LIMITING LAWS FOR DIVERGENT SPIKED EIGENVALUES AND LARGEST NON-SPIKED EIGENVALUE OF SAMPLE COVARIANCE MATRICES

By T. Tony Cai[*,‡] Xiao Han[†,§]
and Guangming Pan[†,§]

University of Pennsylvania[‡] and Nanyang Technological University[§]

We study the asymptotic distributions of the spiked eigenvalues and the largest nonspiked eigenvalue of the sample covariance matrix under a general covariance model with divergent spiked eigenvalues, while the other eigenvalues are bounded but otherwise arbitrary. The limiting normal distribution for the spiked sample eigenvalues is established. It has distinct features that the asymptotic mean relies on not only the population spikes but also the nonspikes and that the asymptotic variance in general depends on the population eigenvectors. In addition, the limiting Tracy-Widom law for the largest nonspiked sample eigenvalue is obtained.

Estimation of the number of spikes and the convergence of the leading eigenvectors are also considered. The results hold even when the number of the spikes diverges. As a key technical tool, we develop a Central Limit Theorem for a type of random quadratic forms where the random vectors and random matrices involved are dependent. This result can be of independent interest.

1. Introduction. Covariance matrix plays a fundamental role in multivariate analysis and high-dimensional statistics. There has been significant recent interest in studying the properties of the leading eigenvalues and their corresponding eigenvectors of the sample covariance matrix, especially in the high-dimensional setting. See, for example, [29, 24, 2, 36, 32, 30, 31, 12, 14, 15]. These problems are not only of interest in their own right they also have close connections to other important statistical problems such as principal component analysis and testing for the covariance structure of high-dimensional data.

Principal component analysis (PCA) is a widely used technique for a range of purposes, including dimension reduction, data visualization, clustering,

and feature extraction [1, 23]. PCA is particularly well suited for the settings where the signal of interest lies in a much lower dimensional subspace and it has been applied in a broad range of fields such as genomics, image recognition, data compression, and financial econometrics. For example, the widely used factor models in financial econometrics typically assume that a small number of unknown common factors drive the asset returns [17]. In PCA, the leading eigenvalues and eigenvectors of the population covariance matrix need to be estimated from data and are conventionally estimated by their empirical counterparts. It is thus important to understand the spectral properties of the sample covariance matrix.

1.1. *The Problem.* To be concrete, consider the data matrix $\mathbf{Y} = \mathbf{\Gamma X}$ where $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_n)$ is a $(p + l) \times n$ random matrix whose entries are independent with zero mean and unit variance and $\mathbf{\Gamma}$ is a $p \times (p + l)$ deterministic matrix with $l/p \to 0$. Let $\mathbf{\Sigma} = \mathbf{\Gamma \Gamma}^\mathsf{T}$ be the population covariance matrix. The sample covariance matrix is defined as

$$(1.1) \qquad \mathbf{S}_n = \frac{1}{n}\mathbf{Y}\mathbf{Y}^\mathsf{T} = \frac{1}{n}\mathbf{\Gamma X X}^\mathsf{T}\mathbf{\Gamma}^\mathsf{T}.$$

Denote the singular value decomposition (SVD) of $\mathbf{\Gamma}$ by

$$(1.2) \qquad \mathbf{\Gamma} = \mathbf{V}\Lambda^{\frac{1}{2}}\mathbf{U},$$

where $\mathbf{V}$ and $\mathbf{U}$ are $p \times p$ and $p \times (p + l)$ orthogonal matrices respectively ($\mathbf{V}\mathbf{V}^\mathsf{T} = \mathbf{U}\mathbf{U}^\mathsf{T} = \mathbf{I}$), and $\mathbf{\Lambda}$ is a diagonal matrix consisting in descending order of the eigenvalues $\mu_1 \geq \cdots \geq \mu_p$ of $\mathbf{\Sigma}$.

In statistical applications such as PCA, one is most interested in the setting where there is a clear separation between a few leading eigenvalues and the rest. In this case, the leading principal components account for a large proportion of the total variability of the data. We consider in the present paper the setting where there are K spiked eigenvalues that are separated from the rest. More specifically, we assume that $\mu_1 \geq \cdots \geq \mu_K$ tend to infinity, while the other eigenvalues $\mu_{K+1} \geq \cdots \geq \mu_p$ are bounded but otherwise arbitrary. Write

$$(1.3) \qquad \mathbf{\Lambda} = \begin{pmatrix} \mathbf{\Lambda}_S & 0 \\ 0 & \mathbf{\Lambda}_P \end{pmatrix},$$

where $\mathbf{\Lambda}_S = \mathsf{diag}(\mu_1, ..., \mu_K)$ and $\mathbf{\Lambda}_P = \mathsf{diag}(\mu_{K+1}, ..., \mu_p)$.

A typical example of (1.3) is the factor model

$$(1.4) \qquad \mathbf{Y} = \mathbf{\Lambda}_1\mathbf{F} + \mathbf{T}\mathbf{Z} = (\mathbf{\Lambda}_1 \quad \mathbf{T}) \begin{pmatrix} \mathbf{F} \\ \mathbf{Z} \end{pmatrix}$$

where $\mathbf{\Lambda}_1$ is $p \times K$-dimensional factor loading, $\mathbf{F}$ is the corresponding $K \times n$ factor, $\mathbf{T}$ is $p \times p$ matrix and $\mathbf{Z}$ is the idiosyncratic noise matrix. A common assumption is that the singular values of the factor component $\mathbf{\Lambda}_1\mathbf{F}$ are significantly larger than those of the noise component (otherwise the signals are overwhelmed by noise). Indeed, [33] considered the weak factor model to test the number of factors, where the leading eigenvalues contributed by the factor component are of order p^θ for some $\theta \in (0,1)$. [4] and [21] assume that the leading eigenvalues of the pervasive factor model are of order p. Here $\mathbf{\Gamma} = (\mathbf{\Lambda}_1 \quad \mathbf{T})$ is not a square matrix, and thus it is necessary to consider the setting where $\mathbf{\Gamma}$ is rectangular.

A second example is the covariance matrix $\mathbf{\Sigma}$ used in the intraclass correlation model, where the covariance matrix is of the form

$$\mathbf{\Sigma} = (1 - \rho)\mathbf{I}_p + \rho\mathbf{e}\mathbf{e}^\mathsf{T}.$$

Here $\mathbf{I}_p$ is a $p \times p$ identity matrix, $\mathbf{e} = (1, 1, ..., 1)^\mathsf{T}$ and $0 < \rho < 1$. It is easy to see that the leading eigenvalue of $\mathbf{\Sigma}$ is $p\rho + (1-\rho)$, while the other eigenvalues are equal to $(1 - \rho)$, i.e. $K = 1$, $\mathbf{\Lambda_S} = p\rho + (1 - \rho)$ and $\mathbf{\Lambda_P} = (1 - \rho)\mathbf{I}_{p-1}$ in (1.3). One can refer to [31] for more discussions about this model.

