The Annals of Statistics
2019, Vol. 47, No. 6, 3009–3031
https://doi.org/10.1214/18-AOS1713

DISTRIBUTED ESTIMATION OF PRINCIPAL EIGENSPACES

BY JIANQING FAN*,[1], DONG WANG*, KAIZHENG WANG* AND ZIWEI ZHU†

Princeton University and University of Michigan†*

Principal component analysis (PCA) is fundamental to statistical machine learning. It extracts latent principal factors that contribute to the most variation of the data. When data are stored across multiple machines, however, communication cost can prohibit the computation of PCA in a central location and distributed algorithms for PCA are thus needed. This paper proposes and studies a distributed PCA algorithm: each node machine computes the top K eigenvectors and transmits them to the central server; the central server then aggregates the information from all the node machines and conducts a PCA based on the aggregated information. We investigate the bias and variance for the resulting distributed estimator of the top K eigenvectors. In particular, we show that for distributions with symmetric innovation, the empirical top eigenspaces are unbiased, and hence the distributed PCA is "unbiased." We derive the rate of convergence for distributed PCA estimators, which depends explicitly on the effective rank of covariance, eigengap, and the number of machines. We show that when the number of machines is not unreasonably large, the distributed PCA performs as well as the whole sample PCA, even without full access of whole data. The theoretical results are verified by an extensive simulation study. We also extend our analysis to the heterogeneous case where the population covariance matrices are different across local machines but share similar top eigenstructures.

1. Introduction. Principal component analysis (PCA) (Pearson (1901), Hotelling (1933)) is one of the most fundamental tools in statistical machine learning. The past century has witnessed great efforts on establishing consistency and asymptotic distribution of empirical eigenvalues and eigenvectors. The early classical work of Anderson (1963) studied the asymptotic normality of eigenvalues and eigenvectors of sample covariances from multivariate Gaussian distribution with dimension d fixed and sample size n going to infinity. Recent focus moves on to the high-dimensional regimes, that is, both n and d go to infinity. A partial list of such literatures are Baik, Ben Arous and Péché (2005), Johnstone (2001), Johnstone and Lu (2009), Jung and Marron (2009), Onatski (2012), Paul (2007), Shen et al. (2016), Wang and Fan (2017). As demonstrated by these papers, asymptotic behaviors of empirical eigenvalues and eigenvectors depend on the scaling of

Received February 2017; revised January 2018.

[1]Supported by NSF Grants DMS-1662139 and DMS-1712591, and NIH Grant R01-GM072611-12.

MSC2010 subject classifications. Primary 62H25; secondary 62E10.

Key words and phrases. Distributed learning, PCA, one-shot approach, communication efficiency, unbiasedness of empirical eigenspaces, heterogeneity.

n, d and also the spikiness of the covariance. When $n \ll d$, the empirical leading eigenvector $\widehat{\mathbf{v}}_1$ is inconsistent in estimating the true top eigenvector $\mathbf{v}_1$ unless the top eigenvalue λ_1 diverges fast. This phenomenon inspires another line of research on sparse PCA where certain sparsity on top eigenvectors is imposed to overcome the noise accumulation due to high dimensionality; see, for example, Johnstone and Lu (2009), Shen, Shen and Marron (2013), Vu and Lei (2013), Cai, Ma and Wu (2013). Besides the asymptotic study, there are also nonasymptotic results on PCA, for example, Nadler (2008) and Reiss and Wahl (2016).

With rapid developments of information and technology, massive datasets are now ubiquitous. Statistical analysis such as regression or PCA on such enormous data is unprecedentedly desirable. However, large data sets are usually scattered across distant places such that to fuse or aggregate them is extremely difficult due to communication cost, privacy, data security and ownerships, among others. Consider giant IT companies that collect data simultaneously from places all around the world. Constraints on communication budget and network bandwidth make it nearly impossible to aggregate and maintain global data in a single data center. Another example is that health records are scattered across many hospitals or countries. It is hard to process the data in a central location due to privacy and ownership concerns. To resolve these issues, efforts have been made to exploiting distributed computing architectures and developing distributed estimators or testing statistics based on data scattered around different locations. A typical distributed statistical method first calculates local statistics based on each subdataset and then combines all the subsample-based statistics to produce an aggregated statistic. Such distributed methods fully adapt to the parallel data collection procedures, and thus significantly reduce the communication cost. Many distributed regression methods follow this fashion (Zhang, Duchi and Wainwright (2013), Chen and Xie (2014), Battey et al. (2015), Lee et al. (2017), Mücke and Blanchard (2018), Guo, Lin and Zhou (2017)). The last two papers study distributed kernel regression with spectral regularization using eigendecomposition of Gram matrices, which is relevant to but different from our distributed PCA.

Among all the efforts toward creating accurate and efficient distributed statistical methods, there has been rapid advancement on distributed PCA over the past two decades. Unlike the traditional PCA where we have the complete data matrix $\mathbf{X} \in \mathbb{R}^{N \times d}$ with d features of N samples at one place, the distributed PCA needs to handle data that are partitioned and stored across multiple servers. There are two data partition regimes: "horizontal" and "vertical." In the horizontal partition regime, each server contains all the features of a subset of subjects, while in the vertical partition regime, each server has a subset of features of all the subjects. To conduct distributed PCA in the horizontal regime, Qu et al. (2002) proposes that each server computes several top eigenvalues and eigenvectors on its local data and then sends them to the central server that aggregates the information together. Yet there is no theoretical guarantee on the approximation error of the proposed algorithm. Liang et al. (2014), Kannan, Vempala and Woodruff (2014)

and Boutsidis, Woodruff and Zhong (2016) aim to find a good rank-K approximation $\widehat{\mathbf{X}}$ of $\mathbf{X}$. To assess the approximation quality, they compare $\|\widehat{\mathbf{X}} - \mathbf{X}\|_F$ against $\min_{\mathrm{rank}(\mathbf{B})\le K} \|\mathbf{B} - \mathbf{X}\|_F$ and study the excess risk. For the distributed PCA in the vertical data partition regime, there is also a great amount of literature, for example, Kargupta et al. (2001), Li, Scaglione and Manton (2011), Bertrand and Moonen (2014), Schizas and Aduroja (2015), etc. This line of research is often motivated from sensor networks and signal processing where the vertically partitioned data are common. Our work focuses on the horizontal partition regime, that is, we have partitions over the samples rather than the features.

Despite these achievements, very few papers establish rigorous statistical error analysis of the proposed distributed PCA methods. To our best knowledge, the only works that provide statistical analysis so far are El Karoui and d'Aspremont (2010) and Chen et al. (2016). To estimate the leading singular vectors of a large target matrix, both papers propose to aggregate singular vectors of multiple random approximations of the original matrix. El Karoui and d'Aspremont (2010) adopts sparse approximation of the matrix by sampling the entries, while Chen et al. (2016) uses Gaussian random sketches. The works are related to ours, since we can perceive subdatasets in the distributed PCA problem as random approximations. However, our analysis is more general, since it does not rely on any matrix incoherence assumption as required by El Karoui and d'Aspremont (2010) and it explicitly characterizes how the probability distribution affects the final statistical error in finite sample error bounds. Besides, our aggregation algorithm is much simpler than the one in Chen et al. (2016). The manuscript Garber, Shamir and Srebro (2017) came out after we submitted the first draft of our work. The authors focused on estimation of the first principal component rather than the multidimensional eigenspaces, based on very different approaches.

We propose a distributed algorithm with only one-shot communication to solve for the top K eigenvectors of the population covariance matrix $\mathbf{\Sigma}$ when samples are scattered across m servers. We first calculate for each subset of data ℓ its top K eigenvectors $\{\widehat{\mathbf{V}}_K^{(\ell)} = (\widehat{\mathbf{v}}_1^{(\ell)}, \ldots, \widehat{\mathbf{v}}_K^{(\ell)})\}_{\ell=1}^m$ of the sample covariance matrix there, then compute the average of projection matrices of the eigenspaces $\widetilde{\mathbf{\Sigma}} = (1/m)\sum_{i=1}^m \widehat{\mathbf{V}}_K^{(\ell)} \widehat{\mathbf{V}}_K^{(\ell)^T}$, and finally take the top K eigenvectors of $\widetilde{\mathbf{\Sigma}}$ as the final estimator $\widetilde{\mathbf{V}}_K = (\tilde{\mathbf{v}}_1^{(\ell)}, \ldots, \tilde{\mathbf{v}}_K^{(\ell)})$. The communication cost of this method is of order $O(mKd)$. We establish rigorous nonasymptotic analysis of the statistical error $\|\widetilde{\mathbf{V}}_K \widetilde{\mathbf{V}}_K^T - \mathbf{V}_K \mathbf{V}_K^T\|_F$, and show that as long as we have a sufficiently large number of samples in each server, $\widetilde{\mathbf{V}}_K$ enjoys the same statistical error rate as the standard PCA over the full sample. The eigenvalues of $\mathbf{\Sigma}$ are easily estimated once we get good estimators of the eigenvectors, using another round of communication.

The rest of the paper is organized as follows. In Section 2, we introduce the problem setup of the distributed PCA. In Section 3, we elucidate our distributed algorithm for estimating the top K eigenvectors. Section 4 develops the statistical error rates of the aggregated estimator. The results are extended to heterogeneous

samples in Section 5. Finally, in Section 6 we present extensive simulation results to validate our theories.

2. Problem setup. We first collect all the notation that will be used. By convention, we use regular letters for scalars and bold letters for both matrices and vectors. We denote the set $\{1, 2, 3, \ldots, d\}$ by $[d]$ for convenience. For two scalar sequences $\{a_n\}_{n\geq 1}$ and $\{b_n\}_{n\geq 1}$, we say $a_n \gtrsim b_n$ ($a_n \lesssim b_n$) if there exists a universal constant $C > 0$ such that $a_n \geq Cb_n$ ($a_n \leq Cb_n$), and $a_n \asymp b_n$ if both $a_n \gtrsim b_n$ and $a_n \lesssim b_n$ hold. For a random variable $X \in \mathbb{R}$, we define $\|X\|_{\psi_2} = \sup_{p\geq 1}(\mathbb{E}|X|^p)^{\frac{1}{p}}/\sqrt{p}$ and define $\|X\|_{\psi_1} = \sup_{p\geq 1}(\mathbb{E}|X|^p)^{\frac{1}{p}}/p$. Please refer to Vershynin (2012) for equivalent definitions of ψ_2-norm and ψ_1-norm. For two random variables X and Y, we use $X \overset{d}{=} Y$ to denote that X and Y have identical distributions. Define $\mathbf{e}_i$ to be the unit vector whose components are all zero except that the ith component equals 1. For $q \geq r$, $\mathcal{O}_{q\times r}$ denotes the space of $q \times r$ matrices with orthonormal columns. For a matrix $\mathbf{A} \in \mathbb{R}^{n\times d}$, we use $\|\mathbf{A}\|_F$, $\|\mathbf{A}\|_*$ and $\|\mathbf{A}\|_2$ to denote the Frobenius norm, nuclear norm and spectral norm of $\mathbf{A}$, respectively. $\mathrm{Col}(\mathbf{A})$ represents the linear space spanned by column vectors of $\mathbf{A}$. We denote the Moore–Penrose pseudo inverse of a matrix $\mathbf{A} \in \mathbb{R}^{d\times d}$ by $\mathbf{A}^\dagger$. For a symmetric matrix $\mathbf{A}$, we use $\lambda_j(\mathbf{A})$ to refer to its jth largest eigenvalue.

