The Annals of Statistics
2017, Vol. 45, No. 3, 1342–1374
DOI: 10.1214/16-AOS1487
© Institute of Mathematical Statistics, 2017

ASYMPTOTICS OF EMPIRICAL EIGENSTRUCTURE FOR HIGH DIMENSIONAL SPIKED COVARIANCE

BY WEICHEN WANG* AND JIANQING FAN[1],*,†

Princeton University and Fudan University†*

We derive the asymptotic distributions of the spiked eigenvalues and eigenvectors under a generalized and unified asymptotic regime, which takes into account the magnitude of spiked eigenvalues, sample size and dimensionality. This regime allows high dimensionality and diverging eigenvalues and provides new insights into the roles that the leading eigenvalues, sample size and dimensionality play in principal component analysis. Our results are a natural extension of those in [*Statist. Sinica* **17** (2007) 1617–1642] to a more general setting and solve the rates of convergence problems in [*Statist. Sinica* **26** (2016) 1747–1770]. They also reveal the biases of estimating leading eigenvalues and eigenvectors by using principal component analysis, and lead to a new covariance estimator for the approximate factor model, called Shrinkage Principal Orthogonal complEment Thresholding (S-POET), that corrects the biases. Our results are successfully applied to outstanding problems in estimation of risks for large portfolios and false discovery proportions for dependent test statistics and are illustrated by simulation studies.

1. Introduction.

Principal Component Analysis (PCA) is a powerful tool for dimension reduction and data visualization. Its theoretical properties such as consistency and asymptotic distributions of the empirical eigenvalues and eigenvectors are challenging especially in high dimensional regime. For the past half century, substantial amount of efforts have been devoted to understanding empirical eigenstructure. An early effort came from Anderson (1963) who established the asymptotic normality of sample eigenvalues and eigenvectors under the classical regime with large sample size n and fixed dimensionality p. However, when dimensionality diverges at the same rate as the sample size, sample covariance matrix is a notoriously bad estimator with dramatically different eigenstructure from the population covariance. A lot of recent literature makes the endeavor to understand the behavior of the empirical eigenvalues and eigenvectors under high dimensional regime where both n and p go to infinity. See, for example, Bai (1999), Baik, Ben Arous and Péché (2005), Johnstone and Lu (2009), Onatski (2012), Paul

Received September 2015; revised June 2016.

[1] Supported in part by NSF Grants DMS-1206464 and DMS-1406266 and NIH Grants R01-GM072611-11 and NIH R01GM100474-04.

MSC2010 subject classifications. Primary 62H25; secondary 62H10.

Key words and phrases. Asymptotic distributions, principal component analysis, diverging eigenvalues, approximate factor model, relative risk management, false discovery proportion.

(2007), Shen et al. (2016) and many related papers. For additional developments and references, see Bai and Silverstein (2010).

Among different structures of population covariance, the spiked covariance model is of great interest. It typically assumes several eigenvalues larger than the remaining, and focuses on recovering only these leading eigenvalues and their associated eigenvectors. The spiked part is of importance, as we are usually interested in the directions that explain the most variations of the data. In this paper, we consider a high dimensional spiked covariance model with the leading eigenvalues larger than the rest. We provide new understanding on how the spiked empirical eigenvalues and eigenvectors fluctuate around their theoretical counterparts and what their asymptotic biases are. Three quantities play an essential role in determining the asymptotic behavior of empirical eigenstructure: the sample size n, the dimensionality p and the magnitude of leading eigenvalues $\{\lambda_j\}_{j=1}^{m}$. The natural question to ask is how the asymptotics of empirical engen-structure depends on the interplay of those quantities. We will give a unified answer to this important question in the principal component analysis. Theoretical properties of PCA have been investigated from three different perspectives: (i) random matrix theories, (ii) sparse PCA and (iii) approximate factor model.

The first angle to analyze PCA is through random matrix theories, where it is typically assumed $p/n \to \gamma \in (0, \infty)$ with bounded spike sizes. It is well known that if the true covariance matrix is identity, the empirical spectral distribution converges almost surely to the Marcenko–Pastur distribution [Bai (1999)] and when $\gamma < 1$ the largest and smallest eigenvalues converge almost surely to $(1+\sqrt{\gamma})^2$ and $(1-\sqrt{\gamma})^2$, respectively [Bai and Yin (1993), Johnstone (2001)]. If the true covariance structure takes the form of a spiked matrix, Baik, Ben Arous and Péché (2005) showed that the asymptotic distribution of the top empirical eigenvalue exhibits an $n^{2/3}$ scaling when the eigenvalue lies below a threshold $1+\sqrt{\gamma}$, and an $n^{1/2}$ scaling when it is above the threshold (named BBP phase transition after the authors). The phase transition is further studied by Benaych-Georges and Nadakuditi (2011) and Bai and Yao (2012) under more general assumptions. For the case where we have the regular scaling, Paul (2007) investigated the asymptotic behavior of the corresponding empirical eigenvectors and showed that the major part of an eigenvector is normally distributed with a regular scaling $n^{1/2}$. The convergence of principal component scores under this regime was considered by Lee, Zou and Wright (2010). The same random matrix regime has also been considered by Onatski (2012) in studying the principal component estimator for high-dimensional factor models. More recently, Koltchinskii and Lounici (2014, 2017) revealed a profound link of concentration bounds of empirical eigenstructure with the effective rank defined as $\bar{r} = \operatorname{tr}(\boldsymbol{\Sigma})/\lambda_1$ [Vershynin (2010)]. Their results extend the regime of bounded eigenvalues to a more general setting, although the asymptotic results in most cases still rely on the assumption $\bar{r} = o(n)$, which essentially requires a low dimensionality, that is, $p/n \to 0$, if λ_1 is bounded. In this paper, we consider the general regime of bounded $p/(n\lambda_1)$, which implies $\bar{r} = O(n)$ and allows diverging λ_1. More discussions will be given in Section 3.

A second line of efforts is through sparse PCA. According to Johnstone and Lu (2009), PCA does not generate consistent estimators for leading eigenvectors if $p/n \to \gamma \in (0, 1)$ with bounded eigenvalues. This motivates the development of sparse PCA, which leverages the extra assumption on the sparsity of eigenvectors. A large amount of literature has contributed to the topic of sparse PCA, for example, Amini and Wainwright (2008), Berthet and Rigollet (2013), Birnbaum et al. (2013), Ma (2013), Vu and Lei (2012). Specifically, Vu and Lei (2012) derived optimal bound for the minimax estimation error of the first sparse leading eigenvector, while Cai, Ma and Wu (2015) conducted a more thorough study on the minimax optimal rates for estimating top eigenvalues and eigenvectors of spiked covariance matrices with jointly k-sparse eigenvectors. This type of work typically assumes bounded eigenvalues, which ignore the contributions of the strong signals from the data in many real applications. To make the problem solvable, sparsity assumptions on the eigenvectors are imposed. In contrast, driven by applications such as genomics, economics and finance, this paper studies the contributions of the diverging eigenvalues (signals) to the estimation of their associated eigenvectors, without relying on sparsity assumptions on the eigenvectors.

In order to illustrate the third perspective, let us briefly review the approximate factor model [Bai (2003), Fan, Liao and Mincheva (2013)] and see how the spiked eigenvalues arise naturally from the model. Consider the following data generating model:

$$\mathbf{y}_t = \mathbf{B}\mathbf{f}_t + \boldsymbol{\varepsilon}_t \qquad \text{for } t = 1, \dots, T,$$

where $\mathbf{y}_t$ is a p-dimensional vector observed at time t, $\mathbf{f}_t \in \mathbb{R}^m$ is the vector of latent factors that drive the cross-sectional dependence at time t, $\mathbf{B}$ is the matrix of the corresponding factor loading coefficients, and $\boldsymbol{\varepsilon}_t$ is the idiosyncratic part that cannot be explained by the factors. Assume without loss of generality that $\operatorname{var}(\mathbf{f}_t) = \mathbf{I}_m$, the $m \times m$ identity matrix. Then the model implies $\boldsymbol{\Sigma} = \operatorname{var}(\mathbf{y}_t) = \mathbf{B}\mathbf{B}' + \boldsymbol{\Sigma}_\varepsilon$, where $\boldsymbol{\Sigma}_\varepsilon = \operatorname{var}(\boldsymbol{\varepsilon})$. It admits a low-rank plus sparse structure when $\boldsymbol{\Sigma}_\varepsilon$ is assumed to be sparse [Fan, Fan and Lv (2008), Fan, Liao and Mincheva (2013)]. The recovery of the low-rank and sparse matrices was considered thoroughly by Candès et al. (2011) and Chandrasekaran et al. (2011) under the incoherence condition in the noiseless setting and by Agarwal, Negahban and Wainwright (2012) in the noisy case. If the factor loadings $\{\mathbf{b}_i\}_{i \le p}$ (the transpose of rows of $\mathbf{B}$) are i.i.d. samples from a population with mean zero and covariance $\boldsymbol{\Sigma}_b$ (this is a pervasive assumption commonly used in the factor models [Fan, Liao and Mincheva (2013)]), then by the law of large numbers, $p^{-1}\mathbf{B}'\mathbf{B} = p^{-1}\sum_{i=1}^{p} \mathbf{b}_i\mathbf{b}_i' \to \boldsymbol{\Sigma}_b$, as $p \to \infty$. In other words, the eigenvalues of $\mathbf{B}\mathbf{B}'$ are approximately

$$p\lambda_1(\boldsymbol{\Sigma}_b)\bigl(1 + o(1)\bigr), \dots, p\lambda_m(\boldsymbol{\Sigma}_b)\bigl(1 + o(1)\bigr), 0, \dots, 0,$$

where $\lambda_j(\mathbf{\Sigma}_b)$ is the jth largest eigenvalue of $\mathbf{\Sigma}_b$. Then, by Weyl's theorem, we conclude that the eigenvalues of $\mathbf{\Sigma}$

$$\lambda_j = p\lambda_j(\mathbf{\Sigma}_b)\bigl(1+o(1)\bigr), \qquad \text{for } j=1,\dots,m, \tag{1.1}$$

and the remaining are bounded, if $\|\mathbf{\Sigma}_\varepsilon\|$ is bounded. Therefore, the factor model implies a spiked covariance with diverging leading eigenvalues. Fan, Liao and Mincheva (2013) showed that if the leading eigenvalues grow linearly with the dimension, then the corresponding eigenvectors can be consistently estimated as long as sample size goes to infinity. See Section 4 for more details.

Deviating from the classical random matrix and sparse PCA literature, we consider the high dimensional regime, allowing $p/n \to \infty$. To take into account the contributions of the signals for PCA, we also allow $\lambda_j \to \infty$ for the first m leading eigenvalues. This leads to the third perspective for understanding PCA from this high dimensional setting. Shen et al. (2016) adopted this point of view and considered the regime of $p/(n\lambda_j) \to \gamma_j$ where $0 \le \gamma_j < \infty$ for leading eigenvalues. This is more general than the bounded eigenvalue condition. Specifically if eigenvalues are bounded, we require the ratio p/n converges to a bounded constant as in the random matrix regime. On the other hand, if the dimension is much larger than the sample size, we offset the dimensionality by assuming increased signals or sample size, without additional sparse eigenvector assumption as in sparse PCA regime. In particular, as shown in (1.1), the strong (or pervasive) factors considered in financial applications corresponds to $\gamma_j = 0$ with the leading eigenvalues $\lambda_j \asymp p$; see, for example, Bai (2003), Bai and Ng (2002), Fan, Liao and Mincheva (2013), Fan, Liao and Wang (2016), Stock and Watson (2002). The weak or semistrong factors considered by De Mol, Giannone and Reichlin (2008) and Onatski (2012) also imply bounded $p/(n\lambda_1)$, with p/n bounded and $\lambda_j \asymp p^\theta$ for some $\theta \in [0, 1)$.

Hall, Marron and Neeman (2005), Jung and Marron (2009) started the research of high dimension low sample size (HDLSS) regime. With n fixed, Jung and Marron (2009) concluded that consistency of leading eigenvalues and eigenvectors is granted if $\lambda_j \asymp p^\theta$ for $\theta > 1$, which also corresponds to $\gamma_j = 0$. Shen et al. (2016) revealed an interesting fact that when $\gamma_j \neq 0$, spiked sample eigenvalues almost surely converge to a biased quantity of the true eigenvalues; furthermore, the corresponding sample eigenvectors show an asymptotic conical structure. However, their work focuses only on the consistency problem. In this study, we will consider the same regime as theirs, but focus more on the rates of convergence and the asymptotic distributions of the empirical eigenstructure, and under more relaxed conditions. Our results can be viewed as a natural extension of Paul (2007) to the high dimensional setting.

We would like to emphasize more on the scope and importance of our contributions here. First, the regime we consider in this paper is $p/(n\lambda_j) \to \gamma_j \in [0, \infty)$ for $j \le m$, which permits high dimensionality $p/n \to \infty$ and diverging eigenvalues without specifying their divergence rates. As we have argued, this encompasses many situations considered in the existing literature. It puts into the same framework of the typical random matrix regime with bounded eigenvalues and HDLSS analysis with fixed sample size. Second, the contributions of diverging eigenvalues are indeed recognized and accounted for in our theoretical developments. This avoids the restrictive assumptions on sparse eigenvectors. PCA without sparsity assumptions has been widely used in the diverging fields such as population association study [Yamaguchi-Kabata et al. (2008)], genome-wide association study [Ringnér (2008)], microarray data [Landgrebe, Wurst and Welzl (2002), Price et al. (2006)], fMRI data [Thomas, Harshman and Menon (2002)] and financial returns [Chamberlain and Rothschild (1983), Chen and Shimerda (1981)]. Our efforts contribute to theoretical understanding of why such a plain PCA works in these diverse fields. Finally, by allowing certain generality, we gain theoretical insights into how n, p and signal strength λ_j interplay.

The results are useful in two ways. On the one hand, they help quantify the biases of empirical eigenstructure and explain where they come from. Specifically, in Theorem 3.1, the bias of the jth sample eigenvalue ($j \le m$) is quantified by $p/(n\lambda_j)$, which is also showed by Yata and Aoshima (2012, 2013) under different assumptions of the spiked covariance model. Our novel contribution lies in Theorem 3.2, revealing the bias of the jth sample eigenvector ($j \le m$). In (3.7), we provide a decomposition of each empirical eigenvector into a spiked part, which converges to the true eigenvector with a deflation factor also quantified by $p/(n\lambda_j)$, and a nonspiked part, which creates a random bias distributed uniformly on an ellipse. More details will be presented in Section 3. On the other hand, the theoretical results provide new technical tools for analyzing factor models, which motivate the study. As we have seen, although it is natural to assume eigenvalues grow linearly with dimension, the assumption imposes a strong signal. Note that when $p/(n\lambda_j) \to 0$, no biases will occur. So in Section 4, we consider to relax the order of spikes to slightly faster than $\sqrt{p}$. By correcting the biases, we propose a new method called Shrinkage Principal Orthogonal complEment Thresholding (S-POET) and employ it to two applications: risk assessment of large portfolios [Fan, Liao and Shi (2015), Pesaran and Zaffaroni (2008)] and false discovery proportion estimation for dependent test statistics [Fan, Han and Gu (2012), Leek and Storey (2008)]. Existing methodologies for those two problems reply on rather strong signal level, but we are able to relax it with the help of S-POET.

The paper is organized as follows. Section 2 introduces the notation, assumptions, and an important fact which serves as basis of our proofs. Sections 3.1 and 3.2 devote to the theoretical results for the sample eigenvalues and eigenvectors of the spiked covariance matrix. In Section 4, we discuss several applications of the theories in the previous section. Simulations are conducted in Section 5 to demonstrate the theoretical results at the finite sample and the performance of S-POET. Section 6 provides concluding remarks. The proofs for Section 3 are provided in the Appendix and those for Section 4 are relegated to the supplementary material [Wang and Fan (2017)].

2. Assumptions and a simple fact. Assume that $\{\mathbf{Y}_i\}_{i=1}^n$ is a sequence of i.i.d. random variables with zero mean and covariance matrix $\mathbf{\Sigma}_{p\times p}$. Let $\lambda_1, \dots, \lambda_p$ be the eigenvalues of $\mathbf{\Sigma}$ in descending order. We consider the spiked covariance model as follows.