We study in the present paper the asymptotic distributions of the leading eigenvalues and the largest nonspiked eigenvalue of the sample covariance matrix $\mathbf{S}_n$, under the general spiked covariance matrix model given in (1.2) and (1.3) with divergent spiked eigenvalues $\mu_1 \geq \cdots \geq \mu_K$. In many statistical applications, determining the number of principal components is an important problem. In addition, properties of the eigenvectors associated with the spiked eigenvalues are of significant interest. In this paper, we also consider estimation of the number of spikes as well as the convergence of the leading eigenvectors.

The model defined through (1.2) and (1.3) belongs to the class of spiked covariance matrix models. Johnstone [29] was the first to introduce a specific spiked covariance matrix model, where the population covariance matrix is diagonal and is of the form

(1.5) $$\mathbf{\Sigma} = \mathsf{diag}(\mu_1^2, ..., \mu_K^2, 1, ..., 1)$$

with $\mu_1 > \mu_2 \cdots \geq \mu_K > 1$. [29] established the limiting Tracy-Widom distribution for the maximum eigenvalue of the real Wishart matrices when p and n are comparable. The spiked covariance matrix model (1.5) has been extended in various directions. So far the focus has mostly been on the settings of bounded spiked eigenvalues with all the nonspiked eigenvalues being equal to 1. See more discussion in Section 1.3.

1.2. *Our contributions.* In this paper, we first establish the limiting normal distribution for the spiked eigenvalues of the sample covariance matrix

$\mathbf{S}_n$. The limiting distribution has a distinct feature. Unlike in the more conventional settings, the asymptotic variance in general depends on the population eigenvectors. More precisely, the variance of a spiked sample eigenvalue depends on the right singular vector matrix $\mathbf{U}$ defined in the SVD (1.2) (but not the left singular vector matrix $\mathbf{V}$). The limiting distribution of the spiked sample eigenvalues also precisely characterizes the dependence on the corresponding spiked population eigenvalues as well as the nonspiked ones. New technical tools are needed to establish the result. In particular, we develop a Central Limit Theorem (CLT) for a type of random quadratic forms where the random vectors and random matrices involved are dependent. This result can be of independent interest. In addition, we establish the limiting Tracy-Widom law for the largest nonspiked eigenvalue of $\mathbf{S}_n$. We also consider the properties of the leading principal components and show that they are consistent estimators of their population counterparts under the L_2 loss. An important improvement of our paper over many known results in the literature is that our results hold even when the number of the spikes diverges as $n, p \to \infty$, and we allow the nonspiked eigenvalues to be unequal.

The limiting distributions for the spiked eigenvalues and the largest nonspiked eigenvalue have important applications. In particular, based on our theoretical results, we propose an algorithm for estimating the number of the spikes, which is of interest in many statistical applications.

1.3. *Background and related work.* Since the seminal work of Johnstone [29], the special spiked covariance matrix model (1.5) has been studied much further and the model has been extended in several directions. See, for example [2, 3, 8, 13, 30, 31, 36, 38, 41, 14, 15]. We discuss briefly here some of these results. This review is by no means exhaustive.

Paul [36] showed that if $p/n \to \gamma \in (0,1)$ as $n \to \infty$, and the largest eigenvalue μ_1 of $\mathbf{\Sigma}$ satisfies $\mu_1 \leq 1 + \sqrt{\gamma}$, then the leading sample principal eigenvector $\widehat{\mathbf{v}}_1$ is asymptotically almost surely orthogonal to the leading population eigenvector $\mathbf{v}_1$, i.e., $|\mathbf{v}_1'\widehat{\mathbf{v}}_1| \to 0$ almost surely. Thus, in this case, $\widehat{\mathbf{v}}_1$ is not useful at all as an estimate of $\mathbf{v}_1$. Even when $\mu_1 > 1 + \sqrt{\gamma}$, the angle between $\mathbf{v}_1$ and $\widehat{\mathbf{v}}_1$ still does not converge to zero unless $\mu_1 \to \infty$.

Baik and Silverstein [2] considered a case where the covariance matrix

$$(1.6) \qquad\qquad \mathbf{\Sigma} = \mathbf{V} \begin{pmatrix} \Lambda_S & 0 \\ 0 & \mathbf{I} \end{pmatrix} \mathbf{V}^\intercal$$

with Λ_S being a diagonal matrix of fixed rank and $\mathbf{V}$ a unitary matrix. It is shown that the spiked eigenvalues tend to some limits in probability, assuming that the spectral norm of Λ_S is bounded and $\lim_{n\to\infty} \frac{p}{n} = \gamma \in (0, \infty)$. Bai and Yao [8] further showed that the spiked eigenvalues converge in distribution to Gaussian distribution or the eigenvalues of a finite dimensional

matrix with i.i.d. Gaussian entries. Baik, et al. [3] investigated the asymptotic behavior of the largest eigenvalue when the entries of $\mathbf{X}$ follow the standard complex Gaussian distribution and observed a phase transition phenomenon that the asymptotic distribution depends on the scale of the spiked population eigenvalues. El Karoui [20] proved that the largest eigenvalue tends to Type-2 Tracy-Widom distribution for the general $\mathbf{\Sigma}$ without the spiked part, i.e., $\Lambda_S = 0$. Shi [39] discussed asymptotics of the components of the sample eigenvectors corresponding to the sample spiked eigenvalues for $\mathbf{\Sigma} = \begin{pmatrix} \Lambda_S & 0 \\ 0 & \mathbf{I} \end{pmatrix}$ under some moment conditions on $\mathbf{X}$. Recently, Bloemendal et al. [13] obtained the precise large deviation of the spiked eigenvalues and non-spiked eigenvalues under a more general model than (1.6). We should note that the above results except [20] are only for the case of bounded spiked eigenvalues with the nonspiked eigenvalues all being equal to 1.

Jung and Marron [31] and Shen et al. [38] considered the model

$$(1.7) \qquad \mathbf{Y} = \mathbf{V}\Lambda^{\frac{1}{2}}\mathbf{X},$$

where the entries of $\mathbf{X}$ are i.i.d. standard normal random variables, and $\Lambda = \mathsf{diag}(\mu_1, ..., \mu_K, \mu_{K+1}, \cdots, \mu_p)$ is the diagonal matrix consisting of the population eigenvalues, and $\mathbf{V}$ is an orthogonal matrix. [31] and [38] showed the almost sure convergence of the spiked eigenvalues when the spiked population eigenvalues satisfy that $p/(\mu_j n), j = 1, \cdots, K$, tend to nonnegative constants and $\mu_{K+1}, \cdots, \mu_p$ are approximately equal to one. The almost sure convergence of the eigenvectors associated with the spikes is also studied.