Suppose we have N i.i.d. random samples $\{\mathbf{X}_i\}_{i=1}^N \subseteq \mathbb{R}^d$ with $\mathbb{E}\mathbf{X}_1 = \mathbf{0}$ and covariance matrix $\mathbb{E}(\mathbf{X}_1\mathbf{X}_1^T) = \mathbf{\Sigma}$. By spectral decomposition, $\mathbf{\Sigma} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T$, where $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_d)$ with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$ and $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_d) \in \mathcal{O}_{d\times d}$. For a given $K \in [d]$, let $\mathbf{V}_K = (\mathbf{v}_1, \ldots, \mathbf{v}_K)$. Our goal is to estimate $\mathrm{Col}(\mathbf{V}_K)$, that is, the linear space spanned by the top K eigenvectors of $\mathbf{\Sigma}$. To ensure the identifiability of $\mathrm{Col}(\mathbf{V}_K)$, we assume $\Delta := \lambda_K - \lambda_{K+1} > 0$ and define $\kappa := \lambda_1/\Delta$ to be the condition number. Let $r = r(\mathbf{\Sigma}) := \mathrm{Tr}(\mathbf{\Sigma})/\lambda_1$ be the effected rank of $\mathbf{\Sigma}$.

The standard way of estimating $\mathrm{Col}(\mathbf{V}_K)$ is to use the top K eigenspace of the sample covariance $\widehat{\mathbf{\Sigma}} = \frac{1}{N}\sum_{i=1}^N \mathbf{X}_i\mathbf{X}_i^T$. Let $\widehat{\mathbf{\Sigma}} = \widehat{\mathbf{V}}\widehat{\mathbf{\Lambda}}\widehat{\mathbf{V}}^T$ be spectral decomposition of $\widehat{\mathbf{\Sigma}}$, where $\widehat{\mathbf{\Lambda}} = \mathrm{diag}(\widehat{\lambda}_1, \ldots, \widehat{\lambda}_d)$ with $\widehat{\lambda}_1 \geq \cdots \geq \widehat{\lambda}_d$ and $\widehat{\mathbf{V}} = (\widehat{\mathbf{v}}_1, \ldots, \widehat{\mathbf{v}}_d)$. We use the empirical top K eigenspace $\mathrm{Col}(\widehat{\mathbf{V}}_K)$, where $\widehat{\mathbf{V}}_K = (\widehat{\mathbf{v}}_1, \ldots, \widehat{\mathbf{v}}_K)$, to estimate the eigenspace $\mathrm{Col}(\mathbf{V}_K)$. To measure the statistical error, we adopt $\rho(\widehat{\mathbf{V}}_K, \mathbf{V}_K) := \|\widehat{\mathbf{V}}_K\widehat{\mathbf{V}}_K^T - \mathbf{V}_K\mathbf{V}_K^T\|_F$, which is the Frobenius norm of the difference between projection matrices of two spaces and is a well-defined distance between linear subspaces. In fact, $\rho(\mathbf{V}_K, \widehat{\mathbf{V}}_K)$ is equivalent to the so-called $\sin\mathbf{\Theta}$ distance. Denote the singular values of $\widehat{\mathbf{V}}_K^T\mathbf{V}_K$ by $\{\sigma_i\}_{i=1}^K$ in descending order. Recall that $\mathbf{\Theta}(\widehat{\mathbf{V}}_K, \mathbf{V}_K) = \mathrm{diag}(\theta_1, \ldots, \theta_K)$, the principal angles between $\mathrm{Col}(\mathbf{V}_K)$ and $\mathrm{Col}(\widehat{\mathbf{V}}_K)$, are defined as $\mathrm{diag}(\cos^{-1}\sigma_1, \ldots, \cos^{-1}\sigma_K)$. Then we define $\sin\mathbf{\Theta}(\widehat{\mathbf{V}}_K, \mathbf{V}_K)$ to be $\mathrm{diag}(\sin\theta_1, \ldots, \sin\theta_K)$. Note that

$$\begin{aligned}\rho^2(\mathbf{V}_K, \widehat{\mathbf{V}}_K) &= \|\mathbf{V}_K\mathbf{V}_K^T\|_F^2 + \|\widehat{\mathbf{V}}_K\widehat{\mathbf{V}}_K^T\|_F^2 - 2\,\mathrm{Tr}(\mathbf{V}_K\mathbf{V}_K^T\widehat{\mathbf{V}}_K\widehat{\mathbf{V}}_K^T)\\ &= 2K - 2\|\widehat{\mathbf{V}}_K^T\mathbf{V}_K\|_F^2 = 2\sum_{i=1}^K(1-\sigma_i^2) = 2\sum_{i=1}^K \sin^2\theta_i \qquad (2.1)\end{aligned}$$

$$= 2\|\sin\boldsymbol{\Theta}(\widehat{\mathbf{V}}_K, \mathbf{V}_K)\|_F^2.$$

Therefore, $\rho(\mathbf{V}_K, \widehat{\mathbf{V}}_K)$ and $\|\sin\boldsymbol{\Theta}(\mathbf{V}_K, \widehat{\mathbf{V}}_K)\|_F$ are equivalent.

Now consider the estimation of top K eigenspace under the distributed data setting, where our $N = m \cdot n$ samples are scattered across m machines with each machine storing n samples.[2] Application of standard PCA here requires data or covariance aggregation, thus leads to huge communication cost for high-dimensional big data. In addition, for the areas such as genetic, biomedical studies and customer services, it is hard to communicate raw data because of privacy and ownership concerns. To address these problems, we need to avoid naive data aggregation and design a communication-efficient and privacy-preserving distributed algorithm for PCA. In addition, this new algorithm should be statistically accurate in the sense that it enjoys the same statistical error rate as the full sample PCA.

Throughout the paper, we assume that all the random samples $\{\mathbf{X}_i\}_{i=1}^N$ are i.i.d. sub-Gaussian. We adopt the definition of sub-Gaussian random vectors in Koltchinskii and Lounici (2017) and Reiss and Wahl (2016) as specified below, where M is assumed to be a constant. It is not hard to show that the following definition is equivalent to the definition $\|(\boldsymbol{\Sigma}^{1/2})^\dagger \mathbf{X}\|_{\psi_2} \le M$ used in Vershynin (2012), Wang and Fan (2017), and many other authors.

DEFINITION 2.1. We say the random vector $\mathbf{X} \in \mathbb{R}^d$ is sub-Gaussian if there exists $M > 0$ such that $\|\mathbf{u}^T\mathbf{X}\|_{\psi_2} \le M\sqrt{\mathbb{E}(\mathbf{u}^T\mathbf{X})^2}$, $\forall \mathbf{u} \in \mathbb{R}^d$.

We emphasize here that the global i.i.d. assumption on $\{\mathbf{X}_i\}_{i=1}^N$ can be further relaxed. In fact, our statistical analysis only requires the following three conditions: (i) within each server ℓ, data are i.i.d.; (ii) across different servers, data are independent; (iii) the covariance matrices of the data in each server $\{\boldsymbol{\Sigma}^{(\ell)}\}_{\ell=1}^m$ share similar top K eigenspaces. We will further study this heterogeneous regime in Section 5. To avoid future confusion, unless specified, we always assume i.i.d. data across servers.

3. Methodology. We now introduce our distributed PCA algorithm. For $\ell \in [m]$, let $\{\mathbf{X}_i^{(\ell)}\}_{i=1}^n$ denote the samples stored on the ℓth machine. We specify the distributed in Algorithm 1.

In other words, each server first calculates the top K eigenvectors of the local sample covariance matrix, and then transmits these eigenvectors $\{\widehat{\mathbf{V}}_K^{(\ell)}\}_{\ell=1}^m$ to a central server, where the estimators get aggregated. This procedure has similar spirit as distributed estimation based on one-shot averaging in Zhang, Duchi and Wainwright (2013), Battey et al. (2015), Lee et al. (2017), among others. To see this,

[2]Note that here for simplicity we assume the subsample sizes are homogeneous. We can easily extend our analysis to the case of heterogeneous subsample sizes with similar theoretical results.

Algorithm 1 Distributed PCA

1. On each server, compute locally the K leading eigenvectors $\widehat{\mathbf{V}}_K^{(\ell)} = (\widehat{\mathbf{v}}_1^{(\ell)}, \ldots, \widehat{\mathbf{v}}_K^{(\ell)}) \in \mathbb{R}^{d\times K}$ of the sample covariance matrix $\widehat{\mathbf{\Sigma}}^{(\ell)} = (1/n)\sum_{i=1}^n \mathbf{X}_i^{(\ell)}\mathbf{X}_i^{(\ell)^T}$. Send $\widehat{\mathbf{V}}_K^{(\ell)}$ to the central processor.
2. On the central processor, compute $\widetilde{\mathbf{\Sigma}} = (1/m)\sum_{\ell=1}^m \widehat{\mathbf{V}}_K^{(\ell)}\widehat{\mathbf{V}}_K^{(\ell)^T}$, and its K leading eigenvectors $\{\tilde{\mathbf{v}}_j\}_{j=1}^K$. Output: $\widetilde{\mathbf{V}}_K = (\tilde{\mathbf{v}}_1, \ldots, \tilde{\mathbf{v}}_K) \in \mathbb{R}^{d\times K}$.

we recall the SDP formulation of the eigenvalue problem. Let $\widehat{\mathbf{V}}_K = (\widehat{\mathbf{v}}_1, \ldots, \widehat{\mathbf{v}}_K)$ contain the K leading eigenvectors of $\widehat{\mathbf{\Sigma}} = \frac{1}{m}\sum_{\ell=1}^m \widehat{\mathbf{\Sigma}}^{(\ell)}$. Lemma 3 in the Supplementary Material (Fan et al. (2019)) asserts that $\widehat{\mathbf{P}}_K = \widehat{\mathbf{V}}_K\widehat{\mathbf{V}}_K^T$ solves the SDP:

$$\begin{aligned} &\min_{\mathbf{P}\in S^{d\times d}} -\operatorname{Tr}(\mathbf{P}^T\widehat{\mathbf{\Sigma}}) \\ &\text{s.t.} \quad \operatorname{Tr}(\mathbf{P}) \le K, \|\mathbf{P}\|_2 \le 1, \mathbf{P} \succeq 0. \end{aligned} \tag{3.1}$$

Here, $S^{d\times d}$ refers to the set of $d\times d$ symmetric matrices. In the traditional setting, we have access to all the data, and $\widehat{\mathbf{P}}_K$ is a natural estimator for $\mathbf{V}_K\mathbf{V}_K^T$. In the distributed setting, each machine can only access $\widehat{\mathbf{\Sigma}}^{(\ell)}$. Consequently, it solves a local version of (3.1):

$$\begin{aligned} &\min_{\mathbf{P}\in S^{d\times d}} -\operatorname{Tr}(\mathbf{P}^T\widehat{\mathbf{\Sigma}}^{(\ell)}) \\ &\text{s.t.} \quad \operatorname{Tr}(\mathbf{P}) \le K, \|\mathbf{P}\|_2 \le 1, \mathbf{P} \succeq 0. \end{aligned} \tag{3.2}$$

The optimal solution is $\widehat{\mathbf{P}}_K^{(\ell)} = \widehat{\mathbf{V}}_K^{(\ell)}\widehat{\mathbf{V}}_K^{(\ell)T}$. Since the loss function in (3.1) is the average of local loss functions in (3.2), we can intuitively average the optimal solutions $\widehat{\mathbf{P}}_K^{(\ell)}$ to approximate $\widehat{\mathbf{P}}_K$. However, the average $\frac{1}{m}\sum_{\ell=1}^m \widehat{\mathbf{P}}_K^{(\ell)}$ may no longer be a rank-K projection matrix. Hence a rounding step is needed, extracting the leading eigenvectors of that average to get a projection matrix.