ASSUMPTION 2.1. $\lambda_1 > \lambda_2 > \cdots > \lambda_m > \lambda_{m+1} \geq \cdots \geq \lambda_p > 0$, where the nonspiked eigenvalues are bounded, that is, $c_0 \leq \lambda_j \leq C_0$, $j > m$ for constants $c_0, C_0 > 0$ and the spiked eigenvalues are well separated, that is, $\exists \delta_0 > 0$ such that $\min_{j\leq m}(\lambda_j - \lambda_{j+1})/\lambda_j \geq \delta_0$.

The eigenvalues are divided into the spiked ones and bounded nonspiked ones. We do not need to specify the order of the leading eigenvalues nor require them to diverge. Thus, our results in Section 3 are applicable to both bounded and diverging leading eigenvalues. For simplicity, we only consider distinguishable eigenvalues (multiplicity 1) for the largest m eigenvalues and a fixed number m, independent of n and p.

The factor model $\mathbf{y} = \mathbf{Bf} + \boldsymbol{\varepsilon}$ with pervasive factors considered in Fan, Liao and Mincheva (2013) implies a spiked covariance model with $\lambda_j \asymp p$ in (1.1) and satisfies the above assumption. For the interplay of the sample size n, dimension p and the spikes λ_j's, the following relationship is assumed as in Shen et al. (2016).

ASSUMPTION 2.2. Assume $p > n$. For the spiked part $1 \leq j \leq m$, $c_j = p/(n\lambda_j)$ is bounded, and for the nonspiked part, $(p-m)^{-1}\sum_{j=m+1}^p \lambda_j = \bar{c} + o(n^{-1/2})$.

We allow $p/n \to \infty$ in any manner, though λ_j also needs to grow fast enough to ensure bounded c_j. In particular, $c_j = o(1)$ is allowed as in the factor model. We do not assume the nonspiked eigenvalues are identical, as in most spiked covariance model literature [e.g., Johnstone and Lu (2009), Paul (2007)].

By spectral decomposition, $\mathbf{\Sigma} = \mathbf{\Gamma}\mathbf{\Lambda}\mathbf{\Gamma}'$, where the orthonormal matrix $\mathbf{\Gamma}$ is constructed by the eigenvectors of $\mathbf{\Sigma}$ and $\mathbf{\Lambda} = \operatorname{diag}(\lambda_1, \dots, \lambda_p)$. Let $\mathbf{X}_i = \mathbf{\Gamma}'\mathbf{Y}_i$. Since the empirical eigenvalues are invariant and the empirical eigenvectors are equivariant under an orthonormal transformation, we focus the analysis on the transformed domain of $\mathbf{X}_i$ and then translate the results into those of the original data. Note that $\operatorname{var}(\mathbf{X}_i) = \mathbf{\Lambda}$. Let $\mathbf{Z}_i = \mathbf{\Lambda}^{-1/2}\mathbf{X}_i$ be the elementwise standardized random vector.

ASSUMPTION 2.3. $\{\mathbf{Z}_i\}_{i=1}^n$ are i.i.d. copies of $\mathbf{Z}$. The standardized random vector $\mathbf{Z} = (Z_1, \dots, Z_p)$ is sub-Gaussian with independent entries of mean zero and variance one. The sub-Gaussian norms of all components are uniformly bounded: $\max_j \|Z_j\|_{\psi_2} \leq C_0$, where $\|Z_j\|_{\psi_2} = \sup_{q\geq 1} q^{-1/2}(E|Z_j|^q)^{1/q}$.

Since $\operatorname{Var}(\mathbf{X}_i) = \operatorname{diag}(\lambda_1, \lambda_2, \dots, \lambda_p)$, the first m population eigenvectors are simply unit vectors $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_m$ with only one nonvanishing element. Denote the n by p transformed data matrix by $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n)'$. Then the sample covariance matrix is

$$\widehat{\boldsymbol{\Sigma}}_{p\times p} = \frac{1}{n}\mathbf{X}'\mathbf{X} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i',$$

whose eigenvalues are denoted as $\widehat{\lambda}_1, \widehat{\lambda}_2, \dots, \widehat{\lambda}_p$ ($\widehat{\lambda}_j = 0$ for $j > n$) with corresponding eigenvectors $\widehat{\boldsymbol{\xi}}_1, \widehat{\boldsymbol{\xi}}_2, \dots, \widehat{\boldsymbol{\xi}}_p$. Note that the empirical eigenvectors of data $\mathbf{Y}_i$'s are $\widehat{\boldsymbol{\xi}}_j^{(Y)} = \boldsymbol{\Gamma}\widehat{\boldsymbol{\xi}}_j$.

Let $\mathbf{Z}_j$ be the jth column of the standardized $\mathbf{X}$. Then each $\mathbf{Z}_j$ has i.i.d. sub-Gaussian entries with zero mean and unit variance. Exchanging the roles of rows and columns, we get the n by n Gram matrix

$$\widetilde{\boldsymbol{\Sigma}}_{n\times n} = \frac{1}{n}\mathbf{X}\mathbf{X}' = \frac{1}{n}\sum_{j=1}^{p}\lambda_j\mathbf{Z}_j\mathbf{Z}_j',$$

with the same nonzero eigenvalues $\widehat{\lambda}_1, \widehat{\lambda}_2, \dots, \widehat{\lambda}_n$ as $\widehat{\boldsymbol{\Sigma}}$ and the corresponding eigenvectors $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_n$. It is well known that for $i = 1, 2, \dots, n$

$$\widehat{\boldsymbol{\xi}}_i = (n\widehat{\lambda}_i)^{-1/2}\mathbf{X}'\mathbf{u}_i \quad \text{and} \quad \mathbf{u}_i = (n\widehat{\lambda}_i)^{-1/2}\mathbf{X}\widehat{\boldsymbol{\xi}}_i, \tag{2.1}$$

while the other eigenvectors of $\widehat{\boldsymbol{\Sigma}}$ constitute a $(p-n)$-dimensional orthogonal complement of $\widehat{\boldsymbol{\xi}}_1, \dots, \widehat{\boldsymbol{\xi}}_n$.

By using this simple fact, for the specific case with $c_0 = C_0 = 1$ in Assumption 2.1, $\lambda_j = 1$ for $j > m$ in Assumption 2.2, and Gaussian data in Assumption 2.3, Shen et al. (2016) showed that

$$\frac{\widehat{\lambda}_j}{\lambda_j} \xrightarrow{\text{a.s.}} 1 + c_j, \qquad 1 \le j \le m,$$

and

$$\left|\langle\widehat{\boldsymbol{\xi}}_j, \mathbf{e}_j\rangle\right| \xrightarrow{\text{a.s.}} (1+c_j)^{-\frac{1}{2}},$$

where $\langle\mathbf{a}, \mathbf{b}\rangle$ denotes the inner product of two vectors. However, they fail to establish any results on convergence rates or asymptotic distributions of the empirical eigenstructure. This motivates the current paper.

The aim of this paper is to establish the asymptotic normality of the empirical eigenvalues and eigenvectors under more relaxed conditions. Our results are a natural extension of Paul (2007) to a more general setting with new insights, where the asymptotic normality of sample eigenvectors is derived using complicated random matrix techniques for Gaussian data under the regime of $p/n \to \gamma \in [0, 1)$. In comparison, our proof, based on the relationship (2.1), is much simpler and insightful for understanding the behavior of high dimensional PCA.

Here are some notation that we will use in the paper. For a general matrix $\mathbf{M}$, we denote its matrix entrywise max norm as $\|\mathbf{M}\|_{\max} = \max_{i,j}\{|M_{i,j}|\}$ and define the quantities $\|\mathbf{M}\| = \lambda_{\max}^{1/2}(\mathbf{M}'\mathbf{M})$, $\|\mathbf{M}\|_F = (\sum_{i,j} M_{i,j}^2)^{1/2}$, $\|\mathbf{M}\|_\infty = \max_i \sum_j |M_{i,j}|$ to be its spectral, Frobenius and induced ℓ_∞ norms. If $\mathbf{M}$ is symmetric, we define $\lambda_j(\mathbf{M})$ to be the jth largest eigenvalue of $\mathbf{M}$ and $\lambda_{\max}(\mathbf{M})$, $\lambda_{\min}(\mathbf{M})$ to be the maximal and minimal eigenvalues, respectively. We denote $\operatorname{tr}(\mathbf{M})$ as the trace of $\mathbf{M}$. For any vector $\mathbf{v}$, its ℓ_2 norm is represented by $\|\mathbf{v}\|$ while ℓ_1 norm is written as $\|\mathbf{v}\|_1$. We use $\operatorname{diag}(\mathbf{v})$ to denote the diagonal matrix with the same diagonal entries as $\mathbf{v}$. For two random vectors $\mathbf{a}$, $\mathbf{b}$ of the same length, we say $\mathbf{a} = \mathbf{b} + O_P(\delta)$ if $\|\mathbf{a} - \mathbf{b}\| = O_P(\delta)$ and $\mathbf{a} = \mathbf{b} + o_P(\delta)$ if $\|\mathbf{a} - \mathbf{b}\| = o_P(\delta)$. We denote $\mathbf{a} \overset{d}{\Rightarrow} \mathcal{L}$ for some distribution $\mathcal{L}$ if there exists $\mathbf{b} \sim \mathcal{L}$ such that $\mathbf{a} = \mathbf{b} + o_P(1)$. Throughout the paper, C is a generic constant that may differ from line to line.

3. Asymptotic behavior of empirical eigenstructure.

3.1. *Asymptotic normality of empirical eigenvalues.* Let us first study the behavior of the m leading empirical eigenvalues of $\widehat{\boldsymbol{\Sigma}}$. Denote by $\lambda_j(\mathbf{A})$ the jth largest eigenvalue of matrix $\mathbf{A}$ and recall that $\widehat{\lambda}_j = \lambda_j(\widehat{\boldsymbol{\Sigma}})$. We have the following asymptotic normality of $\widehat{\lambda}_j$.

THEOREM 3.1. *Under Assumptions 2.1–2.3, $\{\widehat{\lambda}_j\}_{j=1}^m$'s have independent limiting distributions. In addition,*

$$\sqrt{n}\left\{\frac{\widehat{\lambda}_j}{\lambda_j} - \left(1 + \bar{c}c_j + O_P\left(\lambda_j^{-1}\sqrt{p/n}\right)\right)\right\} \overset{d}{\Rightarrow} N(0, \kappa_j - 1), \tag{3.1}$$

where κ_j is the kurtosis of X_j.

The theorem shows that the bias of $\widehat{\lambda}_j/\lambda_j$ is $\bar{c}c_j + O_P(\lambda_j^{-1}\sqrt{p/n})$. The second term is dominated by the first term since $p > n$ and it is of order $o_P(n^{-1/2})$ if $\sqrt{p} = o(\lambda_j)$. The latter assumption is satisfied by the strong factor model in Fan, Liao and Mincheva (2013) and a part of weak or semistrong factor model in Onatski (2012). The theorem reveals the bias is controlled by a term of rate $p/(n\lambda_j)$. To get the asymptotically unbiased estimate, it requires $c_j = p/(n\lambda_j) \to 0$ for $j \le m$. This result is more general than that of Shen et al. (2016) and sheds a similar light to that of Koltchinskii and Lounici (2017, 2014), that is, $\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|/\|\boldsymbol{\Sigma}\| \to 0$ almost surely if and only if the effective rank $\bar{r} = \operatorname{tr}(\boldsymbol{\Sigma})/\lambda_1$ is of order $o(n)$, which is true when $c_1 = o(1)$. Our result here holds for each individual spike. Yata and Aoshima (2012, 2013) employed a similar technical trick and gave a comprehensive study on the asymptotic consistency and distributions of the eigenvalues. They got similar results under different conditions from ours. Our framework is more general here. If $c_j \nrightarrow 0$, bias reduction can also be made; see Section 4.2, where an estimator for $\bar{c}$ is proposed. Under the bounded spiked covariance model considered in Baik, Ben Arous and Péché (2005), Johnstone and Lu (2009) and Paul (2007), it is assumed $\lambda_j = c_0 = C_0$, $j > m$ so that $\bar{c} = c_0$, the minimum eigenvalue of the population covariance matrix. Our result is also consistent with Anderson's (1963) result that

$$\sqrt{n}(\widehat{\lambda}_j - \lambda_j) \overset{d}{\Rightarrow} N(0, 2\lambda_j^2),$$

for Gaussian data and fixed p and λ_j's, where the nonspiked part does not exist, and thus the bias $O_P(\lambda_j^{-1}\sqrt{p/n})$ disappears. The proof is relegated to the Appendix.

3.2. *Behavior of empirical eigenvectors.* Let us consider the asymptotic distribution of the empirical eigenvectors $\widehat{\boldsymbol{\xi}}_j$'s corresponding to $\widehat{\lambda}_j$, $j = 1, 2, \dots, m$. As in Paul (2007), each $\widehat{\boldsymbol{\xi}}_j$ is divided into two parts corresponding to the spiked and nonspiked components, that is, $\widehat{\boldsymbol{\xi}}_j = (\widehat{\boldsymbol{\xi}}_{jA}', \widehat{\boldsymbol{\xi}}_{jB}')'$ where $\widehat{\boldsymbol{\xi}}_{jA}$ is of length m.

THEOREM 3.2. *Under Assumptions 2.1–2.3, we have*:

(i) *For the spiked part, if* $m = 1$,

$$\frac{2(1+\bar{c}c_1)}{\bar{c}c_1}\sqrt{n}\left(\sqrt{1+\bar{c}c_1}\widehat{\xi}_{1A} - 1 + O_P\left(\sqrt{\frac{p}{n\lambda_1^2}}\right)\right) \overset{d}{\Rightarrow} N(0, \kappa_1 - 1), \tag{3.2}$$

while if $m > 1$,

$$\sqrt{n}\left(\frac{\widehat{\boldsymbol{\xi}}_{jA}}{\|\widehat{\boldsymbol{\xi}}_{jA}\|} - \mathbf{e}_{jA} + O_P\left(\sqrt{\frac{p}{n\lambda_j^2}}\right)\right) \overset{d}{\Rightarrow} N_m(\mathbf{0}, \mathbf{\Sigma_j}), \tag{3.3}$$

for $j = 1, 2, \dots, m$, *with*

$$\mathbf{\Sigma}_j = \sum_{k \in [m] \backslash j} a_{jk}^2 \mathbf{e}_{kA}\mathbf{e}_{kA}',$$

where $[m] = \{1, \dots, m\}$, $\mathbf{e}_{kA}$ *is the first* m *elements of the unit vector* $\mathbf{e}_k$, *and* $a_{jk} = \lim_{\lambda_j,\lambda_k} \sqrt{\lambda_j\lambda_k}/(\lambda_j - \lambda_k)$, *which is assumed to exist.*

(ii) *For the nonspiked part, if we further assume the data is Gaussian, there exists* $p - m$ *dimensional vector* $\mathbf{h}_0 \sim \mathrm{Unif}(B_{p-m}(1))$ *such that*

$$\left\|\mathbf{D}_0\frac{\widehat{\boldsymbol{\xi}}_{jB}}{\|\widehat{\boldsymbol{\xi}}_{jB}\|} - \mathbf{h}_0\right\| = O_P\left(\sqrt{\frac{n}{p}}\right) + o_P\left(\frac{1}{\sqrt{n}}\right), \tag{3.4}$$

where $\mathbf{D}_0 = \mathrm{diag}(\sqrt{\bar{c}/\lambda_{m+1}}, \dots, \sqrt{\bar{c}/\lambda_p})$ *is a diagonal matrix and* $\mathrm{Unif}(B_k(r))$ *denotes the uniform distribution over the centered sphere of radius* r. *In addition, the max norm of* $\widehat{\boldsymbol{\xi}}_{jB}$ *satisfies*

$$\|\widehat{\boldsymbol{\xi}}_{jB}\|_{\max} = O_P\big(p/(n\lambda_j^{3/2}) + \sqrt{\log p/(n\lambda_j)}\big). \tag{3.5}$$

(iii) *Furthermore*, $\|\widehat{\boldsymbol{\xi}}_{jA}\| = (1+\bar{c}c_j)^{-1/2} + O_P(\lambda_j^{-1}\sqrt{p/n} + p/(n^{3/2}\lambda_j))$ *and* $\|\widehat{\boldsymbol{\xi}}_{jB}\| = (\frac{\bar{c}c_j}{1+\bar{c}c_j})^{1/2} + O_P(\sqrt{1/\lambda_j} + \sqrt{p/(n^2\lambda_j)})$. *Together with* (i), *this implies the inner product between empirical eigenvector and the population one converges to* $(1+\bar{c}c_j)^{-1/2}$ *in probability and*

$$\langle \widehat{\boldsymbol{\xi}}_j, \mathbf{e}_j \rangle - \frac{1}{\sqrt{1+\bar{c}c_j}} = O_P\big(\lambda_j^{-1}\sqrt{p/n} + p/\big(n^{3/2}\lambda_j\big)\big) + O_P\big(n^{-1}\big)I_{\{m>1\}}. \tag{3.6}$$

In the above theory, we assume $a_{jk} = \lim_{\lambda_j,\lambda_k} \frac{\sqrt{\lambda_j\lambda_k}}{\lambda_j-\lambda_k}$ exists. This is not restrictive if eigenvalues are well separated, that is, $\min_{j\neq k\leq m}|\lambda_j - \lambda_k|/\lambda_j \geq \delta_0$ from Assumption 2.1. The assumption obviously holds for the pervasive factor model, in which $a_{jk} = \sqrt{\lambda_j(\boldsymbol{\Sigma}_b)\lambda_k(\boldsymbol{\Sigma}_b)}/(\lambda_k(\boldsymbol{\Sigma}_b) - \lambda_j(\boldsymbol{\Sigma}_b))$.

