sigma}_1^3 K}{\delta^2 n}(\log d)^2$. Therefore, under the condition that

$$n \gg \frac{\kappa_1^4\lambda_1 d r^2(\log d)^4}{p\sigma^2}\left(\kappa_1\frac{d}{p} \wedge \frac{\lambda_1}{\sigma^2}L\right) \quad \text{and} \quad \frac{\widetilde{\sigma}_1^6 K^2}{\delta^4\sigma^4 n}(\log d)^4 \ll \Big(\frac{\Delta}{\lambda_1}\Big)^2,$$

we have $d^2 r_1(d)^4(\log d)^4\big(p^2\Delta^4\eta_2(d)\big)^{-1} = o(1)$ and $dr_2(d)^2\big(Lp\Delta^2\eta_2(d)\big)^{-1} = o(1)$.

Now we need to verify Assumption 5. It can be seen that the randomness of the leading term comes from $\mathbf{\Omega}$ and $\mathbf{E}_0$ both. We will first establish the results conditional on $\mathbf{\Omega}$. In fact, we will first show a more general CLT that will also cover the case of the leading term under the regime $Lp \gg d$. More specifically, we will show that for any matrix $\mathbf{A} \in \mathbb{R}^{d\times K}$ that satisfies the following two conditions: (1) $\sigma_{\max}(\mathbf{A})/\sigma_{\min}(\mathbf{A}) \le C|\lambda_1|/\Delta$; (2) $\lambda_K\big(\mathbf{\Sigma}_i\big) \ge cn^{-1}\sigma^4\big(\sigma_{\min}(\mathbf{A})\big)^2$, where $C, c > 0$ are fixed constants irrelevant to $\mathbf{A}$ and we abuse the notation by denoting $\mathbf{\Sigma}_i := \mathrm{Cov}(\mathbf{A}^\top\mathbf{E}_0\mathbf{P}_\perp\mathbf{e}_i)$, it holds that

$$\mathbf{\Sigma}_i^{-1/2}\mathbf{A}^\top\mathbf{E}_0\mathbf{P}_\perp\mathbf{e}_i \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{I}_K). \tag{B.30}$$

Now for any matrix $\mathbf{A} \in \mathbb{R}^{d\times K}$ satisfying the aforementioned conditions, to show that $\mathbf{A}^\top\mathbf{E}_0\mathbf{P}_\perp\mathbf{e}_i$ is asymptotically normal, we only need to show that $\mathbf{a}^\top\mathbf{\Sigma}_i^{-1/2}\mathbf{A}^\top\mathbf{E}_0\mathbf{P}_\perp\mathbf{e}_i \xrightarrow{d} \mathcal{N}(0,1)$ for any $\mathbf{a} \in \mathbb{R}^K$ with $\|\mathbf{a}\|_2 = 1$. We can write

$$\begin{aligned}\mathbf{a}^\top\mathbf{\Sigma}_i^{-1/2}\mathbf{A}^\top\mathbf{E}_0\mathbf{P}_\perp\mathbf{e}_i &= \frac{1}{n}\sum_{i=1}^n \mathbf{a}^\top\mathbf{\Sigma}_i^{-1/2}\mathbf{A}^\top\{\boldsymbol{X}_i\boldsymbol{X}_i^\top - \mathbf{\Sigma} - \widetilde{\mathbf{c}}_0^\top(\boldsymbol{Z}_i\boldsymbol{Z}_i^\top - \mathbf{\Lambda}_d)\widetilde{\mathbf{c}}_0\mathbf{I}_d\}\mathbf{P}_\perp\mathbf{e}_i \\ &= \frac{1}{n}\sum_{i=1}^n\Big\{\mathbf{a}^\top\mathbf{\Sigma}_i^{-1/2}\mathbf{A}^\top(\boldsymbol{X}_i\boldsymbol{X}_i^\top - \mathbf{\Sigma})\mathbf{P}_\perp\mathbf{e}_i - \widetilde{\mathbf{c}}_0^\top(\boldsymbol{Z}_i\boldsymbol{Z}_i^\top - \mathbf{\Lambda}_d)\widetilde{\mathbf{c}}_0(\mathbf{a}^\top\mathbf{\Sigma}_i^{-1/2}\mathbf{A}^\top\mathbf{P}_\perp\mathbf{e}_i)\Big\}.\end{aligned}$$

We let $x_i = \mathbf{a}^\top\mathbf{\Sigma}_i^{-1/2}\mathbf{A}^\top(\boldsymbol{X}_i\boldsymbol{X}_i^\top - \mathbf{\Sigma})\mathbf{P}_\perp\mathbf{e}_i$ and $y_i = \widetilde{\mathbf{c}}_0^\top(\boldsymbol{Z}_i\boldsymbol{Z}_i^\top - \mathbf{\Lambda}_d)\widetilde{\mathbf{c}}_0(\mathbf{a}^\top\mathbf{\Sigma}_i^{-1/2}\mathbf{A}^\top\mathbf{P}_\perp\mathbf{e}_i)$. For $\mathbf{\Sigma}_i$, we have that $\|\mathbf{\Sigma}_i^{-1/2}\|_2 \le \sigma_{\min}(\mathbf{\Sigma}_i)^{-1/2} \le \eta_2(d)^{-1/2}/\sigma_{\min}(\mathbf{A}) \le \sqrt{n}/\big(\sigma^2\sigma_{\min}(\mathbf{A})\big)$. Then we have

$$\begin{aligned}&\mathbb{E}|x_i|^3 \lesssim \mathbb{E}|\mathbf{a}^\top\mathbf{\Sigma}_i^{-1/2}\mathbf{A}^\top\boldsymbol{X}_i\boldsymbol{X}_i^\top\mathbf{P}_\perp\mathbf{e}_i|^3 \le \sqrt{\mathbb{E}|\mathbf{a}^\top\mathbf{\Sigma}_i^{-1/2}\mathbf{A}^\top\boldsymbol{X}_i|^6\mathbb{E}|\mathbf{e}_i^\top\mathbf{P}_\perp\boldsymbol{X}_i|^6} \\ &\quad\lesssim \|\mathbf{\Sigma}_i^{-1/2}\|_2^3\sqrt{(\lambda_1+\sigma^2)^3\sigma^6\|\mathbf{A}\|_2^6}, \\ &\mathbb{E}|y_i|^3 \lesssim (\mathbf{a}^\top\mathbf{\Sigma}_i^{-1/2}\mathbf{A}^\top\mathbf{P}_\perp\mathbf{e}_i)^3\mathbb{E}|\widetilde{\mathbf{c}}_0^\top\boldsymbol{Z}_i\boldsymbol{Z}_i^\top\widetilde{\mathbf{c}}_0|^3 \le \|\mathbf{\Sigma}_i^{-1/2}\mathbf{A}\|_2^3\mathbb{E}|\widetilde{\mathbf{c}}_0^\top\boldsymbol{Z}_i|^6 \\ &\quad\lesssim \|\mathbf{\Sigma}_i^{-1/2}\|^3(\lambda_1+\sigma^2)^3\|\mathbf{A}\|_2^3, \\ &\mathbb{E}|x_i - y_i|^3 \lesssim \mathbb{E}|x_i|^3 + \mathbb{E}|y_i|^3 \lesssim \|\mathbf{\Sigma}_i^{-1/2}\|_2^3\Big(\sqrt{(\lambda_1+\sigma^2)^3\sigma^6\|\mathbf{A}\|_2^6} + (\lambda_1+\sigma^2)^3\|\mathbf{A}\|_2^3\Big) \\ &\lesssim n^{3/2}(\lambda_1+\sigma^2)^3\|\mathbf{A}\|_2^3/\big(\sigma^2\sigma_{\min}(\mathbf{A})\big)^3.\end{aligned}$$

Thus

$$\frac{\sum_{i=1}^n\mathbb{E}|x_i - y_i|^3}{\mathrm{Var}\Big\{\sum_{i=1}^n(x_i - y_i)\Big\}^{3/2}} \lesssim \frac{n(\lambda_1+\sigma^2)^3\|\mathbf{A}\|_2^3}{n^{3/2}\sigma^6\sigma_{\min}(\mathbf{A})^3} \lesssim \frac{(\lambda_1+\sigma^2)^3\lambda_1^3}{\sqrt{n}\sigma^6\Delta^3} = o(1).$$

Thus the Lyapunov's condition is met and (B.30) holds. Then we take $\mathbf{A} = \mathbf{\Omega}\mathbf{B}_{\mathbf{\Omega}}$, and define the following event

$$\mathcal{A}_{\mathbf{\Omega}} = \Big\{ 1/2 \le \sigma_{\min}(\widetilde{\mathbf{\Omega}}^{(\ell)}/\sqrt{p}) \le \sigma_{\max}(\widetilde{\mathbf{\Omega}}^{(\ell)}/\sqrt{p}) \le 3/2, \quad \forall \ell \in [L] \Big\}$$
$$\cap \Big\{ \frac{1}{2}\sqrt{\frac{d}{p}} \le \sigma_{\min}(\mathbf{\Omega}) \le \sigma_{\max}(\mathbf{\Omega}) \le \frac{3}{2}\sqrt{\frac{d}{p}}, \quad \forall \ell \in [L] \Big\}.$$

Then from previous results we know that $\mathbb{P}((\mathcal{A}_{\mathbf{\Omega}})^c) = o(1)$, and under the event $\mathcal{A}_{\mathbf{\Omega}}$ we have

$$\frac{\sigma_{\max}(\mathbf{\Omega}\mathbf{B}_{\mathbf{\Omega}})}{\sigma_{\min}(\mathbf{\Omega}\mathbf{B}_{\mathbf{\Omega}})} \le 9\lambda_1/\Delta, \quad \lambda_K(\mathbf{\Sigma}_i) \ge \frac{\sigma^4}{2n}\big(\sigma_{\min}(\mathbf{\Omega}\mathbf{B}_{\mathbf{\Omega}})\big)^2.$$

Thus from the above proof, for any vector $\mathbf{t} \in \mathbb{R}^K$, we have $\mathbb{P}\big(\mathbf{\Sigma}_i^{-1/2}\mathbf{V}(\mathbf{E}_0)^\top \mathbf{e}_i \le \mathbf{t} | \mathcal{A}_{\mathbf{\Omega}}\big) - \Phi(\mathbf{t}) = o(1)$, where $\Phi(\cdot)$ is the CDF for $\mathcal{N}(0, \mathbf{I}_K)$. Then we have

$$\begin{aligned}
&\mathbb{P}\big(\mathbf{\Sigma}_i^{-1/2}\mathbf{V}(\mathbf{E}_0)^\top \mathbf{e}_i \le \mathbf{t}\big) = \mathbb{E}\Big(\mathbb{P}\big(\mathbf{\Sigma}_i^{-1/2}\mathbf{V}(\mathbf{E}_0)^\top \mathbf{e}_i \le \mathbf{t} | \mathbf{\Omega}\big)\Big) \\
&\quad = \mathbb{P}\big(\mathbf{\Sigma}_i^{-1/2}\mathbf{V}(\mathbf{E}_0)^\top \mathbf{e}_i \le \mathbf{t} | \mathbf{\Omega} \in \mathcal{A}_{\mathbf{\Omega}}\big)\mathbb{P}(\mathcal{A}_{\mathbf{\Omega}}) + \mathbb{P}\big(\mathbf{\Sigma}_i^{-1/2}\mathbf{V}(\mathbf{E}_0)^\top \mathbf{e}_i \le \mathbf{t} | \mathbf{\Omega} \in \mathcal{A}_{\mathbf{\Omega}}^c\big)\mathbb{P}(\mathcal{A}_{\mathbf{\Omega}}^c) \\
&\quad = \big(\Phi(\mathbf{t}) + o(1)\big)\big(1 - o(1)\big) + o(1) = \Phi(\mathbf{t}) + o(1).
\end{aligned}$$

Hence we have that Assumption 5 holds and (18) follows. Next we need to show that the result also holds for $\widetilde{\mathbf{\Sigma}}_i$. From previous discussion we already know that $\|\mathbf{\Sigma}_i - \widetilde{\mathbf{\Sigma}}_i\|_2 = o_P\big(\lambda_K(\widetilde{\mathbf{\Sigma}}_i)\big)$, then by Lemma 13 in Chen et al. [10] we have that $\|\widetilde{\mathbf{\Sigma}}_i^{-1/2}\mathbf{\Sigma}_i^{1/2} - \mathbf{I}_d\|_2 = O_P\big(\|\widetilde{\mathbf{\Sigma}}_i^{-1/2}\|_2 \|\mathbf{\Sigma}_i^{1/2} - \widetilde{\mathbf{\Sigma}}_i^{1/2}\|_2\big) = O_P\big(\lambda_K(\widetilde{\mathbf{\Sigma}}_i)^{-1}\|\mathbf{\Sigma}_i - \widetilde{\mathbf{\Sigma}}_i\|_2\big) = o_P(1)$. Then by Slutsky's Theorem, we have

$$\widetilde{\mathbf{\Sigma}}_i^{-1/2}(\widetilde{\mathbf{V}}^{\mathrm{F}}\mathbf{H} - \mathbf{V})^\top \mathbf{e}_i = (\widetilde{\mathbf{\Sigma}}_i^{-1/2}\mathbf{\Sigma}_i^{1/2})\mathbf{\Sigma}_i^{-1/2}(\widetilde{\mathbf{V}}^{\mathrm{F}}\mathbf{H} - \mathbf{V})^\top \mathbf{e}_i \xrightarrow{d} \mathcal{N}(0, \mathbf{I}_K).$$

Finally, we move on to verify the validity of the estimator $\widehat{\mathbf{\Sigma}}_i$ for the asymptotic covariance matrix. From Lemma 7 in Fan et al. [15], it can be seen that with probability $1 - o(1)$, $\mathbf{H}$ is orthonormal. When $\mathbf{H}$ is orthonormal, by Slutsky's Theorem we have that

$$\mathbf{H}\widetilde{\mathbf{\Sigma}}_i^{-1/2}(\widetilde{\mathbf{V}}^{\mathrm{F}}\mathbf{H} - \mathbf{V})^\top \mathbf{e}_i = \mathbf{H}\widetilde{\mathbf{\Sigma}}_i^{-1/2}\mathbf{H}^\top(\widetilde{\mathbf{V}}^{\mathrm{F}} - \mathbf{V}\mathbf{H}^\top)^\top \mathbf{e}_i \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{I}_K),$$

where it can be seen that $\mathbf{H}\widetilde{\mathbf{\Sigma}}_i^{-1/2}\mathbf{H}^\top = (\mathbf{H}\widetilde{\mathbf{\Sigma}}_i\mathbf{H}^\top)^{-1/2}$. Therefore, it suffices to show that $\|\widehat{\mathbf{\Sigma}}_i - \mathbf{H}\widetilde{\mathbf{\Sigma}}_i\mathbf{H}^\top\|_2 = o_P\big(\lambda_K(\widetilde{\mathbf{\Sigma}}_i)\big)$, and the results will hold by Slutsky's Theorem. Recall from the proof of Corollary 4.2, we have the following bounds

$$\|\mathbf{\Sigma} - \widehat{\mathbf{\Sigma}}\|_2 = O_P\Big((\lambda_1 + \sigma^2)\sqrt{\frac{r}{n}}\Big), \quad |\widehat{\sigma}^2 - \sigma^2| = O_P(\widetilde{\sigma}_1\sqrt{\frac{K}{n}}),$$

We will bound the components of $\|\widehat{\mathbf{\Sigma}}_i - \widetilde{\mathbf{\Sigma}}_i\|_2$ respectively. We have

$$\begin{aligned}
\|\sigma^2\mathbf{\Sigma} - \widehat{\sigma}^2\widehat{\mathbf{\Sigma}}\|_2 &\lesssim |\widehat{\sigma}^2 - \sigma^2| \|\mathbf{\Sigma}\|_2 + \sigma^2\|\mathbf{\Sigma} - \widehat{\mathbf{\Sigma}}\|_2 = O_P\Big(\widetilde{\sigma}_1(\lambda_1 + \sigma^2)\sqrt{\frac{K}{n}}\Big) \\
&+ O_P\Big(\sigma^2(\lambda_1 + \sigma^2)\sqrt{\frac{r}{n}}\Big) = O_P\Big(\sigma^2(\lambda_1 + \sigma^2)\sqrt{\frac{r}{n}}\Big),
\end{aligned}$$

Also, from proof of Theorem 4.10, we have that with high probability

$$\|\widetilde{\mathbf{V}}^{\mathrm{F}}\mathbf{H} - \mathbf{V}\|_2 = \|\widetilde{\mathbf{V}}^{\mathrm{F}} - \mathbf{V}\mathbf{H}^\top\|_2 \lesssim \|\mathbf{E}_0\|_2\|\mathbf{\Omega}\|_2\|\mathbf{B}_{\mathbf{\Omega}}\|_2/L = O_P(\kappa_1\sqrt{\frac{dr}{npL}}),$$

and $\|\widehat{\mathbf{\Sigma}}^{\rm tr} - \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top\|_2 = O_P\Big((\lambda_1+\sigma^2)\sqrt{\frac{r}{n}}\Big)$, where $\widehat{\mathbf{\Sigma}}^{\rm tr} = \widehat{\mathbf{\Sigma}} - \widehat{\sigma}^2\mathbf{I}_d$.Then with high probability, for all $\ell \in [L]$ we have that

$$\big\|(\widetilde{\mathbf{V}}^{\rm F\top}\widehat{\mathbf{\Sigma}}^{\rm tr} - \mathbf{H}\mathbf{\Lambda}\mathbf{V}^\top)\mathbf{\Omega}^{(\ell)}/\sqrt{p}\big\|_2 \lesssim \sqrt{\frac{d}{p}}(\|\widehat{\mathbf{\Sigma}}^{\rm tr} - \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top\|_2 + \lambda_1\|\widetilde{\mathbf{V}}^{\rm F} - \mathbf{V}\mathbf{H}^\top\|_2)$$
$$= O_P\left(\kappa_1\lambda_1\sqrt{\frac{d^2r}{np^2L}}\right) = o_P(\Delta),$$

and thus by Theorem 3.3 in Stewart [35], with high probability for all $\ell \in [L]$ we have that

$$\|\widehat{\mathbf{B}}^{(\ell)} - \mathbf{B}^{(\ell)}\mathbf{H}^\top\|_2 = \big\|(\widetilde{\mathbf{V}}^{\rm F\top}\widehat{\mathbf{\Sigma}}^{\rm tr}\mathbf{\Omega}^{(\ell)}/\sqrt{p})^\dagger - (\mathbf{H}\mathbf{\Lambda}\mathbf{V}^\top\mathbf{\Omega}^{(\ell)}/\sqrt{p})^\dagger\big\|_2$$
$$= O_P\bigg(\Delta^{-2}\kappa_1\lambda_1\sqrt{\frac{d^2r}{np^2L}}\bigg),$$

and in turn we have $\|\widehat{\mathbf{B}}_{\mathbf{\Omega}} - \mathbf{B}_{\mathbf{\Omega}}\mathbf{H}^\top\|_2 = O_P\bigg(\Delta^{-2}\kappa_1\lambda_1\sqrt{\frac{d^2r}{np^2L}}\bigg)\sqrt{L} = O_P\bigg(\Delta^{-2}\kappa_1\lambda_1\sqrt{\frac{d^2r}{np^2}}\bigg)$.

Thus combining the above results, under the condition that $\frac{\lambda_1\kappa_1^4}{\sigma^2}\sqrt{\frac{d^2r}{np^2L}} = o(1)$, following basic algebra we have

$$\|\widehat{\mathbf{\Sigma}}_i - \mathbf{H}\widetilde{\mathbf{\Sigma}}_i\mathbf{H}^\top\|_2 \lesssim O_P\Big(\sigma^2(\lambda_1+\sigma^2)\sqrt{\frac{r}{n}}\Big)\frac{d}{nLp\Delta^2} + O_P\bigg(\frac{d\sqrt{L}}{nL^2p\Delta^3}\sigma^2(\sigma^2+\lambda_1)\kappa_1\lambda_1\sqrt{\frac{d^2r}{np^2}}\bigg)$$
$$= O_P\Big(\frac{\lambda_1^2}{\Delta^2\sigma^2}(\lambda_1+\sigma^2)\sqrt{\frac{r}{n}}\Big)\frac{d\sigma^4}{nLp\lambda_1^2} + O_P\Big(\frac{\lambda_1\kappa_1^4}{\sigma^2}\sqrt{\frac{d^2r}{np^2L}}\Big)\frac{d\sigma^4}{nLp\lambda_1^2} = o_P\big(\lambda_K(\widetilde{\mathbf{\Sigma}}_i)\big).$$

Therefore, by Slutsky's Theorem, under the event $\mathcal{B} := \{\mathbf{H} \text{ is orthonormal}\}$, for any vector $\mathbf{t} \in \mathbb{R}^K$, we have that $\mathbb{P}(\widehat{\mathbf{\Sigma}}_i^{-1/2}(\widetilde{\mathbf{V}}^{\rm F} - \mathbf{V}\mathbf{H}^\top)^\top\mathbf{e}_i \le \mathbf{t}|\mathcal{B}) - \Phi(\mathbf{t}) = o(1)$, and thus

$$\mathbb{P}(\widehat{\mathbf{\Sigma}}_i^{-1/2}(\widetilde{\mathbf{V}}^{\rm F} - \mathbf{V}\mathbf{H}^\top)^\top\mathbf{e}_i \le \mathbf{t}) = \mathbb{P}(\widehat{\mathbf{\Sigma}}_i^{-1/2}(\widetilde{\mathbf{V}}^{\rm F} - \mathbf{V}\mathbf{H}^\top)^\top\mathbf{e}_i \le \mathbf{t}|\mathcal{B})\mathbb{P}(\mathcal{B})$$
$$+ \mathbb{P}(\widehat{\mathbf{\Sigma}}_i^{-1/2}(\widetilde{\mathbf{V}}^{\rm F} - \mathbf{V}\mathbf{H}^\top)^\top\mathbf{e}_i \le \mathbf{t}|\mathcal{B}^c)\mathbb{P}(\mathcal{B}^c)$$
$$= \mathbb{P}\big(\widehat{\mathbf{\Sigma}}_i^{-1/2}(\widetilde{\mathbf{V}}^{\rm F} - \mathbf{V}\mathbf{H}^\top)^\top\mathbf{e}_i \le \mathbf{t}|\mathcal{B}\big)\big(1 - o(1)\big) + o(1) = \Phi(\mathbf{t}) + o(1).$$

Hence the claim follows.