Wang and Fan [41] further developed the asymptotic distribution for each λ_j $(j = 1, ..., K)$ of the model (1.7) under a more general setting, which allows $\mu_{K+1}, ..., \mu_p$ to be any bounded number, $p/(\mu_j n)$ to be bounded, $\frac{\mu_j}{\mu_{j+1}} \geq c$ for some constant $c > 1$, $j = 1, \cdots, K$ and the entries of $\mathbf{X}$ to be i.i.d. sub-Gaussian random variables. The asymptotic behaviors of the corresponding eigenvectors are also discussed in [41]. Here we would like to point out that [41] did not provide the limits in probability of the spikes unless $\frac{\sqrt{p}}{\sqrt{n}\mu_j} = o(1)$ and the joint distribution of $\{\lambda_j\}$, $j = 1, ..., K$. To the best of our knowledge, the asymptotic behavior of the spiked eigenvalues for general $\mu_{K+1}, ..., \mu_p$ when $p/(\mu_j n)$, $j = 1, \cdots, K$, converge to positive constants is still open for the model (1.1). More recently, [42] considered a similar spiked model with the population eigenvalues $\mu_j = \alpha_j d^{\alpha_j}, j = 1, \cdots, K$. They proposed a bias corrected estimator of eigenvalues when either $p \to \infty, n \to \infty$ and $\alpha_j > 1/2$ or $p^{2-4\alpha_j}/n \to 0$ and $0 < \alpha_j \leq 1/2$.

Note that [31], [38] and [41] swapped the roles of the sample size n and the dimension p so that they essentially studied the matrix $\mathbf{X}^\intercal\Lambda\mathbf{X}$. This is equivalent to assuming that the population covariance matrix is diagonal.

Indeed, as will be seen later, in general the asymptotic variance of the spiked eigenvalues depends on the population eigenvectors. This phenomenon does not occur under the previously studied model.

As we mentioned before, an important application of (1.1) is high dimensional factor model (1.4). There is a significant interest in the estimation of Λ_1 and $\mathbf{F}$. See for example [5], [6], [35], [40] and [41]. The asymptotic properties of their respective estimators have been investigated in these papers under different conditions. Another important problem is to determine the number of factors K in (1.4). There are several popular procedures available to estimate K, including PC_p and IC_p ([4] and [6]), AIC and BIC ([9]), and the spectral method ([33] and [34]).

1.4. *Organization of the paper.*   The rest of the paper is organized as follows. Section 2 establishes the limiting normal distribution for the spiked eigenvalues and the limiting Tracy-Widom distribution for the largest non-spiked eigenvalue of the sample covariance matrix $\mathbf{S}_n$. An algorithm for identifying the number of spikes is developed in Section 3. Section 4 considers the properties of the principal components and shows that the sample eigenvectors corresponding to the spiked eigenvalues are consistent estimators of the population eigenvectors in terms of the L_2 norm. Most of the results developed for $\mathbf{S}_n$ also hold for the centralized sample covariance matrices and this is discussed in Section 5. Section 6 investigates the numerical performance through simulations and an application of a factor model. The proof of Theorem 2.4 is given in Section 7 and the proof of the other results, including Theorems 2.1-2.3, 2.5 and 4.1, and other technical results, are provided in the supplementary material [16].

2. Asymptotics for Spiked Eigenvalues and Largest Nonspiked Eigenvalue of $\mathbf{S}_n$. We investigate in this section the limiting laws for the leading eigenvalues and the largest nonspiked eigenvalue of the sample covariance matrix $\mathbf{S}_n$ under the general spiked covariance matrix model (1.2) and (1.3) with divergent spiked eigenvalues $\mu_1 \geq \cdots \geq \mu_K$, while the other eigenvalues are bounded but otherwise arbitrary. We begin with the notation that will be used throughout the rest of the paper.

For two sequences of positive numbers a_n and b_n, we write $a_n \gtrsim b_n$ when $a_n \geq c b_n$ for some absolute constant $c > 0$, and $a_n \lesssim b_n$ when $b_n \gtrsim a_n$. Alternatively we denote $a_n \gtrsim b_n$ by $a_n = \Omega(b_n)$. We write $a_n \sim b_n$ when both $a_n \gtrsim b_n$ and $a_n \lesssim b_n$ hold. Moreover, we write $a_n \ll b_n$ when $a_n/b_n \to 0$. Then we say $a_n = O(b_n)$ or $b_n = \Omega(a_n)$. For a sequence of random variables A_n, if A_n converges to b in probability, then we write $A_n \overset{i.p.}{\to} b$. We say an event $\mathcal{A}_n$ holds with high probability if $\mathbb{P}(\mathcal{A}_n) \geq 1 - O(n^{-l})$ for some constant $l > 0$. Denote the j-th largest eigenvalue of a symmetric matrix $\mathbf{M}$ by $\lambda_j(\mathbf{M})$ and the largest singular value by $\|\mathbf{M}\|$. Set $\|\mathbf{M}\|_F = \sqrt{tr(\mathbf{MM^\intercal})}$.

For simplicity, denote by $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_K \geq \cdots \geq \lambda_p$ the ordered eigenvalues of the sample covariance matrix $\mathbf{S}_n$, and denote by $\mu_1 \geq \mu_2 \geq ... \geq \mu_K \geq \cdots \geq \mu_p$ the ordered eigenvalues of the population covariance matrix $\mathbf{\Sigma}$. Throughout this paper c and C are constants that may vary from place to place.

To investigate the sample covariance matrix $\mathbf{S}_n = \frac{1}{n}\mathbf{\Gamma}\mathbf{X}\mathbf{X}^\mathsf{T}\mathbf{\Gamma}^\mathsf{T}$ with the population covariance matrix $\mathbf{\Sigma}$ specified in (1.2) and (1.3) we make the following assumptions.

ASSUMPTION 1. $\{\mathbf{x}_j = (\mathbf{x}_{1j}, \cdots, \mathbf{x}_{p+l,j})^\mathsf{T}, j = 1, ..., n\}$ are i.i.d. random vectors. $\{\mathbf{x}_{ij}: i = 1, ..., p+l, j = 1, ..., n\}$ are independent random variables such that $\mathbb{E}\mathbf{x}_{ij} = 0$, $\mathbb{E}|\mathbf{x}_{ij}|^2 = 1$, $\mathbb{E}|\mathbf{x}_{ij}|^4 = \gamma_{4i}$ and $\sup_i \gamma_{4i} \leq C$.

ASSUMPTION 2. $p \gtrsim n$ and the K largest population eigenvalues μ_i are such that $d_i \equiv \frac{p}{n\mu_i} \to 0$, $i = 1, 2, ..., K$. And for $i = K+1, ..., p$, μ_i are bounded by C. Moreover, $\frac{K}{n^{1/6}} \to 0$ and $K^2 d_K \to 0$.

Assumption 2′. $\frac{p}{n} \to 0$, $\mu_i \gg 1$, $i = 1, ..., K$ and $K \ll \min\{p, n^{1/6}\}$.

That is to say, we focus on the matrix $\mathbf{S}_n$ with the population covariance matrix $\mathbf{\Sigma} = \mathbf{\Gamma}\mathbf{\Gamma}^\mathsf{T}$ satisfying Assumption 2 or 2′.