Here is another way of understanding the aggregation procedure. Given a collection of estimators $\{\widehat{\mathbf{V}}_K^{(\ell)}\}_{\ell=1}^m \subseteq \mathcal{O}_{d\times K}$ and the loss $\rho(\cdot,\cdot)$, we want to find the center $\mathbf{U} \in \mathcal{O}_{d\times K}$ that minimizes the sum of squared losses $\sum_{\ell=1}^m \rho^2(\mathbf{U}, \widehat{\mathbf{V}}_K^{(\ell)})$. Lemma 4 in the Supplementary Material indicates that $\mathbf{U} = \widetilde{\mathbf{V}}_K$ is an optimal solution. Therefore, our distributed PCA estimator $\widetilde{\mathbf{V}}_K$ is a generalized "center" of individual estimators.

It is worth noting that in this algorithm, we do not really need to compute $\{\widehat{\mathbf{\Sigma}}^{(\ell)}\}_{\ell=1}^m$ and $\widetilde{\mathbf{\Sigma}}$. $\{\widehat{\mathbf{V}}_K^{(\ell)}\}_{\ell=1}^m$ and $\widetilde{\mathbf{V}}_K$ can be derived from top-K SVD of data matrices. This is far more expeditious than the entire SVD and highly scalable, by using, for example, the power method (Golub and Van Loan (2012)). In regard to the estimation of the top eigenvalues of $\mathbf{\Sigma}$, we can send the aggregated eigenvectors $\{\tilde{\mathbf{v}}_j\}_{j=1}^K$ back to the m servers, where each one computes

$\{\lambda_j^{(\ell)}\}_{j=1}^K = \{\tilde{\mathbf{v}}_j^T \widehat{\mathbf{\Sigma}}^{(\ell)} \tilde{\mathbf{v}}_j\}_{j=1}^K$. Then the central server collects all of the eigenvalues and delivers the average eigenvalues $\{\tilde{\lambda}_j\}_{j=1}^K = \{\frac{1}{m}\sum_{\ell=1}^m \lambda_j^{(\ell)}\}_{j=1}^K$ as the estimators of all eigenvalues.

As we can see, the communication cost of the proposed distributed PCA algorithm is of order $O(mKd)$. In contrast, to share all the data or entire covariance, the communication cost will be of order $O(md\min(n,d))$. Since in most cases $K = o(\min(n,d))$, our distributed PCA requires much less communication cost than naive data aggregation.

4. Statistical error analysis. Algorithm 1 delivers $\widetilde{\mathbf{V}}_K$ to estimate the top K eigenspace of $\mathbf{\Sigma}$. In this section, we analyze the statistical error of $\widetilde{\mathbf{V}}_K$, that is, $\rho(\widetilde{\mathbf{V}}_K, \mathbf{V}_K)$. The main message is that $\widetilde{\mathbf{V}}_K$ enjoys the same statistical error rate as the full sample counterpart $\widehat{\mathbf{V}}_K$ as long as the subsample size n is sufficiently large.

We first conduct a bias and variance decomposition of $\rho(\widetilde{\mathbf{V}}_K, \mathbf{V}_K)$, which serves as the key step in establishing our theoretical results. Recall that $\widetilde{\mathbf{\Sigma}} = (1/m)\sum_{\ell=1}^m \widehat{\mathbf{V}}_K^{(\ell)}\widehat{\mathbf{V}}_K^{(\ell)T}$ and $\widetilde{\mathbf{V}}_K$ consists of the top K eigenvectors of $\widetilde{\mathbf{\Sigma}}$. Define $\mathbf{\Sigma}^* := \mathbb{E}(\widehat{\mathbf{V}}_K^{(\ell)}\widehat{\mathbf{V}}_K^{(\ell)T})$ and denote its top K eigenvectors by $\mathbf{V}_K^* = (\mathbf{v}_1^*, \ldots, \mathbf{v}_K^*) \in \mathbb{R}^{d\times K}$. When the number of machines goes to infinity, $\widetilde{\mathbf{\Sigma}}$ converges to $\mathbf{\Sigma}^*$, and naturally we expect $\mathrm{Col}(\widetilde{\mathbf{V}}_K)$ to converge to $\mathrm{Col}(\mathbf{V}_K^*)$ as well. This line of thinking inspires us to decompose the statistical error $\rho(\widetilde{\mathbf{V}}_K, \mathbf{V}_K)$ into the following bias and sample variance terms:

$$\rho(\widetilde{\mathbf{V}}_K, \mathbf{V}_K) \le \underbrace{\rho(\widetilde{\mathbf{V}}_K, \mathbf{V}_K^*)}_{\text{sample variance term}} + \underbrace{\rho(\mathbf{V}_K^*, \mathbf{V}_K)}_{\text{bias term}}. \tag{4.1}$$

The first term is stochastic and the second term is deterministic. Here, we elucidate on why we call $\rho(\widetilde{\mathbf{V}}_K, \mathbf{V}_K^*)$ the sample variance term and $\rho(\mathbf{V}_K^*, \mathbf{V}_K)$ the bias term, respectively.

1. Sample variance term $\rho(\widetilde{\mathbf{V}}_K, \mathbf{V}_K^*)$:

By Davis–Kahan's theorem (Theorem 2 in Yu, Wang and Samworth (2015)) and (2.1), we have

$$\rho(\widetilde{\mathbf{V}}_K, \mathbf{V}_K^*) \lesssim \frac{\|\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\|_F}{\lambda_K(\mathbf{\Sigma}^*) - \lambda_{K+1}(\mathbf{\Sigma}^*)}. \tag{4.2}$$

As we can see, $\rho(\widetilde{\mathbf{V}}_K, \mathbf{V}_K^*)$ depends on how the average $\widetilde{\mathbf{\Sigma}} = \frac{1}{m}\sum_{\ell=1}^m \widehat{\mathbf{V}}_K^{(\ell)}\widehat{\mathbf{V}}_K^{(\ell)T}$ concentrates to its mean $\mathbf{\Sigma}^*$. This explains why we call $\rho(\widetilde{\mathbf{V}}_K, \mathbf{V}_K^*)$ the sample variance term. We will show in the sequel that for sub-Gaussian random samples, $\{\|\widehat{\mathbf{V}}_K^{(\ell)}\widehat{\mathbf{V}}_K^{(\ell)T} - \mathbf{\Sigma}^*\|_F\}_{\ell=1}^m$ and $\|\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\|_F$ are subexponential random variables and under appropriate regularity assumptions,

$$\big\|\|\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\|_F\big\|_{\psi_1} \lesssim \frac{1}{\sqrt{m}}\big\|\|\widehat{\mathbf{V}}_K^{(1)}\widehat{\mathbf{V}}_K^{(1)T} - \mathbf{\Sigma}^*\|_F\big\|_{\psi_1}. \tag{4.3}$$

If we regard ψ_1-norm as a proxy for standard deviation, this result is a counterpart to the formula for the standard deviation of the sample mean under the context of matrix concentration. By (4.3), the average of projection matrices $\widetilde{\mathbf{\Sigma}}$ enjoys a similar square-root convergence, so does $\rho(\widetilde{\mathbf{V}}_K, \mathbf{V}_K^*)$.

2. Bias term $\rho(\mathbf{V}_K^*, \mathbf{V}_K)$:

The error $\rho(\mathbf{V}_K^*, \mathbf{V}_K)$ is deterministic and independent of how many machines we have, and is therefore called the bias term. We will show this bias term is exactly zero when the random sample has a symmetric innovation (to be defined later). In general, we will show that the bias term is negligible in comparison with the sample variance term when the number of nodes m is not unreasonably large.

In the following subsections, we will analyze the sample variance term and bias term, respectively, and then combine these results to obtain the convergence rate for $\rho(\widetilde{\mathbf{V}}_K, \mathbf{V}_K)$.

4.1. *Analysis of the sample variance term.* To analyze $\rho(\widetilde{\mathbf{V}}_K, \mathbf{V}_K^*)$, as shown by (4.2), we need to derive the order of the numerator $\|\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\|_F$ and denominator $\lambda_K(\mathbf{\Sigma}^*) - \lambda_{K+1}(\mathbf{\Sigma}^*)$. We first focus on the matrix concentration term $\|\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\|_F = \|\frac{1}{m}\sum_{\ell=1}^m(\widehat{\mathbf{V}}_K^{(\ell)}\widehat{\mathbf{V}}_K^{(\ell)T} - \mathbf{\Sigma}^*)\|_F$. Note that $\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}^*$ is an average of m centered random matrices. To establish the correspondent concentration inequality, we first investigate each individual term in the average, that is, $\widehat{\mathbf{V}}_K^{(\ell)}\widehat{\mathbf{V}}_K^{(\ell)T} - \mathbf{\Sigma}^*$ for $\ell \in [m]$. In the following lemma, we show that when random samples are sub-Gaussian, $\|\widehat{\mathbf{V}}_K^{(\ell)}\widehat{\mathbf{V}}_K^{(\ell)T} - \mathbf{\Sigma}^*\|_F$ is subexponential and we can give an explicit upper bound of its ψ_1-norm.

LEMMA 1. *Suppose that on the ℓth server we have n i.i.d. sub-Gaussian random samples $\{\mathbf{X}_i\}_{i=1}^n$ in $\mathbb{R}^d$ with covariance matrix $\mathbf{\Sigma}$. There exists a constant $C > 0$ such that when $n \geq r$,*

$$\big\|\|\widehat{\mathbf{V}}_K^{(\ell)}\widehat{\mathbf{V}}_K^{(\ell)T} - \mathbf{\Sigma}^*\|_F\big\|_{\psi_1} \leq C\kappa\sqrt{\frac{Kr}{n}}.$$

Note that here we use the Frobenius norm to measure the distance between two matrices. Therefore, it is equivalent to treat $\{\widehat{\mathbf{V}}_K^{(\ell)}\widehat{\mathbf{V}}_K^{(\ell)T}\}_{\ell=1}^K$ and $\mathbf{\Sigma}^*$ as d^2-dimensional vectors and apply the concentration inequality for random vectors to bound $\|\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\|_F$. As we will demonstrate in the proof of Theorem 1, $\|\|\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\|_F\|_{\psi_1} \lesssim \frac{1}{\sqrt{m}}\|\|\widehat{\mathbf{V}}_K^{(\ell)}\widehat{\mathbf{V}}_K^{(\ell)T} - \mathbf{\Sigma}^*\|_F\|_{\psi_1}$.

With regard to $\lambda_K(\mathbf{\Sigma}^*) - \lambda_{K+1}(\mathbf{\Sigma}^*)$, when the individual node has enough samples, $\widehat{\mathbf{V}}_K^{(\ell)}$ and $\mathbf{V}_K$ will be close to each other and so will $\mathbf{\Sigma}^* = \mathbb{E}(\widehat{\mathbf{V}}_K^{(\ell)}\widehat{\mathbf{V}}_K^{(\ell)T})$ and $\mathbf{V}_K\mathbf{V}_K^T$. Given $\lambda_K(\mathbf{V}_K\mathbf{V}_K^T) = 1$ and $\lambda_{K+1}(\mathbf{V}_K\mathbf{V}_K^T) = 0$, we accordingly expect $\lambda_K(\mathbf{\Sigma}^*)$ and $\lambda_{K+1}(\mathbf{\Sigma}^*)$ be separated by a positive constant as well.

All the arguments above lead to the following theorem on $\rho(\widetilde{\mathbf{V}}_K, \mathbf{V}_K^*)$.