B.8. Proof of Corollary 4.12. We will verify that Assumptions 1, 2, 3 and 5 hold. First, it is not hard to see that there exists some orthonormal matrix $\mathbf{O} \in \mathbb{R}^{K\times K}$ such that $\mathbf{V} = \mathbf{F}^*\mathbf{C}^{-1}\mathbf{O}$, where $\mathbf{C} = \mathrm{diag}(\sqrt{d_1}, \ldots, \sqrt{d_K})$. From the problem setting of Example 3 we also know that there exists a constant $C > 0$ such that

$$C^{-1}K\max_k d_k \le K\min_k d_k \le d \le K\max_k d_k, \quad d_1 \asymp \ldots \asymp d_K \asymp d/K,$$

and thus that $\sqrt{d/K} \lesssim \sigma_K(\mathbf{C}) \le \|\mathbf{C}\|_2 \lesssim \sqrt{d/K}$. Then $\|\mathbf{V}\|_{2,\infty} \lesssim \sqrt{\frac{K}{d}}\|\mathbf{F}^*\|_{2,\infty} = \sqrt{\frac{K}{d}}$. Thus Assumption 3 holds with $\mu = O(1)$.

From the proof of Corollary 4.2 we know that Assumption 1 is satisfied. Besides, recall from Remark 17, under the condition that $\sqrt{K/d}\log d = O(1)$, with probability at least $1 - d^{-10}$ we have that $\|\mathbf{E}\|_2 \lesssim d\Delta_0/\sqrt{K} + \sqrt{dn}\log d := r_1'(d)$, which is sharper than $r_1(d)\log d$. Since $\mathbf{E}_b = 0$, we have $r_2(d) = 0$ and

Assumption 2 holds trivially. Now we move on to study the minimum covariance eigenvalue rate $\eta_2(d)$. From the proof of Corollary 4.2, we know that

$$\mathbf{E} = \mathbf{E}_0 = \mathbf{F}^*\boldsymbol{\Theta}^{*\top}\mathbf{Z} + \mathbf{Z}^\top\boldsymbol{\Theta}^*\mathbf{F}^{*\top} + \mathbf{Z}^\top\mathbf{Z} - n\mathbf{I}_d = \sum_{j=1}^n \big\{\mathbf{Q}_j\mathbf{Z}_{j.}^\top + \mathbf{Z}_{j.}\mathbf{Q}_j^\top + \mathbf{Z}_{j.}\mathbf{Z}_{j.}^\top - \mathbf{I}_d\big\},$$

where $\mathbf{Q}_j = \mathbf{F}^*\boldsymbol{\Theta}_{j.}^* \in \mathbb{R}^d$ with $\boldsymbol{\Theta}_{j.}^*$ the j-th row of $\boldsymbol{\Theta}^*$, $\mathbf{Z}_{j.}$ is the j-th row of $\mathbf{Z}$ and $\mathbf{Z}_{j.} \overset{\text{i.i.d}}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$. Then for $i \in [d]$, we have

$$\begin{aligned}
\mathrm{Cov}(\mathbf{E}_0\mathbf{P}_\perp\mathbf{e}_i) &= \mathrm{Cov}\Big(\sum_{j=1}^n \big\{\mathbf{Q}_j\mathbf{Z}_{j.}^\top + \mathbf{Z}_{j.}\mathbf{Q}_j^\top + \mathbf{Z}_{j.}\mathbf{Z}_{j.}^\top - \mathbf{I}_d\big\}\mathbf{P}_\perp\mathbf{e}_i\Big)\\
&= \sum_{j=1}^n \mathrm{Cov}\Big(\big\{\mathbf{Q}_j\mathbf{Z}_{j.}^\top + \mathbf{Z}_{j.}\mathbf{Q}_j^\top + \mathbf{Z}_{j.}\mathbf{Z}_{j.}^\top - \mathbf{I}_d\big\}\mathbf{P}_\perp\mathbf{e}_i\Big)\\
&= \sum_{j=1}^n \mathrm{Cov}\Big(\big\{\mathbf{Q}_j\mathbf{Z}_{j.}^\top + \mathbf{Z}_{j.}\mathbf{Z}_{j.}^\top - \mathbf{I}_d\big\}\mathbf{P}_\perp\mathbf{e}_i\Big).
\end{aligned}$$

where the last equality is due to the fact that $\mathbf{P}_\perp\mathbf{Q}_j = \mathbf{P}_\perp\mathbf{F}^*\boldsymbol{\Theta}_{j.}^* = \mathbf{0}$. Now for $j \in [n]$, we calculate $\mathrm{Cov}\Big(\big\{\mathbf{Q}_j\mathbf{Z}_{j.}^\top + \mathbf{Z}_{j.}\mathbf{Z}_{j.}^\top - \mathbf{I}_d\big\}\mathbf{P}_\perp\mathbf{e}_i\Big)$. Following basic algebra, we have that

$$\begin{aligned}
&\mathrm{Cov}\Big(\big\{\mathbf{Q}_j\mathbf{Z}_{j.}^\top + \mathbf{Z}_{j.}\mathbf{Z}_{j.}^\top - \mathbf{I}_d\big\}\mathbf{P}_\perp\mathbf{e}_i\Big)\\
=&\mathbb{E}\Big(\big\{\mathbf{Q}_j\mathbf{Z}_{j.}^\top + \mathbf{Z}_{j.}\mathbf{Z}_{j.}^\top\big\}\mathbf{P}_\perp\mathbf{e}_i\mathbf{e}_i^\top\mathbf{P}_\perp\big\{\mathbf{Z}_{j.}\mathbf{Q}_j^\top + \mathbf{Z}_{j.}\mathbf{Z}_{j.}^\top\big\}\Big) - \mathbf{P}_\perp\mathbf{e}_i\mathbf{e}_i^\top\mathbf{P}_\perp\\
=& \|\mathbf{P}_\perp\mathbf{e}_i\|_2^2(\mathbf{Q}_j\mathbf{Q}_j^\top + \mathbf{I}_d) + 2\mathbf{P}_\perp\mathbf{e}_i\mathbf{e}_i^\top\mathbf{P}_\perp - \mathbf{P}_\perp\mathbf{e}_i\mathbf{e}_i^\top\mathbf{P}_\perp\\
=& \|\mathbf{P}_\perp\mathbf{e}_i\|_2^2(\mathbf{Q}_j\mathbf{Q}_j^\top + \mathbf{I}_d) + \mathbf{P}_\perp\mathbf{e}_i\mathbf{e}_i^\top\mathbf{P}_\perp,
\end{aligned}$$

and thus

$$\begin{aligned}
\mathrm{Cov}(\mathbf{E}_0\mathbf{P}_\perp\mathbf{e}_i) &= \sum_{j=1}^n \|\mathbf{P}_\perp\mathbf{e}_i\|_2^2(\mathbf{Q}_j\mathbf{Q}_j^\top + \mathbf{I}_d) + \mathbf{P}_\perp\mathbf{e}_i\mathbf{e}_i^\top\mathbf{P}_\perp\\
&= \|\mathbf{P}_\perp\mathbf{e}_i\|_2^2\Big(\sum_{j=1}^n \mathbf{Q}_j\mathbf{Q}_j^\top + n\mathbf{I}_d\Big) + n\mathbf{P}_\perp\mathbf{e}_i\mathbf{e}_i^\top\mathbf{P}_\perp\\
&= \|\mathbf{P}_\perp\mathbf{e}_i\|_2^2(\mathbf{F}^*\boldsymbol{\Theta}^{*\top}\boldsymbol{\Theta}^*\mathbf{F}^{*\top} + n\mathbf{I}_d) + n\mathbf{P}_\perp\mathbf{e}_i\mathbf{e}_i^\top\mathbf{P}_\perp.
\end{aligned}$$

Then since $\|\mathbf{P}_\perp\mathbf{e}_i\|_2 = 1 - K/d = 1 - o(1)$, we have that $\lambda_d\big(\mathrm{Cov}(\mathbf{E}_0\mathbf{P}_\perp\mathbf{e}_i)\big) \gtrsim n$, and hence we have $\eta_2(d) \asymp dn/(\lambda_1^2 Lp)$. Then under the condition that $n \gg d^3L/p$ and $\Delta_0^2 \gg K(\log d)^2\sqrt{dnL/p}$, we have that

$$\frac{r_1'(d)^4}{\eta_2(d)} \lesssim \frac{\lambda_1^2 Lp}{d}\left(\frac{d^4\Delta_0^4}{K^2 n} + d^2 n(\log d)^4\right) \ll \frac{\lambda_1^2 p^2 d^2\Delta_0^4}{K^2 d^2} \asymp \frac{p^2\Delta^4}{d^2}, \quad \frac{d^2 r_1'(d)^4}{p^2\Delta^4\eta_2(d)} = o(1).$$

Now we move on to check Assumption 5. Similar as in the proof of Corollary 4.11, we will first show the results conditional on $\boldsymbol{\Omega}$ by establishing a more general CLT . More specifically, we will show that for any $\mathbf{a} \in \mathbb{R}^K$ with $\|\mathbf{a}\|_2 = 1$, and $\mathbf{A} \in \mathbb{R}^{d\times K}$ such that $\lambda_K\big(\boldsymbol{\Sigma}_i\big) \ge cn\sigma_{\min}(\mathbf{A})^2$ and $\sigma_{\max}(\mathbf{A})/\sigma_{\min}(\mathbf{A}) \le C$, where $C, c > 0$ are constants irrelevant to $\mathbf{A}$, we have $\mathbf{a}^\top\boldsymbol{\Sigma}_i^{-1/2}\mathbf{A}^\top\mathbf{E}_0\mathbf{P}_\perp\mathbf{e}_i \xrightarrow{d} \mathcal{N}(0,1)$. We know that $\|\mathbf{Q}\|_{2,\infty} = \max_{j\in[n]}\|\mathbf{Q}_j\|_2 \le \|\mathbf{F}^*\|_2\|\boldsymbol{\Theta}^*\|_{2,\infty} \lesssim \mu_\theta\Delta_0\sqrt{\frac{d}{n}}$, and $\mathbf{a}^\top\boldsymbol{\Sigma}_i^{-1/2}\mathbf{A}^\top\mathbf{E}_0\mathbf{P}_\perp\mathbf{e}_i = \sum_{j=1}^n\left\{\mathbf{a}^\top\boldsymbol{\Sigma}_i^{-1/2}\mathbf{A}^\top(\mathbf{Q}_j\mathbf{Z}_{j\cdot}^\top + \mathbf{Z}_{j\cdot}\mathbf{Z}_{j\cdot}^\top - \mathbf{I}_d)\mathbf{P}_\perp\mathbf{e}_i\right\}$, and we denote

$$x_j = \mathbf{a}^\top\boldsymbol{\Sigma}_i^{-1/2}\mathbf{A}^\top\mathbf{Q}_j\mathbf{Z}_{j\cdot}^\top\mathbf{P}_\perp\mathbf{e}_i, \quad y_j = \mathbf{a}^\top\boldsymbol{\Sigma}_i^{-1/2}\mathbf{A}^\top(\mathbf{Z}_{j\cdot}\mathbf{Z}_{j\cdot}^\top - \mathbf{I}_d)\mathbf{P}_\perp\mathbf{e}_i.$$

Then we have

$$\begin{aligned}\mathbb{E}|x_j+y_j|^3 &\lesssim \mathbb{E}|x_j|^3 + \mathbb{E}|y_j|^3 \lesssim \|\boldsymbol{\Sigma}_i^{-1/2}\mathbf{A}^\top\mathbf{Q}_j\|_2^3 + \|\boldsymbol{\Sigma}_i^{-1/2}\mathbf{A}^\top\|_2^3 \\ &\le \|\boldsymbol{\Sigma}_i^{-1/2}\|_2^3\|\mathbf{A}\|_2^3(\|\mathbf{Q}\|_{2,\infty}^3+1) \lesssim n^{-3/2}\Big\{\frac{\|\mathbf{A}\|_2}{\sigma_{\min}(\mathbf{A})}\Big\}^3\Big\{\mu_\theta^3\Delta_0^3\Big(\frac{d}{n}\Big)^{3/2}+1\Big\} \\ &\lesssim n^{-3/2}\mu_\theta^3\Delta_0^3\Big(\frac{d}{n}\Big)^{3/2} + n^{-3/2}.\end{aligned}$$

Then

$$\frac{\sum_{j=1}^n \mathbb{E}|x_j+y_j|^3}{\mathrm{Var}\left\{\sum_{j=1}^n(x_j+y_j)\right\}^{3/2}} = \sum_{j=1}^n \mathbb{E}|x_j+y_j|^3 \lesssim n^{-2}\mu_\theta^3\Delta_0^3 d^{3/2} + n^{-1/2}.$$

Then under the condition that $\Delta_0^2 \ll n^{4/3}/(\mu_\theta^2 d)$, we have that

$$n^{-2}\mu_\theta^3\Delta_0^3 d^{3/2} = o(1) \quad \text{and} \quad \left(\sum_{j=1}^n \mathbb{E}|x_j+y_j|^3\right)\mathrm{Var}\Big(\sum_{j=1}^n(x_j+y_j)\Big)^{-3/2} = o(1).$$

Thus the Lyapunov's condition is met and the CLT holds. Also recall from previous arguments, there exists a fixed constant $C>0$ such that with high probability we have

$$\frac{\sigma_{\max}(\boldsymbol{\Omega}\mathbf{B}_{\boldsymbol{\Omega}})}{\sigma_{\min}(\boldsymbol{\Omega}\mathbf{B}_{\boldsymbol{\Omega}})} \le 9\left(\frac{\sigma_1(\boldsymbol{\Theta}^*)}{\sigma_K(\boldsymbol{\Theta}^*)}\right)^2 \le C, \quad \lambda_K(\boldsymbol{\Sigma}_i) \ge \frac{n}{2}\big(\sigma_{\min}(\boldsymbol{\Omega}\mathbf{B}_{\boldsymbol{\Omega}})\big)^2.$$

Then by taking $\mathbf{A} = \boldsymbol{\Omega}\mathbf{B}_{\boldsymbol{\Omega}}$ and following similar steps as in the proof of Corollary 4.11, we know that Assumption 5 is satisfied. Then by Theorem 4.10, (17) holds.

We move on to prove (20). It suffices to show that $\|\boldsymbol{\Sigma}_i - \widetilde{\boldsymbol{\Sigma}}_i\|_2 = o_P\big(\lambda_K(\widetilde{\boldsymbol{\Sigma}}_i)\big)$. When $\Delta_0^2 \ll n$ and $K \ll d$, with high probability we have

$$\begin{aligned}\|\boldsymbol{\Sigma}_i - \widetilde{\boldsymbol{\Sigma}}_i\|_2 &\lesssim \frac{d}{L\Delta^2 p}\big\{(n+\Delta)(1-\|\mathbf{P}_\perp\mathbf{e}_i\|_2^2) + n\|\mathbf{P}_\perp\mathbf{e}_i\mathbf{e}_i^\top\mathbf{P}_\perp - \mathbf{e}_i\mathbf{e}_i^\top\|_2\big\} \\ &+ \frac{n}{L\Delta^2}\|\boldsymbol{\Omega}\|_{2,\infty}^2 \lesssim \frac{d}{L\Delta^2 p}\Big(\frac{Kn}{d} + \Delta_0^2 + n\sqrt{\frac{K}{d}}\Big) + \frac{n}{\Delta^2} = o\Big(\frac{dn}{L\lambda_1^2 p}\Big) = o_P\big(\lambda_K(\widetilde{\boldsymbol{\Sigma}}_i)\big).\end{aligned}$$

Thus (20) holds.

Last we verify the validity of $\widehat{\boldsymbol{\Sigma}}_i$. Similar as in the proof of Corollary 4.11, it suffices to show that $\|\widehat{\boldsymbol{\Sigma}}_i - \mathbf{H}\widetilde{\boldsymbol{\Sigma}}_i\mathbf{H}^\top\|_2 = o_P\big(\lambda_K(\widetilde{\boldsymbol{\Sigma}}_i)\big)$. Recall with high probability $\|\widehat{\mathbf{M}} - \mathbf{M}\|_2 \lesssim r_1'(d) = d\Delta_0/\sqrt{K} + \sqrt{dn}\log d$.

Also, from the proof of Theorem 4.10, we have that

$$\|\widetilde{\mathbf{V}}^{\mathrm{F}}\mathbf{H} - \mathbf{V}\|_2 = \|\widetilde{\mathbf{V}}^{\mathrm{F}} - \mathbf{V}\mathbf{H}^\top\|_2 = \frac{1}{L}O_P(\|\widehat{\mathbf{M}} - \mathbf{M}\|_2\|\boldsymbol{\Omega}\|_2\|\mathbf{B}_{\boldsymbol{\Omega}}\|_2) = O_P\Big(\sqrt{\frac{d}{pL}}\frac{r_1'(d)}{\Delta}\Big),$$

Then with high probability, for all $\ell \in [L]$ we have that

$$\begin{aligned}&\big\|(\widetilde{\mathbf{V}}^{\mathrm{F}\top}\widehat{\mathbf{M}} - \mathbf{H}\boldsymbol{\Lambda}\mathbf{V}^\top)\boldsymbol{\Omega}^{(\ell)}/\sqrt{p}\big\|_2 \lesssim \sqrt{\frac{d}{p}}(\|\widehat{\mathbf{M}} - \mathbf{M}\|_2 + \lambda_1\|\widetilde{\mathbf{V}}^{\mathrm{F}} - \mathbf{V}\mathbf{H}^\top\|_2) \\ &= O_P\left(\sqrt{\frac{d^2}{p^2 L}}r_1'(d)\right) = o_P(\Delta),\end{aligned}$$

and thus by Theorem 3.3 in Stewart [35], we have that

$$\|\widehat{\mathbf{B}}^{(\ell)} - \mathbf{B}^{(\ell)}\mathbf{H}^\top\|_2 = \left\|(\widetilde{\mathbf{V}}^{\mathrm{F}\top}\widehat{\mathbf{M}}\boldsymbol{\Omega}^{(\ell)}/\sqrt{p})^\dagger - (\mathbf{H}\boldsymbol{\Lambda}\mathbf{V}^\top\boldsymbol{\Omega}^{(\ell)}/\sqrt{p})^\dagger\right\|_2 = O_P\left(\sqrt{\frac{d^2}{p^2L}\frac{r_1'(d)}{\Delta^2}}\right),$$

and in turn we have $\|\widehat{\mathbf{B}}_{\boldsymbol{\Omega}} - \mathbf{B}_{\boldsymbol{\Omega}}\mathbf{H}^\top\|_2 = O_P\left(\sqrt{\frac{d^2}{p^2L}\frac{r_1'(d)}{\Delta^2}}\right)\sqrt{L} = O_P\left(\frac{dr_1'(d)}{p\Delta^2}\right)$.

Therefore, under the condition that $\Delta_0^2 \ll KLp^2n^2/d^4$, we have

$$\|\widehat{\boldsymbol{\Sigma}}_i - \mathbf{H}\widetilde{\boldsymbol{\Sigma}}_i\mathbf{H}^\top\|_2 \lesssim \frac{d}{L\Delta^2 p}\|\widehat{\mathbf{M}} - \mathbf{M}\|_2 + (n+\lambda_1)\frac{d}{pL\Delta}O_P\left(\frac{d}{p\sqrt{L}}\frac{r_1'(d)}{\Delta^2}\right)$$

$$= o_P\big(\frac{dn}{L\Delta^2 p}\big) = o_P\big(\lambda_K(\widetilde{\boldsymbol{\Sigma}}_i)\big).$$

Thus the claim follows.