REMARK 1. Here the requirement about the order of K comes from the fact that the study of the spiked eigenvalues essentially boils down to an $K \times K$ matrix. In order to allow K to tend to infinity we have to analyze the convergence rate of each entry of the matrix. One can see (10.15) in [16] for more details.

Note that we do not assume that p and n are of the same order. The following theorems hold either under Assumption 2 or Assumption 2′ except Theorem 2.5. We only give the proofs under Assumption 2. The proofs under Assumption 2′ are similar and thus we omit them.

ASSUMPTION 3. There exists a positive constant c not depending on n such that $\frac{\mu_{i-1}}{\mu_i} \geq c > 1$, $i = 1, 2, ..., K$.

2.1. *Asymptotic behavior of the spiked sample eigenvalues.* Our first result gives the limits in probability for the spiked eigenvalues of $\mathbf{S}_n$, $\lambda_1 \geq ... \geq \lambda_K$.

THEOREM 2.1. *Suppose that Assumption 1 holds. Moreover, either Assumption 2 or Assumption 2′ holds. Then*

$$(2.1) \qquad \frac{\lambda_i}{\mu_i} - 1 = O_p(d_i + \frac{K^4}{n} + \frac{1}{\mu_i}),$$

uniformly for all $i = 1, ..., K$.

REMARK 2. As mentioned in the introduction, PCA is an important statistical tool for analyzing high-dimensional data. Several recent results on high-dimensional PCA are quite relevant to Theorem 2.1. Recently [9] considered AIC and BIC criteria for selecting the number of significant components in high dimensional PCA when p and n are comparable. Comparing to the paper [9], Theorem 2.1 here covers Lemma 2.2(i) of [9] and we allow K to tend to infinity. Their assumption $\mu_{K+1} = \cdots = \mu_p = 1$ is also relaxed to bounded eigenvalues here. In addition, checking the proof of Theorems 3.3 and 3.4 of [9], we find that for general population covariance matrices, their criteria $\tilde{A}_j$ and $\tilde{B}_j$ for estimating the number of spikes may not work since the proof highly depends on the assumption $\mu_{K+1} = \cdots = \mu_p = 1$, as demonstrated in Table 4 given in Section 6. In addition, Theorem 2.1 also covers part of Theorem 3.1 in [38] where it assumes normality for the data.

Note that $\frac{\lambda_i}{\mu_i} \overset{i.p.}{\to} 1$ does not imply that λ_i is a good estimator of μ_i due to the fact that μ_i tends to infinity. Moreover, Theorem 2.1 does not precisely characterize how the nonspiked population eigenvalues affect the spiked sample eigenvalues. To see this, it is helpful to make a comparison with the conventional setting studied in [2]. Consider the model (1.6) and recall the assumptions of [2] that $1 + \sqrt{\gamma} < \mu_i = O(1)$ and $\gamma = \lim_{n\to\infty} \frac{p}{n} \in (0, \infty)$. It was shown in [2] that

$$(2.2) \qquad\qquad \lambda_i \overset{a.s.}{\to} \mu_i + \frac{\gamma\mu_i}{\mu_i - 1}.$$

So the effect of the population eigenvalues on the corresponding sample eigenvalues can be precisely characterized in the setting considered in [2]. On the other hand, one cannot see the effect of the nonspiked population eigenvalues on the spiked sample eigenvalues from (2.2). Note that if there are no spikes, then all the sample eigenvalues are not bigger than $(1+\sqrt{\gamma})^2 + c$ for any positive constant c with probability one. When there are sufficiently large spikes, the sample spikes are pulled outside of the boundary $(1 + \sqrt{\gamma})^2$ due to the population spikes with probability one. Moreover, (2.2) precisely quantifies the effect of the population spike. In view of this, one would ask whether there is a similar phenomenon for unbounded spikes. Indeed, it is natural to imagine that for the case $\mu_i \to \infty$, the term $\frac{\gamma\mu_i}{\mu_i-1}$ will not disappear and thus one needs to subtract it from λ_i in order to obtain the CLT. Surprisingly, a more precise limit of λ_i turns out to be determined not only by μ_i but also the nonspiked eigenvalues. This is very different from (2.2) and can be seen clearly from (2.9) below.

We now characterize how the population eigenvalues including spiked eigenvalues and non-spiked eigenvalues affect the sample spiked eigenval-

ues. To this end, corresponding to (1.3), partition $\mathbf{U}$ as $\mathbf{U} = \begin{pmatrix} \mathbf{U}_1 \\ \mathbf{U}_2 \end{pmatrix}$, where $\mathbf{U}_1$ is the $K \times (p + l)$ submatrix of $\mathbf{U}$, and define

$$(2.3) \qquad \boldsymbol{\Sigma}_1 = \mathbf{U}_2^\mathsf{T} \boldsymbol{\Lambda}_P \mathbf{U}_2.$$

For any distribution function H, its Stieltjes transform is defined by

$$m_H(z) = \int \frac{1}{\lambda - z} dH(\lambda), \quad \text{for all } z \in \mathbb{C}^+.$$

For any $\theta \neq 0$, let $\tilde{m}_\theta(z)$ be the unique solution to the following equation

$$(2.4) \quad \tilde{m}_\theta(z) = -\left(z - \frac{1}{n} tr \left[(\mathbf{I} + \tilde{m}_\theta(z) \frac{\boldsymbol{\Sigma}_1}{\theta})^{-1} \frac{\boldsymbol{\Sigma}_1}{\theta} \right] \right)^{-1}, \quad z \in \mathbb{C}^+,$$

where $\mathbb{C}^+$ denotes the complex upper half plane and $\boldsymbol{\Sigma}_1$ is defined in (2.3). Indeed, as will be seen, for $\theta \gg \frac{p}{n}$,

$$\tilde{m}_\theta(z) + \frac{1}{n} \mathbb{E} tr(z\mathbf{I} - \frac{1}{n\theta} \mathbf{X}^\mathsf{T} \boldsymbol{\Sigma}_1 \mathbf{X})^{-1} \to 0,$$

for $z \in \mathbb{C}^+$ by a slight modification of the proof of Appendix 7.2. One can also refer to (1.6) of [11] or (6.12)-(6.15) of [7] for (2.4). One may see below that $\tilde{m}_\theta(z)$ describes the collective contribution of the nonspiked eigenvalues of $\boldsymbol{\Sigma}$ to the spiked sample eigenvalues.

By (2.4), we set θ_i to be the solution to

$$(2.5) \qquad \tilde{m}_{\theta_i}(1) + \frac{\theta_i}{\mu_i} = 0,$$

where $\tilde{m}_{\theta_i}(1) = \lim\limits_{z \in \mathbb{C}^+ \to 1} \tilde{m}_{\theta_i}(z)$. It turns out that θ_i instead of μ_i is the more precise limit of the spiked sample eigenvalues λ_i. From (2.5) one can see that θ_i depends on μ_i as well as the nonspiked part $\boldsymbol{\Sigma}_1$. Indeed, this point can be seen more clearly from (2.9) below. A similar dependence of θ_i on μ_i as well as the nonspiked part $\boldsymbol{\Sigma}_1$ has appeared in [35], where a different factor model is considered.