THEOREM 1. *Suppose* $\mathbf{X}_1, \ldots, \mathbf{X}_N$ *are i.i.d. sub-Gaussian random vectors in* $\mathbb{R}^d$ *with covariance matrix* $\mathbf{\Sigma}$ *and they are scattered across* m *machines. If* $n \geq r$ *and* $\|\mathbf{\Sigma}^* - \mathbf{V}_K \mathbf{V}_K^T\|_2 \leq 1/4$, *then*

$$\left\|\rho(\widetilde{\mathbf{V}}_K, \mathbf{V}_K^*)\right\|_{\psi_1} \leq C\kappa\sqrt{\frac{Kr}{N}},$$

where C *is some universal constant.*

4.2. *Analysis of the bias term.* In this section, we study the bias term $\rho(\mathbf{V}_K^*, \mathbf{V}_K)$ in (4.1). We first focus on a special case where the bias term is exactly zero. For a random vector $\mathbf{X}$ with covariance $\mathbf{\Sigma} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T$, let $\mathbf{Z} = \mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{V}^T\mathbf{X}$. We say $\mathbf{X}$ has symmetric innovation if $\mathbf{Z} \overset{d}{=} (\mathbf{I}_d - 2\mathbf{e}_j\mathbf{e}_j^T)\mathbf{Z}, \forall j \in [d]$. In other words, flipping the sign of one component of $\mathbf{Z}$ will not change the distribution of $\mathbf{Z}$. Note that if $\mathbf{Z}$ has density, this is equivalent to say that its density function has the form $p(|z_1|, |z_2|, \ldots, |z_d|)$. All elliptical distributions centered at the origin belong to this family. In addition, if $\mathbf{Z}$ has symmetric and independent entries, $\mathbf{X}$ has also symmetric innovation. It turns out that when the random samples have symmetric innovation, $\mathbf{\Sigma}^* := \mathbb{E}(\widehat{\mathbf{V}}_K^{(\ell)}\widehat{\mathbf{V}}_K^{(\ell)T})$ and $\mathbf{\Sigma}$ share exactly the same set of eigenvectors. When we were completing the paper, we noticed that Chen et al. (2016) had independently established a similar result for the Gaussian case.

DEFINITION 4.1. Let $\mathcal{V}$ be a K-dimensional linear subspace of $\mathbb{R}^d$. For a subspace estimator represented by $\widehat{\mathbf{V}} \in \mathcal{O}_{d\times K}$, we say it is *unbiased* for $\mathcal{V}$ if and only if the top K eigenspace of $\mathbb{E}(\widehat{\mathbf{V}}\widehat{\mathbf{V}}^T)$ is $\mathcal{V}$.

If $\widehat{\mathbf{V}}_K^{(\ell)}$ is unbiased for $\mathrm{Col}(\mathbf{V}_K)$, then $\rho(\mathbf{V}_K^*, \mathbf{V}_K) = 0$ and we will only have the sample variance term in (4.1). In that case, aggregating $\{\widehat{\mathbf{V}}_K^{(\ell)}\}_{\ell=1}^m$ reduces variance and yields a better estimator $\widetilde{\mathbf{V}}_K$. Theorem 2 shows that this is the case so long as the distribution has symmetric innovation and the sample size is large enough.

THEOREM 2. *Suppose on the* ℓ*th server we have* n *i.i.d. random samples* $\{\mathbf{X}_i\}_{i=1}^n$ *with covariance* $\mathbf{\Sigma}$. *If* $\{\mathbf{X}_i\}_{i=1}^n$ *have symmetric innovation, then* $\mathbf{V}^T\mathbf{\Sigma}^*\mathbf{V}$ *is diagonal, that is,* $\mathbf{\Sigma}^*$ *and* $\mathbf{\Sigma}$ *share the same set of eigenvectors. Furthermore, if* $\|\mathbf{\Sigma}^* - \mathbf{V}_K\mathbf{V}_K^T\|_2 < 1/2$, *then* $\{\widehat{\mathbf{V}}_K^{(\ell)}\}_{\ell=1}^m$ *are unbiased for* $\mathrm{Col}(\mathbf{V}_K)$ *and* $\rho(\mathbf{V}_K^*, \mathbf{V}_K) = 0$.

It is worth pointing out that distributed PCA is closely related to aggregation of random sketches of a matrix (Halko, Martinsson and Tropp (2011), Tropp et al. (2016)). To approximate the subspace spanned by the K leading left singular vectors of a large matrix $\mathbf{A} \in \mathbb{R}^{d_1\times d_2}$, we could construct a suitable random matrix $\mathbf{Y} \in \mathbb{R}^{d_2\times n}$ with $n \geq K$, and use the left singular subspace of $\mathbf{AY} \in \mathbb{R}^{d_1\times n}$ as an

estimator. $\mathbf{AY}$ is called a random sketch of $\mathbf{A}$. It has been shown that to obtain reasonable statistical accuracy, n can be much smaller than $\min(d_1, d_2)$ as long as $\mathbf{A}$ is approximately low rank. Hence it is much cheaper to compute SVD on $\mathbf{AY}$ than on $\mathbf{A}$. When we want to aggregate a number of such subspace estimators, a smart choice of the random matrix ensemble for $\mathbf{Y}$ is always preferable. It follows from Theorem 2 that if we let $\mathbf{Y}$ have i.i.d. columns from a distribution with symmetric innovation (e.g., Gaussian distribution or independent entries), then the subspace estimators are unbiased, which facilitates aggregation.

Here, we explain why we need the condition $\|\mathbf{\Sigma}^* - \mathbf{V}_K\mathbf{V}_K^T\|_2 < 1/2$ to achieve zero bias. First of all, the condition is similar to a bound on the "variance" of the random matrix $\widehat{\mathbf{V}}_K^{(\ell)}$ whose covariance $\mathbf{\Sigma}^*$ is under investigation. As demonstrated above, with the symmetric innovation, $\mathbf{\Sigma}^*$ has the same set of eigenvectors as $\mathbf{\Sigma}$, but we still cannot guarantee that the top K eigenvectors of $\mathbf{\Sigma}^*$ match with those of $\mathbf{\Sigma}$. For example, the $(K+1)$-th eigenvector of $\mathbf{\Sigma}$ might be the Kth eigenvector of $\mathbf{\Sigma}^*$. In order to ensure the top K eigenspace of $\mathbf{\Sigma}^*$ is exactly the same as that of $\mathbf{\Sigma}$, we require $\widehat{\mathbf{V}}_K^{(\ell)}$ to not deviate too far from $\mathbf{V}_K$ so that $\mathbf{\Sigma}^*$ is close enough to $\mathbf{V}_K\mathbf{V}_K^T$. Both Theorems 1 and 2 require control of $\|\mathbf{\Sigma}^* - \mathbf{V}_K\mathbf{V}_K^T\|_2$, which will be studied shortly.

For general distributions, the bias term is not necessarily zero. However, it turns out that when the subsample size is large enough, the bias term $\rho(\mathbf{V}_K^*, \mathbf{V}_K)$ is of high order compared with the statistical error of $\widehat{\mathbf{V}}_K^{(\ell)}$ on the individual subsample. By the decomposition (4.1) and Theorem 1, we can therefore expect the aggregated estimator $\widetilde{\mathbf{V}}_K$ to enjoy sharper statistical error rate than PCA on the individual subsample. In other words, the aggregation does improve the statistical efficiency. A similar phenomenon also appears in statistical error analysis of the average of the debiased Lasso estimators in Battey et al. (2015) and Lee et al. (2017). Recall that in sparse linear regression, the Lasso estimator $\widehat{\boldsymbol{\beta}}$ satisfies that $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 = O_P(\sqrt{s\log d/n})$, where $\boldsymbol{\beta}^*$ is the true regression vector, s is the number of nonzero coefficients of $\boldsymbol{\beta}^*$ and d is the dimension. The debiasing step reduces the bias of $\widehat{\boldsymbol{\beta}}$ to the order $O_P(s\log d/n)$, which is negligible when m is not too large, compared with the statistical error of $\widehat{\boldsymbol{\beta}}$, and thus enables the average of the debiased Lasso estimators to enhance the statistical efficiency.

Below we present Lemma 2, a high-order Davis–Kahan theorem that explicitly characterizes the linear term and high-order error on top K eigenspace due to matrix perturbation. This is a genuine generalization of the former high-order perturbation theorems on a single eigenvector, for example, Lemma 1 in Kneip and Utikal (2001) and Theorem 2 in El Karoui and d'Aspremont (2010). An elegant result on eigenspace perturbation is Lemma 2 in Koltchinskii and Lounici (2016). Our error bound uses Frobenius norm while theirs uses spectral norm. Besides, when the top K eigenspace is of interest, the upper bound in Lemma 2 in Koltchinskii and Lounici (2016) contains an extra factor $1 + (\lambda_1 - \lambda_K)/\Delta$. Hence we have better dependence on problem parameters. Other related works in the

literature consider asymptotic expansions of perturbation (Kato (1966), Vaccaro (1994), Xu (2002)), and singular space of a matrix contaminated by Gaussian noise (Wang (2015)). Our result is both nonasymptotic and deterministic. It serves as the core of bias analysis.

LEMMA 2. *Let* $\mathbf{A}, \widehat{\mathbf{A}} \in \mathbb{R}^{d\times d}$ *be symmetric matrices with eigenvalues* $\lambda_1 \geq \cdots \geq \lambda_d$, *and* $\widehat{\lambda}_1 \geq \cdots \geq \widehat{\lambda}_d$, *respectively. Let* $\{\mathbf{u}_j\}_{j=1}^d$, $\{\widehat{\mathbf{u}}_j\}_{j=1}^d$ *be two orthonormal bases of* $\mathbb{R}^d$ *such that* $\mathbf{A}\mathbf{u}_j = \lambda_j \mathbf{u}_j$ *and* $\widehat{\mathbf{A}}\widehat{\mathbf{u}}_j = \widehat{\lambda}_j \widehat{\mathbf{u}}_j$ *for all* $j \in [d]$. *Fix* $s \in \{0, 1, \ldots, d-K\}$ *and assume that* $\Delta = \min\{\lambda_s - \lambda_{s+1}, \lambda_{s+K} - \lambda_{s+K+1}\} > 0$, *where* $\lambda_0 = +\infty$ *and* $\lambda_{d+1} = -\infty$. *Define* $\mathbf{U} = (\mathbf{u}_{s+1}, \ldots, \mathbf{u}_{s+K})$, $\widehat{\mathbf{U}} = (\widehat{\mathbf{u}}_{s+1}, \ldots, \widehat{\mathbf{u}}_{s+K})$. *Define* $\mathbf{E} = \widehat{\mathbf{A}} - \mathbf{A}$, $S = \{s+1, \ldots, s+K\}$, $\mathbf{G}_j = \sum_{i\notin S}(\lambda_i - \lambda_{s+j})^{-1}\mathbf{u}_i\mathbf{u}_i^T$ *for* $j \in [K]$, *and*

$$f : \mathbb{R}^{d\times K} \to \mathbb{R}^{d\times K} \quad (\mathbf{w}_1, \ldots, \mathbf{w}_K) \mapsto (-\mathbf{G}_1\mathbf{w}_1, \ldots, -\mathbf{G}_K\mathbf{w}_K).$$

When $\varepsilon = \|\mathbf{E}\|_2/\Delta \leq 1/10$, *we have*

$$\|\widehat{\mathbf{U}}\widehat{\mathbf{U}}^T - \mathbf{U}\mathbf{U}^T - [f(\mathbf{E}\mathbf{U})\mathbf{U}^T + \mathbf{U}f(\mathbf{E}\mathbf{U})^T]\|_F \leq 24\sqrt{K}\varepsilon^2.$$