B.9. Proof of Theorem 4.5. We will first decompose $\widetilde{\mathbf{V}}^{\mathrm{F}}\mathbf{H} - \mathbf{V} = (\widetilde{\mathbf{V}}^{\mathrm{F}}\mathbf{H} - \widetilde{\mathbf{V}}\mathbf{H}_1\mathbf{H}_0) + (\widetilde{\mathbf{V}}\mathbf{H}_1\mathbf{H}_0 - \widehat{\mathbf{V}}\mathbf{H}_0) + (\widehat{\mathbf{V}}\mathbf{H}_0 - \mathbf{V})$. We will show that when L is sufficiently large the first two terms are negligible, and we will consider the third term $\widehat{\mathbf{V}}\mathbf{H}_0 - \mathbf{V}$ first. We will first study $\|\widehat{\mathbf{V}}\mathbf{H}_0 - \mathbf{V} - \mathbf{P}_\perp\mathbf{E}_0\mathbf{V}\boldsymbol{\Lambda}^{-1}\|_{2,\infty}$ by conducting decomposition of the error term similar to the proof of Lemma 8 in Fan et al. [15]. For the convenience of notations, we let $\mathbf{P} = \mathbf{V}^\top\mathbf{V}$ for short. If we define $\widehat{\mathbf{H}}_0 = \widehat{\mathbf{V}}^\top\mathbf{V}$, we can decompose

$$\widehat{\mathbf{V}}\mathbf{H}_0 - \mathbf{V} - \mathbf{P}_\perp\mathbf{E}_0\mathbf{V}\boldsymbol{\Lambda}^{-1}$$

$$= \mathbf{P}_\perp\widehat{\mathbf{V}}\widehat{\mathbf{H}}_0 - \mathbf{P}_\perp\mathbf{E}_0\mathbf{V}\boldsymbol{\Lambda}^{-1} + \mathbf{P}_\perp\widehat{\mathbf{V}}(\mathbf{H}_0 - \widehat{\mathbf{H}}_0) + (\mathbf{P}\widehat{\mathbf{V}}\mathbf{H}_0 - \mathbf{V}).$$

Under the condition that $\|\mathbf{E}\|_2/\Delta = O_P\big(r_1(d)/\Delta\big) = o_P(1)$, we have that $\mathbf{H}_0$ is a full-rank orthonormal matrix with probability $1 - o(1)$. Then we have with probability $1 - o(1)$ that

$$\|\mathbf{P}_\perp\widehat{\mathbf{V}}(\mathbf{H}_0 - \widehat{\mathbf{H}}_0)\|_{2,\infty} = \|(\mathbf{I} - \mathbf{V}\mathbf{V}^\top)(\widehat{\mathbf{V}}\mathbf{H}_0 - \mathbf{V})\mathbf{H}_0^\top(\mathbf{H}_0 - \widehat{\mathbf{H}}_0)\|_{2,\infty}$$

$$\le \|(\widehat{\mathbf{V}}\mathbf{H}_0 - \mathbf{V})\mathbf{H}_0^\top(\mathbf{H}_0 - \widehat{\mathbf{H}}_0)\|_{2,\infty} + \|\mathbf{V}\mathbf{V}^\top(\widehat{\mathbf{V}}\mathbf{H}_0 - \mathbf{V})\mathbf{H}_0^\top(\mathbf{H}_0 - \widehat{\mathbf{H}}_0)\|_{2,\infty}$$

$$\le \|\widehat{\mathbf{V}}\mathbf{H}_0 - \mathbf{V}\|_{2,\infty}\|\mathbf{H}_0 - \widehat{\mathbf{H}}_0\|_2 + \|\mathbf{V}\|_{2,\infty}\|\widehat{\mathbf{V}}\mathbf{H}_0 - \mathbf{V}\|_2\|\mathbf{H}_0 - \widehat{\mathbf{H}}_0\|_2$$

$$\lesssim \Big(r_3(d) + \sqrt{\frac{\mu K}{d}}\frac{\|\mathbf{E}\|_2}{\Delta}\Big)\|\mathbf{H}_0 - \widehat{\mathbf{H}}_0\|_2.$$

From Lemma 7 in Fan et al. [15], we know that $\|\mathbf{H}_0 - \widehat{\mathbf{H}}_0\|_2 \lesssim \|\widehat{\mathbf{V}}\widehat{\mathbf{V}}^\top - \mathbf{V}\mathbf{V}^\top\|_2^2 \lesssim (\|\mathbf{E}\|_2/\Delta)^2 = O_P(r_1(d)^2/\Delta^2)$, and thus we have

$$\|\mathbf{P}_\perp\widehat{\mathbf{V}}(\mathbf{H}_0 - \widehat{\mathbf{H}}_0)\|_{2,\infty} = O_P\left(\Big(r_3(d) + \sqrt{\frac{\mu K}{d}}\frac{r_1(d)}{\Delta}\Big)r_1(d)^2/\Delta^2\right).$$

We move on to bound $\|\mathbf{P}\widehat{\mathbf{V}}\mathbf{H}_0 - \mathbf{V}\|_{2,\infty}$,

$$\|\mathbf{P}\widehat{\mathbf{V}}\mathbf{H}_0 - \mathbf{V}\|_{2,\infty} = \|\mathbf{V}(\widehat{\mathbf{H}}_0^\top\mathbf{H}_0 - \mathbf{I}_K)\|_{2,\infty} \le \|\mathbf{V}\|_{2,\infty}\|\mathbf{H}_0 - \widehat{\mathbf{H}}_0\|_2$$

$$= O_P\left(\sqrt{\frac{\mu K}{d}}r_1(d)^2/\Delta^2\right).$$

Finally, we consider the term $\mathbf{P}_\perp\widehat{\mathbf{V}}\widehat{\mathbf{H}}_0 - \mathbf{P}_\perp\mathbf{E}_0\mathbf{V}\boldsymbol{\Lambda}^{-1}$. We can decompose

$$\mathbf{P}_\perp\widehat{\mathbf{V}}\widehat{\mathbf{H}}_0 - \mathbf{P}_\perp\mathbf{E}_0\mathbf{V}\boldsymbol{\Lambda}^{-1} = \mathbf{P}_\perp\widehat{\mathbf{V}}\widehat{\mathbf{H}}_0\boldsymbol{\Lambda}\boldsymbol{\Lambda}^{-1} - \mathbf{P}_\perp\mathbf{E}_0\mathbf{V}\boldsymbol{\Lambda}^{-1}$$

$$= \mathbf{P}_\perp\big(\mathbf{E}\widehat{\mathbf{V}}\widehat{\mathbf{H}}_0 - \mathbf{E}_0\mathbf{V} + \widehat{\mathbf{V}}(\boldsymbol{\Lambda} - \widehat{\boldsymbol{\Lambda}})\widehat{\mathbf{H}}_0 + \widehat{\mathbf{V}}(\widehat{\mathbf{H}}_0\boldsymbol{\Lambda} - \boldsymbol{\Lambda}\widehat{\mathbf{H}}_0)\big)\boldsymbol{\Lambda}^{-1}.$$

We bound the three terms separately, with high probability

$$
\begin{aligned}
&\|\mathbf{P}_\perp(\mathbf{E}\widehat{\mathbf{V}}\widehat{\mathbf{H}}_0-\mathbf{E}_0\mathbf{V})\mathbf{\Lambda}^{-1}\|_{2,\infty}\le\|\mathbf{P}_\perp\mathbf{E}_0(\widehat{\mathbf{V}}\widehat{\mathbf{H}}_0-\mathbf{V})\mathbf{\Lambda}^{-1}\|_{2,\infty}+\|\mathbf{P}_\perp\mathbf{E}_b\widehat{\mathbf{V}}\widehat{\mathbf{H}}_0\mathbf{\Lambda}^{-1}\|_{2,\infty}\\
&\quad\le\|\mathbf{E}_0(\widehat{\mathbf{V}}\widehat{\mathbf{H}}_0-\mathbf{V})\mathbf{\Lambda}^{-1}\|_{2,\infty}+\|\mathbf{V}\mathbf{V}^\top\mathbf{E}_0(\widehat{\mathbf{V}}\widehat{\mathbf{H}}_0-\mathbf{V})\mathbf{\Lambda}^{-1}\|_{2,\infty}+\|\mathbf{E}_b\|_2/\Delta\\
&\quad\le\|\mathbf{E}_0(\widehat{\mathbf{V}}\widehat{\mathbf{H}}_0-\mathbf{V})\|_{2,\infty}/\Delta+\|\mathbf{V}\|_{2,\infty}\|\mathbf{E}_0\|_2\|\widehat{\mathbf{V}}\widehat{\mathbf{H}}_0-\mathbf{V}\|_2/\Delta+r_2(d)/\Delta\\
&\quad=O_P\left(r_4(d,\mathbf{\Lambda})/\Delta+\sqrt{\frac{\mu K}{d}}r_1(d)^2/\Delta^2+r_2(d)/\Delta\right).
\end{aligned}
$$

As for $\mathbf{P}_\perp\widehat{\mathbf{V}}(\widehat{\mathbf{\Lambda}}_K-\mathbf{\Lambda})\widehat{\mathbf{H}}_0\mathbf{\Lambda}^{-1}$, we have

$$
\begin{aligned}
&\|\mathbf{P}_\perp\widehat{\mathbf{V}}(\widehat{\mathbf{\Lambda}}_K-\mathbf{\Lambda})\widehat{\mathbf{H}}_0\mathbf{\Lambda}^{-1}\|_{2,\infty}\le\|(\widehat{\mathbf{V}}\mathbf{H}_0-\mathbf{V})\mathbf{H}_0^\top(\widehat{\mathbf{\Lambda}}_K-\mathbf{\Lambda})\widehat{\mathbf{H}}_0\mathbf{\Lambda}^{-1}\|_{2,\infty}\\
&\qquad+\|\mathbf{V}\mathbf{V}^\top(\widehat{\mathbf{V}}\mathbf{H}_0-\mathbf{V})\mathbf{H}_0^\top(\widehat{\mathbf{\Lambda}}_K-\mathbf{\Lambda})\widehat{\mathbf{H}}_0\mathbf{\Lambda}^{-1}\|_{2,\infty}\\
&\quad\le\|\widehat{\mathbf{V}}\mathbf{H}_0-\mathbf{V}\|_{2,\infty}\|\mathbf{E}_0\|_2/\Delta+\|\mathbf{V}\|_{2,\infty}\|\widehat{\mathbf{V}}\mathbf{H}_0-\mathbf{V}\|_2\|\mathbf{E}_0\|_2/\Delta\\
&\quad=O_P\left\{r_3(d)r_1(d)/\Delta+\sqrt{\frac{\mu K}{d}}r_1(d)^2/\Delta^2\right\},
\end{aligned}
$$

and finally

$$
\begin{aligned}
&\|\mathbf{P}_\perp\widehat{\mathbf{V}}(\mathbf{\Lambda}\widehat{\mathbf{H}}_0-\widehat{\mathbf{H}}_0\mathbf{\Lambda})\mathbf{\Lambda}^{-1}\|_{2,\infty}\le\|(\widehat{\mathbf{V}}\mathbf{H}_0-\mathbf{V})\mathbf{H}_0^\top(\mathbf{\Lambda}\widehat{\mathbf{H}}_0-\widehat{\mathbf{H}}_0\mathbf{\Lambda})\mathbf{\Lambda}^{-1}\|_{2,\infty}\\
&\qquad+\|\mathbf{V}\mathbf{V}^\top(\widehat{\mathbf{V}}\mathbf{H}_0-\mathbf{V})\mathbf{H}_0^\top(\mathbf{\Lambda}\widehat{\mathbf{H}}_0-\widehat{\mathbf{H}}_0\mathbf{\Lambda})\mathbf{\Lambda}^{-1}\|_{2,\infty}\\
&\quad=O_P\left(\big(r_3(d)+\sqrt{\frac{\mu K}{d}}r_1(d)/\Delta\big)\|\mathbf{\Lambda}\widehat{\mathbf{H}}_0-\widehat{\mathbf{H}}_0\mathbf{\Lambda}\|_2/\Delta\right)\\
&\quad=O_P\left(\big(r_3(d)+\sqrt{\frac{\mu K}{d}}r_1(d)/\Delta\big)r_1(d)/\Delta\right),
\end{aligned}
$$

where the last inequality is due to the fact that

$$
\begin{aligned}
\|\mathbf{\Lambda}\widehat{\mathbf{H}}_0-\widehat{\mathbf{H}}_0\mathbf{\Lambda}\|_2&=\|\mathbf{\Lambda}\widehat{\mathbf{V}}^\top\mathbf{V}\mathbf{V}^\top-\widehat{\mathbf{V}}^\top\mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top\|_2\\
&=\|\mathbf{\Lambda}\widehat{\mathbf{V}}^\top\mathbf{V}\mathbf{V}^\top-\widehat{\mathbf{V}}^\top\mathbf{M}\mathbf{V}\mathbf{V}^\top\|_2\\
&\le\|\mathbf{\Lambda}\widehat{\mathbf{V}}^\top\mathbf{V}\mathbf{V}^\top-\widehat{\mathbf{V}}^\top\widehat{\mathbf{M}}\mathbf{V}\mathbf{V}^\top\|_2+\|\widehat{\mathbf{V}}^\top\mathbf{E}\mathbf{V}\mathbf{V}^\top\|_2\\
&=\|(\mathbf{\Lambda}-\widehat{\mathbf{\Lambda}})\widehat{\mathbf{V}}^\top\mathbf{V}\mathbf{V}^\top\|_2+\|\widehat{\mathbf{V}}^\top\mathbf{E}\mathbf{V}\mathbf{V}^\top\|_2\le2\|\mathbf{E}\|_2.
\end{aligned}
$$

Thus in summary, we have

$$
\|\widehat{\mathbf{V}}\mathbf{H}_0-\mathbf{V}-\mathbf{P}_\perp\mathbf{E}_0\mathbf{V}\mathbf{\Lambda}^{-1}\|_{2,\infty}=O_P\left\{\frac{r_3(d)r_1(d)}{\Delta}+\sqrt{\frac{\mu K}{d}}\frac{r_1(d)^2}{\Delta^2}+\frac{r_2(d)+r_4(d)}{\Delta}\right\},
$$

Now we move on to bound $\|\widetilde{\mathbf{V}}\mathbf{H}_1\mathbf{H}_0-\widehat{\mathbf{V}}\mathbf{H}_0\|_2$. By Theorem 4.1, we know that

$$
\begin{aligned}
\|\widetilde{\mathbf{V}}\mathbf{H}_1\mathbf{H}_0-\widehat{\mathbf{V}}\mathbf{H}_0\|_2\le\|\widetilde{\mathbf{V}}\mathbf{H}_1-\widehat{\mathbf{V}}\|_2&\lesssim\|\widetilde{\mathbf{V}}\widetilde{\mathbf{V}}^\top-\widehat{\mathbf{V}}\widehat{\mathbf{V}}^\top\|_2\\
&\le\|\widetilde{\mathbf{V}}\widetilde{\mathbf{V}}^\top-\mathbf{V}'\mathbf{V}'^\top\|_2+\|\mathbf{V}'\mathbf{V}'^\top-\widehat{\mathbf{V}}\widehat{\mathbf{V}}^\top\|_2\\
&\le\|\widetilde{\mathbf{V}}\widetilde{\mathbf{V}}^\top-\mathbf{V}'\mathbf{V}'^\top\|_\mathrm{F}+\|\mathbf{V}'\mathbf{V}'^\top-\widehat{\mathbf{V}}\widehat{\mathbf{V}}^\top\|_2\\
&=O_P\Big(\frac{1}{\sqrt{d}}\frac{r_1(d)^2}{\Delta^2}+\sqrt{\frac{Kd}{\Delta^2pL}}r_1(d)\Big).
\end{aligned}
$$

Finally, we consider $\|\widetilde{\mathbf{V}}^{\mathrm{F}}\mathbf{H} - \widetilde{\mathbf{V}}\mathbf{H}_1\mathbf{H}_0\|_2$. From the proof of Theorem 4.1, we know that

$$\|\widetilde{\mathbf{V}}^{\mathrm{F}}\mathbf{H} - \widetilde{\mathbf{V}}\mathbf{H}_1\mathbf{H}_0\|_2 \le \|\widetilde{\mathbf{V}}^{\mathrm{F}}\mathbf{H}_2 - \widetilde{\mathbf{V}}\|_2 \lesssim \|\widetilde{\mathbf{V}}^{\mathrm{F}}\widetilde{\mathbf{V}}^{\mathrm{F}\top} - \widetilde{\mathbf{V}}\widetilde{\mathbf{V}}^{\top}\|_2$$
$$= O_P\big(\mathbb{E}(\|\widetilde{\mathbf{V}}^{\mathrm{F}}\widetilde{\mathbf{V}}^{\mathrm{F}\top} - \widetilde{\mathbf{V}}\widetilde{\mathbf{V}}^{\top}\|_2^2|\widetilde{\boldsymbol{\Sigma}})^{1/2}\big) \lesssim \sqrt{\frac{d}{p'}}\frac{\|\widetilde{\boldsymbol{\Sigma}} - \mathbf{V}\mathbf{V}^{\top}\|_2^q}{\big(1 - \|\widetilde{\boldsymbol{\Sigma}} - \mathbf{V}\mathbf{V}^{\top}\|_2\big)^q}.$$

From the proof of Theorem 4.1, we know that with probability converging to 1, there exists some constant $\eta > 0$ such that $\|\widetilde{\boldsymbol{\Sigma}} - \mathbf{V}\mathbf{V}^{\top}\|_2 \le \eta r_1(d)\log d\sqrt{d/p}/\Delta = o(1)$, and thus that

$$\|\widetilde{\mathbf{V}}^{\mathrm{F}}\mathbf{H}_2 - \widetilde{\mathbf{V}}\|_2 = O_P\left(\sqrt{\frac{d}{p'}}\left(2\eta r_1(d)\log d\sqrt{\frac{d}{\Delta^2 p}}\right)^q\right).$$

When we choose q to be large enough, i.e.,

$$q \ge 2 + \frac{\log(Ld)}{\log\log d} \gg 1 + \frac{\log\left(\log d\sqrt{Ld/(Kp')}\right)}{\log\left((2\eta\log d)^{-1}\Delta/r_1(d)\sqrt{p/d}\right)},$$

we have $\|\widetilde{\mathbf{V}}^{\mathrm{F}}\mathbf{H}_2 - \widetilde{\mathbf{V}}\|_2 = O_P(\sqrt{\frac{Kd}{\Delta^2 pL}}r_1(d))$. Therefore, if we denote

$$r(d) := \Delta^{-1}\Big(\sqrt{\frac{Kd}{pL}}r_1(d) + r_3(d)r_1(d) + \sqrt{\frac{\mu K}{d\Delta^2}}r_1(d)^2 + r_2(d) + r_4(d)\Big),$$

we can write

$$\widetilde{\mathbf{V}}^{\mathrm{F}}\mathbf{H} - \mathbf{V} = \mathbf{P}_{\perp}\mathbf{E}_0\mathbf{V}\boldsymbol{\Lambda}^{-1} + \mathbf{R}(d),$$

where $\|\mathbf{R}(d)\|_{2,\infty} = O_P\big(r(d)\big)$. Then under the condition that $\eta_1(d)^{-1/2}r(d) = o(1)$, we have that $\|\mathbf{R}(d)\|_{2,\infty} = o_P\Big(\sigma_{\min}\big(\boldsymbol{\Sigma}_i\big)\Big)^{1/2}$. Thus by Assumption 5,

$$\boldsymbol{\Sigma}_i^{-1/2}(\widetilde{\mathbf{V}}^{\mathrm{F}}\mathbf{H} - \mathbf{V})^{\top}\mathbf{e}_i = \boldsymbol{\Sigma}_i^{-1/2}(\boldsymbol{\Lambda}^{-1}\mathbf{V}^{\top}\mathbf{E}_0\mathbf{P}_{\perp}\mathbf{e}_i) + o_P(1) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{I}_K).$$

B.10. Proof of Corollary 4.6. We define $\mathbf{E}_0$ and $\mathbf{E}_b$ the same as in the proof of Corollary 4.11. Then Assumptions 1 and 2 are satisfied as been proven for Corollary 4.11. As for Assumption 5, we have shown that under the condition that $\kappa_1^3(\lambda_1/\sigma^2)^3 = o(\sqrt{n})$, the results (B.30) holds for any matrix $\mathbf{A} \in \mathbb{R}^{d\times K}$ such that $\sigma_{\max}(\mathbf{A})/\sigma_{\min}(\mathbf{A}) \le C|\lambda_1|/\Delta$ and $\lambda_K\big(\mathrm{Cov}(\mathbf{A}^{\top}\mathbf{E}_0\mathbf{P}_{\perp}\mathbf{e}_i)\big) \ge cn^{-1}\sigma^4\big(\sigma_{\min}(\mathbf{A})\big)^2$ in the proof of Corollary 4.11. Under the regime $Lp \gg d$, the leading term $\mathbf{V}(\mathbf{E}_0) = \mathbf{P}_{\perp}\mathbf{E}_0\mathbf{V}\boldsymbol{\Lambda}^{-1}$, and it can be seen that

$$\sigma_{\max}(\mathbf{V}\boldsymbol{\Lambda}^{-1})/\sigma_{\min}(\mathbf{V}\boldsymbol{\Lambda}^{-1}) = \sigma_{\max}(\boldsymbol{\Lambda})/\sigma_{\min}(\boldsymbol{\Lambda}) \le |\lambda_1|/\Delta,$$

and if we can show that $\eta_1(d) \ge (2n)^{-1}\lambda_1^{-2}\sigma^4$, we have $\lambda_K(\boldsymbol{\Sigma}_i) \ge \eta_1(d) = (2n)^{-1}\sigma^4\big(\sigma_{\min}(\mathbf{V}\boldsymbol{\Lambda}^{-1})\big)^2$ and Assumption 5 is satisfied. Thus we only need to verify Assumption 4 and the conditions for $\eta_1(d)$. Recall from the proof of Corollary 4.11 we have the following rates

$$r_1(d) = (\lambda_1 + \sigma^2)\sqrt{\frac{r}{n}}, \quad r_2(d) \asymp \frac{\widetilde{\sigma}_1^3 K}{\delta^2 n}(\log d)^2,$$

and we can further derive that the following bounds hold with high probability

$$\|\widehat{\mathbf{V}}\,\mathrm{sgn}(\widehat{\mathbf{V}}^{\top}\mathbf{V}) - \mathbf{V}\|_{2,\infty} \le \|\widehat{\mathbf{V}}\,\mathrm{sgn}(\widehat{\mathbf{V}}^{\top}\mathbf{V}) - \mathbf{V}\|_2 \lesssim \|\mathbf{E}_0\|_2/\Delta \lesssim r_1(d)\log d/\Delta;$$
$$\|\mathbf{E}_0(\widehat{\mathbf{V}}(\widehat{\mathbf{V}}^{\top}\mathbf{V}) - \mathbf{V})\|_{2,\infty} \lesssim \|\mathbf{E}_0\|_2\|\widehat{\mathbf{V}}\,\mathrm{sgn}(\widehat{\mathbf{V}}^{\top}\mathbf{V}) - \mathbf{V}\|_2 \lesssim r_1(d)^2(\log d)^2/\Delta.$$

Thus we know $r_3(d) \asymp \kappa_1 \log d\sqrt{r/n}$ and $r_4(d) \asymp r_1(d)^2(\log d)^2/\Delta = \kappa_1(\lambda_1+\sigma^2)(\log d)^2 r/n$.