ASSUMPTION 4. Assume that the following limits exist:

$$\sigma_i = \lim\limits_{p \to \infty} \sqrt{\sum_{s=1}^{p+l} (\gamma_{4s} - 3) u_{is}^4 + 2}, \quad \sigma_{ij} = \lim\limits_{p \to \infty} \sum_{s=1}^{p+l} (\gamma_{4s} - 3) u_{is}^2 u_{js}^2, \ i, j \leq K.$$

REMARK 3. If $\max_{1 \leq t \leq K, 1 \leq s \leq p+l} |u_{ts}| \to 0$ then $\sigma_i = \sqrt{2}$ and $\sigma_{ij} = 0$. Furthermore, if $\mathbf{U}_1$ is a random unitary matrix independent of $\mathbf{X}$ with the

condition $\max_{1 \leq t \leq K, 1 \leq s \leq p+l} |u_{ts}| \xrightarrow{i.p.} 0$, then

$$|\sum_{s=1}^{p+l}(\gamma_{4s}-3)u_{is}^2 u_{js}^2| \leq \max_s |\gamma_{4s}-3| \max_{1 \leq t \leq K, 1 \leq s \leq p+l} |u_{ts}|^2 \sum_{s=1}^{p+l}|u_{js}|^2 \xrightarrow{i.p.} 0,$$

$$\sum_{s=1}^{p+l}(\gamma_{4s}-3)u_{is}^4 \xrightarrow{i.p.} 0.$$

Therefore $\sigma_i \xrightarrow{i.p.} \sqrt{2}$ and $\sigma_{ij} \xrightarrow{i.p.} 0$. In addition, when $\mathbf{U}$ is haar distributed, then $\sigma_i \xrightarrow{i.p.} \sqrt{2}$ and $\sigma_{ij} \xrightarrow{i.p.} 0$(e.g. see [28]).

We are ready to state the asymptotic distribution of the spiked eigenvalues of $\mathbf{S}_n$. Let $\mathbf{u}_i^{\mathsf{T}}$ be the i-th row of $\mathbf{U}$ with u_{ij} being the (i,j)-th entry of $\mathbf{U}$.

THEOREM 2.2. *Suppose that Assumptions 1, 3, and 4 hold. Moreover, either Assumption 2 or Assumption 2$'$ hold. Then for all $i = 1, 2, ..., K$,*

$$(2.6) \qquad \sqrt{n}\frac{\lambda_i - \theta_i}{\theta_i} \xrightarrow{D} N\left(0, \sigma_i^2\right).$$

Moreover, for any fixed $r \geq 2$

$$(2.7) \qquad \left(\sqrt{n}\frac{\lambda_1 - \theta_1}{\theta_1}, ..., \sqrt{n}\frac{\lambda_r - \theta_r}{\theta_r}\right) \xrightarrow{D} N\left(0, \mathbf{\Sigma}^{(r)}\right),$$

where $\mathbf{\Sigma}^{(r)} = (\mathbf{\Sigma}_{ij}^{(r)})$ with

$$\mathbf{\Sigma}_{ij}^{(r)} = \begin{cases} \sigma_i^2, & i = j \\ \sigma_{ij}, & i \neq j, \end{cases}$$

It follows from (2.4) and (2.5) that $\tilde{m}_{\theta_i}(1) \to -1$. Therefore $\frac{\theta_i}{\mu_i} \to 1$. However, we can not replace θ_i by μ_i in (2.7) directly because the convergence rate of $\frac{\theta_i}{\mu_i}$ to 1 is unknown. Indeed, by (2.4), we have

$$(2.8) \qquad \theta = -\frac{\theta}{\tilde{m}_\theta(1)} + \frac{p-K}{n}\int \frac{t dF_{\mathbf{\Lambda}_P}(t)}{1 + t\tilde{m}_\theta(1)\theta^{-1}},$$

where $F_{\mathbf{\Lambda}_P}$ is the empirical spectral distribution of $\mathbf{\Lambda}_P$. Here for any $n \times n$ symmetric matrix $\mathbf{A}$ with real eigenvalues, the empirical spectral distribution (ESD) of $\mathbf{A}$ is defined as

$$F_{\mathbf{A}}(x) = \frac{1}{n}\sum_{i=1}^{n} I_{\{\lambda_i(\mathbf{A}) \leq x\}}.$$

Together with (2.5), we conclude that

$$(2.9) \qquad \theta_i = \mu_i(1 + \frac{p-K}{n} \int \frac{t}{\mu_i - t} dF_{\boldsymbol{\Lambda}_P}(t)).$$

By the Taylor's expansion we have

$$(2.10) \qquad \frac{\theta_i}{\mu_i} = 1 + ff_i + O(\frac{p}{n\mu_i^2}),$$

where

$$f = \frac{1}{p-K} \sum_{j=K+1}^{p} \mu_j \quad \text{and} \quad f_i = \frac{p-K}{n\mu_i}.$$

In particular, for the special case $\mu_{K+1} = ... = \mu_p = 1$, (2.9) yields that

$$(2.11) \qquad \theta_i = \mu_i(1 + \frac{p-K}{n(\mu_i - 1)}).$$

It is interesting to note that, although here the spiked eigenvalues $\mu_1, \cdots, \mu_K$ are divergent, this is consistent with the right hand side of (2.2), which is for the conventional setting of bounded spiked eigenvalues. It then follows from (2.10) that

$$(2.12) \qquad \sqrt{n}\Big(\frac{\lambda_i}{\mu_i} - 1 - ff_i + O(\frac{p}{n\mu_i^2})\Big) \xrightarrow{D} N\left(0, \sigma_i^2\right).$$

REMARK 4. We note that Assumption 4 is not needed if we consider the individual asymptotic distribution of the spiked sample eigenvalues. To see this, it suffices to normalize $(\lambda_i - \theta_i)/\theta_i$ by $\sigma_i = \sqrt{\sum_{j=1}^{p+l}(\gamma_{4j} - 3)u_{ij}^4 + 2}$. Moreover, the joint distribution of $\frac{\lambda_i - \theta_i}{\sigma_i \theta_i}$, $i = 1, ..., r$ tends to the normal distribution with the covariance matrix being the correlation matrix corresponding to $\boldsymbol{\Sigma}^{(r)}$.

REMARK 5. It is helpful to compare the above theorem with Theorem 3.1 of [41]. Besides the difference between the models in (1.2) and (1.7), one of the key differences is that σ_i^2 in (2.12) depends on the entries of the eigenvector matrix $\mathbf{U}$ while the variance in Theorem 3.1 of [41] does not depend on it. This is due to the fact that [41] assumes that $\mathbf{U} = \mathbf{I}$. Secondly, Theorem 3.1 of [41] involves $O_p(\frac{\sqrt{p}}{\sqrt{n}\mu_i})$ which reduces to $O(\frac{p}{n\mu_i^2})$ (essentially $O(\frac{1}{\mu_i})$) in (2.12) by dropping the additional $\frac{\sqrt{p}}{\sqrt{n}}$. Thirdly we also allow K to diverge. Fourthly [41] assumes x_{ij} to be subGaussian random variables while Theorem 2.2 holds under the bounded fourth moment assumption.