Similar to Taylor expansion, the difference is decomposed into the linear leading term and residual of higher order with respect to the perturbation. Here, we only present a version that is directly applicable to bias analysis. Stronger results are summarized in Lemma 6 in the Supplementary Material, which refines the Davis–Kahan theorem by showing that under certain conditions, we have $\|\widehat{\mathbf{U}}\widehat{\mathbf{U}}^T - \mathbf{U}\mathbf{U}^T\|_F \asymp \|f(\mathbf{E}\mathbf{U})\|_F$ and

$$\|\widehat{\mathbf{U}}\widehat{\mathbf{U}}^T - \mathbf{U}\mathbf{U}^T - [f(\mathbf{E}\mathbf{U})\mathbf{U}^T + \mathbf{U}f(\mathbf{E}\mathbf{U})^T]\|_F \lesssim \varepsilon\|f(\mathbf{E}\mathbf{U})\|_F.$$

Hence it may be of independent interest in perturbation analysis of spectral projectors. Now we apply Lemma 2 to the context of principal eigenspace estimation. Let $\mathbf{A} = \mathbf{\Sigma}$, $\widehat{\mathbf{A}} = \widehat{\mathbf{\Sigma}}^{(1)}$ and $S = [K]$. It thus follows that $\mathbf{U} = \mathbf{V}_K$, $\widehat{\mathbf{U}} = \widehat{\mathbf{V}}_K^{(1)}$ and $\mathbf{E} = \widehat{\mathbf{\Sigma}}^{(1)} - \mathbf{\Sigma}$. From the second inequality in Lemma 2, we can conclude that the bias term $\rho(\mathbf{V}_K^*, \mathbf{V}_K)$ is a high-order term compared with the linear leading term. More specifically, the Davis–Kahan theorem helps us control the bias as follows:

$$\rho(\mathbf{V}_K^*, \mathbf{V}_K) \lesssim \|\mathbf{\Sigma}^* - \mathbf{V}_K\mathbf{V}_K^T\|_F = \|\mathbb{E}[\widehat{\mathbf{V}}_K^{(1)}\widehat{\mathbf{V}}_K^{(1)T} - \mathbf{V}_K\mathbf{V}_K^T]\|_F.$$

By the facts that $\mathbb{E}(\mathbf{E}) = 0$ and f is linear, we have

$$\rho(\mathbf{V}_K^*, \mathbf{V}_K) = \|\mathbb{E}[\widehat{\mathbf{V}}_K^{(1)}\widehat{\mathbf{V}}_K^{(1)T} - (\mathbf{V}_K\mathbf{V}_K^T + f(\mathbf{E}\mathbf{V}_K)\mathbf{V}_K^T + \mathbf{V}_K f(\mathbf{E}\mathbf{V}_K)^T)]\|_F.$$

By Jensen's inequality, the right-hand side above is further bounded by

$$\mathbb{E}\|\widehat{\mathbf{V}}_K^{(1)}\widehat{\mathbf{V}}_K^{(1)T} - (\mathbf{V}_K\mathbf{V}_K^T + f(\mathbf{E}\mathbf{V}_K)\mathbf{V}_K^T + \mathbf{V}_K f(\mathbf{E}\mathbf{V}_K)^T)\|_F. \tag{4.4}$$

When n is large enough, the typical size of $\varepsilon = \|\mathbf{E}\|_2/\Delta$ is small, and Lemma 1 controls it tail and all of the moments. Together with Lemma 2, this fact implies

that (4.4) has roughly the same order as $\sqrt{K} \cdot \mathbb{E}\varepsilon^2$, which should be much smaller than the typical size of $\sqrt{K}\varepsilon$, that is, the upper bound for $\rho(\widehat{\mathbf{V}}_K^{(1)}, \mathbf{V}_K)$ given by Davis–Kahan theorem. The following theorem makes our hand-waving analysis rigorous.

THEOREM 3. *There are constants C_1 and C_2 such that when $n \geq r$,*

$$\rho(\mathbf{V}_K^*, \mathbf{V}_K) \leq C_1 \|\mathbf{\Sigma}^* - \mathbf{V}_K \mathbf{V}_K^T\|_F \leq C_2 \kappa^2 \sqrt{K} r/n.$$

As a by-product, we get $\|\mathbf{\Sigma}^* - \mathbf{V}_K \mathbf{V}_K^T\|_2 \lesssim \kappa^2 \sqrt{K} r/n$. Hence when $n \geq C\kappa^2 \sqrt{K} r$ for some large enough C, the assumptions in Theorems 1 and 2 on $\|\mathbf{\Sigma}^* - \mathbf{V}_K \mathbf{V}_K^T\|_2$ are guaranteed to hold.

4.3. *Properties of distributed PCA.* We now combine the results we obtained in the previous two subsections to derive the statistical error rate of $\widetilde{\mathbf{V}}_K$. We first present a theorem under the setting of global i.i.d. data and discuss its optimality.

THEOREM 4. *Suppose we have N i.i.d. sub-Gaussian random samples with covariance $\mathbf{\Sigma}$. They are scattered across m servers, each of which stores n samples. There exist constants C, C_1, C_2, C_3 and C_4 such that the followings hold when $n \geq C\kappa^2 \sqrt{K} r$:*

1. *Symmetric innovation*:

$$\|\rho(\widetilde{\mathbf{V}}_K, \mathbf{V}_K)\|_{\psi_1} \leq C_1 \kappa \sqrt{\frac{Kr}{N}}. \tag{4.5}$$

2. *General distribution*:

$$\|\rho(\widetilde{\mathbf{V}}_K, \mathbf{V}_K)\|_{\psi_1} \leq C_1 \kappa \sqrt{\frac{Kr}{N}} + C_2 \kappa^2 \frac{\sqrt{K} r}{n}. \tag{4.6}$$

Furthermore, if we further assume $m \leq C_3 n/(\kappa^2 r)$,

$$\|\rho(\widetilde{\mathbf{V}}_K, \mathbf{V}_K)\|_{\psi_1} \leq C_4 \kappa \sqrt{\frac{Kr}{N}}. \tag{4.7}$$

As we can see, with appropriate scaling conditions on n, m and d, $\widetilde{\mathbf{V}}_K$ can achieve the statistical error rate of order $\kappa\sqrt{Kr/N}$. The result is applicable to the whole sample or traditional PCA, in which $m = 1$. Hence the distributed PCA and the traditional PCA share the same error bound as long as the technical conditions are satisfied.

In the second part of Theorem 4, the purpose of setting restrictions on n and m is to ensure that the distributed PCA algorithm delivers the same statistical rate as the centralized PCA which uses all the data. In the boundary case where $n \asymp \kappa^2 \sqrt{K} r$,

the bias of the local empirical eigenspace is of constant order. Since our aggregation cannot kill bias, there is no hope to achieve the centralized rate unless the number of machines is of constant order so that the centralized PCA has constant error, too. Besides, our result says that when n is large, we can tolerate more data splits (larger m) for achieving the centralized statistical rate.

We now illustrate our result through a simple spiked covariance model introduced by Johnstone (2001). Assume that $\mathbf{\Lambda} = \operatorname{diag}(\lambda, \underbrace{1, \ldots, 1}_{d-1})$, where $\lambda > 1$, and we are interested in the first eigenvector of $\mathbf{\Sigma}$. Note that $K = 1, r = \operatorname{Tr}(\mathbf{\Sigma})/\|\mathbf{\Sigma}\|_2 = (\lambda + d - 1)/\lambda \asymp d/\lambda$ when $\lambda = O(d)$, and $\kappa = \lambda/(\lambda - 1) \asymp: 1$. It is easy to see from (4.5) or (4.7) that

$$\|\rho(\widetilde{\mathbf{V}}_1, \mathbf{V}_1)\|_{\psi_1} \lesssim \kappa\sqrt{\frac{r}{N}} \lesssim \sqrt{\frac{d}{N\lambda}}.$$

Without loss of generality, we could always assume that the direction of $\widetilde{\mathbf{V}}_1$ is chosen such that $\widetilde{\mathbf{V}}_1^T\mathbf{V}_1 \geq 0$, that is, $\widetilde{\mathbf{V}}_1$ is aligned with $\mathbf{V}_1$. Note that

$$\begin{aligned}\rho^2(\widetilde{\mathbf{V}}_1, \mathbf{V}_1) &= \|\widetilde{\mathbf{V}}_1\widetilde{\mathbf{V}}_1^T - \mathbf{V}_1\mathbf{V}_1^T\|_F^2 = 2(1 - \widetilde{\mathbf{V}}_1^T\mathbf{V}_1)(1 + \widetilde{\mathbf{V}}_1^T\mathbf{V}_1)\\ &\geq 2(1 - \widetilde{\mathbf{V}}_1^T\mathbf{V}_1) = \|\widetilde{\mathbf{V}}_1 - \mathbf{V}_1\|_2^2.\end{aligned}$$

Hence

$$\mathbb{E}\|\widetilde{\mathbf{V}}_1 - \mathbf{V}_1\|_2^2 \lesssim \|\rho(\widetilde{\mathbf{V}}_1, \mathbf{V}_1)\|_{\psi_1}^2 \lesssim \frac{d}{N\lambda}. \tag{4.8}$$

We now compare this rate with the previous results under the spiked model. In Paul and Johnstone (2012), the authors derived the ℓ_2 risk of the empirical eigenvectors when random samples are Gaussian. It is not hard to derive from Theorem 1 therein that given N i.i.d. d-dimensional Gaussian samples, when N, d and λ go to infinity,

$$\mathbb{E}\|\widehat{\mathbf{V}}_1 - \mathbf{V}_1\|_2^2 \asymp \frac{d}{N\lambda},$$

where $\widehat{\mathbf{V}}_1$ is the empirical leading eigenvector with $\widehat{\mathbf{V}}_1^T\mathbf{V}_1 \geq 0$. We see from (4.8) that the aggregated estimator $\widetilde{\mathbf{V}}_1$ performs as well as the full sample estimator $\widehat{\mathbf{V}}_1$ in terms of the mean squared error. See Wang and Fan (2017) for generalization of the results for spiked covariance.

In addition, our result is consistent with the minimax lower bound developed in Cai, Ma and Wu (2013). For $\lambda > 0$ and fixed $c \geq 1$, define

$$\Theta = \{\mathbf{\Sigma} \succeq 0 : \lambda + 1 \leq \lambda_K \leq \lambda_1 \leq c\lambda + 1, \lambda_j = 1 \text{ for } K + 1 \leq j \leq d\}.$$

Assume that $K \leq d/2$ and $1 \lesssim d/\lambda \lesssim N$. Theorem 8 in Cai, Ma and Wu (2013) shows that under the Gaussian distribution with $\mathbf{\Sigma} \in \Theta$, the minimax lower bound of $\mathbb{E}\rho^2(\widehat{\mathbf{V}}, \mathbf{V}_K)$ satisfies

$$\inf_{\widehat{\mathbf{V}}} \sup_{\mathbf{\Sigma}\in\Theta} \mathbb{E}\rho^2(\widehat{\mathbf{V}}, \mathbf{V}_K) \gtrsim \min\left\{K, (d - K), \frac{K(\lambda + 1)(d - K)}{N\lambda^2}\right\} \gtrsim \frac{Kd}{N\lambda}. \tag{4.9}$$

Based on $r = \mathrm{Tr}(\mathbf{\Sigma})/\|\mathbf{\Sigma}\|_2 \le (cK\lambda + d)/(c\lambda + 1) \lesssim Kd/\lambda$ and $\kappa \le c \lesssim 1$, our (4.5) gives an upper bound

$$\mathbb{E}\rho^2(\widetilde{\mathbf{V}}_1, \mathbf{V}_1) \lesssim \kappa^2 \frac{Kr}{n} \lesssim \frac{Kd}{N\lambda},$$

which matches the lower bound in (4.9).