Theorem 3.2 is an extension of random matrix results into high dimensional regime. Its proof sheds light on how to use the smaller $n\times n$ matrix $\widetilde{\boldsymbol{\Sigma}}$ as a tool to understand the behavior of the larger $p\times p$ covariance matrix $\widehat{\boldsymbol{\Sigma}}$. Specifically, we start from $\widetilde{\boldsymbol{\Sigma}}\mathbf{u}_j = \widehat{\lambda}_j\mathbf{u}_j$ or identity (A.3) and then use the simple fact (2.1) to get a relationship (A.4) of eigenvector $\widehat{\boldsymbol{\xi}}_j$. Then (A.4) is rearranged as (A.6) which gives a clear separation of the dominating term, that is asymptotically normal, and the error term. This makes the whole proof much simpler in comparison with Paul (2007) who showed a similar type of result through a complicated representation of $\widehat{\boldsymbol{\xi}}_j$ and $\widehat{\lambda}_j$ under more restricted assumptions. From this simple trick, we can understand deeply how some important high and low dimensional quantities link together and differ from each other.

Several remarks are in order. First, since $\widehat{\boldsymbol{\xi}}_j^{(\mathbf{Y})} = \boldsymbol{\Gamma}\widehat{\boldsymbol{\xi}}_j$ is the jth empirical eigenvector based on observed data $\mathbf{Y}$, we have decomposition

$$\widehat{\boldsymbol{\xi}}_j^{(\mathbf{Y})} = \boldsymbol{\Gamma}_A\widehat{\boldsymbol{\xi}}_{jA} + \boldsymbol{\Gamma}_B\widehat{\boldsymbol{\xi}}_{jB}, \tag{3.7}$$

where $\boldsymbol{\Gamma} = (\boldsymbol{\Gamma}_A, \boldsymbol{\Gamma}_B)$. Note that $\boldsymbol{\Gamma}_A\widehat{\boldsymbol{\xi}}_{jA}$ converges to the true eigenvector deflated by a factor of $\sqrt{1+\bar{c}c_j}$ with the convergence rate $O_P(\sqrt{p/(n\lambda_j^2)} + p/(n^{3/2}\lambda_j) + n^{-1/2})$ while $\boldsymbol{\Gamma}_B\widehat{\boldsymbol{\xi}}_{jB}$ creates a random bias, which is distributed uniformly on an ellipse of $(p-m)$ dimension and projected into the p dimensional space spanned by $\boldsymbol{\Gamma}_B$. The two parts intertwine in such a way that correction for the biases of estimating eigenvectors is almost impossible. More details are discussed in Section 4 for the factor models. Second, it is clearly as in the eigenvalue case, the bias term $\lambda_j^{-1}\sqrt{p/n}$ in (i) of Theorem 3.2 disappears when $\sqrt{p} = o(\lambda_j)$. In particular, for the stronger factor given by (1.1), $\widehat{\boldsymbol{\xi}}_j^{(\mathbf{Y})}$ is a consistent estimator. Third, the situations $m=1$ and $m>1$ have slight difference in that multiple spikes could interact with each other. Especially this is reflected in the convergence of the angle between the empirical eigenvector and its population counterpart: the angle converges to $(1+\bar{c}c_j)^{-1/2}$ with an extra rate $O_P(1/n)$ which stems from estimating

$\widehat{\xi}_{jk}$ for $j \neq k \leq m$ [see proof of Theorem 3.2(iii)]. The difference will only be seen when the spike magnitude is higher than the order $\sqrt{pn} \vee pn^{-1/2}$. We will verify this by a simple simulation in Section 5. Finally, it is the first time that the max norm bound of the nonspiked part is derived. This bound will be useful for analyzing factor models in Section 4.

Theorem 3.2 again implies the results of Shen et al. (2016). It also generalizes the asymptotic distribution of nonspiked part from pure orthogonal invariant case of Paul (2007) to a more general setting. In particular, when $p/n \to \infty$, the asymptotic distribution of the normalized nonspiked component is not uniform over a sphere any more, but over an ellipse. In addition, our result can be compared with the low dimensional case, where Anderson (1963) showed that

$$\sqrt{n}(\widehat{\boldsymbol{\xi}}_j - \mathbf{e}_j) \overset{d}{\Rightarrow} N_p\left(\mathbf{0}, \sum_{k \in [m] \setminus j} \frac{\lambda_j \lambda_k}{(\lambda_j - \lambda_k)^2} \mathbf{e}_k \mathbf{e}_k'\right), \tag{3.8}$$

for fixed p and λ_j's. Under our assumptions, since the spiked eigenvalues may go to infinity, the constants in the asymptotic covariance matrix are replaced by the limits a_{jk}'s. Similar to the behavior of eigenvalues, the spiked part $\widehat{\boldsymbol{\xi}}_{jA}$ preserves the normality property except for a bias factor $1/(1 + \bar{c}c_j)$ caused by the high dimensionality. Also, recent work of Koltchinskii and Lounici (2014) provides general asymptotic results for the empirical eigenvectors from a spectral projector point of view, but they mainly focus on the regime of $p/n\lambda_j \to 0$ or $\bar{r} = o(n)$. Last but not least, it has been shown by Johnstone and Lu (2009) that PCA generates consistent eigenvector estimation if and only if $p/n \to 0$ when the spike sizes are fixed. This motivates the study of sparse PCA. We take the spike magnitude into account and provide additional insights by showing that PCA consistently estimate eigenvalues and eigenvectors if and only if $p/(n\lambda_j) \to 0$. This explains why Fan, Liao and Mincheva (2013) can consistently estimate the eigenvalues and eigenvectors while Johnstone and Lu (2009) cannot.

4. Applications to factor models.

In this section, we propose a method named Shrinkage Principal Orthogonal complEment Thresholding (S-POET) for estimating large covariance matrices induced by the approximate factor models. The estimator is based on correction of the bias of the empirical eigenvalues as specified in (3.1). We derive for the first time the bound for the relative estimation errors of covariance matrices under the spectral norm. The results are then applied to assessing large portfolio risks and estimating false discovery proportions, where the conditions in existing literature are significantly relaxed.

4.1. *Approximate factor models.*

Factor models have been widely used in various disciplines. For example, it is used to extract information from financial market for sufficient forecasting of other time series [Fan, Xue and Yao (2015), Stock and Watson (2002)] and to adjust heterogeneity for biological data aggregation of multiple sources [Fan et al. (2016), Leek et al. (2010)]. Consider the approximate factor model

$$y_{it} = \mathbf{b}_i'\mathbf{f}_t + u_{it}, \tag{4.1}$$

where y_{it} is the observed data for the ith ($i = 1, \dots, p$) individual (e.g., returns of stocks) or component (e.g., expressions of genes) at time $t = 1, \dots, T$; $\mathbf{f}_t$ is an $m \times 1$ vector of latent common factors and $\mathbf{b}_i$ is the factor loadings for the ith individual or component; u_{it} is the idiosyncratic error, uncorrelated with the common factors. In genomics application, t can also index repeated experiments. For simplicity, we assume there is no time dependency.

The factor model can be written into a matrix form as follows:

$$\mathbf{Y} = \mathbf{B}\mathbf{F}' + \mathbf{U}, \tag{4.2}$$

where $\mathbf{Y}_{p\times T}, \mathbf{B}_{p\times m}, \mathbf{F}_{T\times m}, \mathbf{U}_{p\times T}$ are respectively the matrix form of the observed data, the factor loading matrix, the factor matrix, and the error matrix. For identifiability, we impose the condition that $\operatorname{cov}(\mathbf{f}_t) = \mathbf{I}$. Thus, the covariance matrix is given by

$$\mathbf{\Sigma} = \mathbf{B}\mathbf{B}' + \mathbf{\Sigma}_u, \tag{4.3}$$

where $\mathbf{\Sigma}_u$ is the covariance matrix of the idiosyncratic error at any time t.

Under the assumption that $\mathbf{\Sigma}_u = (\sigma_{u,ij})_{i,j\le p}$ is sparse with its eigenvalues bounded away from zero and infinity, the population covariance exhibits a low-rank plus sparse structure. The sparsity is measured by the following quantity:

$$m_p = \max_{i\le p} \sum_{j\le p} |\sigma_{u,ij}|^q,$$

for some $q \in [0, 1]$ [Bickel and Levina (2008)]. In particular, with $q = 0$, m_p equals the maximum number of nonzero elements in each row of $\mathbf{\Sigma}_u$.

In order to estimate the true covariance matrix with the above factor structure, Fan, Liao and Mincheva (2013) proposed a method called "POET" to recover the unknown factor matrix as well as the factor loadings. The idea is simply to first decompose the sample covariance matrix into the spiked and nonspiked part and estimate them separately. Specifically, define $\widehat{\mathbf{\Sigma}} = T^{-1}\mathbf{Y}\mathbf{Y}'$ using the observed data and let $\{\widehat{\lambda}_j\}$ and $\{\widehat{\boldsymbol{\xi}}_j\}$ be its corresponding eigenvalues and eigenvectors. They define

$$\widehat{\mathbf{\Sigma}}^{\top} = \sum_{j=1}^{m} \widehat{\lambda}_j \widehat{\boldsymbol{\xi}}_j \widehat{\boldsymbol{\xi}}_j' + \widehat{\mathbf{\Sigma}}_u^{\top}, \tag{4.4}$$

where $\widehat{\mathbf{\Sigma}}_u^{\top}$ is the matrix after applying thresholding method [Bickel and Levina (2008)] to $\widehat{\mathbf{\Sigma}}_u = \widehat{\mathbf{\Sigma}} - \sum_{j=1}^{m} \widehat{\lambda}_j \widehat{\boldsymbol{\xi}}_j \widehat{\boldsymbol{\xi}}_j'$.

They showed that the above estimation procedure is equivalent to the least square approach that minimizes

$$(\widehat{\mathbf{B}}, \widehat{\mathbf{F}}) = \arg\min_{\mathbf{B},\mathbf{F}} \|\mathbf{Y} - \mathbf{B}\mathbf{F}'\|_F^2 \qquad \text{s.t. } \frac{1}{T}\mathbf{F}'\mathbf{F} = \mathbf{I}_m, \mathbf{B}'\mathbf{B} \text{ is diagonal.} \tag{4.5}$$

The columns of $\widehat{\mathbf{F}}/\sqrt{T}$ are the eigenvectors corresponding to the m largest eigenvalues of the $T \times T$ matrix $T^{-1}\mathbf{Y}'\mathbf{Y}$ and $\widehat{\mathbf{B}} = T^{-1}\mathbf{Y}\widehat{\mathbf{F}}$. After $\mathbf{B}$ and $\mathbf{F}$ are estimated, the sample covariance of $\widehat{\mathbf{U}} = \mathbf{Y} - \widehat{\mathbf{B}}\widehat{\mathbf{F}}'$ can be formed: $\widehat{\mathbf{\Sigma}}_u = T^{-1}\widehat{\mathbf{U}}\widehat{\mathbf{U}}'$. Finally, thresholding is applied to $\widehat{\mathbf{\Sigma}}_u$ to generate $\widehat{\mathbf{\Sigma}}_u^{\top} = (\widehat{\sigma}_{u,ij}^{\top})_{p\times p}$, where

$$\widehat{\sigma}_{u,ij}^{\top} = \begin{cases} \widehat{\sigma}_{u,ij}, & i = j, \\ s_{ij}(\widehat{\sigma}_{u,ij}) I(|\widehat{\sigma}_{u,ij}| \geq \tau_{ij}), & i \neq j. \end{cases} \tag{4.6}$$

Here, $s_{ij}(\cdot)$ is the generalized shrinkage function [Antoniadis and Fan (2001), Rothman, Levina and Zhu (2009)] and $\tau_{ij} = \tau(\widehat{\sigma}_{u,ii}\widehat{\sigma}_{u,jj})^{1/2}$ is the entry-dependent threshold. The above adaptive threshold corresponds to applying thresholding with parameter τ to the correlation matrix of $\widehat{\mathbf{\Sigma}}_u$. The positive parameter τ will be determined later.

Fan, Liao and Mincheva (2013) showed that under Assumptions B.1–B.4 listed in Appendix B in the supplementary material [Wang and Fan (2017)],

$$\|\widehat{\mathbf{\Sigma}}^{\top} - \mathbf{\Sigma}\|_{\Sigma,F} = O_P\left(\frac{\sqrt{p}\log p}{T} + m_p\left(\frac{\log p}{T} + \frac{1}{p}\right)^{(1-q)/2}\right), \tag{4.7}$$

where $\|\mathbf{A}\|_{\Sigma,F} = p^{-1/2}\|\mathbf{\Sigma}^{-1/2}\mathbf{A}\mathbf{\Sigma}^{-1/2}\|_F$ and $\|\cdot\|_F$ is the Frobenius norm. Note that

$$\|\widehat{\mathbf{\Sigma}}^{\top} - \mathbf{\Sigma}\|_{\Sigma,F} = p^{-1/2}\|\mathbf{\Sigma}^{-1/2}\widehat{\mathbf{\Sigma}}^{\top}\mathbf{\Sigma}^{-1/2} - \mathbf{I}_p\|_F,$$

which measures the relative error in Frobenius norm. A more natural metric is relative error under the spectral norm $\|\mathbf{A}\|_{\Sigma} = \|\mathbf{\Sigma}^{-1/2}\mathbf{A}\mathbf{\Sigma}^{-1/2}\|$, which cannot be obtained by using the technical device of Fan, Liao and Mincheva (2013). Note $\|\mathbf{A}\|_{\Sigma,F} \leq \|\mathbf{A}\|_{\Sigma}$. Via our new results in the last section, we will establish a result under those two relative norms, under weaker conditions than their pervasiveness assumption. Note that the relative error convergence is particularly meaningful for spiked covariance matrix, as eigenvalues are in different scales.

4.2. *Shrinkage POET under relative spectral norm.* The discussion above reveals several drawbacks of POET. First, the spike size has to be of order p which rules out relatively weak factors. Second, it is well known that the empirical eigenvalues are inconsistent if the spiked eigenvalues do not significantly dominate the nonspiked part. Therefore, a proper correction or shrinkage is needed. See Donoho, Gavish and Johnstone (2014) for optimal shrinkage of empirical eigenvalues.

Regarding to the first drawback, we relax the assumption $\|p^{-1}\mathbf{B}'\mathbf{B} - \mathbf{\Omega}_0\| = o(1)$ in Assumption B.1 to the following weaker assumption.

ASSUMPTION 4.1. $\|\mathbf{\Lambda}_A^{-1/2}\mathbf{B}'\mathbf{B}\mathbf{\Lambda}_A^{-1/2} - \mathbf{\Omega}_0\| = o(1)$ for some $\mathbf{\Omega}_0$ with eigenvalues bounded from above and below, where $\mathbf{\Lambda}_A = \operatorname{diag}(\lambda_1, \dots, \lambda_m)$. In addition, we assume $\lambda_m \to \infty$, λ_1/λ_m is bounded from above and below.

This assumption does not require the first m eigenvalues of $\mathbf{\Sigma}$ to take on any specific rate. They can still be much smaller than p, although for simplicity we require them to diverge and share the same diverging rate. Since $\|\mathbf{\Sigma}_u\|$ is assumed to be bounded, the assumption $\lambda_m \to \infty$ is also imposed to avoid the issue of identifiability. When λ_m does not diverge, more sophisticated condition is needed for identifiability [Chandrasekaran et al. (2011)].