From the proof of Corollary 4.11, we know that $\mathbf{\Sigma}_i = n^{-1}\mathbf{\Lambda}^{-1}\mathbf{V}^\top\mathbf{\Sigma}_i^0\mathbf{V}\mathbf{\Lambda}^{-1}$, where

$$\mathbf{\Sigma}_i^0 = \Big\{\sigma^2\|\mathbf{P}_\perp\mathbf{e}\|_2^2\mathbf{\Sigma} + 3\sigma^4\mathbf{P}_\perp\mathbf{e}_i\mathbf{e}_i^\top\mathbf{P}_\perp - 2\sigma^4\rho\|\mathbf{P}_\perp\mathbf{e}\|_2\big[(\mathbf{P}_\perp)_{[:,S]}\mathbf{u}_{K+1}\mathbf{e}_i^\top\mathbf{P}_\perp + \mathbf{P}_\perp\mathbf{e}_i(\mathbf{u}_{K+1})^\top(\mathbf{P}_\perp)_{[S,:]}\big]\Big\}.$$

Similar as in the proof of Corollary 4.11, we will first define $\mathbf{\Sigma}_i'$ as following

$$\mathbf{\Sigma}_i' = \frac{1}{n}\mathbf{\Lambda}^{-1}\mathbf{V}^\top\Big\{\sigma^2\mathbf{\Sigma} + 3\sigma^4\mathbf{e}_i\mathbf{e}_i^\top - 2\sigma^4\rho\|\mathbf{P}_\perp\mathbf{e}\|_2\big((\mathbf{I}_d)_{[:,S]}\mathbf{u}_{K+1}\mathbf{e}_i^\top + \mathbf{e}_i\mathbf{u}_{K+1}^\top(\mathbf{I}_d)_{[S,:]}\big)\Big\}\mathbf{V}\mathbf{\Lambda}^{-1}.$$

Then following similar arguments as in the proof of Corollary 4.11, we have that

$$\|\mathbf{\Sigma}_i - \mathbf{\Sigma}_i'\|_2 = O\Big(\frac{K\sigma^4}{n\Delta^2}\sqrt{\frac{\mu}{d}}\Big) = O\Big(\frac{K\lambda_1^2}{\Delta^2}\sqrt{\frac{\mu}{d}}\Big)\frac{\sigma^4}{n\lambda_1^2} = o\Big(\frac{\sigma^4}{n\lambda_1^2}\Big).$$

Besides, under the condition that $\mu^2\kappa_1^4K^3 \ll d^2$ we have

$$\|\mathbf{\Sigma}_i' - \widetilde{\mathbf{\Sigma}}_i\|_2 \lesssim \frac{\sqrt{K}\sigma^4}{n\Delta^2}\|\mathbf{V}\|_{2,\infty}^2 \lesssim \frac{\mu K\sqrt{K}\sigma^4}{dn\Delta^2} = O\left(\frac{\mu\kappa_1^2K\sqrt{K}}{d}\right)\frac{\sigma^4}{n\lambda_1^2} = o\Big(\frac{\sigma^4}{n\lambda_1^2}\Big).$$

Then we know that $\lambda_K\big(\mathbf{\Sigma}_i\big) \ge \frac{\sigma^4}{2n\lambda_1^2} + \frac{\sigma^2}{2n\lambda_1}$ and we can take $\eta_1(d) = \frac{\sigma^4}{2n\lambda_1^2} + \frac{\sigma^2}{2n\lambda_1}$. Thus Assumption 5 holds. Then by plugging in the above rates, we can derive the rate $r(d)$ as

$$\begin{aligned} r(d) &= \sqrt{\frac{Kd}{pL}}\frac{r_1(d)}{\Delta} + r_3(d)r_1(d)/\Delta + \sqrt{\frac{\mu K}{d}}r_1(d)^2/\Delta^2 + \big(r_2(d)+r_4(d)\big)/\Delta \\ &\lesssim \kappa_1\sqrt{\frac{Kdr}{npL}} + \frac{\kappa_1^2(\log d)^2 r}{n} + \frac{\widetilde{\sigma}_1^3K}{\delta^2 n\Delta}(\log d)^2. \end{aligned}$$

Then under the condition that $L \gg \frac{Kdr}{p}\kappa_1^2(\frac{\lambda_1}{\sigma^2})$, $n \gg \kappa_1^4(\log d)^4r^2(\frac{\lambda_1}{\sigma^2})$ and $K(\frac{\widetilde{\sigma}_1}{\delta})^2 \ll \kappa_1 r$, we have $\eta_1(d)^{-1/2}r(d) = o(1)$, and hence the condition for $\eta_1(d)$ is satisfied and (8) holds. Also recall from the above proof that $\|\widetilde{\mathbf{\Sigma}}_i - \mathbf{\Sigma}_i\|_2 = o\big(\lambda_K(\mathbf{\Sigma}_i)\big)$, and (9) holds.

Now we verify the validity of $\widehat{\mathbf{\Sigma}}_i$. Similar as in the proof of Corollary 4.11, it suffices to show that $\|\widehat{\mathbf{\Sigma}}_i - \mathbf{H}\widetilde{\mathbf{\Sigma}}_i\mathbf{H}^\top\|_2 = o_P\big(\lambda_K(\widetilde{\mathbf{\Sigma}}_i)\big)$, and the results will hold by Slutsky's Theorem. From proof of Corollary 4.11, we have

$$\|\widehat{\mathbf{\Sigma}}^{\mathrm{tr}} - \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top\|_2 = O_P\Big((\lambda_1+\sigma^2)\sqrt{\frac{r}{n}}\Big).$$

Also, we know that with high probability

$$\begin{aligned} \|\widetilde{\mathbf{V}}^{\mathrm{F}} - \mathbf{V}\mathbf{H}^\top\|_2 = \|\widetilde{\mathbf{V}}^{\mathrm{F}}\mathbf{H} - \mathbf{V}\|_2 &\lesssim \sqrt{\frac{Kd}{\Delta^2pL}}r_1(d)\log d + r_1(d)\log d/\Delta \\ &\lesssim r_1(d)\log d/\Delta \lesssim \kappa_1\log d\sqrt{\frac{r}{n}}. \end{aligned}$$

Then we have

$$\begin{aligned} \|\widetilde{\mathbf{\Lambda}} - \mathbf{H}\mathbf{\Lambda}\mathbf{H}^\top\|_2 &\le \|\widetilde{\mathbf{V}}^{\mathrm{F}\top}(\widehat{\mathbf{\Sigma}}^{\mathrm{tr}} - \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top)\widetilde{\mathbf{V}}^{\mathrm{F}}\|_2 + \|(\widetilde{\mathbf{V}}^{\mathrm{F}} - \mathbf{V}\mathbf{H}^\top)^\top(\mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top)\widetilde{\mathbf{V}}^{\mathrm{F}}\|_2 \\ &+ \|\mathbf{H}\mathbf{V}^\top(\mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top)(\widetilde{\mathbf{V}}^{\mathrm{F}} - \mathbf{V}\mathbf{H}^\top)\|_2 = O_P\Big(\lambda_1\kappa_1\log d\sqrt{\frac{r}{n}}\Big). \end{aligned}$$

Then if we denote $\mathbf{D}_{\mathbf{\Lambda}} = (\widetilde{\mathbf{\Lambda}} - \mathbf{H}\mathbf{\Lambda}\mathbf{H}^\top)\mathbf{H}\mathbf{\Lambda}^{-1}\mathbf{H}^\top$, we have that $\|\mathbf{D}_{\mathbf{\Lambda}}\|_2 = O_P(\kappa_1^2 \log d\sqrt{\frac{r}{n}}) = o_P(1)$, and thus we have

$$\|\widetilde{\mathbf{\Lambda}}^{-1} - \mathbf{H}\mathbf{\Lambda}^{-1}\mathbf{H}^\top\|_2 = \|(\mathbf{H}\mathbf{\Lambda}\mathbf{H}^\top + \widetilde{\mathbf{\Lambda}} - \mathbf{H}\mathbf{\Lambda}\mathbf{H}^\top)^{-1} - (\mathbf{H}\mathbf{\Lambda}\mathbf{H}^\top)^{-1}\|_2$$

$$= \left\|\mathbf{H}\mathbf{\Lambda}^{-1}\mathbf{H}^\top\left[(\mathbf{I}_K + \mathbf{D}_{\mathbf{\Lambda}})^{-1} - \mathbf{I}_K\right]\right\|_2 \le \|\mathbf{\Lambda}^{-1}\|_2 \|\sum_{i=1}^{\infty}(-\mathbf{D}_{\mathbf{\Lambda}})^i\|_2$$

$$= O_P\left(\kappa_1^2 \log d\sqrt{\frac{r}{n}}\right)\Delta^{-1},$$

and furthermore, we have

$$\|\widetilde{\mathbf{\Lambda}}^{-2} - \mathbf{H}\mathbf{\Lambda}^{-2}\mathbf{H}^\top\|_2 \lesssim \|\mathbf{\Lambda}^{-1}\|_2\|\widetilde{\mathbf{\Lambda}}^{-1} - \mathbf{H}\mathbf{\Lambda}^{-1}\mathbf{H}^\top\|_2 = O_P\left(\kappa_1^2 \log d\sqrt{\frac{r}{n}}\right)\Delta^{-2}.$$

Then following basic algebra, under the condition that $n \gg \kappa_1^4(\log d)^4 r^2(\lambda_1/\sigma^2)^2$ we have

$$\|\mathbf{H}\widetilde{\mathbf{\Sigma}}_i\mathbf{H}^\top - \widehat{\mathbf{\Sigma}}_i\|_2 = \frac{1}{n}\|\mathbf{H}(\sigma^2\mathbf{\Lambda}^{-1} + \sigma^4\mathbf{\Lambda}^{-2})\mathbf{H}^\top - (\widehat{\sigma}^2\widetilde{\mathbf{\Lambda}}^{-1} + \widehat{\sigma}^4\widetilde{\mathbf{\Lambda}}^{-2})\|_2$$

$$\le \frac{1}{n}\left(\|\sigma^2\mathbf{H}\mathbf{\Lambda}^{-1}\mathbf{H}^\top - \widehat{\sigma}^2\widetilde{\mathbf{\Lambda}}^{-1}\|_2 + \|\sigma^4\mathbf{H}\mathbf{\Lambda}^{-2}\mathbf{H}^\top - \widehat{\sigma}^4\widetilde{\mathbf{\Lambda}}^{-2}\|_2\right)$$

$$= O_P\left(\kappa_1^2\log d\frac{\sigma^2}{n\Delta}\sqrt{\frac{r}{n}}\right) + O_P\left(\frac{\widetilde{\sigma}_1}{n\Delta}\sqrt{\frac{K}{n}}\right) + O_P\left(\kappa_1^2\log d\frac{\sigma^4}{n\Delta^2}\sqrt{\frac{r}{n}}\right) + O_P\left(\frac{\widetilde{\sigma}_1\sigma^2}{n\Delta^2}\sqrt{\frac{K}{n}}\right)$$

$$= O_P\left(\kappa_1^2\log d(\frac{\Delta}{\sigma^2})\sqrt{\frac{r}{n}}\right)\frac{\sigma^4}{n\Delta^2} = O_P\left(\kappa_1^2\log d(\frac{\lambda_1}{\sigma^2})(\frac{\lambda_1}{\Delta})\sqrt{\frac{r}{n}}\right)\frac{\sigma^4}{n\lambda_1^2}$$

$$= O_P\left(\kappa_1^3\log d(\frac{\lambda_1}{\sigma^2})\sqrt{\frac{r}{n}}\right)\frac{\sigma^4}{n\lambda_1^2} = o_P\big(\lambda_K(\widetilde{\mathbf{\Sigma}}_i)\big).$$

Therefore, by Slutsky's Theorem, the claim follows.

B.11. Proof of Corollary 4.7. The proof for the case where no self-loops are present is almost identical to the case where there are self-loops except for some modifications. We will first prove the results for the case when self-loops are present, then in the end we will discuss how to modify the proof for the case where self-loops are absent.

We only need to verify that Assumptions 1 to 5 hold. Recall from the proof of Corollary 4.2 that we have $\|\|\mathbf{E}\|_2\|_{\psi_1} \lesssim r_1(d) = \sqrt{d\theta}$, and thus we know that Assumption 1 is satisfied. Also Assumption 2 holds trivially due to the unbiasedness of $\mathbf{E}$. We will then verify Assumption 3 holds under the model. We know that $\mathbf{\Theta\Pi}$ and $\mathbf{V}$ share the same column space, and thus there exists a non-singular matrix $\mathbf{C} \in \mathbb{R}^{K\times K}$ such that $\mathbf{\Theta\Pi} = \mathbf{VC}$ and $\mathbf{V} = \mathbf{\Theta\Pi C}^{-1}$. Then we can see that $\sigma_{\min}(\mathbf{C}) = \sigma_{\min}(\mathbf{\Theta\Pi}) \gtrsim \sqrt{d\theta/K}$, and $\|\mathbf{C}^{-1}\|_2 \lesssim \sqrt{K/d\theta}$. Hence we have $\|\mathbf{V}\|_{2,\infty} \le \|\mathbf{\Theta\Pi}\|_{2,\infty}\|\mathbf{C}^{-1}\|_2 \lesssim \sqrt{\theta}\sqrt{K/d\theta} = \sqrt{K/d}$. Thus we can see that Assumption 3 is satisfied with $\mu = O(1)$.

Now we move on to verify Assumption 4. Recall from the proof of Corollary 4.2 that $\Delta \gtrsim d\theta/K$, $\|\mathbf{M}\|_2 \lesssim Kd\theta$, $\mathbf{M}_{ij} \asymp \theta$ and $\max_{ij}\mathbb{E}(\mathbf{E}_{ij}^2) \lesssim \theta$. By Theorem 4.2.1 in Chen et al. [11], we have that with probability $1 - O(d^{-5})$,

$$\|\widehat{\mathbf{V}}\operatorname{sgn}(\widehat{\mathbf{V}}^\top\mathbf{V}) - \mathbf{V}\|_{2,\infty} \lesssim \frac{K^3\sqrt{K} + K\sqrt{K\log d}}{d\sqrt{\theta}}, r_3(d) \asymp \frac{K^3\sqrt{K} + K\sqrt{K\log d}}{d\sqrt{\theta}},$$

and by the proof of Theorem 4.2.1 in [11], we further have that with probability $1 - O(d^{-7})$,

$$\|\mathbf{E}(\widehat{\mathbf{V}}(\widehat{\mathbf{V}}^\top\mathbf{V}) - \mathbf{V})\|_{2,\infty} \lesssim \frac{K\sqrt{K\theta\log d}}{d\theta}\|\mathbf{E}\|_2 + r_3(d)(\log d + \sqrt{d\theta})$$

$$\lesssim r_3(d)(\log d + \sqrt{d\theta}) + K\sqrt{K\log d/d}$$
$$\lesssim \frac{K^3\sqrt{K} + K\sqrt{K\log d}}{\sqrt{d}}, r_4(d) \asymp \frac{K^3\sqrt{K} + K\sqrt{K\log d}}{\sqrt{d}}.$$

Thus Assumption 4 is met and now we move on to study the order of $\eta_1(d)$. Before we continue with the proof, we state the following basic lemma that helps study the operator norm of a covariance matrix.

LEMMA B.5. $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^d$ *are two random vectors, then we have*

$$\|\operatorname{Cov}(\mathbf{x}_1, \mathbf{x}_2)\|_2 = \|\operatorname{Cov}(\mathbf{x}_2, \mathbf{x}_1)\|_2 \leq \sqrt{\|\operatorname{Cov}(\mathbf{x}_1)\|_2 \|\operatorname{Cov}(\mathbf{x}_2)\|_2},$$

and

$$\|\operatorname{Cov}(\mathbf{x}_1 + \mathbf{x}_2)\|_2 \leq 2\|\operatorname{Cov}(\mathbf{x}_1)\|_2 + 2\|\operatorname{Cov}(\mathbf{x}_2)\|_2.$$

The proof of Lemma B.5 can be found in Supplementary Materials C.4. With the help of Lemma B.5, we first decompose $\mathbf{E} = \mathbf{E}_1 + \mathbf{E}_2$, where $\mathbf{E}_1 = [\mathbf{E}_{ij}\mathbb{I}\{i \leq j\}]_{i,j}$ is composed of the diagonal and upper triangular entries of $\mathbf{E}$ and $\mathbf{E}_2 = [\mathbf{E}_{ij}\mathbb{I}\{i > j\}]_{i,j}$ is composed of the off-diagonal lower triangular entries of $\mathbf{E}$. Then it can be seen that both $\mathbf{E}_1$ and $\mathbf{E}_2$ have independent entries. Now for $i \in [d]$, we can write

$$\mathbf{E}\mathbf{P}_\perp \mathbf{e}_i = \mathbf{E}\mathbf{e}_i - \mathbf{E}\mathbf{V}\mathbf{V}^\top \mathbf{e}_i = \mathbf{E}\mathbf{e}_i - (\mathbf{E}_1\mathbf{V}\mathbf{V}^\top \mathbf{e}_i + \mathbf{E}_2\mathbf{V}\mathbf{V}^\top \mathbf{e}_i).$$

Then we study the covariance of the three terms separately. We have

$$\operatorname{Cov}(\mathbf{E}\mathbf{e}_i) = \operatorname{Cov}(\mathbf{E}_{.i}) = \operatorname{diag}\big(\mathbf{M}_{1i}(1-\mathbf{M}_{1i}), \ldots, \mathbf{M}_{di}(1-\mathbf{M}_{di})\big);$$
$$\operatorname{Cov}(\mathbf{E}_1\mathbf{V}\mathbf{V}^\top \mathbf{e}_i) = \operatorname{diag}\Big(\big[\sum_{k=1}^d \mathbf{M}_{jk}(1-\mathbf{M}_{jk})(\mathbf{P}_{\mathbf{V}}\mathbf{e}_i)_k^2 \mathbb{I}\{j \leq k\}\big]_{j=1}^d\Big);$$
$$\operatorname{Cov}(\mathbf{E}_2\mathbf{V}\mathbf{V}^\top \mathbf{e}_i) = \operatorname{diag}\Big(\big[\sum_{k=1}^d \mathbf{M}_{jk}(1-\mathbf{M}_{jk})(\mathbf{P}_{\mathbf{V}}\mathbf{e}_i)_k^2 \mathbb{I}\{j > k\}\big]_{j=1}^d\Big).$$

Then we have $\theta \lesssim \lambda_d\big(\operatorname{Cov}(\mathbf{E}\mathbf{e}_i)\big) \leq \|\operatorname{Cov}(\mathbf{E}\mathbf{e}_i)\|_2 \leq \max_{ij} \mathbb{E}(\mathbf{E}_{ij}^2) \lesssim \theta$ and

$$\|\operatorname{Cov}(\mathbf{E}_1\mathbf{V}\mathbf{V}^\top \mathbf{e}_i)\|_2 \leq \max_{j\in[d]} \sum_{k=1}^d \mathbf{M}_{jk}(1-\mathbf{M}_{jk})(\mathbf{P}_{\mathbf{V}}\mathbf{e}_i)_k^2 \mathbb{I}\{j \leq k\}$$
$$\leq \max_{jk} \mathbb{E}(\mathbf{E}_{jk})^2 \sum_{k=1}^d (\mathbf{P}_{\mathbf{V}}\mathbf{e}_i)_k^2 \lesssim \theta \|\mathbf{P}_{\mathbf{V}}\mathbf{e}_i\|_2^2 \leq \theta \|\mathbf{V}\|_{2,\infty}^2 \leq \frac{\theta K}{d},$$

and very similarly we also have $\|\operatorname{Cov}(\mathbf{E}_2\mathbf{V}\mathbf{V}^\top \mathbf{e}_i)\|_2 \lesssim \theta K/d$. Thus by Lemma B.5, we know that $\|\operatorname{Cov}(\mathbf{E}_1\mathbf{V}\mathbf{V}^\top \mathbf{e}_i + \mathbf{E}_2\mathbf{V}\mathbf{V}^\top \mathbf{e}_i)\|_2 \lesssim \theta K/d$ and

$$\|\operatorname{Cov}(\mathbf{E}_1\mathbf{V}\mathbf{V}^\top \mathbf{e}_i + \mathbf{E}_2\mathbf{V}\mathbf{V}^\top \mathbf{e}_i, \mathbf{E}\mathbf{e}_i)\|_2 \lesssim \sqrt{\theta^2 K/d} = \theta\sqrt{K/d}.$$

Therefore, we can write

$$\|\operatorname{Cov}(\mathbf{E}\mathbf{P}_\perp \mathbf{e}_i) - \operatorname{Cov}(\mathbf{E}\mathbf{e}_i)\|_2 \leq 2\|\operatorname{Cov}(\mathbf{E}_1\mathbf{V}\mathbf{V}^\top \mathbf{e}_i + \mathbf{E}_2\mathbf{V}\mathbf{V}^\top \mathbf{e}_i, \mathbf{E}\mathbf{e}_i)\|_2$$
$$+ \|\operatorname{Cov}(\mathbf{E}_1\mathbf{V}\mathbf{V}^\top \mathbf{e}_i + \mathbf{E}_2\mathbf{V}\mathbf{V}^\top \mathbf{e}_i)\|_2 \lesssim \theta\sqrt{K/d}.$$

Thus we have $\lambda_d\big(\operatorname{Cov}(\mathbf{E}\mathbf{P}_\perp \mathbf{e}_i)\big) \geq \lambda_d\big(\operatorname{Cov}(\mathbf{E}\mathbf{e}_i)\big) - \|\operatorname{Cov}(\mathbf{E}\mathbf{P}_\perp \mathbf{e}_i) - \operatorname{Cov}(\mathbf{E}\mathbf{e}_i)\|_2 \gtrsim \theta$, and we have $\eta_1(d) \asymp \lambda_1^{-2}\theta$. Therefore, when $\theta = K^2 d^{-1/2+\epsilon}$ for some constant $\epsilon > 0$, $p = \Omega(\sqrt{d})$ and $L \gg K^5 d^2/p$,

$K = o(d^{1/18})$, we have that

$$r(d) = \Delta^{-1}\Big(\sqrt{\frac{Kd}{pL}}r_1(d) + r_3(d)r_1(d) + \sqrt{\frac{\mu K}{d\Delta^2}}r_1(d)^2 + r_2(d) + r_4(d)\Big)$$
$$\lesssim \frac{K^4\sqrt{K} + K^2\sqrt{K\log d}}{d^{3/2}\theta} + K\sqrt{\frac{K}{\theta pL}} \ll \frac{1}{Kd\sqrt{\theta}} \lesssim \eta_1(d)^{1/2}.$$

Thus $\eta_1(d)^{-1/2}r(d) = o(1)$ and the condition for the asymptotic covariance matrix is satisfied. Now we need to verify Assumption 5, and similar as in the proof of Corollary 4.11, we can verify the following more general result.