REMARK 6. We would compare the above theorem with Theorem 2 in [5] which deals with the factor model (1.4). Recall the estimator of Λ_1 in [5], i.e. the estimator $\tilde{\Lambda}_1$ is such that $\tilde{\Lambda}_1^T \tilde{\Lambda}_1 = diag(\lambda_1, ..., \lambda_r)$ essentially. Hence, we have proved the central limit theorem for $\tilde{\Lambda}_1^T \tilde{\Lambda}_1$ (when r is fixed). While Theorem 2 of [5] shows the CLT of $(\tilde{\Lambda}_1)_i$, where $(\tilde{\Lambda}_1)_i$ is the i-th row of $\tilde{\Lambda}_1$. Checking the dimension condition in Theorem 2 of [5], the CLT holds for $d_i \to 0$ and $\mu_i \sim n$, which is a special case of Assumption 2.

In view of (2.10) we need to estimate f and f_i in practice. A natural estimator of f_i is $\frac{p-K}{n\lambda_i}$ by Theorem 2.1. For f, one can use

$$(2.13) \qquad \hat{f} = \frac{\frac{1}{n} tr(\Gamma \mathbf{X}\mathbf{X}^\mathsf{T}\Gamma) - \sum_{i=1}^{K}\lambda_i}{p - K - pK/n}$$

which was proposed in [41]. When $p \sim n$, by Proposition 1 in the next section, K can be estimated accurately.

Moreover, Theorem 2.2 can be extended to the case when the population eigenvalues μ_i have multiplicity more than one.

ASSUMPTION 5. Suppose that $K \ll n^{1/6}$, $\alpha_\mathcal{L} = \mu_K = ... = \mu_{K-n_\mathcal{L}} < \alpha_{\mathcal{L}-1} = \mu_{K-n_\mathcal{L}+1}... < \alpha_1 = \mu_{n_1} = ... = \mu_1$, and there exists a constant c such that $\frac{\alpha_{i-1}}{\alpha_i} \geq c > 1$, $i = 1, 2, ..., \mathcal{L}$. Moreover, $n_1,..., n_\mathcal{L}$ are finite.

ASSUMPTION 6. Suppose that the following limits exist

$$G(r_i, k_1, k_2, l_1, l_2) = \lim_{n\to\infty} n^2 \times \text{Cov}(\mathbf{u}_{r_i+k_1}^\mathsf{T}\mathbf{x}_1\mathbf{u}_{r_i+l_1}^\mathsf{T}\mathbf{x}_1, \mathbf{u}_{r_i+k_2}^\mathsf{T}\mathbf{x}_1\mathbf{u}_{r_i+l_2}^\mathsf{T}\mathbf{x}_1).$$

If either the fourth moments $\gamma_{4s} = 3$, $s = 1, ..., p + l$ or the entries of the population eigenvectors satisfy $\min\limits_{r\in\{k_1,k_2,l_1,l_2\}} \max\limits_{j} |u_{r_i+r,j}| = o(1)$, then

$$g(r_i, k_1, k_2, l_1, l_2) = \begin{cases} 1 & \text{if } k_1 = k_2 \text{ and } l_1 = l_2 \text{ or } k_1 = l_2 \text{ and } l_1 = k_2 \\ 0 & \text{otherwise.} \end{cases}$$

Then we have the following result.

THEOREM 2.3. *Suppose that Assumptions 1, 5 and 6 hold. Moreover, either Assumption 2 or Assumption 2′ holds. Let*

$$\theta_i = \alpha_i\Big(1 + \frac{p-K}{n}\int \frac{t}{\alpha_i - t}dF_{\mathbf{\Lambda}_P}(t)\Big).$$

Let $r_i = \sum_{j=0}^{i-1} n_j$, for $i = 1, 2, ..., \mathcal{L}$. Then

$$(2.14) \qquad \frac{\sqrt{n}}{\theta_i}(\lambda_{r_i+1} - \theta_i, \lambda_{r_i+2} - \theta_i, ..., \lambda_{r_i+n_i} - \theta_i) \xrightarrow{D} \mathcal{R}_i,$$

where $\mathcal{R}_i$ are the eigenvalues of $n_i \times n_i$ Gaussian matrix $\mathfrak{S}_i$ with $\mathbb{E}\mathfrak{S}_i = 0$ and the covariance of the $(\mathfrak{S}_i)_{k_1,l_1}$ and $(\mathfrak{S}_i)_{k_2,l_2}$ being $G(r_i, k_1, k_2, l_1, l_2)$.

The proof of Theorem 2.2 requires new technical tools. The following CLT for a type of random quadratic forms, where the random vectors and random matrices involved are dependent, plays a key role in the proof. This result can be of independent interest.

THEOREM 2.4. *Suppose that Assumption 1 holds and the spectral norm of Σ_1 is bounded. In addition, suppose that there exists an orthogonal unit vector $\mathbf{w}_1$ such that $\mathbf{w}_1^\mathsf{T}\mathbf{U}_2^\mathsf{T} = 0$. If $\frac{\theta}{\frac{p+l}{n}} \to \infty$ and $\theta \to \infty$, then*

$$(2.15) \quad \frac{\sqrt{n}}{\tilde{\sigma}_1}\Big(\mathbf{w}_1^\mathsf{T}\mathbf{X}(n\mathbf{I} - \mathbf{X}^\mathsf{T}\frac{\mathbf{\Sigma}_1}{\theta}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{w}_1 + \tilde{m}_\theta(1)\Big) \overset{D}{\to} N\,(0,1)\,.$$

Moreover, if there exists another unit vector $\mathbf{w}_2$ such that $\mathbf{w}_2^\mathsf{T}\mathbf{U}_2^\mathsf{T} = 0$ and $\mathbf{w}_1^\mathsf{T}\mathbf{w}_2 = 0$, we have

$$(2.16) \qquad \frac{\sqrt{n}}{\tilde{\sigma}_{12}}\mathbf{w}_1^\mathsf{T}\mathbf{X}(n\mathbf{I} - \mathbf{X}^\mathsf{T}\frac{\mathbf{\Sigma}_1}{\theta}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{w}_2 \overset{D}{\to} N\,(0,1)$$

where $\tilde{\sigma}_1^2 = \sum_{j=1}^{p+l}[(\gamma_{4j} - 3)w_{1j}^4] + 2$, $\tilde{\sigma}_{12}^2 = \sum_{s=1}^{p+l}[(\gamma_{4s} - 3)w_{1s}^2w_{2s}^2] + 1$ and w_{ij} is the j-th element of $\mathbf{w}_i$, $i = 1,2$.