In order to handle the second drawback, we propose the Shrinkage POET (S-POET) method. Inspired by (3.1), the shrinkage POET modifies the first part in POET estimator (4.4) as follows:

$$\widehat{\mathbf{\Sigma}}^S = \sum_{j=1}^m \widehat{\lambda}_j^S \widehat{\boldsymbol{\xi}}_j \widehat{\boldsymbol{\xi}}_j' + \widehat{\mathbf{\Sigma}}_u^\top, \tag{4.8}$$

where $\widehat{\lambda}_j^S = \max\{\widehat{\lambda}_j - \bar{c}p/n, 0\}$, a simple soft thresholding correction. Obviously if $\widehat{\lambda}_j$ is sufficiently large, $\widehat{\lambda}_j^S/\lambda_j = \widehat{\lambda}_j/\lambda_j - \bar{c}c_j = 1 + o_P(1)$. Since $\bar{c}$ is unknown, a natural estimator $\widehat{c}$ is such that the total of the eigenvalues remains unchanged:

$$\operatorname{tr}(\widehat{\mathbf{\Sigma}}) = \sum_{j=1}^m (\widehat{\lambda}_j - \widehat{c}p/n) + (p-m)\widehat{c}$$

or $\widehat{c} = (\operatorname{tr}(\widehat{\mathbf{\Sigma}}) - \sum_{j=1}^m \widehat{\lambda}_j)/(p - m - pm/n)$. It has been shown by Lemma 7 of Yata and Aoshima (2012) that

$$(\widehat{c} - \bar{c})\frac{p}{n\lambda_j} = O_P\left(\frac{\operatorname{tr}(\widehat{\mathbf{\Sigma}}) - \sum_{j=1}^m \widehat{\lambda}_j}{(n-m)\lambda_m} - \frac{\bar{c}p}{n\lambda_m}\right) = O_P(n^{-1}).$$

Thus, replacing $\bar{c}$ by $\widehat{c}$, we have $\widehat{\lambda}_j^S/\lambda_j - 1 = O_P(\lambda_j^{-1}\sqrt{p/n} + n^{-1/2})$, that is, the estimation error in $\widehat{c}$ is negligible. From Lemma 3.1, we can easily obtain the asymptotic normality, $\sqrt{n}(\widehat{\lambda}_j^S/\lambda_j - 1) \overset{d}{\Rightarrow} N(0, \kappa_j - 1)$ if $\sqrt{p} = o(\lambda_j)$.

To get the convergence of relative errors under the spectral norm, we also need the following additional assumptions.

ASSUMPTION 4.2. (i) $\{\mathbf{u}_t, \mathbf{f}_t\}_{t\geq 1}$ are independently and identically distributed with $\mathbb{E}[u_{it}] = \mathbb{E}[u_{it}f_{jt}] = 0$ for all $i \leq p$, $j \leq m$ and $t \leq T$.

(ii) There exist positive constants c_1 and c_2 such that $\lambda_{\min}(\mathbf{\Sigma}_u) > c_1$, $\|\mathbf{\Sigma}_u\|_\infty < c_2$, and $\min_{i,j} \operatorname{Var}(u_{it}u_{jt}) > c_1$.

(iii) There exist positive constants r_1, r_2, b_1 and b_2 such that for $s > 0, i \leq p, j \leq m$,

$$\mathbb{P}(|u_{it}| > s) \leq \exp(-(s/b_1)^{r_1}) \quad \text{and} \quad \mathbb{P}(|f_{jt}| > s) \leq \exp(-(s/b_2)^{r_2}).$$

(iv) There exists $M > 0$ such that for all $i \le p$, $j \le m$, $|b_{ij}| \le M\sqrt{\lambda_j/p}$.

(v) $\sqrt{p}(\log T)^{1/r_2} = o(\lambda_m)$.

The first three conditions are common in factor model literature. If we write $\mathbf{B} = (\widetilde{\mathbf{b}}_1, \dots, \widetilde{\mathbf{b}}_m)$, by Weyl's inequality we have $\max_{1\le j\le m} \|\widetilde{\mathbf{b}}_j\|^2/\lambda_j \le 1 + \|\boldsymbol{\Sigma}_u\|/\lambda_j = 1 + o(1)$. Thus, it is reasonable to assume the magnitude $|b_{ij}|$ of factor loadings is of order $\sqrt{\lambda_j/p}$ in the fourth condition. The last condition is imposed to ease technical presentation.

Now we are ready to investigate $\|\widehat{\boldsymbol{\Sigma}}^S - \boldsymbol{\Sigma}\|_\Sigma$. Suppose the SVD decomposition of $\boldsymbol{\Sigma}$,

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Gamma}_{p\times m} & \boldsymbol{\Omega}_{p\times(p-m)} \end{pmatrix} \begin{pmatrix} \boldsymbol{\Lambda}_{m\times m} & \\ & \boldsymbol{\Theta}_{(p-m)\times(p-m)} \end{pmatrix} \begin{pmatrix} \boldsymbol{\Gamma}' \\ \boldsymbol{\Omega}' \end{pmatrix}.$$

Then obviously

$$\begin{aligned} &\|\widehat{\boldsymbol{\Sigma}}^S - \boldsymbol{\Sigma}\|_\Sigma \\ &\quad\le \|\boldsymbol{\Sigma}^{-\frac{1}{2}}(\widehat{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\Lambda}}^S\widehat{\boldsymbol{\Gamma}}' - \mathbf{B}\mathbf{B}')\boldsymbol{\Sigma}^{-\frac{1}{2}}\| + \|\boldsymbol{\Sigma}^{-\frac{1}{2}}(\widehat{\boldsymbol{\Sigma}}_u^\top - \boldsymbol{\Sigma}_u)\boldsymbol{\Sigma}^{-\frac{1}{2}}\| \\ &\quad =: \Delta_L + \Delta_S \end{aligned} \tag{4.9}$$

and

$$\Delta_S \le \|\boldsymbol{\Sigma}^{-1}\|\|\widehat{\boldsymbol{\Sigma}}_u^\top - \boldsymbol{\Sigma}_u\| \le C\|\widehat{\boldsymbol{\Sigma}}_u^\top - \boldsymbol{\Sigma}_u\|. \tag{4.10}$$

It can be shown

$$\begin{aligned} \Delta_L &= \left\| \begin{pmatrix} \boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{\Gamma}' \\ \boldsymbol{\Theta}^{-\frac{1}{2}}\boldsymbol{\Omega}' \end{pmatrix} (\widehat{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\Lambda}}^S\widehat{\boldsymbol{\Gamma}}' - \mathbf{B}\mathbf{B}') \begin{pmatrix} \boldsymbol{\Gamma}\boldsymbol{\Lambda}^{-\frac{1}{2}} & \boldsymbol{\Omega}\boldsymbol{\Theta}^{-\frac{1}{2}} \end{pmatrix} \right\| \\ &\le \Delta_{L1} + \Delta_{L2} + 2\Delta_{L3}, \end{aligned} \tag{4.11}$$

where $\Delta_{L1} = \|\boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{\Gamma}'(\widehat{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\Lambda}}^S\widehat{\boldsymbol{\Gamma}}' - \mathbf{B}\mathbf{B}')\boldsymbol{\Gamma}\boldsymbol{\Lambda}^{-\frac{1}{2}}\|$, $\Delta_{L2} = \|\boldsymbol{\Theta}^{-\frac{1}{2}}\boldsymbol{\Omega}'(\widehat{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\Lambda}}^S\widehat{\boldsymbol{\Gamma}}' - \mathbf{B}\mathbf{B}')\boldsymbol{\Omega}\boldsymbol{\Theta}^{-\frac{1}{2}}\|$ and $\Delta_{L3} = \|\boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{\Gamma}'(\widehat{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\Lambda}}^S\widehat{\boldsymbol{\Gamma}}' - \mathbf{B}\mathbf{B}')\boldsymbol{\Omega}\boldsymbol{\Theta}^{-\frac{1}{2}}\|$. Thus, in order to find the convergence rate of relative spectral norm, we need to consider the terms Δ_{L1}, Δ_{L2}, Δ_{L3} and Δ_S separately. Notice that Δ_{L1} measures the relative error of the estimated spiked eigenvalues, Δ_{L2} reflects the goodness of the estimated eigenvectors, Δ_{L3} is the cross term and Δ_S controls the error of estimating the sparse idiosyncratic covariance matrix. To bound the relative Frobenius norm $\|\widehat{\boldsymbol{\Sigma}}^S - \boldsymbol{\Sigma}\|_{\Sigma,F}$, we define similar quantities $\widetilde{\Delta}_{L1}$, $\widetilde{\Delta}_{L2}$, $\widetilde{\Delta}_{L3}$, $\widetilde{\Delta}_S$ which replace the spectral norm by Frobenius norm multiplied by $p^{-1/2}$. Note that (4.9)–(4.11) also hold for relative Frobenius norm with $\widetilde{\Delta}_{L1}$, $\widetilde{\Delta}_{L2}$, $\widetilde{\Delta}_{L3}$, $\widetilde{\Delta}_S$. The following theorem reveals the rate of each term. Its proof will be provided in Appendix C of the supplementary material [Wang and Fan (2017)].

THEOREM 4.1. *Under Assumptions 2.1, 2.2, 2.3, 4.1 and 4.2, if* $p\log p > \max\{T(\log T)^{4/r_2}, T(\log(pT))^{2/r_1}\}$, *we have*

$$\widetilde{\Delta}_{L1} \le \Delta_{L1} = O_P(T^{-1/2}),$$

$$\Delta_{L2} + \Delta_{L3} = O_P\left(\frac{p}{T}\right), \qquad \widetilde{\Delta}_{L2} + \widetilde{\Delta}_{L3} = O_P\left(\frac{\sqrt{p}}{T}\right),$$

and by the applying adaptive thresholding estimator (4.6) *with*

$$\tau_{ij} = C\omega_T(\widehat{\sigma}_{u,ii}\widehat{\sigma}_{u,jj})^{1/2}, \quad \text{and} \quad \omega_T = \sqrt{\log p/T} + \sqrt{1/p},$$

we have

$$\widetilde{\Delta}_S \le \Delta_S = O_P(m_p\omega_T^{1-q}).$$

Combining the three terms, $\|\widehat{\mathbf{\Sigma}}^S - \mathbf{\Sigma}\|_{\Sigma} = O_P(p/T + m_p\omega_T^{1-q})$ *and* $\|\widehat{\mathbf{\Sigma}}^S - \mathbf{\Sigma}\|_{\Sigma,F} = O_P(\sqrt{p}/T + m_p\omega_T^{1-q})$.

The relative error convergence characterizes the accuracy of estimation for the spiked covariance matrix. Comparing the rate with (4.7), we are able to improve the relative Frobenius error rate by a $\log p$ term, thanks to the new asymptotic results. Additionally, this is the first time that the relative rate under spectral norm is derived. As long as λ_m grows slightly faster than $\sqrt{p}$, we obtain the same convergence rate for relative Frobenius norm as in the literature. Therefore, we conclude S-POET is effective even under a much weaker signal level.

4.3. *Portfolio risk management.* The risk of a given portfolio with allocation weight $\mathbf{w}$ is conventionally measured by its variance $\mathbf{w}'\mathbf{\Sigma}\mathbf{w}$, where $\mathbf{\Sigma}$ is the volatility (covariance) matrix of the returns of underlying assets. To evaluate portfolio risks, it is necessary to estimate a large covariance matrix $\mathbf{\Sigma}$ and factor models are frequently used to reduce the dimensionality. This was the idea of Fan, Liao and Shi (2015) in which they used POET estimator to estimate $\mathbf{\Sigma}$. However, the basic method for bounding the risk error $|\mathbf{w}'\widehat{\mathbf{\Sigma}}\mathbf{w} - \mathbf{w}'\mathbf{\Sigma}\mathbf{w}|$ in their paper is

$$|\mathbf{w}'\widehat{\mathbf{\Sigma}}\mathbf{w} - \mathbf{w}'\mathbf{\Sigma}\mathbf{w}| \le \|\mathbf{w}\|_1^2\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_{\max}.$$

They assumed that the gross exposure of the portfolio is bounded, that is, $\|\mathbf{w}\|_1 = O(1)$. Technically, when p is large, $\mathbf{w}'\mathbf{\Sigma}\mathbf{w}$ can be small. What an investor cares mostly is the relative risk error $\mathrm{RE}(\mathbf{w}) = |\mathbf{w}'\widehat{\mathbf{\Sigma}}\mathbf{w}/\mathbf{w}'\mathbf{\Sigma}\mathbf{w} - 1|$. Often $\mathbf{w}$ is a data-driven investment strategy, which depends on the past data. Regardless of what $\mathbf{w}$ is

$$\max_{\mathbf{w}} \mathrm{RE}(\mathbf{w}) = \|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_{\mathbf{\Sigma}},$$

which does not converge by Theorem 4.1 for $p > T$. Thus, the question of interest is what kind of portfolio $\mathbf{w}$ will make the relative error converge. Decompose $\mathbf{w}$ as a linear combination of the eigenvectors of $\mathbf{\Sigma}$, namely $\mathbf{w} = (\mathbf{\Gamma}, \mathbf{\Omega})\boldsymbol{\eta}$ and $\boldsymbol{\eta} = (\boldsymbol{\eta}_A', \boldsymbol{\eta}_B')'$. We have the following useful result for risk management.

THEOREM 4.2. *Under Assumptions 2.1, 2.2, 4.1, 4.2 and the factor model (4.1) with Gaussian noises and factors, if there exists $C_1 > 0$ such that $\|\boldsymbol{\eta}_B\|_1 \le C_1$, and assume $\lambda_j \propto p^\alpha$ for $j = 1, \dots, m$ and $T \ge Cp^\beta$ for $\alpha > 1/2, 0 < \beta < 1, \alpha + \beta > 1$, then the relative risk error is of order*

$$\mathrm{RE}(\mathbf{w}) = \left|\frac{\mathbf{w}'\widehat{\boldsymbol{\Sigma}}^S\mathbf{w}}{\mathbf{w}'\boldsymbol{\Sigma}\mathbf{w}} - 1\right| = O_P\left(T^{-\min\{\frac{2(\alpha+\beta-1)}{\beta}, \frac{1}{2}\}} + m_p w_T^{1-q}\right),$$

for $\alpha < 1$. If $\alpha \ge 1$ or there exists $C_2 > 0$ such that $\|\boldsymbol{\eta}_A\| \ge C_2$, $\mathrm{RE}(\mathbf{w}) = O_P(m_p w_T^{1-q})$.

The condition $\|\boldsymbol{\eta}_B\|_1 \le C_1$ is generally weaker than $\|\mathbf{w}\|_1 = O(1)$. It does not limit the total exposure of investor's position, but only put constraint on investment of the nonspiked section. Note that under the conditions of Theorem 4.2, $p/(T\lambda_j) \to 0$, so S-POET and POET are approximately the same. Hence, the stated result is valid for POET, also.

4.4. *Estimation of false discovery proportion.* Another important application of the factor model is the estimation of false discovery proportion. For simplicity, we assume Gaussian data $\mathbf{X}_i \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with an unknown correlation matrix $\boldsymbol{\Sigma}$ and wish to test which coordinates of $\boldsymbol{\mu}$ are nonvanishing. Consider the test statistic $\mathbf{Z} = \sqrt{n}\bar{\mathbf{X}}$ where $\bar{\mathbf{X}}$ is the sample mean of all data. Then $\mathbf{Z} \sim N(\boldsymbol{\mu}^*, \boldsymbol{\Sigma})$ with $\boldsymbol{\mu}^* = \sqrt{n}\boldsymbol{\mu}$. The problem is to test

$$H_{0j} : \mu_j^* = 0 \quad \text{v.s.} \quad H_{1j} : \mu_j^* \neq 0.$$

Define the number of discoveries $R(t) = \#\{j : P_j \le t\}$ and the number of false discoveries $V(t) = \#\{\text{true null} : P_j \le t\}$, where P_j is the p-value associated with the jth test. Note that $R(t)$ is observable while $V(t)$ needs to be estimated. The false discovery proportion (FDP) is defined as $\mathrm{FDP}(t) = V(t)/R(t)$.