For any matrix $\mathbf{A} \in \mathbb{R}^{d\times K}$ that satisfies the following two conditions: (1)$\|\mathbf{A}\|_{2,\infty}/\sigma_{\min}(\mathbf{A}) \leq C\sqrt{\lambda_1^2\mu K/(d\Delta^2)}$; (2) $\lambda_K\big(\mathbf{\Sigma}_i\big) \geq c\theta\big(\sigma_{\min}(\mathbf{A})\big)^2$, where $\mathbf{\Sigma}_i := \mathrm{Cov}(\mathbf{A}^\top\mathbf{E}_0\mathbf{P}_\perp\mathbf{e}_i)$ and $C, c > 0$ are fixed constants independent of $\mathbf{A}$, it holds that

$$\mathbf{\Sigma}_i^{-1/2}\mathbf{A}^\top\mathbf{E}_0\mathbf{P}_\perp\mathbf{e}_i \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{I}_K). \tag{B.31}$$

It can be checked from the previous proof that $\mathbf{V}\mathbf{\Lambda}^{-1}$ satisfies the two conditions. To show (B.31), we need to show that $\mathbf{a}^\top\mathbf{\Sigma}_i^{-1/2}\mathbf{A}^\top\mathbf{E}\mathbf{P}_\perp\mathbf{e}_i \xrightarrow{d} \mathcal{N}(0,1)$ for any $\mathbf{a} \in \mathbb{R}^K$, $\|\mathbf{a}\|_2 = 1$. We will first study the entries of $\mathbf{P}_\perp\mathbf{e}_i$ and $\mathbf{A}\mathbf{\Sigma}_i^{-1/2}\mathbf{a}$. It holds that

$$|(\mathbf{P}_\perp\mathbf{e}_i)_i| = |\big((\mathbf{I}_d - \mathbf{V}\mathbf{V}^\top)\mathbf{e}_i\big)_i| \leq 1 + \|\mathbf{V}\|_{2,\infty}^2 = 1 + o(1);$$
$$\max_{j\neq i}|(\mathbf{P}_\perp\mathbf{e}_i)_j| = \max_{j\neq i}|\mathbf{e}_j^\top\mathbf{e}_i - \mathbf{e}_j^\top\mathbf{V}\mathbf{V}^\top\mathbf{e}_i| \leq 0 + \|\mathbf{V}\|_{2,\infty}^2 = \frac{K}{d};$$
$$\|\mathbf{A}\mathbf{\Sigma}_i^{-1/2}\mathbf{a}\|_\infty \leq \|\mathbf{A}\|_{2,\infty}\|\mathbf{\Sigma}_i^{-1/2}\|_2 \lesssim \theta^{-1/2}\|\mathbf{A}\|_{2,\infty}/\sigma_{\min}(\mathbf{A}) \lesssim K^2\sqrt{\frac{K}{d\theta}}.$$

Then we know that

$$\mathbf{a}^\top\mathbf{\Sigma}_i^{-1/2}\mathbf{A}^\top\mathbf{E}\mathbf{P}_\perp\mathbf{e}_i = \sum_{jk}\mathbf{E}_{jk}(\mathbf{A}\mathbf{\Sigma}_i^{-1/2}\mathbf{a})_j(\mathbf{P}_\perp\mathbf{e}_i)_k = \sum_{j=1}^d\mathbf{E}_{jj}(\mathbf{A}\mathbf{\Sigma}_i^{-1/2}\mathbf{a})_j(\mathbf{P}_\perp\mathbf{e}_i)_j$$
$$+ \sum_{j<k}\mathbf{E}_{jk}\big[(\mathbf{A}\mathbf{\Sigma}_i^{-1/2}\mathbf{a})_j(\mathbf{P}_\perp\mathbf{e}_i)_k + (\mathbf{A}\mathbf{\Sigma}_i^{-1/2}\mathbf{a})_k(\mathbf{P}_\perp\mathbf{e}_i)_j\big].$$

Then for the diagonal entries we have

$$\sum_{j=1}^d\mathbb{E}|\mathbf{E}_{jj}(\mathbf{A}\mathbf{\Sigma}_i^{-1/2}\mathbf{a})_j(\mathbf{P}_\perp\mathbf{e}_i)_j|^3$$
$$= \mathbb{E}|\mathbf{E}_{ii}(\mathbf{A}\mathbf{\Sigma}_i^{-1/2}\mathbf{a})_i(\mathbf{P}_\perp\mathbf{e}_i)_i|^3 + \sum_{j\neq i}\mathbb{E}|\mathbf{E}_{jj}(\mathbf{A}\mathbf{\Sigma}_i^{-1/2}\mathbf{a})_j(\mathbf{P}_\perp\mathbf{e}_i)_j|^3$$
$$\lesssim \theta\|\mathbf{A}\mathbf{\Sigma}_i^{-1/2}\mathbf{a}\|_\infty^3 + d\theta\|\mathbf{A}\mathbf{\Sigma}_i^{-1/2}\mathbf{a}\|_\infty^3\max_{j\neq i}|(\mathbf{P}_\perp\mathbf{e}_i)_j|^3 \lesssim \frac{K^6}{d}\sqrt{\frac{K^3}{d\theta}},$$

and for the off-diagonal entries, when $K = o(d^{1/26})$ it holds that

$$\sum_{j<k}\mathbb{E}\Big|\mathbf{E}_{jk}\big[(\mathbf{A}\mathbf{\Sigma}_i^{-1/2}\mathbf{a})_j(\mathbf{P}_\perp\mathbf{e}_i)_k + (\mathbf{A}\mathbf{\Sigma}_i^{-1/2}\mathbf{a})_k(\mathbf{P}_\perp\mathbf{e}_i)_j\big]\Big|^3 \lesssim d\theta\|\mathbf{A}\mathbf{\Sigma}_i^{-1/2}\mathbf{a}\|_\infty^3$$
$$+ d^2\theta\|\mathbf{A}\mathbf{\Sigma}_i^{-1/2}\mathbf{a}\|_\infty^3\big(\frac{K}{d}\big)^3 \lesssim K^6\sqrt{\frac{K^3}{d\theta}} = o(1).$$

Moreover, since $\mathrm{Var}(\mathbf{a}^\top \boldsymbol{\Sigma}_i^{-1/2}\mathbf{A}^\top\mathbf{E}\mathbf{P}_\perp\mathbf{e}_i)=1$, by the Lyapunov's condition and plugging in $\mathbf{A}=\mathbf{V}\boldsymbol{\Lambda}^{-1}$, Assumption 5 is met and (8) follows.

Now we only need to verify that the result also holds when replacing $\boldsymbol{\Sigma}_i$ by $\widetilde{\boldsymbol{\Sigma}}_i$. From previous discussion we learnt that

$$\begin{aligned}\|\widetilde{\boldsymbol{\Sigma}}_i-\boldsymbol{\Sigma}_i\|_2 &\le \|\mathbf{V}\boldsymbol{\Lambda}^{-1}\|_2^2\|\operatorname{Cov}(\mathbf{E}\mathbf{P}_\perp\mathbf{e}_i)-\operatorname{Cov}(\mathbf{E}\mathbf{e}_i)\|_2\\ &\le \frac{K^2}{d^2\theta}\sqrt{\frac{K}{d}} \lesssim K^4\sqrt{\frac{K}{d}}\lambda_K(\widetilde{\boldsymbol{\Sigma}}_i)=o(\lambda_K(\widetilde{\boldsymbol{\Sigma}}_i)).\end{aligned}$$

Then by Slutsky's Theorem, (11) holds.

Now we verify the validity of $\widehat{\boldsymbol{\Sigma}}_i$. Similar as in the proof of Corollary 4.6, $\mathbf{H}$ is orthonormal with probability $1-o(1)$, and we will start by showing that $\|\widehat{\boldsymbol{\Sigma}}_i-\mathbf{H}\widetilde{\boldsymbol{\Sigma}}_i\mathbf{H}^\top\|_2=o_P\big(\lambda_K(\widetilde{\boldsymbol{\Sigma}}_i)\big)$. From previous discussion we have the following bounds

$$\|\widehat{\mathbf{M}}-\mathbf{M}\|_2=O_P(\sqrt{d\theta}),\quad \|\widetilde{\mathbf{V}}^{\mathrm{F}}-\mathbf{V}\mathbf{H}^\top\|_2=\|\widetilde{\mathbf{V}}^{\mathrm{F}}\mathbf{H}-\mathbf{V}\|_2=O_P(\frac{K}{\sqrt{d\theta}}),$$

and

$$\begin{aligned}\|\widetilde{\mathbf{V}}^{\mathrm{F}}\mathbf{H}-\mathbf{V}\|_{2,\infty} &\le \|\widetilde{\mathbf{V}}^{\mathrm{F}}\mathbf{H}-\widehat{\mathbf{V}}\mathbf{H}_0\|_2+\|\widehat{\mathbf{V}}\mathbf{H}_0-\mathbf{V}\|_{2,\infty}=o_P(\frac{1}{Kd\sqrt{\theta}})\\ &\quad +O_P(\frac{K^3\sqrt{K}+K\sqrt{K\log d}}{d\sqrt{\theta}})=O_P(\frac{K^3\sqrt{K}+K\sqrt{K\log d}}{d\sqrt{\theta}}).\end{aligned}$$

With the help of the above results, we will study the components of $\widehat{\boldsymbol{\Sigma}}_i$ separately. In the following proof, we will base the discussion on the event that $\mathbf{H}$ is orthonormal. We first study $\widetilde{\mathbf{M}}=(\widetilde{\mathbf{V}}^{\mathrm{F}}\widetilde{\mathbf{V}}^{\mathrm{F}\top})\widehat{\mathbf{M}}(\widetilde{\mathbf{V}}^{\mathrm{F}}\widetilde{\mathbf{V}}^{\mathrm{F}\top})=\widetilde{\mathbf{V}}^{\mathrm{F}}\mathbf{H}(\mathbf{H}^\top\widetilde{\mathbf{V}}^{\mathrm{F}\top}\widehat{\mathbf{M}}\widetilde{\mathbf{V}}^{\mathrm{F}}\mathbf{H})\mathbf{H}^\top\widetilde{\mathbf{V}}^{\mathrm{F}\top}$. We have that

$$\begin{aligned}\|\mathbf{H}^\top\widetilde{\mathbf{V}}^{\mathrm{F}\top}\widehat{\mathbf{M}}\widetilde{\mathbf{V}}^{\mathrm{F}}\mathbf{H}-\boldsymbol{\Lambda}\|_2 &\le \|\mathbf{H}^\top\widetilde{\mathbf{V}}^{\mathrm{F}\top}\widehat{\mathbf{M}}\widetilde{\mathbf{V}}^{\mathrm{F}}\mathbf{H}-\mathbf{H}^\top\widetilde{\mathbf{V}}^{\mathrm{F}\top}\mathbf{M}\widetilde{\mathbf{V}}^{\mathrm{F}}\mathbf{H}\|_2\\ &\quad +\|\mathbf{H}^\top\widetilde{\mathbf{V}}^{\mathrm{F}\top}\mathbf{M}(\widetilde{\mathbf{V}}^{\mathrm{F}}\mathbf{H}-\mathbf{V})\|_2+\|(\widetilde{\mathbf{V}}^{\mathrm{F}}\mathbf{H}-\mathbf{V})^\top\mathbf{M}\mathbf{V}\|_2\\ &\le \|\widehat{\mathbf{M}}-\mathbf{M}\|_2+2\|\mathbf{M}\|_2\|\widetilde{\mathbf{V}}^{\mathrm{F}}\mathbf{H}-\mathbf{V}\|_2=O_P(K^2\sqrt{d\theta}).\end{aligned}$$

Then for $i,j\in[d]$, we have

$$\begin{aligned}|\widetilde{\mathbf{M}}_{ij}-\mathbf{M}_{ij}| &= |(\widetilde{\mathbf{V}}^{\mathrm{F}}\mathbf{H})_i^\top(\mathbf{H}^\top\widetilde{\mathbf{V}}^{\mathrm{F}\top}\widehat{\mathbf{M}}\widetilde{\mathbf{V}}^{\mathrm{F}}\mathbf{H})(\widetilde{\mathbf{V}}^{\mathrm{F}}\mathbf{H})_j-\mathbf{M}_{ij}|\\ &\le |(\widetilde{\mathbf{V}}^{\mathrm{F}}\mathbf{H})_i^\top(\mathbf{H}^\top\widetilde{\mathbf{V}}^{\mathrm{F}\top}\widehat{\mathbf{M}}\widetilde{\mathbf{V}}^{\mathrm{F}}\mathbf{H}-\boldsymbol{\Lambda})(\widetilde{\mathbf{V}}^{\mathrm{F}}\mathbf{H})_j|+|(\widetilde{\mathbf{V}}^{\mathrm{F}}\mathbf{H}-\mathbf{V})_i\boldsymbol{\Lambda}(\widetilde{\mathbf{V}}^{\mathrm{F}}\mathbf{H})_j|\\ &\quad +|(\mathbf{V})_i\boldsymbol{\Lambda}(\widetilde{\mathbf{V}}^{\mathrm{F}}\mathbf{H}-\mathbf{V})_j|.\end{aligned}$$

It is not hard to see that

$$\begin{aligned}&|(\widetilde{\mathbf{V}}^{\mathrm{F}}\mathbf{H})_i^\top(\mathbf{H}^\top\widetilde{\mathbf{V}}^{\mathrm{F}\top}\widehat{\mathbf{M}}\widetilde{\mathbf{V}}^{\mathrm{F}}\mathbf{H}-\boldsymbol{\Lambda})(\widetilde{\mathbf{V}}^{\mathrm{F}}\mathbf{H})_j|\lesssim\|\mathbf{H}^\top\widetilde{\mathbf{V}}^{\mathrm{F}\top}\widehat{\mathbf{M}}\widetilde{\mathbf{V}}^{\mathrm{F}}\mathbf{H}-\boldsymbol{\Lambda}\|_2\|\widetilde{\mathbf{V}}^{\mathrm{F}}\mathbf{H}\|_{2,\infty}^2\\ &=O_P(K^2\sqrt{d\theta}\|\widetilde{\mathbf{V}}^{\mathrm{F}}\mathbf{H}\|_{2,\infty}^2)=O_P\left(K^3\sqrt{\frac{\theta}{d}}\right),\\ &|(\widetilde{\mathbf{V}}^{\mathrm{F}}\mathbf{H}-\mathbf{V})_i\boldsymbol{\Lambda}(\widetilde{\mathbf{V}}^{\mathrm{F}}\mathbf{H})_j|+|(\mathbf{V})_i\boldsymbol{\Lambda}(\widetilde{\mathbf{V}}^{\mathrm{F}}\mathbf{H}-\mathbf{V})_j|\\ &=O_P(Kd\theta\|\mathbf{V}\|_{2,\infty}\|\widehat{\mathbf{V}}\mathbf{H}_0-\mathbf{V}\|_{2,\infty})=O_P\left(K^3(K^2+\sqrt{\log d})\sqrt{\frac{\theta}{d}}\right),\end{aligned}$$

and in turn we have the upper bound

$$
\begin{aligned}
|\widetilde{\mathbf{M}}_{ij} - \mathbf{M}_{ij}| &= O_P\left(K^3\sqrt{\frac{\theta}{d}}\right) + O_P\left(K^3(K^2+\sqrt{\log d})\sqrt{\frac{\theta}{d}}\right) \\
&= O_P\Big(\frac{K^3(K^2+\sqrt{\log d})}{\sqrt{d\theta}}\Big)\theta = o_P(\theta) = o_P(\mathbf{M}_{ij}).
\end{aligned}
$$

Thus we have

$$
\|\operatorname{diag}\big([\widetilde{\mathbf{M}}_{ij}(1-\widetilde{\mathbf{M}}_{ij})]_{j=1}^d\big) - \operatorname{diag}\big([\mathbf{M}_{ij}(1-\mathbf{M}_{ij})]_{j=1}^d\big)\|_2 = O_P\Big(\frac{K^3(K^2+\sqrt{\log d})}{\sqrt{d\theta}}\theta\Big).
$$

Then we move on to study $\widetilde{\mathbf{\Lambda}}$. We have

$$
\begin{aligned}
\|\widetilde{\mathbf{\Lambda}} - \mathbf{H}\mathbf{\Lambda}\mathbf{H}^\top\|_2 &\le \|\widetilde{\mathbf{V}}^{\mathrm{F}\top}(\widehat{\mathbf{M}}-\mathbf{M})\widetilde{\mathbf{V}}^{\mathrm{F}}\|_2 + \|(\widetilde{\mathbf{V}}^{\mathrm{F}} - \mathbf{V}\mathbf{H}^\top)^\top\mathbf{M}\widetilde{\mathbf{V}}^{\mathrm{F}}\|_2 \\
&\quad + \|\mathbf{H}\mathbf{V}^\top\mathbf{M}(\widetilde{\mathbf{V}}^{\mathrm{F}} - \mathbf{V}\mathbf{H}^\top)\|_2 = O_P(\sqrt{d\theta}) + O_P(K^2\sqrt{d\theta}) = O_P(K^2\sqrt{d\theta}).
\end{aligned}
$$

Then if we denote $\mathbf{D}_{\mathbf{\Lambda}} = (\widetilde{\mathbf{\Lambda}} - \mathbf{H}\mathbf{\Lambda}\mathbf{H}^\top)\mathbf{H}\mathbf{\Lambda}^{-1}\mathbf{H}^\top$, we have that $\|\mathbf{D}_{\mathbf{\Lambda}}\|_2 = O_P(K^3/\sqrt{d\theta}) = o_P(1)$, and thus we have

$$
\begin{aligned}
\|\widetilde{\mathbf{\Lambda}}^{-1} - \mathbf{H}\mathbf{\Lambda}^{-1}\mathbf{H}^\top\|_2 &= \|(\mathbf{H}\mathbf{\Lambda}\mathbf{H}^\top + \widetilde{\mathbf{\Lambda}} - \mathbf{H}\mathbf{\Lambda}\mathbf{H}^\top)^{-1} - (\mathbf{H}\mathbf{\Lambda}\mathbf{H}^\top)^{-1}\|_2 \\
&= \Big\|\mathbf{H}\mathbf{\Lambda}^{-1}\mathbf{H}^\top\big[(\mathbf{I}_K + \mathbf{D}_{\mathbf{\Lambda}})^{-1} - \mathbf{I}_K\big]\Big\| \le \|\mathbf{\Lambda}^{-1}\|_2\Big\|\sum_{i=1}^{\infty}(-\mathbf{D}_{\mathbf{\Lambda}})^i\Big\|_2 \\
&= O_P\big(K^4/(d\theta)^{3/2}\big).
\end{aligned}
$$

Thus, following basic algebra we have the following bounds

$$
\begin{aligned}
&\|\widetilde{\mathbf{V}}^{\mathrm{F}\top}\operatorname{diag}\big([\widetilde{\mathbf{M}}_{ij}(1-\widetilde{\mathbf{M}}_{ij})]_{j=1}^d\big)\widetilde{\mathbf{V}}^{\mathrm{F}} - \mathbf{H}\mathbf{V}^\top\operatorname{diag}\big([\mathbf{M}_{ij}(1-\mathbf{M}_{ij})]_{j=1}^d\big)\mathbf{V}\mathbf{H}^\top\|_2 \\
&\le \|\widetilde{\mathbf{V}}^{\mathrm{F}\top}\Big(\operatorname{diag}\big([\widetilde{\mathbf{M}}_{ij}(1-\widetilde{\mathbf{M}}_{ij})]_{j=1}^d\big) - \operatorname{diag}\big([\mathbf{M}_{ij}(1-\mathbf{M}_{ij})]_{j=1}^d\big)\Big)\widetilde{\mathbf{V}}^{\mathrm{F}}\|_2 \\
&\quad + 2\|\widetilde{\mathbf{V}}^{\mathrm{F}} - \mathbf{V}\mathbf{H}^\top\|_2\|\operatorname{diag}\big([\mathbf{M}_{ij}(1-\mathbf{M}_{ij})]_{j=1}^d\big)\|_2 = O_P\Big(\frac{K^3(K^2+\sqrt{\log d})}{\sqrt{d\theta}}\Big)\theta,
\end{aligned}
$$

and further, under the condition that $K = o(d^{1/32})$, we have

$$
\begin{aligned}
\|\widehat{\mathbf{\Sigma}}_i - \mathbf{H}\widetilde{\mathbf{\Sigma}}_i\mathbf{H}^\top\|_2 &\lesssim O_P\Big(\frac{K^3(K^2+\sqrt{\log d})}{\sqrt{d\theta}}\theta\Big)\|\widetilde{\mathbf{\Lambda}}^{-1}\|_2^2 + \theta\|\mathbf{\Lambda}^{-1}\|_2\|\widetilde{\mathbf{\Lambda}}^{-1} - \mathbf{H}\mathbf{\Lambda}^{-1}\mathbf{H}^\top\|_2 \\
&= O_P\Big(\frac{K^7(K^2+\sqrt{\log d})}{\sqrt{d\theta}}\Big)\frac{1}{K^2d^2\theta} + O_P\Big(\frac{K^7}{\sqrt{d\theta}}\Big)\frac{1}{K^2d^2\theta} \\
&= O_P\Big(\frac{K^7(K^2+\sqrt{\log d})}{\sqrt{d\theta}}\Big)\frac{1}{K^2d^2\theta} = o_P\big(\lambda_K(\widetilde{\mathbf{\Sigma}}_i)\big).
\end{aligned}
$$

Thus with similar arguments as in the proof of Corollary 4.6, the claim follows.