2.2. *Tracy-Widow law for the largest nonspiked eigenvalue of $\mathbf{S}_n$.* We now turn to the limiting distribution of the largest nonspiked eigenvalue of the sample covariance matrix $\mathbf{S}_n$. The limiting law is of interest in its own right and it is also important for the estimation of the number of the spikes. To this end we introduce additional assumptions.

ASSUMPTION 7. There exist constants c_k such that $\mathbb{E}|\mathbf{x}_{ij}|^k \le c_k$ for all $k \in \mathbb{N}^+$.

ASSUMPTION 8. Let $m_{\Sigma_1}(z)$ be the solution to

$$(2.17) \qquad m_{\Sigma_1}(z) = -\frac{1}{z - \frac{1}{n}tr(\mathbf{I} + m_{\Sigma_1}(z)\Sigma_1)^{-1}\Sigma_1}, \quad z \in \mathbb{C}^+,$$

and define

$$\gamma_+ = \inf\{x \in \mathbb{R}, F_0(x) = 1\},$$

where $F_0(x)$ is the c.d.f. determined by $m_{\Sigma_1}(z)$ (One can also refer to page 4 of [11]). Suppose that

$$(2.18) \qquad \limsup_n \mu_{K+1}d < 1,$$

where $d = -\lim_{z \in \mathbb{C}^+ \to \gamma_+} m_{\Sigma_1}(z)$.

Intuitively, (2.18) restricts the upper bound of μ_{K+1} to ensure λ_{K+1} to be a nonspiked eigenvalue. Denote the i-th largest eigenvalue of $\frac{1}{n}\mathbf{X}^{\mathsf{T}}\mathbf{\Sigma}_1\mathbf{X}$ by ν_i. Note that the limiting law of ν_1 is the Type-1 Tracy-Widom distribution. Recall the definition of l above (1.1). i.e. Γ is a $p \times (p+l)$ matrix and l is the dimensional difference of columns and rows of Γ. By contrast, K is the number of spiked eigenvalues.

THEOREM 2.5. *Suppose Assumptions 1, 7, and 8 hold. In addition, either Assumption 2 or 5 holds. $l \ll n^{1/6}$ and $p \sim n$. For any i satisfying $1 \leq i - K \leq \log n$, we have, with high probability,*

$$|\lambda_i - \nu_{i-K}| \leq n^{-2/3-\epsilon},$$

In particular, λ_{K+1} has limiting Type-1 Tracy-Widom distribution.

REMARK 7. Theorem 2.5 shows that the non-spiked sample eigenvalues $\lambda_{K+1}, \lambda_{K+2},..., \lambda_{K+r}$ share the same asymptotic distribution as $\nu_1, \nu_2,..., \nu_r$ since the fluctuation of $\nu_1, \nu_2,..., \nu_r$ are $n^{-2/3} \gg n^{-2/3-\epsilon}$. Here r is a fixed integer. See [10] and [25] for more details.

3. Estimating The Number of Spiked Eigenvalues. Identifying the number of spikes is an important problem for a range of statistical applications. For example, a critical step in PCA is the determination of the number of the significant principal components. This issue arises in virtually all practical applications where PCA is used. Choosing the number of principal components is often subjective and based on heuristic methods. As an application of the main theorems discussed in the last section, we propose in this section a procedure to identify the number of the spiked eigenvalues.

Suppose that the conditions of Theorem 2.5 hold. Define the asymptotic variance of ν_1 by (see also (3) of [20])

$$(3.1) \qquad \sigma_n^3 = \frac{1}{d^3}(1 + \frac{p-K}{n}\int(\frac{\lambda d}{1-\lambda d})^3 dF_{\mathbf{\Lambda}_P}(\lambda)).$$

By Theorem 2.5, λ_{K+1} has the same asymptotic distribution as ν_1. Together with Theorem 1 of [20], we have

$$(3.2) \qquad n^{2/3}\frac{\lambda_{K+1} - \gamma_+}{\sigma_n} \xrightarrow{D} TW1,$$

where $TW1$ is the Type-1 Tracy-Widom distribution. Onatski [33] also established such a result for the complex case, but Theorem 1 of [33] requires that the spiked eigenvalues are much bigger than $n^{2/3}$ and $p/n = o(1)$. Moreover, the statistics used in [33] does not estimate γ_+ and σ_n, while our approach estimates them.

From (3.2) one can see that the confidence interval of γ_+ is $[\lambda_{K+1} - w^*\sigma_n n^{-2/3}, \lambda_{K+1} + w^*\sigma_n n^{-2/3}]$, where w^* is a suitable critical value from the Type-1 Tracy-Widom distribution. This, together with Theorem 2.2, implies that it suffices to count the number of the eigenvalues of $\mathbf{S}_n$ that lie beyond $(\gamma_+ + w^*\sigma_n n^{-2/3}\log n)$ to estimate the number of spikes K where $\log n$ can be replaced by any number tending to infinity. However, in practice γ_+ and σ_n are unknown and need to be estimated.

We first consider estimation of σ_n. It turns out that

$$(3.3) \qquad \sigma_n = \left(- \lim_{z\to\gamma_+^+} \frac{\int \frac{dF_0(x)}{(x-z)^3}}{(\int \frac{dF_0(x)}{(x-z)^2})^3} \right)^{1/3}.$$

Moreover, one can verify that with high probability

$$(3.4) \qquad \lambda_{K+1} \le \lambda_{n^\alpha} + \log n \times n^{-\frac{2(1-\alpha)}{3}},$$

where α is a constant such that $\alpha \in [1/6, 1)$ (see Section 8 in the supplementary material). In view of (3.4) we estimate $F_0(x)$ by its empirical version $\lambda_{n^\alpha}, \lambda_{n^\alpha+1}, ..., \lambda_n$ in (3.3), i.e. we exclude the first n^α eigenvalues of $\mathbf{S}_n$. Moreover, for γ_+ in (3.3), we use $\lambda_{n^\alpha} + n^{-4/9}$ to replace it. The reason for using $\lambda_{n^\alpha} + n^{-4/9}$ to estimate γ_+ instead of λ_{n^α} is to avoid singularity in $\int \frac{dF_0(x)}{(x-\gamma_+)^3}$. The estimator of σ_n is then given by

$$\hat{\sigma}_n = \left(- \frac{\frac{1}{n-n^\alpha}\sum_{i=n^\alpha}^n \frac{1}{(\lambda_i-z_0)^3}}{(\frac{1}{n-n^\alpha}\sum_{i=n^\alpha}^n \frac{1}{(\lambda_i-z_0)^2})^3} \right)^{1/3}, \quad \text{where} \quad z_0 = \lambda_{n^\alpha} + n^{-4/9}.$$

We next consider estimation of γ_+ defined below (2.17). By the assumption that $K \ll n^{1/6}$, it follows from Theorems 2.2 and 2.5 that $\lambda_{n^{1/6}}$ is not a spiked eigenvalue. Based on this, an upper bound of λ_{K+1} is given in (3.4). Hence we use the following $\hat{p}_0$ as an initial upper bound of λ_{K+1}

$$(3.5) \qquad \hat{p}_0 = \lambda_{n^\alpha} + \log n \times n^{-\frac{2(1-\alpha)}{3}}.$$

Although $\hat{p}_0$ is a good upper bound for λ_{K+1} theoretically, it does not depend on σ_n and hence in practice $\hat{p}_0$ may not work well. Based on (3.2), we propose the following iteration approach to update $\hat{p}_0$. The idea behind the iteration is that even if $\hat{p}_0$ is not larger than λ_{K+1} in practice, $\hat{p}_0$ is still close to λ_{K+1}. Thus by (3.2), there is at least one eigenvalue in the interval $[\hat{p}_0, \hat{p}_0 + w^*m_n\sigma_n n^{-2/3}]$, where $m_n \to \infty$.