Fan and Han (2013) proposed to employ the factor structure

$$\boldsymbol{\Sigma} = \mathbf{B}\mathbf{B}' + \mathbf{A}, \tag{4.12}$$

where $\mathbf{B} = (\sqrt{\lambda_1}\boldsymbol{\xi}_1, \dots, \sqrt{\lambda_m}\boldsymbol{\xi}_m)$. λ_j and $\boldsymbol{\xi}_j$ are respectively the jth eigenvalue and eigenvector of $\boldsymbol{\Sigma}$ as before. Then $\mathbf{Z}$ can be stochastically decomposed as

$$\mathbf{Z} = \boldsymbol{\mu}^* + \mathbf{B}\mathbf{W} + \mathbf{K},$$

where $\mathbf{W} \sim N(\mathbf{0}, \mathbf{I}_m)$ are m common factors and $\mathbf{K} \sim N(\mathbf{0}, \mathbf{A})$, independent of $\mathbf{W}$, are the idiosyncratic errors. For simplicity, assume the maximal number of nonzero elements of each row of $\mathbf{A}$ is bounded. In Fan and Han (2013), they argued that the asymptotic upper bound

$$\mathrm{FDP}_A(t) = \sum_{i=1}^{p}\left[\Phi\left(a_i(z_{t/2} + \eta_i)\right) + \Phi\left(a_i(z_{t/2} - \eta_i)\right)\right]/R(t) \tag{4.13}$$

of FDP(t) should be a realistic target to estimate for dependence tests, where $z_{t/2}$ is the $t/2$-quantile of the standard normal distribution, $a_i = (1 - \|\mathbf{b}_i\|^2)^{-1/2}$, $\eta_i = \mathbf{b}_i'\mathbf{W}$ and $\mathbf{b}_i'$ is the ith row of $\mathbf{B}$.

Realized factors $\mathbf{W}$ and the loading matrix $\mathbf{B}$ are typically unknown. If a generic estimator $\widehat{\mathbf{\Sigma}}$ is provided, then we are able to estimate $\mathbf{B}$, and thus $\mathbf{b}_i$ from its empirical eigenvalues and eigenvectors $\widehat{\lambda}_j$'s and $\widehat{\boldsymbol{\xi}}_j$'s. $\mathbf{W}$ can be estimated by the least-squares estimate $\widehat{\mathbf{W}} = (\widehat{\mathbf{B}}'\widehat{\mathbf{B}})^{-1}\widehat{\mathbf{B}}'\mathbf{Z}$. Fan and Han (2013) proposed the following estimator for FDP$_A(t)$:

$$\widehat{\mathrm{FDP}}_U(t) = \sum_{i=1}^{p}\bigl[\Phi\bigl(\widehat{a}_i(z_{t/2} + \widehat{\eta}_i)\bigr) + \Phi\bigl(\widehat{a}_i(z_{t/2} - \widehat{\eta}_i)\bigr)\bigr]/R(t), \tag{4.14}$$

where $\widehat{a}_i = (1 - \|\widehat{\mathbf{b}}_i\|^2)^{-1/2}$ and $\widehat{\eta}_i = \widehat{\mathbf{b}}_i'\widehat{\mathbf{W}}$. The following assumptions are in their paper.

ASSUMPTION 4.3. There exists a constant $h > 0$ such that (i) $R(t)/p > hp^{-\theta}$ for $h > 0$ and $\theta \geq 0$ as $p \to \infty$ and (ii) $\widehat{a}_i \leq h$, $a_i \leq h$ for all $i = 1, \dots, p$.

They showed that if $\widehat{\mathbf{\Sigma}}$ is based on the POET estimator with a spike size $\lambda_m \asymp p$, under Assumptions B.1–B.4, on the event that Assumption 4.3 holds:

$$\bigl|\widehat{\mathrm{FDP}}_{U,\mathrm{POET}}(t) - \mathrm{FDP}_A(t)\bigr| = O_P\biggl(p^{\theta}\biggl(\sqrt{\frac{\log p}{T}} + \frac{\|\boldsymbol{\mu}^*\|}{\sqrt{p}}\biggr)\biggr). \tag{4.15}$$

Again we can relax the assumption on the spike magnitude from order p to much weaker Assumption 4.1. Since $\mathbf{\Sigma}$ is a correlation matrix, $\lambda_1 \leq \mathrm{tr}(\mathbf{\Sigma}) = p$. This, together with Assumption 4.1, leads us to consider leading eigenvalues of order p^{α} for $1/2 < \alpha \leq 1$.

Now we apply the proposed S-POET method to obtain $\widehat{\mathbf{\Sigma}}^S$ and use it for FDP estimation. The following theorem shows the estimation error.

THEOREM 4.3. *If Assumptions 2.1, 2.2, 4.1 and 4.2 are applied to Gaussian independent data $\mathbf{X}_i \sim N(\boldsymbol{\mu}, \mathbf{\Sigma})$, and $\lambda_j \propto p^{\alpha}$ for $j = 1, \dots, m$, $T \geq Cp^{\beta}$ for $1/2 < \alpha \leq 1, 0 < \beta < 1, \alpha + \beta > 1$, on the event that Assumption 4.3 holds, we have*

$$\bigl|\widehat{\mathrm{FDP}}_{U,\mathrm{SPOET}}(t) - \mathrm{FDP}_A(t)\bigr| = O_P\bigl(p^{\theta}\bigl(\|\boldsymbol{\mu}^*\|p^{-\frac{1}{2}} + T^{-\min\{\frac{\alpha+\beta-1}{\beta}, \frac{1}{2}\}}\bigr)\bigr).$$

Comparing the result with (4.15), this convergence rate attained by S-POET is more general than the rate achieved before. The only difference is the second term, which is $O(T^{-1/2})$ if $\alpha + \frac{1}{2}\beta \geq 1$ and $T^{-(\alpha+\beta-1)/\beta}$ otherwise. So we relax the condition from $\alpha = 1$ in Fan and Han (2013) to $\alpha \in (1/2, 1]$. This means a weaker signal than order p is actually allowed to obtain a consistent estimate of false discovery proportion.

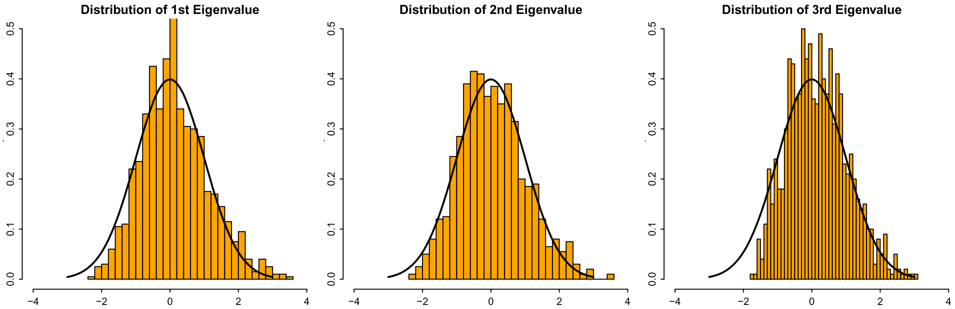

FIG. 1. *Behavior of empirical eigenvalues. The empirical distributions of* $\sqrt{n/2}(\widehat{\lambda}_j/\lambda_j - 1 - c_j)$ *for* $j = 1, 2, 3$ *are compared with their asymptotic distributions* $N(0, 1)$.

5. Simulations.

We conducted simulations to demonstrate the finite sample behavior of empirical eigenstructure, the performance of S-POET, and validity of applying it to estimate false discovery proportion.

5.1. *Eigenstructure.* In this simulation, we set $n = 50$, $p = 500$ and $\mathbf{\Sigma} = \operatorname{diag}(50, 20, 10, 1, \ldots, 1)$, which has three spiked eigenvalues ($m = 3$) $\lambda_1 = 50, \lambda_2 = 20, \lambda_3 = 10$ and correspondingly $c_1 = 0.2, c_2 = 0.5, c_3 = 1$. Data are generated from multivariate Gaussian. The number of simulations is 1000. The histograms of the standardized empirical eigenvalues $\sqrt{n/2}(\widehat{\lambda}_j/\lambda_j - 1 - c_j)$, and their associated asymptotic distributions (standard normal) are plotted in Figure 1. The approximations are very good even for this low sample size $n = 50$.

Figure 2 shows the histograms of $\sqrt{n}(\widehat{\boldsymbol{\xi}}_{jA}/\|\widehat{\boldsymbol{\xi}}_{jA}\| - \mathbf{e}_{jA})$ for the first three elements (the spiked part) of the first three eigenvectors. On the one hand, according to the asymptotic results, the values in the diagonal position should stochastically converge to 0 as observed. On the other hand, plots in the off-diagonal positions should converge in distribution to $N(0, 1)$ after standardization, which is indeed the case. We also report the correlations between the first three elements for the three eigenvectors based on those 1000 repetitions in Table 1. The correlations are all quite close to 0, which is consistent with the theory.

For the normalized nonspiked part $\widehat{\boldsymbol{\xi}}_{jB}/\|\widehat{\boldsymbol{\xi}}_{jB}\|$, it should be distributed uniformly over the unit sphere. This can be tested by the results of Cai, Fan and Jiang (2013). For any n data points $\mathbf{X}_1, \ldots, \mathbf{X}_n$ on a p-dimensional sphere, define the normalized empirical distribution of angles of each pair of vectors as

$$\mu_{n,p} = \frac{1}{\binom{n}{2}} \sum_{1 \le i < j \le n} \delta_{\sqrt{p-2}(\pi/2 - \Theta_{ij})},$$

where $\Theta_{ij} \in [0, \pi]$ is the angle between vectors $\mathbf{X}_i$ and $\mathbf{X}_j$. When the data are generated uniformly from a sphere, $\mu_{n,p}$ converges to the standard normal distri-

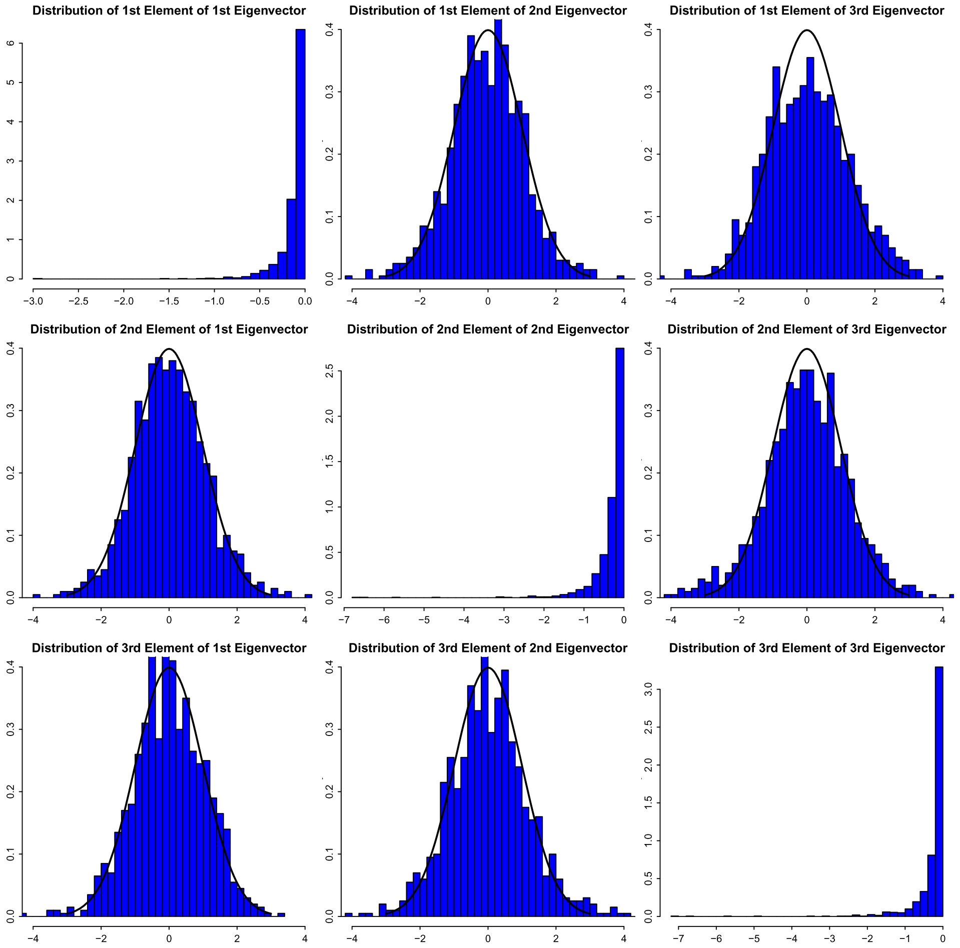

FIG. 2. *Behavior of empirical eigenvectors. The histogram of the kth element of the jth empirical vector is depicted in the location* (k, j) *for* $k,\ j \le 3$. *Off-diagonal plots of values* $\sqrt{n}\widehat{\xi}_{jk}/\|\widehat{\boldsymbol{\xi}}_{jA}\|/\sqrt{\frac{c_j c_k}{(c_j - c_k)^2}}$ *are compared to their asymptotic distributions* $N(0, 1)$ *for* $k \neq j$ *while diagonal plots of values* $\sqrt{n}(\widehat{\xi}_{jj}/\|\widehat{\boldsymbol{\xi}}_{jA}\| - 1)$ *are compared to stochastically* 0.

TABLE 1

The correlations between the first three elements for each of the three empirical eigenvectors based on 1000 *repetitions*

	1st & 2nd elements	**1st & 3rd elements**	**2nd & 3rd elements**
1st Eigenvector	0.00156	−0.00192	−0.04112
2nd Eigenvector	−0.02318	−0.00403	0.01483
3rd Eigenvector	−0.02529	−0.04004	0.12524

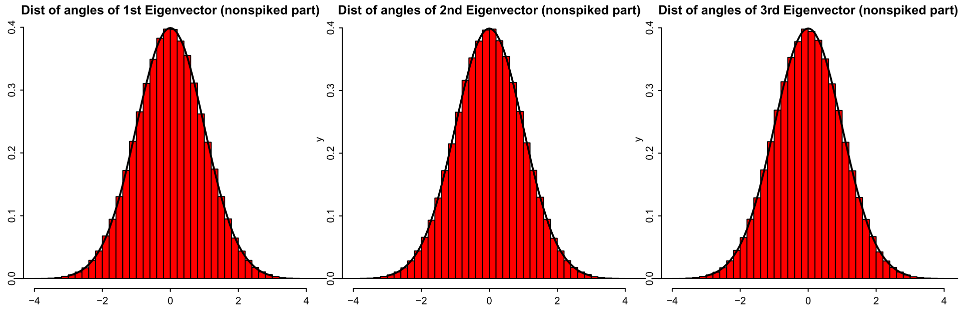

FIG. 3. *The empirical distributions of all pairwise angles of the* 1000 *realized* $\widehat{\boldsymbol{\xi}}_{jB}/\|\widehat{\boldsymbol{\xi}}_{jB}\|$ $(j = 1, 2, 3)$ *compared with their asymptotic distributions* $N(0, 1)$.

bution with probability 1. Figure 3 shows the empirical distributions of all pairwise angles of the realized $\widehat{\boldsymbol{\xi}}_{jB}/\|\widehat{\boldsymbol{\xi}}_{jB}\|$ $(j = 1, 2, 3)$ in 1000 simulations. Since number of such pairwise angels is $\binom{1000}{2}$, the empirical distributions and the asymptotic distributions $N(0, 1)$ are almost identical. The normality holds even for a small subset of the angles.

Lastly, we did simulations to verify the rate difference of $\langle \widehat{\boldsymbol{\xi}}_j, \mathbf{e}_j \rangle$ for $m = 1$ and $m > 1$, revealed in Theorem 3.2(iii). We choose $n = [10 \times 1.2^l]$ for $l = 0, \dots, 9$, $p = [n^3/100]$, where $[\cdot]$ represents rounding. We set $\lambda_j = 1$ for $j \ge 3$ and consider two situations: (1) $\lambda_1 = p, \lambda_2 = 1$, (2) $\lambda_1 = 2\lambda_2 = p$. Under both cases, simulations were carried out 500 times and the corresponding angles of the empirical eigenvector and its truth were calculated for each simulation. The logarithm of the median absolute error of $\langle \widehat{\boldsymbol{\xi}}_1, \mathbf{e}_1 \rangle - 1/\sqrt{1 + c_1}$ was plotted against $\log(n)$. Under the two situations, the rates of convergence are $O_P(n^{-3/2})$ and $O_P(n^{-1})$ respectively. Thus the slope of the curves should be $-3/2$ for a single spike and -1 for two spikes, which is indeed the case as shown in Figure 4.

In short, all the simulation results match well with the theoretical results for the high dimensional regime.