Although the upper bound $\kappa\sqrt{Kr/N}$ established in Theorem 4 is optimal in the minimax sense as discussed above, the nonminimax risk of empirical eigenvectors can be improved when the condition number κ is large. See Vu and Lei (2013), Koltchinskii and Lounici (2016) and Reiss and Wahl (2016) for sharper results. We use (4.5) as a benchmark rate for the centralized PCA only for the sake of simplicity.

Notice that in Theorem 4, the prerequisite for $\widetilde{\mathbf{V}}_K$ to enjoy the sharp statistical error rate is a lower bound on the subsample size n, that is,

$$n \gtrsim \kappa^2\sqrt{Kr}. \tag{4.10}$$

As in the remarks after Lemma 2, this is the condition we used to ensure closeness between $\mathbf{\Sigma}^*$ and $\mathbf{V}_K\mathbf{V}_K^T$. It is natural to ask whether this required sample complexity is sharp, or in other words, is it possible for $\widetilde{\mathbf{V}}_K$ to achieve the same statistical error rate with a smaller sample size on each machine? The answer is no. The following theorem presents a distribution family under which $\mathrm{Col}(\widetilde{\mathbf{V}}_K)$ is even perpendicular to $\mathrm{Col}(\mathbf{V}_K)$ with high probability when n is smaller than the threshold given in (4.10). This means that having a smaller sample size on each machine is too uninformative such that the aggregation step completely fails in improving estimation consistency.

THEOREM 5. *Consider a Bernoulli random variable W with $P(W = 0) = P(W = 1) = 1/2$, a Rademacher random variable $P(Y = 1) = P(Y = -1) = 1/2$, and a random vector $\mathbf{Z} \in \mathbb{R}^{d-1}$ that is uniformly distributed over the $(d-1)$-dimensional unit sphere. For $\lambda \ge 2$, we say a random vector $\mathbf{X} \in \mathbb{R}^d$ follows the distribution $\mathcal{D}(\lambda)$ if*

$$\mathbf{X} \overset{d}{=} \begin{pmatrix} \mathbf{1}_{\{W=0\}}\sqrt{2\lambda}Y \\ \mathbf{1}_{\{W=1\}}\sqrt{2(d-1)}\mathbf{Z} \end{pmatrix}.$$

Now suppose we have $\{\mathbf{X}_i\}_{i=1}^N$ as N i.i.d. random samples of $\mathbf{X}$. They are stored across m servers, each of which has n samples. When $32\log d \le n \le (d-1)/(3\lambda)$, we have

$$P(\widetilde{\mathbf{V}}_1 \perp \mathbf{V}_1) \ge \begin{cases} 1 - d^{-1} & \text{if } m \le d^3, \\ 1 - e^{-d/2} & \text{if } m > d^3. \end{cases}$$

It is easy to verify that $\mathcal{D}(\lambda)$ is symmetric, sub-Gaussian and satisfies $\mathbb{E}\mathbf{X} = \mathbf{0}$ and $\mathbb{E}(\mathbf{X}\mathbf{X}^T) = \operatorname{diag}(\lambda, 1, \ldots, 1)$. Besides, $\kappa = \lambda/(\lambda - 1) \asymp 1$ and $r = (\lambda + d - 1)/\lambda = d/\lambda + 1 - \lambda^{-1} \asymp d/\lambda$ when $2 \le \lambda \lesssim d$. According to (4.10), we require $n \gtrsim d/\lambda$ to achieve the rate as demonstrated in (4.5). Theorem 5 shows that if we have fewer samples than this threshold, the aggregated estimator $\widetilde{\mathbf{V}}_1$ will be perpendicular to the true top eigenvector $\mathbf{V}_1$ with high probability. Therefore, our lower bound for the subsample size n is sharp.

5. Extension to heterogeneous samples. We now relax global *i.i.d.* assumptions in the previous section to the setting of heterogeneous covariance structures across servers. Suppose data on the server ℓ has covariance matrix $\mathbf{\Sigma}^{(\ell)}$, whose top K eigenvalues and eigenvectors are denoted by $\{\lambda_k^{(\ell)}\}_{k=1}^K$ and $\mathbf{V}_K^{(\ell)} = (\mathbf{v}_1^{(\ell)}, \ldots, \mathbf{v}_K^{(\ell)})$, respectively. We will study two specific cases of heterogeneous covariances: one requires all covariances to share exactly the same principal eigenspaces, while the other considers the heterogeneous factor models with common factor eigenstructures.

5.1. *Common principal eigenspaces.* We assume that $\{\mathbf{\Sigma}^{(\ell)}\}_{\ell=1}^m$ share the same top K eigenspace, that is, there exists some $\mathbf{V}_K \in \mathcal{O}_{d\times K}$ such that $\mathbf{V}_K^{(\ell)}\mathbf{V}_K^{(\ell)T} = \mathbf{V}_K\mathbf{V}_K^T$ for all $\ell \in [m]$. The following theorem can be viewed as a generalization of Theorem 4.

THEOREM 6. *Suppose we have in total N sub-Gaussian samples scattered across m servers, each of which stores n i.i.d. samples with covariance $\mathbf{\Sigma}^{(\ell)}$. Assume that $\{\mathbf{\Sigma}^{(\ell)}\}_{\ell=1}^m$ share the same top K eigenspace. For each $\ell \in [m]$, let $S_\ell = \kappa_\ell\sqrt{\frac{Kr_\ell}{N}}$ and $B_\ell = \frac{\kappa_\ell^2\sqrt{K}r_\ell}{n}$, where $r_\ell := \operatorname{Tr}(\mathbf{\Sigma}^{(\ell)})/\lambda_1^{(\ell)}$ and $\kappa_\ell := \lambda_1^{(\ell)}/(\lambda_K^{(\ell)} - \lambda_{K+1}^{(\ell)})$.*

1. *Symmetric innovation*: *There exist some positive constants C and C_1 such that*

$$\big\|\rho(\widetilde{\mathbf{V}}_K, \mathbf{V}_K)\big\|_{\psi_1} \le C_1\sqrt{\frac{1}{m}\sum_{\ell=1}^m S_\ell^2} \tag{5.1}$$

so long as $n \ge C\sqrt{K}\max_{\ell\in[m]}(\kappa_\ell^2 r_\ell)$.

2. *General distribution*: *There exist positive constant C_2 and C_3 such that when $n \ge \max_{\ell\in[m]} r_\ell$,*

$$\big\|\rho(\widetilde{\mathbf{V}}_K, \mathbf{V}_K)\big\|_{\psi_1} \le C_2\sqrt{\frac{1}{m}\sum_{\ell=1}^m S_\ell^2} + \frac{C_3}{m}\sum_{\ell=1}^m B_{(\ell)}. \tag{5.2}$$

5.2. *Heterogeneous factor models.* Suppose on the server ℓ, the data conform to a factor model as below:

$$\mathbf{X}_i^{(\ell)} = \mathbf{B}^{(\ell)}\mathbf{f}_i^{(\ell)} + \mathbf{u}_i^{(\ell)}, \quad i \in [n],$$

where $\mathbf{B}^{(\ell)} \in \mathbb{R}^{d\times K}$ is the loading matrix, $\mathbf{f}_i^{(\ell)} \in \mathbb{R}^K$ is the factor that satisfies $\mathrm{Cov}(\mathbf{f}_i^{(\ell)}) = \mathbf{I}$ and $\mathbf{u}_i^{(\ell)} \in \mathbb{R}^d$ is the residual vector. It is not hard to see that $\mathbf{\Sigma}^{(\ell)} = \mathrm{Cov}(\mathbf{X}_i^{(\ell)}) = \mathbf{B}^{(\ell)}\mathbf{B}^{(\ell)T} + \mathbf{\Sigma}_u^{(\ell)}$, where $\mathbf{\Sigma}_u^{(\ell)}$ is the covariance matrix of $\mathbf{u}_i^{(\ell)}$.

Let $\mathbf{B}^{(\ell)}\mathbf{B}^{(\ell)T} = \mathbf{V}_K^{(\ell)}\mathbf{\Lambda}_K^{(\ell)}\mathbf{V}_K^{(\ell)T}$ be the spectral decomposition of $\mathbf{B}^{(\ell)}\mathbf{B}^{(\ell)T}$. We assume that there exists a projection matrix $\mathbf{P}_K = \mathbf{V}_K\mathbf{V}_K^T$, where $\mathbf{V}_K \in \mathcal{O}_{d\times K}$, such that $\mathbf{V}_K^{(\ell)}\mathbf{V}_K^{(\ell)T} = \mathbf{P}_K$ for all $\ell \in [m]$. In other words, $\{\mathbf{B}^{(\ell)}\mathbf{B}^{(\ell)T}\}_{\ell=1}^m$ share the same top K eigenspace. Given the context of factor models, this implies that the factors have similar impact on the variation of the data across servers. Our goal now is to recover $\mathrm{Col}(\mathbf{V}_K)$ by the distributed PCA approach, namely Algorithm 1.

Recall that $\widehat{\mathbf{\Sigma}}^{(\ell)} = \frac{1}{n}\sum_{i=1}^n \mathbf{X}_i^{(\ell)}\mathbf{X}_i^{(\ell)^T}$ is the sample covariance matrix on the ℓth machine, and $\widehat{\mathbf{V}}_K^{(\ell)} = (\widehat{\mathbf{v}}_1^{(\ell)}, \ldots, \widehat{\mathbf{v}}_K^{(\ell)}) \in \mathcal{O}_{d\times K}$ stores K leading eigenvectors of $\widehat{\mathbf{\Sigma}}^{(\ell)}$. Define $\widetilde{\mathbf{\Sigma}} = \frac{1}{m}\sum_{\ell=1}^m \widehat{\mathbf{V}}_K^{(\ell)}\widehat{\mathbf{V}}_K^{(\ell)T}$, and let $\widetilde{\mathbf{V}}_K \in \mathcal{O}_{d\times K}$ be the top K eigenvectors of $\widetilde{\mathbf{\Sigma}}$. Below we present a theorem that characterizes the statistical performance of the distributed PCA under the heterogeneous factor models.

THEOREM 7. *For each $\ell \in [m]$, let $S_\ell = \kappa_\ell\sqrt{\frac{Kr_\ell}{N}}$ and $B_\ell = \frac{\kappa_\ell^2\sqrt{K}r_\ell}{n}$. There exist some positive constants C_1, C_2 and C_3 such that when $n \geq \max_{\ell\in[m]} r_\ell$,*

$$\big\|\rho(\widetilde{\mathbf{V}}_K, \mathbf{V}_K)\big\|_{\psi_1} \leq C_1\sqrt{\frac{1}{m}\sum_{\ell=1}^m S_\ell^2} + \frac{C_2}{m}\sum_{\ell=1}^m B_\ell + C_3\frac{\sqrt{K}}{m}\sum_{\ell=1}^m \frac{\|\mathbf{\Sigma}_u^{(\ell)}\|_2}{\lambda_K(\mathbf{\Lambda}_K^{(\ell)})}. \tag{5.3}$$

The first two terms in the RHS of (5.3) are similar to those in (5.2), while the third term characterizes the effect of heterogeneity in statistical efficiency of $\widetilde{\mathbf{V}}_K$. When $\|\mathbf{\Sigma}_u^{(\ell)}\|_2$ is small compared with $\lambda_K(\mathbf{\Lambda}_K^{(\ell)})$ as in spiky factor models, $\mathbf{\Sigma}_u^{(\ell)}$ can hardly distort the eigenspace $\mathrm{Col}(\mathbf{V}_K)$, and thus has little influence on the final statistical error of $\widetilde{\mathbf{V}}_K$.