REMARK 19. The inferential results also hold for the case where self-loops are absent. We define $\widehat{\mathbf{M}}' = \mathbf{M} + \mathbf{E} - \operatorname{diag}(\mathbf{E})$ and denote by $\widehat{\mathbf{V}}'$ its K leading eigenvectors. By Weyl's inequality [17] we know that with probability at least $1 - d^{-10}$ we have that $\sigma_K(\widehat{\mathbf{M}}') - \sigma_{K+1}(\widehat{\mathbf{M}}') \ge \Delta - O(\sqrt{d\theta}) \gtrsim d\theta/K$, and hence by Davis-Kahan's Theorem [42] we have

$$
\|\widehat{\mathbf{V}}\widehat{\mathbf{V}}^\top - \widehat{\mathbf{V}}'\widehat{\mathbf{V}}'^\top\|_2 \le r_2(d)/\big(\sigma_K(\widehat{\mathbf{M}}') - \sigma_{K+1}(\widehat{\mathbf{M}}')\big) \lesssim K/d,
$$

with probability at least $1-d^{-10}$. The verification of Assumptions 1, 3 and 5 when self-loops are present can also be applied to the no-self-loop case. For Assumption 2, we can take $\mathbf{E}_0=\mathbf{E}-\operatorname{diag}(\mathbf{E})$ and $\mathbf{E}_b=-\operatorname{diag}(\mathbf{M})$. Then $r_2(d)=\|\operatorname{diag}(\mathbf{M})\|_2\lesssim\theta=o(r_1(d))$ and Assumption 2 is satisfied. As for Assumption 4, by Lemma 7 in Fan et al. [15], we have

$$\|\operatorname{sgn}(\widehat{\mathbf{V}}'^{\top}\mathbf{V})-\widehat{\mathbf{V}}'^{\top}\mathbf{V}\|_2\lesssim\|\widehat{\mathbf{V}}'\widehat{\mathbf{V}}'^{\top}-\mathbf{V}\mathbf{V}^{\top}\|_2^2\lesssim\frac{K^2}{d\theta},\ \|\operatorname{sgn}(\widehat{\mathbf{V}}^{\top}\widehat{\mathbf{V}}')-\widehat{\mathbf{V}}^{\top}\widehat{\mathbf{V}}'\|_2\lesssim\|\widehat{\mathbf{V}}\widehat{\mathbf{V}}^{\top}-\widehat{\mathbf{V}}'\widehat{\mathbf{V}}'^{\top}\|_2^2\lesssim\frac{K^2}{d^2}.$$

With similar arguments as in the self-loop case, for $\widehat{\mathbf{V}}'$ with high probability we have

$$\|\widehat{\mathbf{V}}'\operatorname{sgn}(\widehat{\mathbf{V}}'^{\top}\mathbf{V})-\mathbf{V}\|_{2,\infty}\lesssim\frac{K^3\sqrt{K}+K\sqrt{K\log d}}{d\sqrt{\theta}},\quad \|\mathbf{E}_0(\widehat{\mathbf{V}}'(\widehat{\mathbf{V}}'^{\top}\mathbf{V})-\mathbf{V})\|_{2,\infty}\lesssim\frac{K^3\sqrt{K}+K\sqrt{K\log d}}{\sqrt{d}}.$$

Then for $\widehat{\mathbf{V}}$, with high probability we have that

$$\begin{aligned}
&\|\widehat{\mathbf{V}}\operatorname{sgn}(\widehat{\mathbf{V}}^{\top}\mathbf{V})-\mathbf{V}\|_{2,\infty}\le\|\widehat{\mathbf{V}}(\widehat{\mathbf{V}}^{\top}\mathbf{V})-\mathbf{V}\|_{2,\infty}+\|\widehat{\mathbf{V}}\big(\operatorname{sgn}(\widehat{\mathbf{V}}^{\top}\mathbf{V})-\widehat{\mathbf{V}}^{\top}\mathbf{V}\big)-\mathbf{V}\|_{2,\infty}\\
&\quad\le\|\widehat{\mathbf{V}}\big(\widehat{\mathbf{V}}^{\top}(\mathbf{I}_d-\widehat{\mathbf{V}}'\widehat{\mathbf{V}}'^{\top})\mathbf{V}\big)-\mathbf{V}\|_{2,\infty}+\|\widehat{\mathbf{V}}(\widehat{\mathbf{V}}^{\top}\widehat{\mathbf{V}}'\widehat{\mathbf{V}}'^{\top}\mathbf{V})-\mathbf{V}\|_{2,\infty}+O(\frac{K^2}{d\theta})\|\widehat{\mathbf{V}}\|_{2,\infty}\\
&\quad\le\|\widehat{\mathbf{V}}(\widehat{\mathbf{V}}^{\top}\widehat{\mathbf{V}}')-\widehat{\mathbf{V}}'\|_{2,\infty}+\|\widehat{\mathbf{V}}'(\widehat{\mathbf{V}}'^{\top}\mathbf{V})-\mathbf{V}\|_{2,\infty}+O(\frac{K^2}{d\theta})\|\widehat{\mathbf{V}}\|_{2,\infty}+\|\widehat{\mathbf{V}}\widehat{\mathbf{V}}^{\top}-\widehat{\mathbf{V}}'\widehat{\mathbf{V}}'^{\top}\|_2\|\widehat{\mathbf{V}}\|_{2,\infty}\\
&\quad\le O(\frac{K^2}{d\theta})\Big(\|\mathbf{V}\|_{2,\infty}+\|\widehat{\mathbf{V}}\operatorname{sgn}(\widehat{\mathbf{V}}^{\top}\mathbf{V})-\mathbf{V}\|_{2,\infty}\Big)+\|\widehat{\mathbf{V}}\widehat{\mathbf{V}}^{\top}-\widehat{\mathbf{V}}'\widehat{\mathbf{V}}'^{\top}\|_2+\|\widehat{\mathbf{V}}'(\widehat{\mathbf{V}}'^{\top}\mathbf{V})-\mathbf{V}\|_{2,\infty},
\end{aligned}$$

where in the last two inequalities we use the fact that

$$\|(\mathbf{I}_d-\widehat{\mathbf{V}}\widehat{\mathbf{V}}^{\top})\widehat{\mathbf{V}}'\|_2=\|(\mathbf{I}_d-\widehat{\mathbf{V}}'\widehat{\mathbf{V}}'^{\top})\widehat{\mathbf{V}}\|_2=\|\widehat{\mathbf{V}}_{\perp}^{\top}\widehat{\mathbf{V}}'\|_2=\|\widehat{\mathbf{V}}_{\perp}'^{\top}\widehat{\mathbf{V}}\|_2=\|\widehat{\mathbf{V}}\widehat{\mathbf{V}}^{\top}-\widehat{\mathbf{V}}'\widehat{\mathbf{V}}'^{\top}\|_2,$$

with $\widehat{\mathbf{V}}_{\perp}$ and $\widehat{\mathbf{V}}'_{\perp}$ being the orthogonal complement of $\widehat{\mathbf{V}}$ and $\widehat{\mathbf{V}}'$ respectively. Since $K^2/(d\theta)=o(1)$, for large enough d we further get

$$\begin{aligned}
\frac{1}{2}\|\widehat{\mathbf{V}}\operatorname{sgn}(\widehat{\mathbf{V}}^{\top}\mathbf{V})-\mathbf{V}\|_{2,\infty}&\le\big(1-O\left(K^2/(d\theta)\right)\big)\|\widehat{\mathbf{V}}\operatorname{sgn}(\widehat{\mathbf{V}}^{\top}\mathbf{V})-\mathbf{V}\|_{2,\infty}\\
&\le O(\frac{K^2}{d\theta})\|\mathbf{V}\|_{2,\infty}+O(\frac{K}{d})+\|\widehat{\mathbf{V}}'\operatorname{sgn}(\widehat{\mathbf{V}}'^{\top}\mathbf{V})-\mathbf{V}\|_{2,\infty}+O(\frac{K^2}{d\theta})\|\widehat{\mathbf{V}}'\|_{2,\infty}\\
&\lesssim\frac{K^2}{d\theta}\sqrt{\frac{K}{d}}+\frac{K}{d}+\frac{K^3\sqrt{K}+K\sqrt{K\log d}}{d\sqrt{\theta}}\lesssim\frac{K^3\sqrt{K}+K\sqrt{K\log d}}{d\sqrt{\theta}}.
\end{aligned}$$

Hence $r_3(d)=K\sqrt{K}(K^2+\sqrt{\log d})/(d\sqrt{\theta})$. We also have

$$\begin{aligned}
\|\mathbf{E}(\widehat{\mathbf{V}}(\widehat{\mathbf{V}}^{\top}\mathbf{V})-\mathbf{V})\|_{2,\infty}&\lesssim\|\mathbf{E}_0(\widehat{\mathbf{V}}(\widehat{\mathbf{V}}^{\top}\mathbf{V})-\mathbf{V})\|_{2,\infty}+\frac{r_2(d)r_1(d)}{\Delta}\lesssim\|\mathbf{E}_0(\widehat{\mathbf{V}}'(\widehat{\mathbf{V}}'^{\top}\mathbf{V})-\mathbf{V})\|_{2,\infty}+\frac{r_2(d)r_1(d)}{\Delta}\\
&\lesssim\frac{K^3\sqrt{K}+K\sqrt{K\log d}}{\sqrt{d}}+K\sqrt{\frac{\theta}{d}}\lesssim\frac{K^3\sqrt{K}+K\sqrt{K\log d}}{\sqrt{d}},
\end{aligned}$$

and hence we can take $r_4(d)=K\sqrt{K}(K^2+\sqrt{\log d})/\sqrt{d}$. Now to get a sharper rate for $r(d)$, we take into consideration the diagonal structure of $\mathbf{E}_b$ and derive the following bound

$$\begin{aligned}
\|\mathbf{P}_{\perp}\mathbf{E}_b\widehat{\mathbf{V}}\widehat{\mathbf{H}}_0\mathbf{\Lambda}^{-1}\|_{2,\infty}&\le\|\mathbf{V}\mathbf{V}^{\top}\mathbf{E}_b\widehat{\mathbf{V}}\widehat{\mathbf{H}}_0\mathbf{\Lambda}^{-1}\|_{2,\infty}+\|\mathbf{E}_b\widehat{\mathbf{V}}\widehat{\mathbf{H}}_0\mathbf{\Lambda}^{-1}\|_{2,\infty}\le\frac{r_2(d)\|\mathbf{V}\|_{2,\infty}}{\Delta}+\frac{\|\operatorname{diag}(\mathbf{M})\widehat{\mathbf{V}}\|_{2,\infty}}{\Delta}\\
&\lesssim\frac{K}{d}\sqrt{\frac{K}{d}}+\frac{\|\operatorname{diag}(\mathbf{M})\|_2\|\widehat{\mathbf{V}}\|_{2,\infty}}{\Delta}\lesssim\frac{K}{d}\sqrt{\frac{K}{d}}.
\end{aligned}$$

Then from the proof of Theorem 4.5 we have that

$$r(d) \lesssim \frac{K^4\sqrt{K} + K^2\sqrt{K\log d}}{d^{3/2}\theta} + K\sqrt{\frac{K}{\theta pL}} + \frac{K}{d}\sqrt{\frac{K}{d}} \ll \frac{1}{Kd\sqrt{\theta}},$$

and we are only left to verify the minimum eigenvalue condition of $\widehat{\boldsymbol{\Sigma}}_i$ by showing that the order of $\eta_1(d)$ is the same as when there are self-loops. With the same arguments, we know that

$$\|\operatorname{Cov}(\boldsymbol{\Lambda}^{-1}\mathbf{V}^\top\mathbf{E}'\mathbf{P}_\perp\mathbf{e}_i) - \operatorname{Cov}(\boldsymbol{\Lambda}^{-1}\mathbf{V}^\top\mathbf{E}'\mathbf{e}_i)\| \le O\left(K^4\sqrt{\frac{K}{d}}\right)\frac{1}{K^2d^2\theta}.$$

Besides, we also have

$$\begin{aligned}\|\operatorname{Cov}(\boldsymbol{\Lambda}^{-1}\mathbf{V}^\top\mathbf{E}'\mathbf{e}_i) - \widetilde{\boldsymbol{\Sigma}}_i\|_2 &= \big\|\boldsymbol{\Lambda}^{-1}\mathbf{V}^\top\big(\mathbf{M}_{ii}(1-\mathbf{M}_{ii})\mathbf{e}_i\mathbf{e}_i^\top\big)\mathbf{V}\boldsymbol{\Lambda}^{-1}\big\|_2\\ &\lesssim \mathbf{M}_{ii}\|\boldsymbol{\Lambda}^{-1}\|_2^2\|\mathbf{V}\|_{2,\infty}^2 \lesssim \frac{K^2}{d^2\theta}\frac{K}{d} = O(\frac{K^5}{d})\frac{1}{K^2d^2\theta} = o\big(\lambda_K(\widetilde{\boldsymbol{\Sigma}}_i)\big).\end{aligned}$$

Thus we also have $\|\operatorname{Cov}(\boldsymbol{\Lambda}^{-1}\mathbf{V}^\top\mathbf{E}'\mathbf{P}_\perp\mathbf{e}_i) - \widetilde{\boldsymbol{\Sigma}}_i\|_2 = o\big(\lambda_K(\widetilde{\boldsymbol{\Sigma}}_i)\big)$, and thereby

$$\lambda_K\big(\operatorname{Cov}(\boldsymbol{\Lambda}^{-1}\mathbf{V}^\top\mathbf{E}'\mathbf{P}_\perp\mathbf{e}_i)\big) = \lambda_K\big(\widetilde{\boldsymbol{\Sigma}}_i\big)\big(1+o(1)\big) \gtrsim \frac{\theta}{K^2d^2\theta^2}.$$

Thus we still have $\eta_1(d) = \lambda_1^{-2}\theta$ for the case where self-loops are absent. The condition for $\eta_1(d)$ also holds for the no-self-loop case and both (8) and (11) hold. The verification of (12) is identical to the self-loop case and is hence omitted.

B.12. Proof of Corollary 4.8. From the proof of Corollary 4.12, we have verified Assumptions 1-3. It can be checked that $\mathbf{V}\boldsymbol{\Lambda}^{-1}$ satisfies the two conditions for the general CLT results in the proof of Corollary 4.12, then under the condition that $\Delta_0^2 \ll n^{4/3}/(\mu_\theta^2 d)$, Assumption 5 is also satisfied.

Now we move on to check the conditions for $\eta_1(d)$. Recall from the proof of Corollary 4.12, we have

$$\operatorname{Cov}(\mathbf{E}_0\mathbf{P}_\perp\mathbf{e}_i) = \|\mathbf{P}_\perp\mathbf{e}_i\|_2^2(\mathbf{F}^*\boldsymbol{\Theta}^{*\top}\boldsymbol{\Theta}^*\mathbf{F}^{*\top} + n\mathbf{I}_d) + n\mathbf{P}_\perp\mathbf{e}_i\mathbf{e}_i^\top\mathbf{P}_\perp.$$

Then we have

$$\|\widetilde{\boldsymbol{\Sigma}}_i - \boldsymbol{\Sigma}_i\|_2 \lesssim \frac{K}{d\Delta^2}(n+\Delta) \lesssim O\left(\frac{K}{dn}(n+\Delta)\right)\frac{n}{\lambda_1^2} = o\left(\frac{n}{\lambda_1^2}\right).$$

Besides, it can be seen that $\lambda_K(\widetilde{\boldsymbol{\Sigma}}_i) \ge n/\lambda_1^2 + 1/\lambda_1$, and hence we can take $\eta_1(d) = \lambda_1^{-2}n/2 + \lambda_1^{-1}/2$. Next we move on to verify the statistical rates $r_3(d)$ and $r_4(d)$. By Davis-Kahan's Theorem [42], we have that with high probability

$$\|\widehat{\mathbf{V}}\operatorname{sgn}(\widehat{\mathbf{V}}^\top\mathbf{V}) - \mathbf{V}\|_{2,\infty} \le \|\widehat{\mathbf{V}}\operatorname{sgn}(\widehat{\mathbf{V}}^\top\mathbf{V}) - \mathbf{V}\|_2 \lesssim \|\mathbf{E}\|_2/\Delta \lesssim r_1'(d)/\Delta,$$

where $r_1'(d) = d\Delta_0/\sqrt{K} + \sqrt{dn}\log d$ as defined in the proof of Corollary 4.12, and thus we know that $r_3(d) \asymp r_1'(d)/\Delta$. Besides, with high probability we have

$$\|\mathbf{E}_0(\widehat{\mathbf{V}}(\widehat{\mathbf{V}}^\top\mathbf{V}) - \mathbf{V})\|_{2,\infty} \le \|\mathbf{E}_0(\widehat{\mathbf{V}}(\widehat{\mathbf{V}}^\top\mathbf{V}) - \mathbf{V})\|_2 \lesssim r_1'(d)^2/\Delta,$$

and we have $r_4(d) \asymp r_1'(d)^2/\Delta$. Thus Assumption 4 is satisfied. Then we have

$$\begin{aligned} r(d) &= \sqrt{\frac{Kd}{pL}}\frac{r_1(d)}{\Delta} + r_3(d)r_1(d)/\Delta + \sqrt{\frac{K}{d}}r_1(d)^2/\Delta^2 + \big(r_2(d)+r_4(d)\big)/\Delta\\ &\lesssim \sqrt{\frac{Kd}{pL}}\frac{r_1(d)}{\Delta} + r_1'(d)^2/\Delta^2 \lesssim \frac{K}{\Delta_0^2} + \frac{K^2n(\log d)^2}{d\Delta_0^4} + \sqrt{\frac{Kd}{pL}}\Big(\frac{\sqrt{K}}{\Delta_0} + \frac{K\sqrt{n}}{\sqrt{d}\Delta_0^2}\Big).\end{aligned}$$

Therefore, under the conditions that $\Delta_0^2 \gg K\sqrt{n}(\log d)^2$, $n \gg d^2$ and $L \gg Kd^2/p$, we have $\eta_1(d)^{-1/2}r(d) = o(1)$. Thus by Theorem 4.5, (8) holds. As for (13), from the above arguments we have $\|\widetilde{\boldsymbol{\Sigma}}_i - \boldsymbol{\Sigma}_i\|_2 = o\big(\lambda_K(\boldsymbol{\Sigma}_i)\big)$, and hence (13) holds.

Now we need to check the validity of $\widehat{\boldsymbol{\Sigma}}_i$. Similar as before, it suffices for us to prove that $\|\widehat{\boldsymbol{\Sigma}}_i - \mathbf{H}\widetilde{\boldsymbol{\Sigma}}_i\mathbf{H}^\top\|_2 = o_P\big(\lambda_K(\widetilde{\boldsymbol{\Sigma}}_i)\big)$. From Corollary 4.8, we have that $\|\widehat{\mathbf{M}} - \mathbf{M}\|_2 = O_P(d\Delta_0/\sqrt{K} + \sqrt{dn})$ and $\|\widetilde{\mathbf{V}}^{\mathrm{F}}\mathbf{H} - \mathbf{V}\|_2 = \|\widetilde{\mathbf{V}}^{\mathrm{F}} - \mathbf{V}\mathbf{H}^\top\|_2 = O_P\big(K\sqrt{\frac{n}{d}}/\Delta_0^2\big)$. Then we have

$$\begin{aligned}\|\widetilde{\boldsymbol{\Lambda}} - \mathbf{H}\boldsymbol{\Lambda}\mathbf{H}^\top\|_2 &\le \|\widetilde{\mathbf{V}}^{\mathrm{F}\top}(\widehat{\mathbf{M}} - \mathbf{M})\widetilde{\mathbf{V}}^{\mathrm{F}}\|_2 + \|(\widetilde{\mathbf{V}}^{\mathrm{F}} - \mathbf{V}\mathbf{H}^\top)^\top\mathbf{M}\widetilde{\mathbf{V}}^{\mathrm{F}}\|_2 \\ &\quad + \|\mathbf{H}\mathbf{V}^\top\mathbf{M}(\widetilde{\mathbf{V}}^{\mathrm{F}} - \mathbf{V}\mathbf{H}^\top)\|_2 = O_P\Big(d\Delta_0/\sqrt{K} + \sqrt{dn}\Big).\end{aligned}$$

Then if we denote $\mathbf{D}_{\boldsymbol{\Lambda}} = (\widetilde{\boldsymbol{\Lambda}} - \mathbf{H}\boldsymbol{\Lambda}\mathbf{H}^\top)\mathbf{H}\boldsymbol{\Lambda}^{-1}\mathbf{H}^\top$, we have that

$$\|\mathbf{D}_{\boldsymbol{\Lambda}}\|_2 = O_P\left(K\sqrt{\frac{n}{d}}\Delta_0^{-2}\right) = o_P(1),$$

and thus we have

$$\|\widetilde{\boldsymbol{\Lambda}}^{-1} - \mathbf{H}\boldsymbol{\Lambda}^{-1}\mathbf{H}^\top\|_2 \lesssim \|\boldsymbol{\Lambda}^{-1}\|_2\|\mathbf{D}_{\boldsymbol{\Lambda}}\|_2 = O_P\left(K\sqrt{\frac{n}{d}}\Delta_0^{-2}\right)\Delta^{-1} = o_P(n/\lambda_1^2) = o_P\big(\lambda_K(\widetilde{\boldsymbol{\Sigma}}_i)\big),$$

and furthermore, we have

$$n\|\widetilde{\boldsymbol{\Lambda}}^{-2} - \mathbf{H}\boldsymbol{\Lambda}^{-2}\mathbf{H}^\top\|_2 \lesssim n\|\boldsymbol{\Lambda}^{-1}\|_2\|\widetilde{\boldsymbol{\Lambda}}^{-1} - \mathbf{H}\boldsymbol{\Lambda}^{-1}\mathbf{H}^\top\|_2 = O_P\big(K\sqrt{\frac{n}{d}}/\Delta_0^2\big)n\Delta^{-2} = o_P\big(\lambda_K(\widetilde{\boldsymbol{\Sigma}}_i)\big).$$

Combining the above results, we have $\|\widehat{\boldsymbol{\Sigma}}_i - \mathbf{H}\widetilde{\boldsymbol{\Sigma}}_i\mathbf{H}^\top\|_2 = o_P\big(\lambda_K(\widetilde{\boldsymbol{\Sigma}}_i)\big)$, and hence (13) holds with $\widetilde{\boldsymbol{\Sigma}}_i$ replaced by $\widehat{\boldsymbol{\Sigma}}_i$.

B.13. Proof of Corollary 4.9. Recall that $\widehat{\mathbf{M}} = (1/\widehat{\theta})\mathcal{P}_{\mathcal{S}}(\mathbf{M} + \bar{\mathcal{E}})$ and $\widehat{\mathbf{M}}' = (1/\theta)\mathcal{P}_{\mathcal{S}}(\mathbf{M} + \bar{\mathcal{E}})$ share exactly the same sequence of eigenvectors, and we can treat $\widetilde{\mathbf{V}}^{\mathrm{F}}$ as the FADI estimator applied to $\widehat{\mathbf{M}}'$. We will abuse the notation and denote $\mathbf{E} := \widehat{\mathbf{M}}' - \mathbf{M}$.