5.2. *Performance of S-POET.* We demonstrate the effectiveness of S-POET in comparison with POET. A similar setting to the last section is used, i.e. $m = 3$ and $c_1 = 0.2, c_2 = 0.5, c_3 = 1$. The sample size T ranges from 50 to 150 and $p = [T^{3/2}]$. Note that when $T = 150$, $p \approx 1800$. The spiked eigenvalues are determined from $p/(T\lambda_j) = c_j$ so that λ_j is of order $\sqrt{T}$, which is much smaller than p. For each pair of T and p, the following steps are used to generate observed data from the factor model for 200 times:

(1) Each row of $\mathbf{B}$ is simulated from the standard multivariate normal distribution and the jth column is normalized to have norm λ_j for $j = 1, 2, 3$.

(2) Each row of $\mathbf{F}$ is simulated from standard multivariate normal distribution.

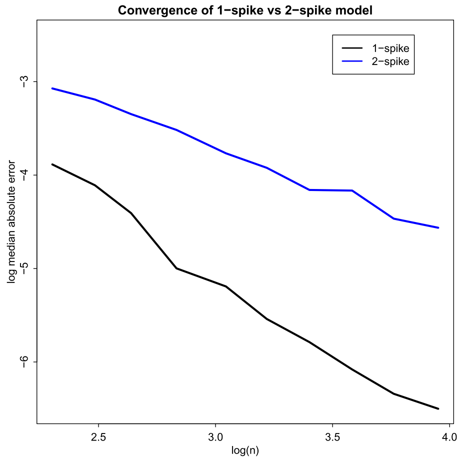

FIG. 4. *Difference of convergence rate of* $\langle \widehat{\boldsymbol{\xi}}_1, \mathbf{e}_1\rangle - 1/\sqrt{1+c_1}$ *for models with a single spike and two spikes. The error should be expected to decrease at the rate of* $n^{-3/2}$ *and* n^{-1}, *respectively.*

(3) Set $\boldsymbol{\Sigma}_u = \mathrm{diag}(\sigma_1^2, \dots, \sigma_p^2)$ where σ_i's are generated from Gamma(α, β) with $\alpha = \beta = 100$ (mean 1, standard deviation 0.1). The idiosyncratic error $\mathbf{U}$ is simulated from $N(\mathbf{0}, \boldsymbol{\Sigma}_u)$.

(4) Compute the observed data $\mathbf{Y} = \mathbf{B}\mathbf{F}' + \mathbf{U}$.

Both S-POET and POET are applied to estimate the covariance matrix $\boldsymbol{\Sigma} = \mathbf{B}\mathbf{B}' + \boldsymbol{\Sigma}_u$. Their mean estimation errors over 200 simulations, measured in relative spectral norm $\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{\boldsymbol{\Sigma}}$, relative Frobenius norm $\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{\boldsymbol{\Sigma},F}$, spectral norm $\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|$ and max norm $\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{\max}$, are reported in Figure 5. The errors for sample covariance matrix are also depicted for comparison. First notice that no matter in what norm, S-POET uniformly outperforms POET and the sample covariance. It affirms the claim that shrinkage of spiked eigenvalues is necessary to maintain good performance when the spikes are not sufficiently large. Since the low rank part is not shrunk for POET, its error under the spectral norm is comparable and even slightly larger than that of the sample covariance matrix. The errors under max norm and relative Frobenius norm as expected decrease as T and p increase. However, the error under the relative spectral norm does not converge: our theory shows it should increase in the order $p/T = \sqrt{T}$.

5.3. *FDP estimation.* In this section, we report simulation results on FDP estimation by using both POET and S-POET. The data are simulated in a similar way as in Section 5.2 with $p = 1000$ and $n = 100$. The first $m = 3$ eigenvalues have spike sizes proportional to $p/\sqrt{n}$ which corresponds to $\alpha = \beta = 2/3$ in Theorem 4.3. The true FDP is calculated by using $\mathrm{FDP}(t) = V(t)/R(t)$ with $t = 0.01$. The approximate FDP, $\mathrm{FDP}_A(t)$, is calculated as in (4.13) with known $\mathbf{B}$ but esti-

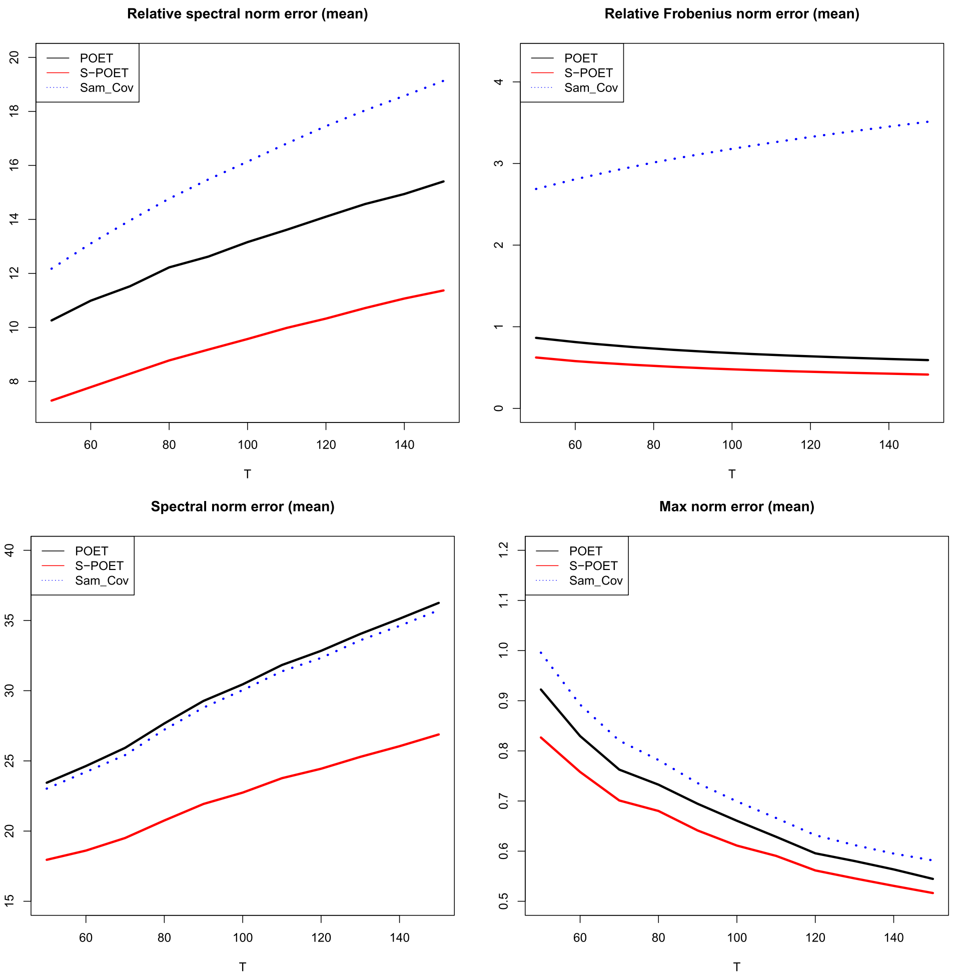

FIG. 5. *Estimation errors of covariance matrix under relative spectral, relative Frobenius, spectral and max norms using S-POET (red), POET (black) and sample covariance (blue).*

mated $\mathbf{W}$ given by $\widehat{\mathbf{W}} = (\mathbf{B}\mathbf{B}')^{-1}\mathbf{B}'\mathbf{Z}$. This $\mathrm{FDP}_A(t)$ based on a known covariance matrix serves as a benchmark for our estimated covariance matrix to compare with. We employed POET and S-POET to get $\widehat{\mathrm{FDP}}_{U,\mathrm{POET}}(t)$ and $\widehat{\mathrm{FDP}}_{U,\mathrm{SPOET}}(t)$.

In Figure 6, three scatter plots are drawn to compare $\mathrm{FDP}_A(t)$, $\widehat{\mathrm{FDP}}_{U,\mathrm{POET}}(t)$ and $\widehat{\mathrm{FDP}}_{U,\mathrm{SPOET}}(t)$ with the true $\mathrm{FDP}(t)$. The points are basically aligned along the 45 degree line, meaning that all of them are quite close to the true FDP. With the semistrong signal $\lambda \propto p/\sqrt{n}$, although much weaker than order p, POET accomplishes the task as well as S-POET. Both estimators perform as well as if we know the covariance matrix $\boldsymbol{\Sigma}$, the benchmark.

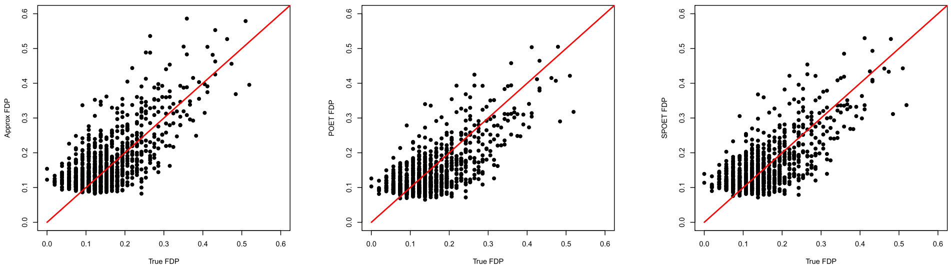

FIG. 6. *Comparison of estimated FDP's with true values. The left plot assumes knowledge of* **B**, *the middle and right ones are corresponding to POET and S-POET methods, respectively. The results are aligned along the* 45-*degree line, indicating the accuracy of the estimated FDP.*

6. Conclusions. In this paper, we studied two closely related problems: the asymptotic behavior of empirical eigenvalues and eigenvectors under a general regime of bounded $p/(n\lambda_j)$ and the large covariance estimation for factor models with relaxed signal level of $\sqrt{p} = o(\lambda_j)$.

The first study provides new technical tools for the derivation of error bounds for large covariance estimation under relative Frobenius norm (with better rate) and relative spectral norm (for the first time). The results motivate the newly proposed covariance estimator S-POET for the second problem by correcting biases of the estimated leading eigenvalues. S-POET is demonstrated to have better sampling properties than POET, and this is convincingly verified in the simulation study. In addition, we are able to apply S-POET to two important applications, risk management and false discovery control, and relax the required signal to $\sqrt{p}$. Those conclusions shed new lights for applications of factor models.

On the other hand, the second problem is a key motivation for us to study the empirical engen-structure in a more general high dimensional regime. We aim to understand why PCA works for pervasive factor models but fails classical random matrix problems, without sparsity assumptions. What are the fundamental limit for PCA in high dimensions? We clearly showed that for both empirical eigenvalues and vectors, consistency is granted once $p/(n\lambda_j) \to 0$. Furthermore, our theories give a fine-grained characterization of the asymptotic behavior under the generalized and unified regime, which includes the situation of bounded eigenvalues, HDLSS and pervasive factor models, especially for empirical eigenvectors. The asymptotic rate of convergence is obtained as long as $p/(n\lambda_j)$ is bounded, while the asymptotic distribution is fully described when $\sqrt{p} = o(\lambda_j)$. Some interesting phenomena, such as interaction between multiple spikes, are also revealed in our results. Our proofs are novel in that we clearly identify terms that keep the low-dimensional asymptotic normality and terms that generate the random biases. In sum, our results serve as a necessary complement of the random matrix literature when the signal diverges with dimensionality.

APPENDIX A: PROOFS FOR SECTION 3

A.1. Proof of Theorem 3.1. We first provide three useful lemmas for the proof. Lemma A.1 provides nonasymptotic upper and lower bound for the eigenvalues of weighted Wishart matrix for sub-Gaussian distributions.

LEMMA A.1. *Let* $\mathbf{A}_1, \dots, \mathbf{A}_n$*'s be n independent p dimensional sub-Gaussian random vectors with zero mean and identity variance, and the sub-Gaussian norms bounded by a constant* C_0. *Then for every* $t \geq 0$, *with probability at least* $1 - 2\exp(-ct^2)$, *one has*

$$\bar{w} - \max\{\delta, \delta^2\} \leq \lambda_p\left(\frac{1}{n}\sum_{i=1}^n w_i \mathbf{A}_i \mathbf{A}_i'\right) \leq \lambda_1\left(\frac{1}{n}\sum_{i=1}^n w_i \mathbf{A}_i \mathbf{A}_i'\right) \leq \bar{w} + \max\{\delta, \delta^2\},$$

where $\delta = C\sqrt{p/n} + t/\sqrt{n}$ *for constants* $C, c > 0$, *depending on* C_0. *Here*, $|w_i|$*'s is bounded for all i and* $\bar{w} = n^{-1}\sum_{i=1}^n w_i$.

The above lemma is the extension of the classical Davidson–Szarek bound [Theorem II.7 of Davidson and Szarek (2001)] to the weighted sample covariance with sub-Gaussian distribution. It was shown by Vershynin (2010) that the conclusion holds with $w_i = 1$ for all i. With similar techniques to those developed in Vershynin (2010), we can obtain the above lemma for general bounded weights. The details are omitted.

Now in order to prove the theorem, let us define two quantities and treat them separately in the following two lemmas. Let

$$\mathbf{A} = n^{-1}\sum_{j=1}^m \lambda_j \mathbf{Z}_j \mathbf{Z}_j' \quad \text{and} \quad \mathbf{B} = n^{-1}\sum_{j=m+1}^p \lambda_j \mathbf{Z}_j \mathbf{Z}_j',$$

where $\mathbf{Z}_j$ is columns of $\mathbf{X}\boldsymbol{\Lambda}^{-\frac{1}{2}}$. Then

$$\tilde{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{j=1}^p \lambda_j \mathbf{Z}_j \mathbf{Z}_j' = \mathbf{A} + \mathbf{B}. \tag{A.1}$$

LEMMA A.2. *Under Assumptions 2.1–2.3, as* $n \to \infty$,

$$\sqrt{n}(\lambda_j(\mathbf{A})/\lambda_j - 1) \overset{d}{\Rightarrow} N(0, \kappa_j - 1) \qquad \textit{for } j = 1, \dots, m.$$

In addition, they are asymptotically independent.

LEMMA A.3. *Under Assumptions 2.1–2.3, for* $j = 1, \dots, m$, *we have*

$$\lambda_k(\mathbf{B})/\lambda_j = \bar{c}c_j + O_P(\lambda_j^{-1}\sqrt{p/n}) + o_P(c_j n^{-\frac{1}{2}}) \qquad \textit{for } k = 1, 2, \dots, n.$$

The proofs of the above two lemmas will be given in Appendix A in the supplementary material [Wang and Fan (2017)].

PROOF OF THEOREM 3.1. By Wely's theorem, $\lambda_j(\mathbf{A}) + \lambda_n(\mathbf{B}) \le \widehat{\lambda}_j \le \lambda_j(\mathbf{A}) + \lambda_1(\mathbf{B})$. Therefore, from Lemma A.3,

$$\frac{\widehat{\lambda}_j}{\lambda_j} = \frac{\lambda_j(\mathbf{A})}{\lambda_j} + \bar{c}c_j + O_P\left(\lambda_j^{-1}\sqrt{\frac{p}{n}}\right) + o_P\left(c_j n^{-1/2}\right).$$

By Lemma A.2 and Slutsky's theorem, we conclude that $\sqrt{n}(\widehat{\lambda}_j/\lambda_j - (1 + \bar{c}c_j + O_P(\lambda_j^{-1}\sqrt{p/n})))$ converges in distribution to $N(0, \kappa_j - 1)$ and the limiting distributions of the first m eigenvalues are independent. □

A.2. Proofs of Theorem 3.2. The proof of Theorem 3.2 is mathematically involved. The basic idea for proving part (i) is outlined in Section 2. We relegate less important technical Lemmas A.4–A.6 to Appendix A in the supplementary material [Wang and Fan (2017)] in order not to distract the readers. The proof of part (ii) utilizes the invariance of standard Gaussian distribution under orthogonal transformations.