6. Simulation study. In this section, we conduct Monte Carlo simulations to validate the statistical error rate of $\widetilde{\mathbf{V}}_K$ that is established in the previous section. We also compare the statistical accuracy of $\widetilde{\mathbf{V}}_K$ and its full sample counterpart $\widehat{\mathbf{V}}_K$, that is, the empirical top K eigenspace based on the full sample covariance. The main message is that our proposed distributed estimator performs equally well as the full sample estimator $\widehat{\mathbf{V}}_K$ when the subsample size n is large enough.

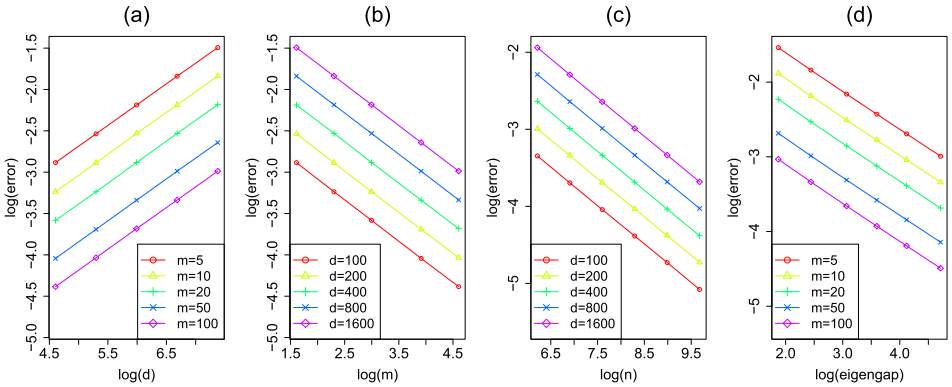

FIG. 1. *Statistical error rate with respect to*: (*a*) *the dimension* d *when* $\lambda = 50$ *and* $n = 2000$; (*b*) *the number of servers* m *when* $\lambda = 50$ *and* $n = 2000$; (*c*) *the subsample size* n *when* $\lambda = 50$ *and* $m = 50$; (*d*) *the eigengap* δ *when* $d = 800$ *and* $n = 2000$.

6.1. *Verification of the statistical error rate.* Consider $\{\mathbf{x}_i\}_{i=1}^N$ i.i.d. following $N(\mathbf{0}, \mathbf{\Sigma})$, where $\mathbf{\Sigma} = \operatorname{diag}(\lambda, \lambda/2, \lambda/4, 1, \ldots, 1)$. Here, the number of spiky eigenvalues $K = 3$ and $\mathbf{V}_K = (\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$. We generate m subsamples, each of which has n samples, and run our proposed distributed PCA algorithm (Algorithm 1) to calculate $\widetilde{\mathbf{V}}_K$. Since the centered multivariate Gaussian distribution is symmetric, according to Theorem 4, when $\lambda = O(d)$ we have

$$\|\rho(\widetilde{\mathbf{V}}_K, \mathbf{V}_K)\|_{\psi_1} = O\left(\frac{C_1\|\mathbf{\Sigma}\|_2}{\lambda_K - \lambda_{K+1}}\sqrt{\frac{Kr(\mathbf{\Sigma})}{N}}\right) = O\left(\sqrt{\frac{d}{mn\delta}}\right), \tag{6.1}$$

where $\delta := \lambda_K - \lambda_{K+1} = \lambda/4 - 1$. Now we provide numerical verification of the order of the number of servers m, the eigengap δ, the subsample size n and dimension d in the statistical error.

Figure 1 presents four plots that demonstrate how $\rho(\widetilde{\mathbf{V}}_K, \mathbf{V}_K)$ changes as d, m, n and δ increases respectively. Each data point on the plots is based on 100 independent Monte Carlo simulations. Figure 1(a) demonstrates how $\rho(\widetilde{\mathbf{V}}_K, \mathbf{V}_K)$ increases with respect to the increasing dimension d when $\lambda = 50$ and $n = 2000$. Each line on the plot represents a fixed number of machines m. Figure 1(b) shows the decay rate of $\rho(\widetilde{\mathbf{V}}_K, \mathbf{V}_K)$ as the number of servers m increases when $\lambda = 50$ and $n = 2000$. Different lines on the plot correspond to different dimensions d. Figure 1(c) demonstrates how $\rho(\widetilde{\mathbf{V}}_K, \mathbf{V}_K)$ decays as the subsample size n increases when $\lambda = 50$ and $m = 50$. Figure 1(d) shows the relationship between $\rho(\widetilde{\mathbf{V}}_K, \mathbf{V}_K)$ and the eigengap δ when $d = 800$ and $n = 2000$. The results from Figures 1(a)–(d) show that $\rho(\widetilde{\mathbf{V}}_K, \mathbf{V}_K)$ is proportion to $d^{\frac{1}{2}}$, $m^{-\frac{1}{2}}$, $n^{-\frac{1}{2}}$ and $\delta^{-\frac{1}{2}}$, respectively when the other three parameters are fixed. These empirical results are all consistent with (6.1).

Figure 1 demonstrates the marginal relationship between $\rho(\widetilde{\mathbf{V}}_K, \mathbf{V}_K)$ and the four parameters m, n, d and δ. Now we study their joint relationship. Inspired by

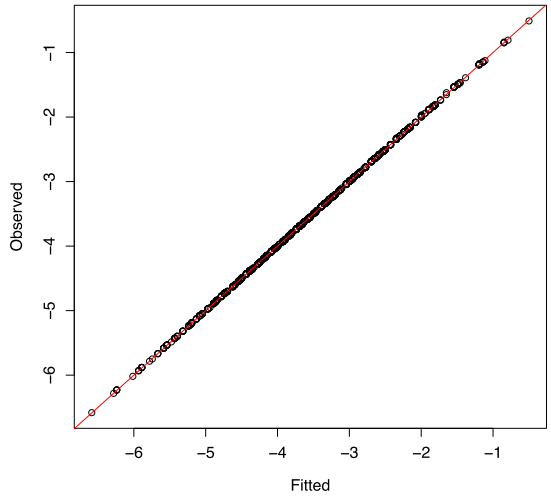

FIG. 2. *Observed and fitted values of* $\log(\rho(\widetilde{\mathbf{V}}_K, \mathbf{V}_K))$.

(6.1), we consider a multiple regression model as follows:

$$\log\bigl(\rho(\widetilde{\mathbf{V}}_K, \mathbf{V}_K)\bigr) = \beta_0 + \beta_1 \log(d) + \beta_2 \log(m) + \beta_3 \log(n) + \beta_4 \log(\delta) + \varepsilon, \tag{6.2}$$

where ε is the error term. We collect all the data points $(d, m, n, \delta, \rho(\widetilde{\mathbf{V}}_K, \mathbf{V}_K))$ from four plots in Figure 1 to fit the regression model (6.2). The fitting result is that $\widehat{\beta}_1 = 0.5043$, $\widehat{\beta}_2 = -0.4995$, $\widehat{\beta}_3 = -0.5011$ and $\widehat{\beta}_4 = -0.5120$ with the multiple $R^2 = 0.99997$. These estimates are quite consistent with the theoretical results in (6.1). Moreover, Figure 2 plots all the observed values of $\log(\rho(\widetilde{\mathbf{V}}_K, \mathbf{V}_K))$ against its fitted values by the linear model (6.2). We can see that the observed and fitted values perfectly match. It indicates that the multiple regression model (6.2) well explains the joint relationship between the statistical error and the four parameters m, n, d and δ.

6.2. *The effects of splitting.* In this section, we investigate how the number of data splits m affects the statistical performance of $\widetilde{\mathbf{V}}_K$ when the total sample size N is fixed. Since $N = mn$, it is easy to see that the larger m is, the smaller n will be, and hence the less computational load there will be on each individual server. In this way, to reduce the time consumption of the distributed algorithm, we prefer more splits of the data. However, per the assumptions of Theorem 4, the subsample size n should be large enough to achieve the optimal statistical performance of $\widetilde{\mathbf{V}}_K$. This motivates us to numerically illustrate how $\rho(\widetilde{\mathbf{V}}_K, \mathbf{V}_K)$ changes as m increases with N fixed.

We adopt the same data generation process as described in the beginning of Section 6.1 with $\lambda = 50$ and $N = 6000$. We split the data into m subsamples where m is chosen to be all the factors of N that are less than or equal to 300. Figure 3

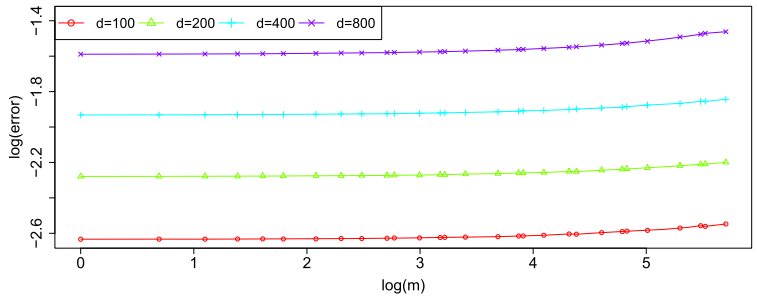

FIG. 3. *Statistical error with respect to the number of machines when the total sample size* $N = 6000$ *is fixed.*

plots $\rho(\widetilde{\mathbf{V}}_K, \mathbf{V}_K)$ with respect to the number of machines m. Each point on the plot is based on 100 simulations. Each line corresponds to a different dimension d.

The results show that when the number of machines is not unreasonably large, or equivalently the number of subsample size n is not small, the statistical error does not depend on the number of machines when N is fixed. This is consistent with (6.1) where the statistical error rate only depends on the total sample size $N = mn$. When the number of machines m is large ($\log m \geq 5$), or the subsample size n is small, we observe slightly growing statistical error of the distributed PCA. This is aligned with the required lower bound of n in Theorem 4 to achieve the optimal statistical performance of $\widetilde{\mathbf{V}}_K$. Note that even when $m = 300$ ($\log(m) \approx 5.7$) and $n = 20$, our distributed PCA performs very well. This demonstrates that distributed PCA is statistically efficient as long as m is within a reasonable range.

6.3. *Comparison between distributed and full sample PCA.* In this subsection, we compare the statistical performance of the following three methods:

1. Distributed PCA (DP);
2. Full sample PCA (FP), that is, the PCA based on the all the samples;
3. Distributed PCA with communication of five additional largest eigenvectors (DP5).

Here, we explain more on the third method DP5. The difference between DP5 and DP is that on each server, DP5 calculates $\widehat{\mathbf{V}}^{(\ell)}_{K+5}$, the top $K + 5$ eigenvectors of $\mathbf{\Sigma}^{(\ell)}$ and send them to the central server, and the central server computes the top K eigenvectors of $(1/m)\sum_{\ell=1}^{m} \widehat{\mathbf{V}}^{(\ell)}_{K+5} \widehat{\mathbf{V}}^{(\ell)T}_{K+5}$ as the final output. Intuitively, DP5 communicates more information of the covariance structure and is designed to guide the spill-over effects of the eigenspace spanned by the top K eigenvalues. In Figure 4, we compare the performance of all the three methods under various scenarios.