To show that (8) holds, we need to verify that Assumptions 1 to 5 hold and the minimum eigenvalue conditions hold for the asymptotic covariance matrix. We know from Corollary 4.2 that Assumption 1 and Assumption 2 are satisfied, and that $r_1(d) = |\lambda_1|\mu K/\sqrt{d\theta} + \sqrt{d\sigma^2/\theta}$ and $r_2(d) = 0$. Define $\widetilde{\sigma} = (|\lambda_1|\mu K/d) \vee \sigma$, we have from the proof of Corollary 4.2 that $\mathrm{Var}(\mathbf{E}_{ij}) \asymp \widetilde{\sigma}^2/\theta$ and $|\mathbf{E}_{ij}| = O(\widetilde{\sigma}\log d/\theta)$ for $i, j \in [d]$. From Theorem 4.2.1 in Chen et al. [11], we have that with probability $1 - O(d^{-5})$

$$\|\widehat{\mathbf{V}}\,\mathrm{sgn}(\widehat{\mathbf{V}}^\top\mathbf{V}) - \mathbf{V}\|_{2,\infty} \lesssim \frac{\kappa_2\widetilde{\sigma}\sqrt{\mu K/\theta} + \widetilde{\sigma}\sqrt{K\log d/\theta}}{\Delta},$$

and thus we know $r_3(d) \asymp \big(\kappa_2\widetilde{\sigma}\sqrt{\mu K/\theta} + \widetilde{\sigma}\sqrt{K\log d/\theta}\big)/\Delta$. Besides, by the proof of Theorem 4.2.1 in Chen et al. [11], with probability $1 - O(d^{-7})$, we have

$$\begin{aligned}\big\|\mathbf{E}\big(\widehat{\mathbf{V}}(\widehat{\mathbf{V}}^\top\mathbf{V}) - \mathbf{V}\big)\big\|_{2,\infty} &\lesssim \frac{\sqrt{dK}\widetilde{\sigma}^2}{\Delta\theta}\big(\sqrt{\log d} + \sqrt{\mu}\big) + \widetilde{\sigma}\sqrt{\frac{d}{\theta}}r_3(d) + \frac{\widetilde{\sigma}}{\Delta}\sqrt{K\frac{\log d}{\theta}}\|\mathbf{E}\|_2 \\ &\lesssim \frac{\sqrt{d}\widetilde{\sigma}^2}{\Delta\theta}\big(\sqrt{K\log d} + \kappa_2\sqrt{\mu K}\big),\end{aligned}$$

and thus $r_4(d) \asymp \frac{\sqrt{d}\widetilde{\sigma}^2}{\Delta\theta}\big(\sqrt{K\log d} + \kappa_2\sqrt{\mu K}\big)$. Therefore, Assumption 4 is met and we have

$$r(d) = \sqrt{\frac{Kd}{pL}\frac{r_1(d)}{\Delta}} + r_3(d)r_1(d)/\Delta + \sqrt{\frac{\mu K}{d}}r_1(d)^2/\Delta^2 + \big(r_2(d) + r_4(d)\big)/\Delta$$

$$\lesssim \left(\frac{\sqrt{d}\widetilde{\sigma}}{\Delta\sqrt{\theta}}\right)\left(\left(\frac{\kappa_2\sqrt{\mu K}+\sqrt{K\log d}}{\Delta}\right)\frac{\widetilde{\sigma}}{\sqrt{\theta}}+\sqrt{\frac{Kd}{pL}}\right).$$

Now we will study the statistical rate $\eta_1(d)$. We know that $\mathbf{E}_{ij}=\mathbf{E}_{ji}$ are i.i.d. across $i\le j$ and $\mathrm{Var}(\mathbf{E}_{ij})\asymp\widetilde{\sigma}^2/\theta$, then by Lemma B.5, with almost identical arguments as in the proof of Corollary 4.7, for $i\in[d]$ we have that $\|\operatorname{Cov}(\mathbf{E}\mathbf{P}_\perp\mathbf{e}_i)-\operatorname{Cov}(\mathbf{E}\mathbf{e}_i)\|_2\lesssim\widetilde{\sigma}^2/\theta\sqrt{\mu K/d}$, and thus $\lambda_d\big(\operatorname{Cov}(\mathbf{E}\mathbf{P}_\perp\mathbf{e}_i)\big)\gtrsim\lambda_d\big(\operatorname{Cov}(\mathbf{E}\mathbf{e}_i)\big)\gtrsim\widetilde{\sigma}^2/\theta$ and we have $\eta_1(d)\asymp\lambda_1^{-2}\theta^{-1}\widetilde{\sigma}^2$. Therefore, under the condition that $L\gg\kappa_2^2Kd^2/p$ and $\widetilde{\sigma}/\Delta\sqrt{d/\theta}\ll\min\left(\left(\kappa_2^2\sqrt{\mu K}+\kappa_2\sqrt{K\log d}\right)^{-1},\sqrt{p/d}\right)$, we have that $\eta_1(d)^{-1/2}r(d)=o(1)$.

Now we move on to verify Assumption 5. More specifically, we will show that the following results hold:

For any matrix $\mathbf{A}\in\mathbb{R}^{d\times K}$ that satisfies the following two conditions: (1)$\|\mathbf{A}\|_{2,\infty}/\sigma_{\min}(\mathbf{A})\le C\sqrt{\lambda_1^2\mu K/(d\Delta^2)}$; (2) $\lambda_K\big(\mathbf{\Sigma}_i\big)\ge c\widetilde{\sigma}^2\theta^{-1}\big(\sigma_{\min}(\mathbf{A})\big)^2$, where $\mathbf{\Sigma}_i:=\operatorname{Cov}(\mathbf{A}^\top\mathbf{E}_0\mathbf{P}_\perp\mathbf{e}_i)$ and $C,c>0$ are fixed constants independent of $\mathbf{A}$, it holds that

$$\mathbf{\Sigma}_i^{-1/2}\mathbf{A}^\top\mathbf{E}_0\mathbf{P}_\perp\mathbf{e}_i\xrightarrow{d}\mathcal{N}(\mathbf{0},\mathbf{I}_K). \tag{B.32}$$

To prove (B.32), it suffices to show that $\mathbf{a}^\top\mathbf{\Sigma}_i^{-1/2}\mathbf{A}^\top\mathbf{E}\mathbf{P}_\perp\mathbf{e}_i\xrightarrow{d}\mathcal{N}(0,1)$ for any $\mathbf{a}\in\mathbb{R}^K$, $\|\mathbf{a}\|_2=1$. We will first study $\mathbf{P}_\perp\mathbf{e}_i$, $\mathbf{A}\mathbf{\Sigma}_i^{-1/2}\mathbf{a}$ and $\max_{ij}\mathbb{E}|\mathbf{E}_{ij}|^3$. It holds that

$$|(\mathbf{P}_\perp\mathbf{e}_i)_i|=|\big((\mathbf{I}_d-\mathbf{V}\mathbf{V}^\top)\mathbf{e}_i\big)_i|\le1+\|\mathbf{V}\|_{2,\infty}^2=1+o(1);$$

$$\max_{j\neq i}|(\mathbf{P}_\perp\mathbf{e}_i)_j|=\max_{j\neq i}|\mathbf{e}_j^\top\mathbf{e}_i-\mathbf{e}_j^\top\mathbf{V}\mathbf{V}^\top\mathbf{e}_i|\le0+\|\mathbf{V}\|_{2,\infty}^2=\frac{\mu K}{d};$$

$$\|\mathbf{A}\mathbf{\Sigma}_i^{-1/2}\mathbf{a}\|_\infty\le\|\mathbf{A}\|_{2,\infty}\|\mathbf{\Sigma}_i^{-1/2}\|_2\lesssim(\widetilde{\sigma}^2/\theta)^{-1/2}\|\mathbf{A}\|_{2,\infty}/\sigma_{\min}(\mathbf{A})\lesssim\kappa_2\sqrt{\frac{\mu K}{d}}\frac{\sqrt{\theta}}{\widetilde{\sigma}};$$

$$\max_{ij}\mathbb{E}|\mathbf{E}_{ij}|^3\lesssim\frac{\|\mathbf{M}\|_{\max}^3}{\theta^3}\theta+\frac{\sigma^3(\log d)^3}{\theta^3}\theta\lesssim\frac{\widetilde{\sigma}^3}{\theta^2}(\log d)^3.$$

Then we know that

$$\mathbf{a}^\top\mathbf{\Sigma}_i^{-1/2}\mathbf{A}^\top\mathbf{E}\mathbf{P}_\perp\mathbf{e}_i=\sum_{jk}\mathbf{E}_{jk}(\mathbf{A}\mathbf{\Sigma}_i^{-1/2}\mathbf{a})_j(\mathbf{P}_\perp\mathbf{e}_i)_k=\sum_{j=1}^d\mathbf{E}_{jj}(\mathbf{A}\mathbf{\Sigma}_i^{-1/2}\mathbf{a})_j(\mathbf{P}_\perp\mathbf{e}_i)_j$$
$$+\sum_{j<k}\mathbf{E}_{jk}\big[(\mathbf{A}\mathbf{\Sigma}_i^{-1/2}\mathbf{a})_j(\mathbf{P}_\perp\mathbf{e}_i)_k+(\mathbf{A}\mathbf{\Sigma}_i^{-1/2}\mathbf{a})_k(\mathbf{P}_\perp\mathbf{e}_i)_j\big].$$

Then for the diagonal entries we have

$$\sum_{j=1}^d\mathbb{E}|\mathbf{E}_{jj}(\mathbf{A}\mathbf{\Sigma}_i^{-1/2}\mathbf{a})_j(\mathbf{P}_\perp\mathbf{e}_i)_j|^3$$
$$=\mathbb{E}|\mathbf{E}_{ii}(\mathbf{A}\mathbf{\Sigma}_i^{-1/2}\mathbf{a})_i(\mathbf{P}_\perp\mathbf{e}_i)_i|^3+\sum_{j\neq i}\mathbb{E}|\mathbf{E}_{jj}(\mathbf{A}\mathbf{\Sigma}_i^{-1/2}\mathbf{a})_j(\mathbf{P}_\perp\mathbf{e}_i)_j|^3$$
$$\lesssim\mathbb{E}|\mathbf{E}_{ii}|^3\|\mathbf{A}\mathbf{\Sigma}_i^{-1/2}\mathbf{a}\|_\infty^3+d\mathbb{E}|\mathbf{E}_{jj}|^3\|\mathbf{A}\mathbf{\Sigma}_i^{-1/2}\mathbf{a}\|_\infty^3\max_{j\neq i}|(\mathbf{P}_\perp\mathbf{e}_i)_j|^3\lesssim\frac{\kappa_2^3K\mu}{d}\sqrt{\frac{\mu K}{d\theta}}(\log d)^3,$$

and for the off-diagonal entries, under the condition $\kappa_2^6K^3\mu^3=o(d^{1/2})$ it holds that

$$\sum_{j<k}\mathbb{E}\Big|\mathbf{E}_{jk}\big[(\mathbf{A}\mathbf{\Sigma}_i^{-1/2}\mathbf{a})_j(\mathbf{P}_\perp\mathbf{e}_i)_k+(\mathbf{A}\mathbf{\Sigma}_i^{-1/2}\mathbf{a})_k(\mathbf{P}_\perp\mathbf{e}_i)_j\big]\Big|^3\lesssim d\frac{\widetilde{\sigma}^3}{\theta^2}\|\mathbf{A}\mathbf{\Sigma}_i^{-1/2}\mathbf{a}\|_\infty^3(\log d)^3$$

$$+ d^2\frac{\widetilde{\sigma}^3}{\theta^2}(\log d)^3\|\mathbf{A}\mathbf{\Sigma}_i^{-1/2}\mathbf{a}\|_\infty^3(\frac{\mu K}{d})^3 \lesssim \kappa_2^3 K\mu\sqrt{\frac{\mu K}{d\theta}}(\log d)^3 = o(1).$$

Moreover, since $\mathrm{Var}(\mathbf{a}^\top\mathbf{\Sigma}_i^{-1/2}\mathbf{A}^\top\mathbf{E}\mathbf{P}_\perp\mathbf{e}_i) = 1$, by the Lyapunov's condition, (B.32) holds and Assumption 5 is satisfied by plugging in $\mathbf{A} = \mathbf{V}\mathbf{\Lambda}^{-1}$. By Theorem 4.5, we have that (8) follows.

To show that (15) holds we need to show that $\|\widetilde{\mathbf{\Sigma}}_i - \mathbf{\Sigma}_i\|_2 = o(\lambda_K(\widetilde{\mathbf{\Sigma}}_i))$. From previous discussion we learnt that

$$\begin{aligned}\|\widetilde{\mathbf{\Sigma}}_i - \mathbf{\Sigma}_i\|_2 &\le \|\mathbf{V}\mathbf{\Lambda}^{-1}\|_2^2\|\operatorname{Cov}(\mathbf{E}\mathbf{P}_\perp\mathbf{e}_i) - \operatorname{Cov}(\mathbf{E}\mathbf{e}_i)\|_2\\ &\le \frac{1}{\Delta^2}\sqrt{\frac{\mu K}{d}}\frac{\widetilde{\sigma}^2}{\theta} \lesssim \kappa_2^2\sqrt{\frac{\mu K}{d}}\lambda_K(\widetilde{\mathbf{\Sigma}}_i) = o(\lambda_K(\widetilde{\mathbf{\Sigma}}_i)).\end{aligned}$$

Then by Slutsky's Theorem, (15) holds.

Last we verify that the distributional convergence still holds when we plug in the estimator $\widehat{\mathbf{\Sigma}}_i$. Similar as in the previous proof, it suffices for us to prove that $\|\widehat{\mathbf{\Sigma}}_i - \mathbf{H}\widetilde{\mathbf{\Sigma}}_i\mathbf{H}^\top\|_2 = o_P\big(\lambda_K(\widetilde{\mathbf{\Sigma}}_i)\big)$. In the following proof, we will base the discussion on the event that $\mathbf{H}$ is orthonormal. We will first bound $\|\widetilde{\mathbf{M}} - \mathbf{M}\|_{\max}$. From previous discussion we have the following bounds

$$\|\widehat{\mathbf{M}}' - \mathbf{M}\|_2 = O_P(\sqrt{d\widetilde{\sigma}^2/\theta}), \quad \|\widetilde{\mathbf{V}}^{\mathrm{F}} - \mathbf{V}\mathbf{H}^\top\|_2 = \|\widetilde{\mathbf{V}}^{\mathrm{F}}\mathbf{H} - \mathbf{V}\|_2 = O_P(\frac{1}{\Delta}\sqrt{d\widetilde{\sigma}^2/\theta}),$$

and

$$\begin{aligned}\|\widetilde{\mathbf{V}}^{\mathrm{F}}\mathbf{H} - \mathbf{V}\|_{2,\infty} &\le \|\widetilde{\mathbf{V}}^{\mathrm{F}}\mathbf{H} - \widehat{\mathbf{V}}\mathbf{H}_0\|_2 + \|\widehat{\mathbf{V}}\mathbf{H}_0 - \mathbf{V}\|_{2,\infty} = o_P(\frac{\widetilde{\sigma}}{|\lambda_1|\sqrt{\theta}})\\ &+ O_P(\frac{\kappa_2\widetilde{\sigma}\sqrt{\mu K/\theta} + \widetilde{\sigma}\sqrt{K\log d/\theta}}{\Delta}) = O_P(\frac{\kappa_2\widetilde{\sigma}\sqrt{\mu K/\theta} + \widetilde{\sigma}\sqrt{K\log d/\theta}}{\Delta}).\end{aligned}$$

Now we can study $\widetilde{\mathbf{M}} = (\widetilde{\mathbf{V}}^{\mathrm{F}}\widetilde{\mathbf{V}}^{\mathrm{F}\top})\widehat{\mathbf{M}}(\widetilde{\mathbf{V}}^{\mathrm{F}}\widetilde{\mathbf{V}}^{\mathrm{F}\top}) = \widetilde{\mathbf{V}}^{\mathrm{F}}\mathbf{H}(\frac{\theta}{\widehat{\theta}}\mathbf{H}^\top\widetilde{\mathbf{V}}^{\mathrm{F}\top}\widehat{\mathbf{M}}'\widetilde{\mathbf{V}}^{\mathrm{F}}\mathbf{H})\mathbf{H}^\top\widetilde{\mathbf{V}}^{\mathrm{F}\top}$. Recall by Hoeffding's inequality [19], with probability $1 - O(d^{-10})$ we have that $|\widehat{\theta} - \theta| \lesssim \frac{\sqrt{\log d}}{d}$ and $|\mathcal{S}| = \Omega(d^2\theta)$, and we have that

$$\begin{aligned}\|\frac{\theta}{\widehat{\theta}}\mathbf{H}^\top\widetilde{\mathbf{V}}^{\mathrm{F}\top}\widehat{\mathbf{M}}'\widetilde{\mathbf{V}}^{\mathrm{F}}\mathbf{H} - \mathbf{\Lambda}\|_2 &\le \|\mathbf{H}^\top\widetilde{\mathbf{V}}^{\mathrm{F}\top}\widehat{\mathbf{M}}'\widetilde{\mathbf{V}}^{\mathrm{F}}\mathbf{H} - \mathbf{H}^\top\widetilde{\mathbf{V}}^{\mathrm{F}\top}\mathbf{M}\widetilde{\mathbf{V}}^{\mathrm{F}}\mathbf{H}\|_2\\ &+ \|\mathbf{H}^\top\widetilde{\mathbf{V}}^{\mathrm{F}\top}\mathbf{M}(\widetilde{\mathbf{V}}^{\mathrm{F}}\mathbf{H} - \mathbf{V})\|_2 + \|(\widetilde{\mathbf{V}}^{\mathrm{F}}\mathbf{H} - \mathbf{V})^\top\mathbf{M}\mathbf{V}\|_2 + O_P\Big(\frac{\sqrt{\log d}}{d\theta}|\lambda_1|\Big)\\ &\lesssim \|\widehat{\mathbf{M}}' - \mathbf{M}\|_2 + 2\|\mathbf{M}\|_2\|\widetilde{\mathbf{V}}^{\mathrm{F}}\mathbf{H} - \mathbf{V}\|_2 + O_P\Big(\frac{\sqrt{\log d}}{d\theta}|\lambda_1|\Big) = O_P(\kappa_2\sqrt{d\widetilde{\sigma}^2/\theta}).\end{aligned}$$

Then for any $i, j \in [d]$, we have

$$\begin{aligned}|\widetilde{\mathbf{M}}_{ij} - \mathbf{M}_{ij}| &= |(\widetilde{\mathbf{V}}^{\mathrm{F}}\mathbf{H})_i^\top(\frac{\theta}{\widehat{\theta}}\mathbf{H}^\top\widetilde{\mathbf{V}}^{\mathrm{F}\top}\widehat{\mathbf{M}}'\widetilde{\mathbf{V}}^{\mathrm{F}}\mathbf{H})(\widetilde{\mathbf{V}}^{\mathrm{F}}\mathbf{H})_j - \mathbf{M}_{ij}|\\ &\le |(\widetilde{\mathbf{V}}^{\mathrm{F}}\mathbf{H})_i^\top(\frac{\theta}{\widehat{\theta}}\mathbf{H}^\top\widetilde{\mathbf{V}}^{\mathrm{F}\top}\widehat{\mathbf{M}}'\widetilde{\mathbf{V}}^{\mathrm{F}}\mathbf{H} - \mathbf{\Lambda})(\widetilde{\mathbf{V}}^{\mathrm{F}}\mathbf{H})_j| + |(\widetilde{\mathbf{V}}^{\mathrm{F}}\mathbf{H} - \mathbf{V})_i\mathbf{\Lambda}(\widetilde{\mathbf{V}}^{\mathrm{F}}\mathbf{H})_j|\\ &+ |(\mathbf{V})_i\mathbf{\Lambda}(\widetilde{\mathbf{V}}^{\mathrm{F}}\mathbf{H} - \mathbf{V})_j| = O_P(\kappa_2\sqrt{d\widetilde{\sigma}^2/\theta}\|\widetilde{\mathbf{V}}^{\mathrm{F}}\mathbf{H}\|_{2,\infty}^2)\\ &+ O_P(|\lambda_1|\|\mathbf{V}\|_{2,\infty}\|\widehat{\mathbf{V}}\mathbf{H}_0 - \mathbf{V}\|_{2,\infty}) = O_P(\frac{\sqrt{d}\widetilde{\sigma}}{\Delta\sqrt{\theta}})\frac{|\lambda_1|\mu K}{d} = O_P\left(\frac{\kappa_2\mu K}{\sqrt{d\theta}}\right)\widetilde{\sigma},\end{aligned}$$

and in turn we have

$$|\widetilde{\mathbf{M}}_{ij}^2 - \mathbf{M}_{ij}^2| \lesssim \frac{|\lambda_1|\mu K}{d}|\widetilde{\mathbf{M}}_{ij} - \mathbf{M}_{ij}| = O_P\left(\frac{\sqrt{d}\widetilde{\sigma}}{\Delta\sqrt{\theta}}\right)\Big(\frac{|\lambda_1|\mu K}{d}\Big)^2, \quad \forall i, j \in [d].$$

Now we move on to bound $\widehat{\sigma}^2$. We know from the setting of Example 4 that ε_{ij}'s are sub-Gaussian with variance proxy of order $O(\sigma^2(\log d)^2)$, and thus

$$
\begin{aligned}
|\widehat{\sigma}^2-\sigma^2| &= \Big|\sum_{(i,j)\in\mathcal{S}}(\mathbf{M}_{ij}+\varepsilon_{ij}-\widetilde{\mathbf{M}}_{ij})^2/|\mathcal{S}|-\sigma^2\Big| = \Big|\sum_{(i,j)\in\mathcal{S}}(\mathbf{M}_{ij}+\varepsilon_{ij}-\mathbf{M}_{ij}+\mathbf{M}_{ij}-\widetilde{\mathbf{M}}_{ij})^2/|\mathcal{S}|-\sigma^2\Big| \\
&\lesssim \Big|\frac{1}{|\mathcal{S}|}\sum_{(i,j)\in\mathcal{S}}\varepsilon_{ij}^2-\sigma^2\Big| + \|\widetilde{\mathbf{M}}-\mathbf{M}\|_{\max}^2 = O_P\big(\frac{\sigma^2(\log d)^2}{\sqrt{|\mathcal{S}|}}\big)+O_P\Big(\frac{\kappa_2^2\mu^2K^2}{d\theta}\Big)\widetilde{\sigma}^2 \\
&= O_P\Big(\frac{(\log d)^2}{d\sqrt{\theta}}\Big)\sigma^2+O_P\Big(\frac{\kappa_2^2\mu^2K^2}{d\theta}\Big)\widetilde{\sigma}^2.
\end{aligned}
$$

Then we have

$$
\begin{aligned}
&\Big|\frac{\widetilde{\mathbf{M}}_{ij}^2(1-\widehat{\theta})}{\widehat{\theta}}+\frac{\widehat{\sigma}^2}{\widehat{\theta}}-\frac{\mathbf{M}_{ij}^2(1-\theta)}{\theta}-\frac{\sigma^2}{\theta}\Big| \lesssim |\widetilde{\mathbf{M}}_{ij}|^2\Big|\frac{1}{\widehat{\theta}}-\frac{1}{\theta}\Big|+\frac{|\widetilde{\mathbf{M}}_{ij}^2-\mathbf{M}_{ij}^2|}{\theta}+\widehat{\sigma}^2\Big|\frac{1}{\widehat{\theta}}-\frac{1}{\theta}\Big|+\frac{|\widehat{\sigma}^2-\sigma^2|}{\theta} \\
&= O_P\left(\frac{\sqrt{d}\widetilde{\sigma}}{\Delta\sqrt{\theta}}\right)\theta^{-1}\Big(\frac{|\lambda_1|\mu K}{d}\Big)^2+O_P\Big(\frac{(\log d)^2}{d\sqrt{\theta}}+\frac{\kappa_2^2\mu^2K^2}{d\theta}\Big)\frac{\widetilde{\sigma}^2}{\theta} = O_P\left(\frac{\sqrt{d}\widetilde{\sigma}}{\Delta\sqrt{\theta}}\right)\frac{\widetilde{\sigma}^2}{\theta},
\end{aligned}
$$

and thus we have that

$$
\|\operatorname{diag}\big([\widetilde{\mathbf{M}}_{ij}^2(1-\widehat{\theta})/\widehat{\theta}+\widehat{\sigma}^2/\widehat{\theta}]_{j=1}^d\big)-\operatorname{diag}\big([\mathbf{M}_{ij}^2(1-\theta)/\theta+\sigma^2/\theta]_{j=1}^d\big)\|_2 = O_P\Big(\frac{\sqrt{d}\widetilde{\sigma}}{\Delta\sqrt{\theta}}\Big)\frac{\widetilde{\sigma}^2}{\theta}.
$$