PROOF OF THEOREM 3.2. (i) Let us start by proving the asymptotic normality of $\widehat{\boldsymbol{\xi}}_{jA}$ for the case $m > 1$. Write

$$\mathbf{X} = (\mathbf{Z}_A\mathbf{\Lambda}_A^{\frac{1}{2}}, \mathbf{Z}_B\mathbf{\Lambda}_B^{\frac{1}{2}}) = (\sqrt{\lambda_1}\mathbf{Z}_1, \dots, \sqrt{\lambda_m}\mathbf{Z}_m, \sqrt{\lambda_{m+1}}\mathbf{Z}_{m+1}, \dots, \sqrt{\lambda_p}\mathbf{Z}_p),$$

where each $\mathbf{Z}_j$ follows a sub-Gaussian distribution with mean $\mathbf{0}$ and identity variance $\mathbf{I}_n$. Then by the eigenvalue relationship of equation (2.1), we have

$$\widehat{\boldsymbol{\xi}}_{jA} = \frac{\mathbf{\Lambda}_A^{\frac{1}{2}}\mathbf{Z}_A'\mathbf{u}_j}{\sqrt{n\widehat{\lambda}_j}} \quad \text{and} \quad \mathbf{u}_j = \frac{\mathbf{X}\widehat{\boldsymbol{\xi}}_j}{\sqrt{n\widehat{\lambda}_j}} = \frac{\mathbf{Z}_A\mathbf{\Lambda}_A^{\frac{1}{2}}\widehat{\boldsymbol{\xi}}_{jA}}{\sqrt{n\widehat{\lambda}_j}} + \frac{\mathbf{Z}_B\mathbf{\Lambda}_B^{\frac{1}{2}}\widehat{\boldsymbol{\xi}}_{jB}}{\sqrt{n\widehat{\lambda}_j}}. \tag{A.2}$$

Recall $\mathbf{u}_j$ is the eigenvector of the matrix $\widetilde{\mathbf{\Sigma}}$, that is, $\frac{1}{n}\mathbf{X}\mathbf{X}'\mathbf{u}_j = \widehat{\lambda}_j\mathbf{u}_j$. Using $\mathbf{X} = (\mathbf{Z}_A\mathbf{\Lambda}_A^{\frac{1}{2}}, \mathbf{Z}_B\mathbf{\Lambda}_B^{\frac{1}{2}})$, we obtain

$$\left(\mathbf{I}_n - \frac{1}{n}\mathbf{Z}_A\frac{\mathbf{\Lambda}_A}{\lambda_j}\mathbf{Z}_A'\right)\mathbf{u}_j = \mathbf{D}\mathbf{u}_j - \Delta\mathbf{u}_j, \tag{A.3}$$

where we denote $\mathbf{D} = (n\lambda_j)^{-1}\mathbf{Z}_B\mathbf{\Lambda}_B\mathbf{Z}_B' - \bar{c}c_j\mathbf{I}_n$, $\Delta = \widehat{\lambda}_j/\lambda_j - (1 + \bar{c}c_j)$. We then left-multiply equation (A.3) by $\mathbf{\Lambda}_A^{\frac{1}{2}}\mathbf{Z}_A'/\sqrt{n\widehat{\lambda}_j}$ and employ relationship (A.2) to replace $\mathbf{u}_j$ by $\widehat{\boldsymbol{\xi}}_{jA}$ and $\widehat{\boldsymbol{\xi}}_{jB}$ as follows:

$$\begin{aligned}\left(\mathbf{I}_m - \frac{\boldsymbol{\Lambda}_A}{\lambda_j}\right)\widehat{\boldsymbol{\xi}}_{jA} = {}& \frac{\boldsymbol{\Lambda}_A^{\frac{1}{2}}(\frac{1}{n}\mathbf{Z}_A'\mathbf{Z}_A - \mathbf{I}_m)\boldsymbol{\Lambda}_A^{\frac{1}{2}}}{\lambda_j}\widehat{\boldsymbol{\xi}}_{jA} + \frac{\boldsymbol{\Lambda}_A^{\frac{1}{2}}\mathbf{Z}_A'\mathbf{D}\mathbf{Z}_A\boldsymbol{\Lambda}_A^{\frac{1}{2}}}{n\widehat{\lambda}_j}\widehat{\boldsymbol{\xi}}_{jA} \\ & + \frac{\boldsymbol{\Lambda}_A^{\frac{1}{2}}\mathbf{Z}_A'\mathbf{D}\mathbf{Z}_B\boldsymbol{\Lambda}_B^{\frac{1}{2}}}{n\widehat{\lambda}_j}\widehat{\boldsymbol{\xi}}_{jB} - \Delta\widehat{\boldsymbol{\xi}}_{jA}.\end{aligned} \tag{A.4}$$

Further, define

$$\mathbf{R} = \sum_{k\in[m]\setminus j} \frac{\lambda_j}{\lambda_j - \lambda_k}\mathbf{e}_{kA}\mathbf{e}_{kA}'.$$

Then we have $\mathbf{R}(\mathbf{I} - \boldsymbol{\Lambda}_A/\lambda_j) = \mathbf{I}_m - \mathbf{e}_{jA}\mathbf{e}_{jA}'$. Note that $\mathbf{R}$ is only well defined if $m > 1$. Therefore, by left multiplying $\mathbf{R}$ to equation (A.4),

$$\begin{aligned}\widehat{\boldsymbol{\xi}}_{jA} - \langle\widehat{\boldsymbol{\xi}}_{jA}, \mathbf{e}_{jA}\rangle\mathbf{e}_{jA} = {}& \mathbf{R}\left(\frac{\boldsymbol{\Lambda}_A}{\lambda_j}\right)^{\frac{1}{2}}\mathbf{K}\left(\frac{\boldsymbol{\Lambda}_A}{\lambda_j}\right)^{\frac{1}{2}}\widehat{\boldsymbol{\xi}}_{jA} \\ & + \mathbf{R}\frac{\boldsymbol{\Lambda}_A^{\frac{1}{2}}\mathbf{Z}_A'\mathbf{D}\mathbf{Z}_B\boldsymbol{\Lambda}_B^{\frac{1}{2}}}{n\widehat{\lambda}_j}\widehat{\boldsymbol{\xi}}_{jB} - \Delta\mathbf{R}\widehat{\boldsymbol{\xi}}_{jA},\end{aligned} \tag{A.5}$$

where $\mathbf{K} = n^{-1}\mathbf{Z}_A'\mathbf{Z}_A - \mathbf{I}_n + \lambda_j(n\widehat{\lambda}_j)^{-1}\mathbf{Z}_A'\mathbf{D}\mathbf{Z}_A$. Dividing both side by $\|\widehat{\boldsymbol{\xi}}_{jA}\|$, we are able to write

$$\frac{\widehat{\boldsymbol{\xi}}_{jA}}{\|\widehat{\boldsymbol{\xi}}_{jA}\|} - \mathbf{e}_{jA} = \mathbf{R}\left(\frac{\boldsymbol{\Lambda}_A}{\lambda_j}\right)^{\frac{1}{2}}\mathbf{K}\left(\frac{\boldsymbol{\Lambda}_A}{\lambda_j}\right)^{\frac{1}{2}}\mathbf{e}_{jA} + \mathbf{r}_n, \tag{A.6}$$

where

$$\begin{aligned}\mathbf{r}_n = {}& \left(\left\langle\frac{\widehat{\boldsymbol{\xi}}_{jA}}{\|\widehat{\boldsymbol{\xi}}_{jA}\|}, \mathbf{e}_{jA}\right\rangle - 1\right)\mathbf{e}_{jA} + \mathbf{R}\left(\frac{\boldsymbol{\Lambda}_A}{\lambda_j}\right)^{\frac{1}{2}}\mathbf{K}\left(\frac{\boldsymbol{\Lambda}_A}{\lambda_j}\right)^{\frac{1}{2}}\left(\frac{\widehat{\boldsymbol{\xi}}_{jA}}{\|\widehat{\boldsymbol{\xi}}_{jA}\|} - \mathbf{e}_{jA}\right) \\ & + \mathbf{R}\frac{\boldsymbol{\Lambda}_A^{\frac{1}{2}}\mathbf{Z}_A'\mathbf{D}\mathbf{Z}_B\boldsymbol{\Lambda}_B^{\frac{1}{2}}}{n\widehat{\lambda}_j}\frac{\widehat{\boldsymbol{\xi}}_{jB}}{\|\widehat{\boldsymbol{\xi}}_{jA}\|} - \Delta\mathbf{R}\left(\frac{\widehat{\boldsymbol{\xi}}_{jA}}{\|\widehat{\boldsymbol{\xi}}_{jA}\|} - \mathbf{e}_{jA}\right).\end{aligned} \tag{A.7}$$

LEMMA A.4. *As $n \to \infty$, $\|\mathbf{r}_n\| = O_P(\lambda_j^{-1}\sqrt{p/n} + 1/n)$.*

By Lemma A.4, $\mathbf{r}_n$ is a smaller order term. Together with $(\boldsymbol{\Lambda}_A/\lambda_j)^{\frac{1}{2}}\mathbf{e}_{jA} = \mathbf{e}_{jA}$, we have

$$\sqrt{n}\left(\frac{\widehat{\boldsymbol{\xi}}_{jA}}{\|\widehat{\boldsymbol{\xi}}_{jA}\|} - \mathbf{e}_{jA} + O_P\left(\sqrt{\frac{p}{n\lambda_j^2}}\right)\right) = \sqrt{n}\mathbf{R}\left(\frac{\Lambda_A}{\lambda_j}\right)^{\frac{1}{2}}\mathbf{K}\mathbf{e}_{jA} + o_P(1). \tag{A.8}$$

Now let us derive normality of the right hand side of (A.2). According to the definition of $\mathbf{R}$,

$$\mathbf{R}\left(\frac{\mathbf{\Lambda}_A}{\lambda_j}\right)^{\frac{1}{2}} = \sum_{k\in[m]\setminus j} \frac{\sqrt{\lambda_j\lambda_k}}{\lambda_j - \lambda_k}\mathbf{e}_{kA}\mathbf{e}'_{kA} \to \sum_{k\in[m]\setminus j} a_{jk}\mathbf{e}_{kA}\mathbf{e}'_{kA}. \tag{A.9}$$

Let $\mathbf{W} = \sqrt{n}\mathbf{K}\mathbf{e}_{jA} = (W_1, \dots, W_m)$ and $\mathbf{W}^{(-j)}$ be the $(m-1)$-dimensional vector without the jth element in $\mathbf{W}$. Since the jth diagonal element of $\mathbf{R}$ is zero, $\mathbf{R}(\mathbf{\Lambda}_A/\lambda_j)^{\frac{1}{2}}\mathbf{W}$ depends only on $\mathbf{W}^{(-j)}$.

LEMMA A.5. $\mathbf{W}^{(-j)} + O_P(\lambda_j^{-1}\sqrt{p/n}) \overset{d}{\Rightarrow} N(\mathbf{0}, \mathbf{I}_{m-1})$.

Therefore, by Lemma A.5 and Slutsky's theorem,

$$\sqrt{n}\mathbf{R}\left(\frac{\mathbf{\Lambda}_A}{\lambda_j}\right)^{\frac{1}{2}}\mathbf{K}\mathbf{e}_{jA} + O_P\left(\sqrt{\frac{p}{n\lambda_j^2}}\right) \overset{d}{\Rightarrow} N_m\left(\mathbf{0}, \sum_{k\in[m]\setminus j} a_{jk}^2\mathbf{e}_{kA}\mathbf{e}'_{kA}\right).$$

Together with (A.2), we concludes (3.3) for the case $m > 1$.

Now let us turn to the case of $m = 1$. Since $\mathbf{R}$ is not defined for $m = 1$, we need to find a different derivation. Equivalently, (A.3) can be written as

$$\frac{1}{n}\mathbf{Z}_1\mathbf{Z}'_1\mathbf{u}_1 + \frac{1}{n\lambda_1}\mathbf{Z}_B\mathbf{\Lambda}_B\mathbf{Z}'_B\mathbf{u}_1 = \frac{\widehat{\lambda}_1}{\lambda_1}\mathbf{u}_1.$$

Left-multiplying $\mathbf{u}'_1$ and using relationship (A.2), we obtain easily

$$\widehat{\xi}^2_{1A} = 1 - \frac{\bar{c}c_1}{\widehat{\lambda}_1/\lambda_1} - \frac{\lambda_1}{\widehat{\lambda}_1}\mathbf{u}'_1\mathbf{D}\mathbf{u}_1 = 1 - \frac{\bar{c}c_1}{\widehat{\lambda}_1/\lambda_1} + O_P\big(\lambda_1^{-1}\sqrt{p/n}\big),$$

where $\mathbf{D}$ is defined as before and $\|\mathbf{D}\| = O_P(\lambda_1^{-1}\sqrt{p/n})$ according to the proof of Lemma A.4. Expanding $\sqrt{1 - \bar{c}c_1/x}$ at the point of $(1 + \bar{c}c_1)$, we have

$$\widehat{\xi}_{1A} = \frac{1}{\sqrt{1+\bar{c}c_1}} + \frac{\bar{c}c_1}{2(1+\bar{c}c_1)^{3/2}}\left(\frac{\widehat{\lambda}_1}{\lambda_1} - (1+\bar{c}c_1)\right) + O_P\left(\sqrt{\frac{p}{n\lambda_1^2}} + c_1 n^{-1}\right).$$

Note that from Lemmas A.2 and A.3, $\widehat{\lambda}_1/\lambda_1 - (1 + \bar{c}c_1) = (\|\mathbf{Z}_1\|^2/n - 1) + O_P(\lambda_1^{-1}(p/n)^{1/2}) + o_P(c_j n^{-1/2})$. Therefore, due to the fact $\sqrt{n}(\|\mathbf{Z}_1\|^2/n - 1)$ is asymptotically $N(0, \kappa_1 - 1)$, we conclude

$$\frac{2(1+\bar{c}c_1)^{3/2}}{\bar{c}c_1}\sqrt{n}\left(\widehat{\xi}_{1A} - \frac{1}{\sqrt{1+\bar{c}c_1}} + O_P\left(\sqrt{\frac{p}{n\lambda_1^2}}\right)\right) \overset{d}{\Rightarrow} N(0, \kappa_1 - 1).$$

This completes the first part of the proof.

(ii) We now prove the conclusion for the nonspiked part $\widehat{\boldsymbol{\xi}}_{jB}$. Recall that $\mathbf{X}_i$ follows $N(\mathbf{0}, \mathbf{\Lambda})$. Consider $\mathbf{X}_i^R = \operatorname{diag}(\mathbf{I}_m, \mathbf{D}_0)\mathbf{X}_i$ where as defined in the theorem

$\mathbf{D}_0 = \operatorname{diag}(\sqrt{\bar{c}/\lambda_{m+1}}, \ldots, \sqrt{\bar{c}/\lambda_p})$. Here, the superscript R indicates rescaled data by $\operatorname{diag}(\mathbf{I}_m, \mathbf{D}_0)$. After rescaling, we have $\mathbf{X}_i^R \sim N(\mathbf{0}, \operatorname{diag}(\mathbf{\Lambda}_A, \bar{c}\mathbf{I}_{p-m}))$. Correspondingly, the $n \times p$ data matrix $\mathbf{X}^R = \mathbf{X}\operatorname{diag}(\mathbf{I}_m, \mathbf{D}_0) = (\mathbf{X}_A, \mathbf{X}_B\mathbf{D}_0)$ where $\mathbf{X}_A = \mathbf{Z}_A\mathbf{\Lambda}_A^{\frac{1}{2}}$ and $\mathbf{X}_B = \mathbf{Z}_B\mathbf{\Lambda}_B^{\frac{1}{2}}$ as the notation before. Assume $\widehat{\boldsymbol{\xi}}_j^R$ and $\mathbf{u}_j^R$ are eigenvectors given by $\widehat{\mathbf{\Sigma}}^R$ and $\widetilde{\mathbf{\Sigma}}^R$ of the rescaled data $\mathbf{X}^R$ and $\widehat{\boldsymbol{\xi}}_j^R = (\widehat{\boldsymbol{\xi}}_{jA}^R, \widehat{\boldsymbol{\xi}}_{jB}^R)'$. It has been proved by Paul (2007) that $\mathbf{h}_0 := \widehat{\boldsymbol{\xi}}_{jB}^R/\|\widehat{\boldsymbol{\xi}}_{jB}^R\|$ is distributed uniformly over the unit sphere and is independent of $\|\widehat{\boldsymbol{\xi}}_{jB}^R\|$ due to the orthogonal invariance of the nonspiked part of $\widehat{\boldsymbol{\xi}}_{jB}^R$. Hence, it only remains to link $\widehat{\boldsymbol{\xi}}_{jB}/\|\widehat{\boldsymbol{\xi}}_{jB}\|$ with $\mathbf{h}_0$.