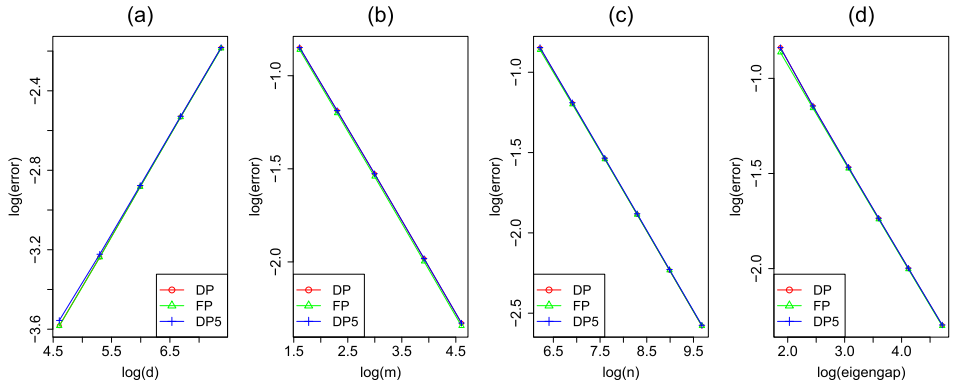

FIG. 4. *Comparison between DP, FP and DP5*: (*a*) $m = 20$, $n = 2000$ *and* $\lambda = 50$; (*b*) $d = 1600$, $n = 1000$ *and* $\lambda = 30$; (*c*) $d = 800$, $m = 5$ *and* $\lambda = 30$; (*d*) $d = 1600$, $m = 10$ *and* $n = 500$.

From Figures 4(a)–(d), we can see that all the three methods have similar finite sample performance. This means that it suffices to communicate K eigenvectors to enjoy the same statistical accuracy as the full sample PCA. For more challenging situations with large $p/(mn\delta)$ ratios, small improvements using FP are visible.

7. Discussion. Our theoretical results are established under sub-Gaussian assumptions of the data. We believe that similar results will hold under distributions with heavier tails than sub-Gaussian tails, or more specifically, with only bounded fourth moment. Typical examples are Student t-distributions with more than four degrees of freedom, Pareto distribution, etc. The only difference is that with heavy-tailed distribution, if the local estimators are still the top eigenspaces of the sample covariance matrix, we will not be able to derive exponential deviation bounds. To establish statistical rate with exponential deviation, special treatments of data, including shrinkage (Fan, Wang and Zhu (2016), Minsker (2018), Wei and Minsker (2017)), are needed, and the bias induced by such treatments should be carefully controlled. This will be an interesting future problem to study.

SUPPLEMENTARY MATERIAL

Supplement to "Distributed estimation of principal eigenspaces" (DOI: 10.1214/18-AOS1713SUPP; .pdf). Proofs of the results in the paper can be found in the Supplementary Material.

REFERENCES

ANDERSON, T. W. (1963). Asymptotic theory for principal component analysis. *Ann. Math. Stat.* **34** 122–148. MR0145620

BAIK, J., BEN AROUS, G. and PÉCHÉ, S. (2005). Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *Ann. Probab.* **33** 1643–1697. MR2165575

BATTEY, H., FAN, J., LIU, H., LU, J. and ZHU, Z. (2015). Distributed estimation and inference with statistical guarantees. Preprint. Available at arXiv:1509.05457.

BERTRAND, A. and MOONEN, M. (2014). Distributed adaptive estimation of covariance matrix eigenvectors in wireless sensor networks with application to distributed PCA. *Signal Process*. **104** 120–135.

BOUTSIDIS, C., WOODRUFF, D. P. and ZHONG, P. (2016). Optimal principal component analysis in distributed and streaming models. In *STOC'*16*—Proceedings of the* 48*th Annual ACM SIGACT Symposium on Theory of Computing* 236–249. ACM, New York. MR3536569

CAI, T. T., MA, Z. and WU, Y. (2013). Sparse PCA: Optimal rates and adaptive estimation. *Ann. Statist*. **41** 3074–3110. MR3161458

CHEN, X. and XIE, M. (2014). A split-and-conquer approach for analysis of extraordinarily large data. *Statist. Sinica* **24** 1655–1684. MR3308656

CHEN, T.-L., CHANG, D. D., HUANG, S.-Y., CHEN, H., LIN, C. and WANG, W. (2016). Integrating multiple random sketches for singular value decomposition. Preprint. Available at arXiv:1608.08285.

EL KAROUI, N. and D'ASPREMONT, A. (2010). Second order accurate distributed eigenvector computation for extremely large matrices. *Electron. J. Stat*. **4** 1345–1385. MR2738536

FAN, J., WANG, W. and ZHU, Z. (2016). Robust low-rank matrix recovery. Preprint. Available at arXiv:1603.08315.

FAN, J., WANG, D., WANG, K. and ZHU, Z. (2019). Supplement to "Distributed estimation of principal eigenspaces." DOI:10.1214/18-AOS1713SUPP.

GARBER, D., SHAMIR, O. and SREBRO, N. (2017). Communication-efficient algorithms for distributed stochastic principal component analysis. Preprint. Available at arXiv:1702.08169.

GOLUB, G. H. and VAN LOAN, C. F. (2012). *Matrix Computations*, 3rd ed. *Johns Hopkins Studies in the Mathematical Sciences*. Johns Hopkins Univ. Press, Baltimore, MD.

GUO, Z.-C., LIN, S.-B. and ZHOU, D.-X. (2017). Learning theory of distributed spectral algorithms. *Inverse Probl*. **33** 074009, 29. MR3672238

HALKO, N., MARTINSSON, P. G. and TROPP, J. A. (2011). Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Rev*. **53** 217–288. MR2806637

HOTELLING, H. (1933). Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol*. **24** 417–441.

JOHNSTONE, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist*. **29** 295–327. MR1863961

JOHNSTONE, I. M. and LU, A. Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *J. Amer. Statist. Assoc*. **104** 682–693. MR2751448

JUNG, S. and MARRON, J. S. (2009). PCA consistency in high dimension, low sample size context. *Ann. Statist*. **37** 4104–4130. MR2572454

KANNAN, R., VEMPALA, S. and WOODRUFF, D. (2014). Principal component analysis and higher correlations for distributed data. In *Conference on Learning Theory* 1040–1057.

KARGUPTA, H., HUANG, W., SIVAKUMAR, K. and JOHNSON, E. (2001). Distributed clustering using collective principal component analysis. *Knowledge and Information Systems* **3** 422–448.

KATO, T. (1966). *Perturbation Theory for Linear Operators*. *Die Grundlehren der Mathematischen Wissenschaften*, *Band* 132. Springer, New York. MR0203473

KNEIP, A. and UTIKAL, K. J. (2001). Inference for density families using functional principal component analysis. *J. Amer. Statist. Assoc*. **96** 519–542. MR1946423

KOLTCHINSKII, V. and LOUNICI, K. (2016). Asymptotics and concentration bounds for bilinear forms of spectral projectors of sample covariance. *Ann. Inst. Henri Poincaré Probab. Stat*. **52** 1976–2013. MR3573302

KOLTCHINSKII, V. and LOUNICI, K. (2017). Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli* **23** 110–133. MR3556768

LEE, J. D., LIU, Q., SUN, Y. and TAYLOR, J. E. (2017). Communication-efficient sparse regression. *J. Mach. Learn. Res.* **18** 1–30. MR3625709

LI, L., SCAGLIONE, A. and MANTON, J. H. (2011). Distributed principal subspace estimation in wireless sensor networks. *IEEE J. Sel. Top. Signal Process.* **5** 725–738.

LIANG, Y., BALCAN, M.-F. F., KANCHANAPALLY, V. and WOODRUFF, D. (2014). Improved distributed principal component analysis. In *Advances in Neural Information Processing Systems* 3113–3121.

MINSKER, S. (2018). Sub-Gaussian estimators of the mean of a random matrix with heavy-tailed entries. *Ann. Statist.* **46** 2871–2903. MR3851758

MÜCKE, N. and BLANCHARD, G. (2018). Parallelizing spectrally regularized kernel algorithms. *J. Mach. Learn. Res.* **19** Paper No. 30, 29. MR3862437

NADLER, B. (2008). Finite sample approximation results for principal component analysis: A matrix perturbation approach. *Ann. Statist.* **36** 2791–2817. MR2485013

ONATSKI, A. (2012). Asymptotics of the principal components estimator of large factor models with weakly influential factors. *J. Econometrics* **168** 244–258. MR2923766

PAUL, D. (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statist. Sinica* **17** 1617–1642. MR2399865

PAUL, D. and JOHNSTONE, I. M. (2012). Augmented sparse principal component analysis for high dimensional data. Preprint. Available at arXiv:1202.1242.

PEARSON, K. (1901). On lines and planes of closest fit to systems of point in space. *Philosophical Magazine Series* 6 **2** 559–572.

QU, Y., OSTROUCHOV, G., SAMATOVA, N. and GEIST, A. (2002). Principal component analysis for dimension reduction in massive distributed data sets. In *IEEE International Conference on Data Mining* (*ICDM*).

REISS, M. and WAHL, M. (2016). Non-asymptotic upper bounds for the reconstruction error of PCA. Preprint. Available at arXiv:1609.03779.

SCHIZAS, I. D. and ADUROJA, A. (2015). A distributed framework for dimensionality reduction and denoising. *IEEE Trans. Signal Process.* **63** 6379–6394. MR3421046

SHEN, D., SHEN, H. and MARRON, J. S. (2013). Consistency of sparse PCA in high dimension, low sample size contexts. *J. Multivariate Anal.* **115** 317–333. MR3004561

SHEN, D., SHEN, H., ZHU, H. and MARRON, J. S. (2016). The statistics and mathematics of high dimension low sample size asymptotics. *Statist. Sinica* **26** 1747–1770. MR3586237

TROPP, J. A., YURTSEVER, A., UDELL, M. and CEVHER, V. (2016). Randomized single-view algorithms for low-rank matrix approximation. Preprint. Available at arXiv:1609.00048.

VACCARO, R. J. (1994). A second-order perturbation expansion for the SVD. *SIAM J. Matrix Anal. Appl.* **15** 661–671. MR1266609

VERSHYNIN, R. (2012). Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing* 210–268. Cambridge Univ. Press, Cambridge. MR2963170

VU, V. Q. and LEI, J. (2013). Minimax sparse principal subspace estimation in high dimensions. *Ann. Statist.* **41** 2905–2947. MR3161452

WANG, R. (2015). Singular vector perturbation under Gaussian noise. *SIAM J. Matrix Anal. Appl.* **36** 158–177. MR3310977

WANG, W. and FAN, J. (2017). Asymptotics of empirical eigenstructure for high dimensional spiked covariance. *Ann. Statist.* **45** 1342–1374. MR3662457

WEI, X. and MINSKER, S. (2017). Estimation of the covariance structure of heavy-tailed distributions. In *Advances in Neural Information Processing Systems* 2855–2864.

XU, Z. (2002). Perturbation analysis for subspace decomposition with applications in subspace-based algorithms. *IEEE Trans. Signal Process.* **50** 2820–2830.

YU, Y., WANG, T. and SAMWORTH, R. J. (2015). A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika* **102** 315–323. MR3371006

ZHANG, Y., DUCHI, J. C. and WAINWRIGHT, M. J. (2013). Divide and conquer kernel ridge regression. In *COLT* 592–617.

J. FAN
D. WANG
K. WANG
DEPARTMENT OF OPERATIONS RESEARCH
AND FINANCIAL ENGINEERING
PRINCETON UNIVERSITY
PRINCETON, NEW JERSEY 08544
USA
E-MAIL: jqfan@princeton.edu
dongwang@princeton.edu
kaizheng@princeton.edu

Z. ZHU
DEPARTMENT OF STATISTICS
UNIVERSITY OF MICHIGAN
ANN ARBOR, MICHIGAN 48109
USA
E-MAIL: ziweiz@umich.edu