Also, we have shown that

$$
\|\widetilde{\mathbf{\Lambda}}-\mathbf{H}\mathbf{\Lambda}\mathbf{H}^\top\|_2 = \|\frac{\theta}{\widehat{\theta}}\mathbf{H}^\top\widetilde{\mathbf{V}}^{\mathrm{F}\top}\widehat{\mathbf{M}}'\widetilde{\mathbf{V}}^{\mathrm{F}}\mathbf{H}-\mathbf{\Lambda}\|_2 = O_P\Big(\kappa_2\frac{\sqrt{d}\widetilde{\sigma}}{\Delta\sqrt{\theta}}\Big)\Delta,
$$

then we have $\|\widetilde{\mathbf{\Lambda}}^{-1}-\mathbf{H}\mathbf{\Lambda}^{-1}\mathbf{H}^\top\|_2 = O_P\Big(\kappa_2\frac{\sqrt{d}\widetilde{\sigma}}{\Delta\sqrt{\theta}}\Big)\frac{1}{\Delta}$, and hence

$$
\begin{aligned}
&\|\widetilde{\mathbf{V}}^{\mathrm{F}}\widetilde{\mathbf{\Lambda}}^{-1}-\mathbf{V}\mathbf{\Lambda}^{-1}\mathbf{H}^\top\|_2 \le \|\widetilde{\mathbf{\Lambda}}^{-1}-\mathbf{H}\mathbf{\Lambda}^{-1}\mathbf{H}^\top\|_2+\|\mathbf{\Lambda}^{-1}\|_2\|\widetilde{\mathbf{V}}^{\mathrm{F}}-\mathbf{V}\mathbf{H}^\top\|_2 \\
&= O_P\Big(\kappa_2\frac{\sqrt{d}\widetilde{\sigma}}{\Delta\sqrt{\theta}}\Big)\frac{1}{\Delta}+O_P\Big(\frac{\sqrt{d}\widetilde{\sigma}}{\Delta\sqrt{\theta}}\Big)\frac{1}{\Delta} = O_P\Big(\kappa_2\frac{\sqrt{d}\widetilde{\sigma}}{\Delta\sqrt{\theta}}\Big)\frac{1}{\Delta}.
\end{aligned}
$$

Then following basic algebra we have that with high probability

$$
\|\widehat{\mathbf{\Sigma}}_i-\mathbf{H}\widetilde{\mathbf{\Sigma}}_i\mathbf{H}^\top\|_2 \lesssim O_P\Big(\frac{\sqrt{d}\widetilde{\sigma}}{\Delta\sqrt{\theta}}\Big)\frac{\widetilde{\sigma}^2}{\Delta^2\theta}+O_P\Big(\kappa_2\frac{\sqrt{d}\widetilde{\sigma}}{\Delta\sqrt{\theta}}\Big)\frac{\widetilde{\sigma}^2}{\Delta^2\theta} = O_P\Big(\kappa_2\frac{\sqrt{d}\widetilde{\sigma}}{\Delta\sqrt{\theta}}\Big)\frac{\widetilde{\sigma}^2}{\Delta^2\theta}.
$$

Then under the condition that $\kappa_2^3\frac{\sqrt{d}\widetilde{\sigma}}{\Delta\sqrt{\theta}} = o(1)$, we have that

$$
\|\widehat{\mathbf{\Sigma}}_i-\mathbf{H}\widetilde{\mathbf{\Sigma}}_i\mathbf{H}^\top\|_2 = O_P(\kappa_2^3\frac{\sqrt{d}\widetilde{\sigma}}{\Delta\sqrt{\theta}})\frac{\widetilde{\sigma}^2}{\lambda_1^2\theta} = o_P\big(\lambda_K(\widetilde{\mathbf{\Sigma}}_i)\big).
$$

APPENDIX C: PROOF OF TECHNICAL LEMMAS

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

C.1. Proof of Lemma B.2. It can be easily seen that

$$
\|\mathbf{\Omega}/\sqrt{p}\|_2 = (\|\mathbf{\Omega}\mathbf{\Omega}^\top/p\|_2)^{1/2} = \Big((d/p)\|\mathbf{\Omega}^\top\mathbf{\Omega}/d\|_2\Big)^{1/2}.
$$

By Lemma 3 in [15], we know that there $\|\|\mathbf{\Omega}^\top\mathbf{\Omega}/d-\mathbf{I}_p\|_2\|_{\psi_1}\lesssim\sqrt{p/d}$, and thus $\|\|\mathbf{\Omega}^\top\mathbf{\Omega}/d\|_2\|_{\psi_1}\lesssim 1+\sqrt{p/d}=O(1)$. Therefore, we have $\|\|\mathbf{\Omega}\mathbf{\Omega}^\top/p\|_2\|_{\psi_1}\lesssim d/p$. By Jensen's inequality, we in turn get $\|\|\mathbf{\Omega}/\sqrt{p}\|_2\|_{\psi_1}\lesssim\sqrt{d/p}$.

C.2. Proof of Lemma B.3. By Proposition 10.4 in [18], we know that for any $t \geq 1$, we have

$$\mathbb{P}\left(\left\|\boldsymbol{\Omega}^{\dagger}\right\|_2 \geq \frac{\mathrm{e}\sqrt{p}}{p-K+1}\cdot t\right) \leq t^{-(p-K+1)}. \tag{C.33}$$

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

$$\mathbb{P}\left(\sqrt{p}\left\|\boldsymbol{\Omega}^{\dagger}\right\|_2 \geq ct\right) \leq t^{-(p-K+1)}. \tag{C.34}$$

Therefore, we have

$$\begin{aligned}
\mathbb{E}\left((\sigma_{\min}(\boldsymbol{\Omega}/\sqrt{p}))^{-a}\right) &= \mathbb{E}\left(\left\|\sqrt{p}\boldsymbol{\Omega}^{\dagger}\right\|_2^a\right) = \int_{u\geq 0}\mathbb{P}\left(\left\|\sqrt{p}\boldsymbol{\Omega}^{\dagger}\right\|_2^a \geq u\right)du \\
&= \int_{0\leq u\leq c^a}\mathbb{P}\left(\left\|\sqrt{p}\boldsymbol{\Omega}^{\dagger}\right\|_2^a \geq u\right)du + \int_{u\geq c^a}\mathbb{P}\left(\left\|\sqrt{p}\boldsymbol{\Omega}^{\dagger}\right\|_2^a \geq u\right)du \\
&\leq c^a + \int_{u\geq c^a}\mathbb{P}\left(\left\|\sqrt{p}\boldsymbol{\Omega}^{\dagger}\right\|_2 \geq u^{1/a}\right)du \leq c^a + \int_{u\geq c^a}\left(u^{1/a}/c\right)^{-(p-K+1)}du \\
&= c^a\left(1+\frac{1}{(p-K+1)/a-1}\right).
\end{aligned}$$

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

C.3. Proof of Lemma B.4. We first consider the probability $\mathbb{P}\left(\|\boldsymbol{\Sigma}'-\mathbf{V}\mathbf{V}\|_2 \geq \varepsilon\right)$. Recall the matrix $\mathbf{Y}^{(\ell)} := \mathbf{V}\mathbf{P}_0\boldsymbol{\Lambda}^0\mathbf{V}^{\top}\boldsymbol{\Omega}^{(\ell)}$. Now by Jensen's inequality and Wedin's Theorem [39], we have

$$\begin{aligned}
\|\boldsymbol{\Sigma}'-\mathbf{V}\mathbf{V}\|_2 &= \|\mathbb{E}\left(\widehat{\mathbf{V}}^{(\ell)}\widehat{\mathbf{V}}^{(\ell)\top}|\widehat{\mathbf{M}}\right)-\mathbf{V}\mathbf{V}\|_2 \leq \mathbb{E}\left(\left\|\widehat{\mathbf{V}}^{(\ell)}\widehat{\mathbf{V}}^{(\ell)\top}-\mathbf{V}\mathbf{V}\right\|_2\Big|\widehat{\mathbf{M}}\right) \\
&\lesssim \mathbb{E}\left(\|\widehat{\mathbf{Y}}^{(\ell)}/\sqrt{p}-\mathbf{Y}^{(\ell)}/\sqrt{p}\|_2/\sigma_K\left(\mathbf{Y}^{(\ell)}/\sqrt{p}\right)|\widehat{\mathbf{M}}\right) \leq \frac{\|\mathbf{E}\|_2}{\Delta}\mathbb{E}\left(\frac{\|\boldsymbol{\Omega}^{(\ell)}/\sqrt{p}\|_2}{\sigma_{\min}\left(\widetilde{\boldsymbol{\Omega}}^{(\ell)}/\sqrt{p}\right)}\Big|\widehat{\mathbf{M}}\right) \\
&= \frac{\|\mathbf{E}\|_2}{\Delta}\mathbb{E}\left(\frac{\|\boldsymbol{\Omega}^{(\ell)}/\sqrt{p}\|_2}{\sigma_{\min}\left(\widetilde{\boldsymbol{\Omega}}^{(\ell)}/\sqrt{p}\right)}\right) \leq \frac{\|\mathbf{E}\|_2}{\Delta}\mathbb{E}\left(\|\boldsymbol{\Omega}^{(\ell)}/\sqrt{p}\|^2\right)^{1/2}\mathbb{E}\left(\left\{\sigma_{\min}(\boldsymbol{\Omega}^{(\ell)}/\sqrt{p})\right\}^{-2}\right)^{1/2} \\
&\lesssim \frac{\|\mathbf{E}\|_2}{\Delta}\|\|\boldsymbol{\Omega}^{(\ell)}/\sqrt{p}\|_2\|_{\psi_1} \lesssim \frac{\|\mathbf{E}\|_2}{\Delta}\sqrt{d/p},
\end{aligned}$$

where the last but one inequality is due to Lemma B.3 under the condition that $p \geq \max(2K, K+3)$, and the last inequality is due to Lemma B.2. Therefore, by Assumption 1, there exists constants $c_0, c_0' > 0$ such that

$$\mathbb{P}\left(\|\boldsymbol{\Sigma}'-\mathbf{V}\mathbf{V}\|_2 \geq \varepsilon\right) \leq \mathbb{P}\left(\frac{\|\mathbf{E}\|_2}{\Delta}\sqrt{d/p} \geq c_0'\varepsilon\right) \leq \exp\left(-c_0\sqrt{\frac{p}{d}}\frac{\Delta\varepsilon}{r_1(d)}\right).$$

Similarly, we consider the probability $\mathbb{P}\left(\|\boldsymbol{\Sigma}'-\widehat{\mathbf{V}}\widehat{\mathbf{V}}^{\top}\|_2 \geq \varepsilon\right)$. By Assumption 1, there exists constants $c_0'', c_0''' > 0$ such that

$$\begin{aligned}
\mathbb{P}\left(\|\boldsymbol{\Sigma}'-\widehat{\mathbf{V}}\widehat{\mathbf{V}}^{\top}\|_2 \geq \varepsilon\right) &\leq \mathbb{P}\left(\|\boldsymbol{\Sigma}'-\mathbf{V}\mathbf{V}\|_2 \geq \varepsilon/2\right) + \mathbb{P}\left(\|\widehat{\mathbf{V}}\widehat{\mathbf{V}}^{\top}-\mathbf{V}\mathbf{V}\|_2 \geq \varepsilon/2\right) \\
&\leq \exp\left(-c_0\sqrt{\frac{p}{d}}\frac{\Delta\varepsilon}{r_1(d)}\right) + \mathbb{P}\left(\frac{\|\mathbf{E}\|_2}{\Delta} \geq c_0'''\varepsilon\right) \leq \exp\left(-c_0\sqrt{\frac{p}{d}}\frac{\Delta\varepsilon}{r_1(d)}\right) + \exp\left(-\frac{c_0''\Delta\varepsilon}{r_1(d)}\right) \\
&\lesssim \exp\left(-c_0\sqrt{\frac{p}{d}}\frac{\Delta\varepsilon}{r_1(d)}\right).
\end{aligned}$$

Therefore, the claim follows.

C.4. Proof of Lemma B.5. We know that $\mathrm{Cov}(\mathbf{x}_1+\mathbf{x}_2) = \mathrm{Cov}(\mathbf{x}_1) + \mathrm{Cov}(\mathbf{x}_2) + \mathrm{Cov}(\mathbf{x}_1, \mathbf{x}_2) + \mathrm{Cov}(\mathbf{x}_2, \mathbf{x}_1)$, where $\mathrm{Cov}(\mathbf{x}_1, \mathbf{x}_2) = \mathbb{E}(\mathbf{x}_1 - \mathbb{E}\mathbf{x}_1)(\mathbf{x}_2 - \mathbb{E}\mathbf{x}_2)^\top$, and

$$\|\mathrm{Cov}(\mathbf{x}_i)\|_2 = \max_{\|\mathbf{v}\|_2=1} \mathbf{v}^\top \mathrm{Cov}(\mathbf{x}_i)\mathbf{v} = \max_{\|\mathbf{v}\|_2=1} \mathrm{Var}\left(\mathbf{v}^\top \mathbf{x}_i\right),$$

for $i = 1, 2$. Therefore, we have

$$\begin{aligned}\|\mathrm{Cov}(\mathbf{x}_1,\mathbf{x}_2)\|_2 &= \max_{\|\mathbf{v}\|_2=1,\|\mathbf{u}\|_2=1} \mathbf{v}^\top \mathrm{Cov}(\mathbf{x}_1,\mathbf{x}_2)\mathbf{u} = \max_{\|\mathbf{v}\|_2=1,\|\mathbf{u}\|_2=1} \mathrm{Cov}(\mathbf{v}^\top\mathbf{x}_1, \mathbf{u}^\top\mathbf{x}_2)\\ &\le \max_{\|\mathbf{v}\|_2=1,\|\mathbf{u}\|_2=1} \sqrt{\mathrm{Var}(\mathbf{v}^\top\mathbf{x}_1)}\sqrt{\mathrm{Var}(\mathbf{v}^\top\mathbf{x}_2)} = \sqrt{\|\mathrm{Cov}(\mathbf{x}_1)\|_2\|\mathrm{Cov}(\mathbf{x}_2)\|_2}\\ &\le \frac{1}{2}\|\mathrm{Cov}(\mathbf{x}_1)\|_2 + \frac{1}{2}\|\mathrm{Cov}(\mathbf{x}_2)\|_2.\end{aligned}$$

Thus we have

$$\begin{aligned}\|\mathrm{Cov}(\mathbf{x}_1+\mathbf{x}_2)\|_2 &\le \|\mathrm{Cov}(\mathbf{x}_1)\|_2 + \|\mathrm{Cov}(\mathbf{x}_2)\|_2 + \|\mathrm{Cov}(\mathbf{x}_1,\mathbf{x}_2)\|_2 + \|\mathrm{Cov}(\mathbf{x}_2,\mathbf{x}_1)\|_2\\ &\le 2\|\mathrm{Cov}(\mathbf{x}_1)\|_2 + 2\|\mathrm{Cov}(\mathbf{x}_2)\|_2.\end{aligned}$$

APPENDIX D: WEDIN'S THEOREM

LEMMA D.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 $n_1 \le n_2$), whose SVDs are given respectively by*

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

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

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

$$\max\left\{\|\mathbf{U}\mathbf{U}^\top - \mathbf{U}^\star\mathbf{U}^{\star\top}\|_2, \|\mathbf{V}\mathbf{V}^\top - \mathbf{V}^\star\mathbf{V}^{\star\top}\|_2\right\} \lesssim \frac{2\|\mathbf{E}\|}{\sigma_r^\star - \sigma_{r+1}^\star}, \tag{D.35}$$

and

$$\max\left\{\|\mathbf{U}\mathbf{U}^\top - \mathbf{U}^\star\mathbf{U}^{\star\top}\|_{\mathrm{F}}, \|\mathbf{V}\mathbf{V}^\top - \mathbf{V}^\star\mathbf{V}^{\star\top}\|_{\mathrm{F}}\right\} \lesssim \frac{2\sqrt{r}\|\mathbf{E}\|}{\sigma_r^\star - \sigma_{r+1}^\star}. \tag{D.36}$$

PROOF. By Wedin's Theorem [39], if $\|\mathbf{E}\|_2 < (1 - 1/\sqrt{2})\left(\sigma_r^\star - \sigma_{r+1}^\star\right)$, (D.35) and (D.36) are true. When $\|\mathbf{E}\|_2 \ge (1 - 1/\sqrt{2})\left(\sigma_r^\star - \sigma_{r+1}^\star\right)$, the RHS of (D.35) are larger than or equal to $2 - \sqrt{2}$, whereas the LHS are bounded by 1. Thus (D.35) follows trivially, and so is (D.36). ☐

APPENDIX E: SUPPLEMENTARY FIGURES

We provide in this section additional figures deferred from the main paper.

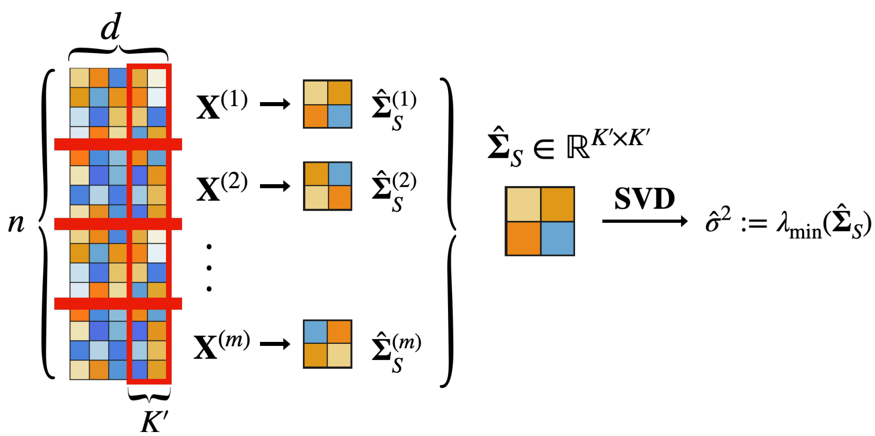


FIGURE 10. *Illustration of Step 0 for Example 1.* $\widehat{\mathbf{\Sigma}}_S^{(j)} = \mathbf{X}_{[:,S]}^{(j)\top}\mathbf{X}_{[:,S]}^{(j)}$ *is calculated by the data columns in the set* S *for the* j*-th split (*$j \in [m]$*), and* $\widehat{\mathbf{\Sigma}}_S = n^{-1}\sum_{j\in[m]}\widehat{\mathbf{\Sigma}}_S^{(j)}$.

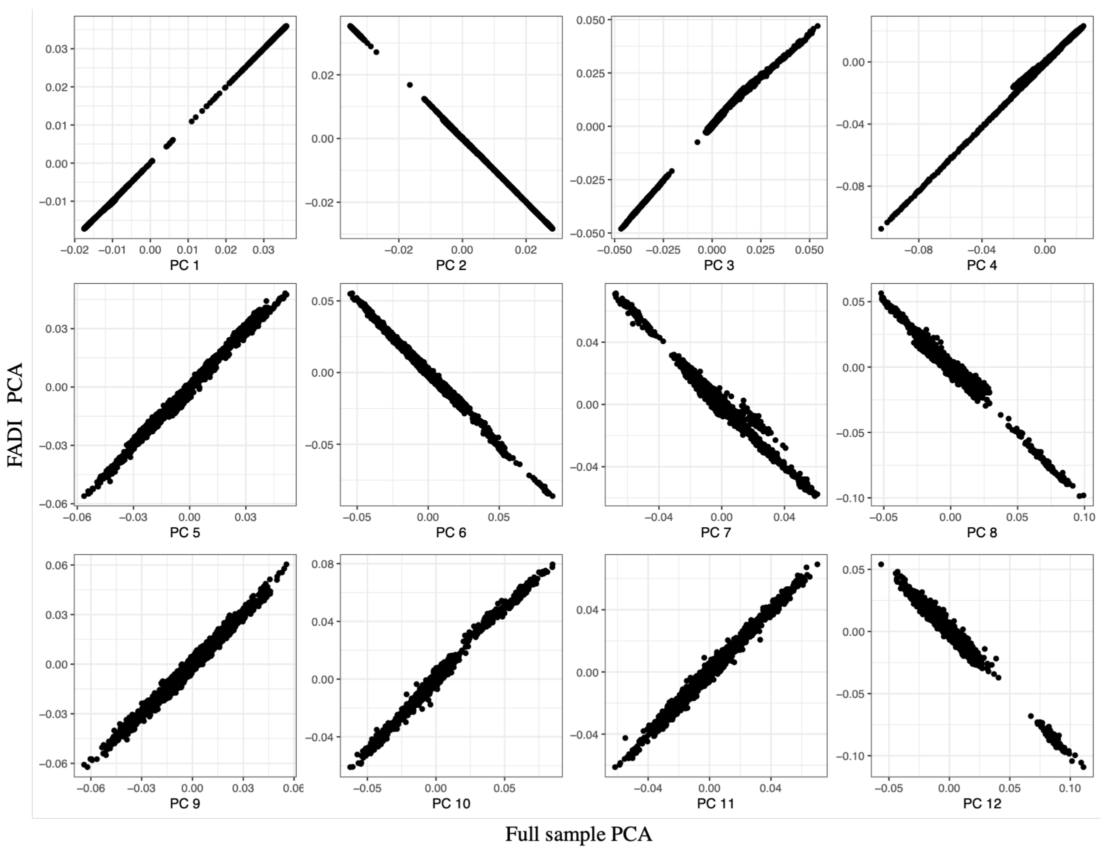


FIGURE 11. *Comparison of the top 12 PCs of the 1000 Genomes Data calculated by full sample traditional PCA and by FADI.*