Note that $\widetilde{\mathbf{\Sigma}} = n^{-1}\mathbf{X}\mathbf{X}'$ and $\widetilde{\mathbf{\Sigma}}^R = n^{-1}\mathbf{X}^R\mathbf{X}^{R\prime}$, so

$$\|\widetilde{\mathbf{\Sigma}} - \widetilde{\mathbf{\Sigma}}^R\| = \left\|\frac{1}{n}\mathbf{X}_B(\mathbf{I} - \mathbf{D}_0^2)\mathbf{X}_B'\right\| = \left\|\frac{1}{n}\sum_{j=m+1}^{p}(\lambda_j - \bar{c})\mathbf{Z}_j\mathbf{Z}_j\right\|,$$

where the last term is of order $O_P(\sqrt{p/n})$ by Lemma A.1. Thus, by the $\sin\theta$ theorem of Davis and Kahan (1970), $\|\mathbf{u}_j - \mathbf{u}_j^R\| = O_P(\lambda_j^{-1}\sqrt{p/n})$. Next, we convert from $\mathbf{u}_j$ to $\widehat{\boldsymbol{\xi}}_{jB}$ using the basic relationship (2.1). We have

$$\begin{aligned}
&\left\|\mathbf{D}_0\frac{\widehat{\boldsymbol{\xi}}_{jB}}{\|\widehat{\boldsymbol{\xi}}_{jB}\|} - \frac{\widehat{\boldsymbol{\xi}}_{jB}^R}{\|\widehat{\boldsymbol{\xi}}_{jB}^R\|}\right\| \\
&\quad = \left\|\frac{\mathbf{D}_0\mathbf{X}_B'\mathbf{u}_j}{\sqrt{n\widehat{\lambda}_j}\|\widehat{\boldsymbol{\xi}}_{jB}\|} - \frac{\mathbf{D}_0\mathbf{X}_B{}'\mathbf{u}_j^R}{\sqrt{n\widehat{\lambda}_j^R}\|\widehat{\boldsymbol{\xi}}_{jB}^R\|}\right\| \\
&\quad \le \left\|\frac{\mathbf{D}_0\mathbf{X}_B'\mathbf{u}_j}{\sqrt{n\lambda_j}}\right\|\left|\sqrt{\frac{\lambda_j}{\widehat{\lambda}_j\|\widehat{\boldsymbol{\xi}}_{jB}\|^2}} - \sqrt{\frac{\lambda_j}{\widehat{\lambda}_j^R\|\widehat{\boldsymbol{\xi}}_{jB}^R\|^2}}\right| + \left\|\frac{\mathbf{D}_0\mathbf{X}_B'}{\sqrt{n\widehat{\lambda}_j^R}\|\widehat{\boldsymbol{\xi}}_{jB}^R\|}\right\|\|\mathbf{u}_j - \mathbf{u}_j^R\| \\
&\quad =: I + II.
\end{aligned}$$

First, it is not hard to see $II = O_P(\lambda_j^{-1}\sqrt{p/n})$ since $\|\mathbf{u}_j - \mathbf{u}_j^R\| = O_P(\lambda_j^{-1}\sqrt{p/n})$, $\|\mathbf{X}_B'/\sqrt{n\lambda_j}\| = O_P(\sqrt{c_j})$, $\lambda_j/\widehat{\lambda}_j^R = O_P(1)$ and $1/\|\widehat{\boldsymbol{\xi}}_{jB}^R\| = O_P(1/\sqrt{c_j})$. The last result is due to the following lemma.

LEMMA A.6. *$\|\widehat{\boldsymbol{\xi}}_{jA}\| = (1 + \bar{c}c_j)^{-1/2} + O_P(\lambda_j^{-1}\sqrt{p/n} + c_jn^{-1/2})$ and $\|\widehat{\boldsymbol{\xi}}_{jB}\| = (\frac{\bar{c}c_j}{1+\bar{c}c_j})^{1/2} + O_P(\sqrt{1/\lambda_j} + \sqrt{c_j}n^{-1/2})$.*

We claim $I = O_P(\sqrt{n/p}) + o_P(n^{-1/2})$. Actually from the proof of Lemma A.6, we have

$$\widehat{\lambda}_j\|\widehat{\boldsymbol{\xi}}_{jB}\|^2/\lambda_j = \bar{c}c_j + O_P(\lambda_j^{-1}\sqrt{p/n}) + o_P(c_jn^{-1/2}).$$

Then some elementary calculation gives the rate of I. Therefore, $\|\mathbf{D}_0\widehat{\boldsymbol{\xi}}_{jB}/\|\widehat{\boldsymbol{\xi}}_{jB}\| - \mathbf{h}_0\| = O_P(\sqrt{n/p}) + o_P(n^{-1/2})$. The conclusion (3.4) follows.

To prove the max norm bound (3.5) of $\|\widehat{\boldsymbol{\xi}}_{jB}\|_{\max}$, we first show $\|\mathbf{h}_0\|_{\max} = O_P(\sqrt{\log p/p})$. Recall that $\mathbf{h}_0$ is uniformly distributed on the unit sphere of dimension $p-m$. This follows easily from its normal representation. Let $\mathbf{G}$ to be $(p-m)$-dimensional multivariate standard normal distributed, then $\mathbf{h}_0 \overset{d}{=} \mathbf{G}/\|\mathbf{G}\|$. It then follows

$$\|\mathbf{h}_0\|_{\max} = \max_{i \le p-m} |G_i|/\|\mathbf{G}\| = O_P(\sqrt{\log p/p}).$$

From the derivation above,

$$\|\widehat{\boldsymbol{\xi}}_{jB}\|_{\max} \le \sqrt{\widehat{\lambda}_j^R/\widehat{\lambda}_j}\|\mathbf{D}_0^{-1}\|\|\widehat{\boldsymbol{\xi}}_{jB}^R\|(II + \|\mathbf{h}_0\|_{\max}),$$

which gives $O_P(\sqrt{c_j}(\sqrt{p/(n\lambda_j^2)} + \sqrt{\log p/p})) = O_P(p/(n\lambda_j^{3/2}) + \sqrt{\log p/(n\lambda_j)})$, given the fact that $\|\widehat{\boldsymbol{\xi}}_{jB}^R\| = O_P(\sqrt{c_j})$ by Lemma A.6. Thus, we are done with the second part of the proof.

(iii) The proof for the convergence of $\|\widehat{\boldsymbol{\xi}}_{jA}\|$ and $\|\widehat{\boldsymbol{\xi}}_{jB}\|$ has been given in Lemma A.6. If $m = 1$, the result for $\|\widehat{\boldsymbol{\xi}}_{jA}\|$ directly gives (3.6) with the same rate. For $m > 1$, from Lemma A.6 we have

$$\|\widehat{\boldsymbol{\xi}}_{jA}\|^2 = (1 + \bar{c}c_j)^{-1} + O_P\Big(\sqrt{p/(n\lambda_j^2)} + c_j n^{-1/2}\Big).$$

On the other hand, from Theorem 3.2(i), $\widehat{\xi}_{jk}^2 = O_P(p/(n\lambda_j^2) + 1/n)$ for $k \ne j \le m$. So $\widehat{\xi}_{j1}^2 = (1 + \bar{c}c_j)^{-1} + O_P(\sqrt{p/(n\lambda_j^2)} + c_j n^{-1/2} + 1/n)$, which implies (3.6). □

Acknowledgments. We would like to acknowledge the Associate Editor and the anonymous reviewers for their constructive comments which lead to a better presentation of the current paper.

SUPPLEMENTARY MATERIAL

Technical proofs [Wang and Fan (2017)] (DOI: 10.1214/16-AOS1487SUPP; .pdf). This document contains technical lemmas for Section 3 and the comparison of assumptions and theoretical proofs for Section 4.

REFERENCES

AGARWAL, A., NEGAHBAN, S. and WAINWRIGHT, M. J. (2012). Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions. *Ann. Statist.* **40** 1171–1197. MR2985947

AMINI, A. A. and WAINWRIGHT, M. J. (2008). High-dimensional analysis of semidefinite relaxations for sparse principal components. In *Information Theory*, 2008. *ISIT* 2008. *IEEE International Symposium on* 2454–2458. IEEE, New York.

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

ANTONIADIS, A. and FAN, J. (2001). Regularization of wavelet approximations. *J. Amer. Statist. Assoc.* **96** 939–967. MR1946364

BAI, Z. D. (1999). Methodologies in spectral analysis of large-dimensional random matrices, a review. *Statist. Sinica* **9** 611–677. MR1711663

BAI, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica* **71** 135–171. MR1956857

BAI, J. and NG, S. (2002). Determining the number of factors in approximate factor models. *Econometrica* **70** 191–221. MR1926259

BAI, Z. and SILVERSTEIN, J. W. (2010). *Spectral Analysis of Large Dimensional Random Matrices*, 2nd ed. Springer, Berlin. MR2567175

BAI, Z. and YAO, J. (2012). On sample eigenvalues in a generalized spiked population model. *J. Multivariate Anal.* **106** 167–177. MR2887686

BAI, Z. D. and YIN, Y. Q. (1993). Limit of the smallest eigenvalue of a large-dimensional sample covariance matrix. *Ann. Probab.* **21** 1275–1294. MR1235416

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

BENAYCH-GEORGES, F. and NADAKUDITI, R. R. (2011). The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices. *Adv. Math.* **227** 494–521. MR2782201

BERTHET, Q. and RIGOLLET, P. (2013). Optimal detection of sparse principal components in high dimension. *Ann. Statist.* **41** 1780–1815. MR3127849

BICKEL, P. J. and LEVINA, E. (2008). Covariance regularization by thresholding. *Ann. Statist.* **36** 2577–2604. MR2485008

BIRNBAUM, A., JOHNSTONE, I. M., NADLER, B. and PAUL, D. (2013). Minimax bounds for sparse PCA with noisy high-dimensional data. *Ann. Statist.* **41** 1055–1084. MR3113803

CAI, T., FAN, J. and JIANG, T. (2013). Distributions of angles in random packing on spheres. *J. Mach. Learn. Res.* **14** 1837–1864. MR3104497

CAI, T., MA, Z. and WU, Y. (2015). Optimal estimation and rank detection for sparse spiked covariance matrices. *Probab. Theory Related Fields* **161** 781–815. MR3334281

CANDÈS, E. J., LI, X., MA, Y. and WRIGHT, J. (2011). Robust principal component analysis? *J. ACM* **58** Art. 11, 37. MR2811000

CHAMBERLAIN, G. and ROTHSCHILD, M. (1983). Funds, factors, and diversification in arbitrage pricing models. *Econometrica* **51** 1305–1324.

CHANDRASEKARAN, V., SANGHAVI, S., PARRILO, P. A. and WILLSKY, A. S. (2011). Rank-sparsity incoherence for matrix decomposition. *SIAM J. Optim.* **21** 572–596. MR2817479

CHEN, K. H. and SHIMERDA, T. A. (1981). An empirical analysis of useful financial ratios. *Financ. Manag.* 51–60.

DAVIDSON, K. R. and SZAREK, S. J. (2001). Local operator theory, random matrices and Banach spaces. In *Handbook of the Geometry of Banach Spaces* **I** 317–366. North-Holland, Amsterdam. MR1863696

DAVIS, C. and KAHAN, W. M. (1970). The rotation of eigenvectors by a perturbation. III. *SIAM J. Numer. Anal.* **7** 1–46. MR0264450

DE MOL, C., GIANNONE, D. and REICHLIN, L. (2008). Forecasting using a large number of predictors: Is Bayesian shrinkage a valid alternative to principal components? *J. Econometrics* **146** 318–328. MR2465176

DONOHO, D. L., GAVISH, M. and JOHNSTONE, I. M. (2014). Optimal shrinkage of eigenvalues in the spiked covariance model. Preprint. Available at arXiv:1311.0851.

FAN, J., FAN, Y. and LV, J. (2008). High dimensional covariance matrix estimation using a factor model. *J. Econometrics* **147** 186–197. MR2472991

FAN, J. and HAN, X. (2013). Estimation of false discovery proportion with unknown dependence. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* To appear. Available at arXiv:1305.7007.

FAN, J., HAN, X. and GU, W. (2012). Estimating false discovery proportion under arbitrary covariance dependence. *J. Amer. Statist. Assoc.* **107** 1019–1035. MR3010887

FAN, J., LIAO, Y. and MINCHEVA, M. (2013). Large covariance estimation by thresholding principal orthogonal complements. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **75** 603–680. MR3091653

FAN, J., LIAO, Y. and SHI, X. (2015). Risks of large portfolios. *J. Econometrics* **186** 367–387. MR3343792

FAN, J., LIAO, Y. and WANG, W. (2016). Projected principal component analysis in factor models. *Ann. Statist.* **44** 219–254. MR3449767

FAN, J., XUE, L. and YAO, J. (2015). Sufficient forecasting using factor models. Preprint. Available at arXiv:1505.07414.

FAN, J., LIU, H., WANG, W. and ZHU, Z. (2016). Heterogeneity adjustment with applications to graphical model inference. Preprint. Available at arXiv:1602.05455.

HALL, P., MARRON, J. S. and NEEMAN, A. (2005). Geometric representation of high dimension, low sample size data. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 427–444. MR2155347

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

KOLTCHINSKII, V. and LOUNICI, K. (2014). Asymptotics and concentration bounds for bilinear forms of spectral projectors of sample covariance. Preprint. Available at arXiv:1408.4643.

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

LANDGREBE, J., WURST, W. and WELZL, G. (2002). Permutation-validated principal components analysis of microarray data. *Genome Biol.* **3** 1–11.

LEE, S., ZOU, F. and WRIGHT, F. A. (2010). Convergence and prediction of principal component scores in high-dimensional settings. *Ann. Statist.* **38** 3605–3629. MR2766862

LEEK, J. T. and STOREY, J. D. (2008). A general framework for multiple testing dependence. *Proc. Natl. Acad. Sci. USA* **105** 18718–18723.

LEEK, J. T., SCHARPF, R. B., BRAVO, H. C., SIMCHA, D., LANGMEAD, B., JOHNSON, W. E., GEMAN, D., BAGGERLY, K. and IRIZARRY, R. A. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.* **11** 733–739.

MA, Z. (2013). Sparse principal component analysis and iterative thresholding. *Ann. Statist.* **41** 772–801. MR3099121

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

PESARAN, M. H. and ZAFFARONI, P. (2008). Optimal asset allocation with factor models for large portfolios.

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. *Nat. Genet.* **38** 904–909.

RINGNÉR, M. (2008). What is principal component analysis? *Nat. Biotechnol.* **26** 303–304.

ROTHMAN, A. J., LEVINA, E. and ZHU, J. (2009). Generalized thresholding of large covariance matrices. *J. Amer. Statist. Assoc.* **104** 177–186. MR2504372

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

STOCK, J. H. and WATSON, M. W. (2002). Forecasting using principal components from a large number of predictors. *J. Amer. Statist. Assoc.* **97** 1167–1179. MR1951271

THOMAS, C. G., HARSHMAN, R. A. and MENON, R. S. (2002). Noise reduction in BOLD-based fMRI using component analysis. *Neuroimage* **17** 1521–1537.

VERSHYNIN, R. (2010). Introduction to the non-asymptotic analysis of random matrices. Preprint. Available at arXiv:1011.3027.

VU, V. Q. and LEI, J. (2012). Minimax rates of estimation for sparse PCA in high dimensions. In *AISTATS* **15** 1278–1286.

WANG, W. and FAN, J. (2017). Supplement to "Asymptotics of empirical eigenstructure for high dimensional spiked covariance." DOI:10.1214/16-AOS1487SUPP.

YAMAGUCHI-KABATA, Y., NAKAZONO, K., TAKAHASHI, A., SAITO, S., HOSONO, N., KUBO, M., NAKAMURA, Y. and KAMATANI, N. (2008). Japanese population structure, based on SNP genotypes from 7003 individuals compared to other ethnic groups: Effects on population-based association studies. *Am. J. Hum. Genet.* **83** 445–456.

YATA, K. and AOSHIMA, M. (2012). Effective PCA for high-dimension, low-sample-size data with noise reduction via geometric representations. *J. Multivariate Anal.* **105** 193–215. MR2877512

YATA, K. and AOSHIMA, M. (2013). PCA consistency for the power spiked model in high-dimensional settings. *J. Multivariate Anal.* **122** 334–354. MR3189327

DEPT. OF OPERATIONS RESEARCH
& FINANCIAL ENGINEERING
SHERRERD HALL
PRINCETON UNIVERSITY
PRINCETON, NEW JERSEY 08544
USA
E-MAIL: weichenw@princeton.edu

DEPT. OF OPERATIONS RESEARCH
& FINANCIAL ENGINEERING
SHERRERD HALL
PRINCETON UNIVERSITY
PRINCETON, NEW JERSEY 08544
USA
AND
SCHOOL OF DATA SCIENCE
FUDAN UNIVERSITY
SHANGHAI
CHINA
E-MAIL: jqfan@princeton.edu