Theorem 2.5 implies that $\hat{K}$ is a good estimator of the number of significant components K.

PROPOSITION 1.   *Under the conditions of Theorem 2.5, we have $\hat{K} = K$ with high probability.*

Algorithm 1

1: Define the initial value $\hat{p}_0$ in (3.5).

2: Suppose that we have $\hat{p}_{m-1}$. If there is at least one eigenvalue of $\mathbf{S}_n$ belonging to $[\hat{p}_{m-1}, \hat{p}_{m-1} + 2.02(\log n)\sigma_n n^{-2/3}]$, where 2.02 is the 99% quantile of Type-1 Tracy-Widom distribution, we renew $\hat{p}_n = \hat{p}_{n-1} + 2.02 \log n \sigma_n n^{-2/3}$. Here $\log n$ can be also replaced by the other number tending to infinity too. Otherwise the iteration stops.

3: After getting $\hat{p}_n$, we return to Step 2 until the iteration stops.

4: Denote the final value of the above iteration by $\hat{p}_{end}$. We define $\hat{K}$ to be the number of eigenvalues larger than $\hat{p}_{end}$.

Identifying The Number of Factors. A closely related problem is the estimation of the number of factors under a factor model, which is widely used in financial econometrics. Consider the factor model

$$(3.6) \qquad \mathbf{y}_t = \Lambda_1 \mathbf{f}_t + \mathbf{T}\varepsilon_t, \quad t = 1, 2, \ldots, n,$$

where Λ_1 is $p \times K$-dimensional factor loading, $\mathbf{f}_t$ is the corresponding K-dimensional factor, $\{\varepsilon_{it} : i = 1, 2, \ldots, p; t = 1, 2, \ldots, n\}$ are the independent idiosyncratic components.

In many applications, the number of factors K is unknown. An important step in factor analysis is to determine the value of K. Let $\mathbf{F} = (\mathbf{f}_1, ..., \mathbf{f}_n)$, $\mathbf{Z} = (\varepsilon_1, ..., \varepsilon_n)$ and $\mathbf{Y} = (\mathbf{y}_1, ..., \mathbf{y}_n)$. Then (3.6) can be rewritten as

$$(3.7) \qquad \mathbf{Y} = \Lambda_1 \mathbf{F} + \mathbf{T}\mathbf{Z} = (\Lambda_1 \quad \mathbf{T}) \begin{pmatrix} \mathbf{F} \\ \mathbf{Z} \end{pmatrix}.$$

Suppose that $\begin{pmatrix} \mathbf{F} \\ \mathbf{Z} \end{pmatrix}$ satisfies Assumptions 1 and 7 and $(\Lambda_1 \quad \mathbf{T})$ satisfies Assumptions 2 and 8. It is easy to conclude that the $(K + 1)$-st largest eigenvalue of $\frac{1}{n}\mathbf{Y}\mathbf{Y}^\intercal$ follows the Type-1 Tracy-Widom distribution asymptotically. The following result is a direct consequence of Proposition 1.

COROLLARY 1. *For the model (3.6), if* $\begin{pmatrix} \mathbf{F} \\ \mathbf{Z} \end{pmatrix}$ *satisfies Assumptions 1 and 7 and* $(\Lambda_1 \quad \mathbf{T})$ *satisfies Assumptions 2 and 8, $K \ll n^{1/6}$ and $p \sim n$, then we have $\hat{K} = K$ with high probability.*

Comparing to the approaches in [4] and [33], here we allow the number of factors K to diverge with n. Moreover, we only assume that the spiked population eigenvalues diverge to infinity, while [4] and [33] assume that they are much larger than $n^{2/3}$ or grow linearly with n.

4. Estimating the Eigenvectors.   As mentioned in the introduction, the leading eigenvectors of the population covariance matrix are of significant interest in PCA and many other statistical applications. They are conventionally estimated by their empirical counterparts.

We consider in this section estimation of the population eigenvectors associated with the spiked population eigenvalues $\mu_1,...,\mu_K$, involved in σ_i^2 in (2.7). To this end, we first characterize the relationship between the sample eigenvectors and the corresponding population eigenvectors. Write the population eigenvectors matrix $\mathbf{V}$ as $\mathbf{V} = (\mathbf{v}_1, \cdots, \mathbf{v}_p)$.

THEOREM 4.1.   *Suppose that the conditions of Theorem 2.2 hold. Let ξ_i be the eigenvector of $\mathbf{S}_n$ corresponding to the eigenvalue λ_i. Then for $1 \leq i \leq K$, we have*

$$(4.1) \qquad\qquad \mathbf{v}_i^{\mathsf{T}} \xi_i \xi_i^{\mathsf{T}} \mathbf{v}_i \xrightarrow{i.p.} 1.$$

Theorem 4.1 also implies that for $i = 1,...,K$, $j = 1,...,p$, $i \neq j$, we have

$$\mathbf{v}_j^{\mathsf{T}} \xi_i \xi_i^{\mathsf{T}} \mathbf{v}_j \xrightarrow{i.p.} 0.$$

One should notice that the convergence is uniformly for $j = 1,...,p$ since $1 = \xi_i^{\mathsf{T}} \xi_i = \sum_{j=1}^{p} \mathbf{v}_j^{\mathsf{T}} \xi_i \xi_i^{\mathsf{T}} \mathbf{v}_j$.

Theorem 4.1 shows that the sample eigenvector ξ_i is a good estimator of $\mathbf{v}_i$ up to a sign difference. An immediate application of Theorem 4.1 is to estimate σ_i^2 for the case when $\mathbf{V} = \mathbf{U}^{\mathsf{T}}$ and $\gamma_{41} = ... = \gamma_{4p} = \gamma_4$ by Corollary 2. This corollary shows that the empirical eigenvector plays an important role in statistical inference of the spiked eigenvalue.

COROLLARY 2.   *Under the conditions of Theorem 4.1, we have*

$$\sum_{j=1}^{p} v_{ij}^4 - \sum_{j=1}^{p} \xi_{ij}^4 \xrightarrow{i.p.} 0.$$

We now consider the extension to the case when the multiplicity of the population eigenvalues μ_i is more than one. Correspondingly the following corollary holds and its proof is the same as that of Theorem 4.1.

COROLLARY 3.   *Recall the definition of r_i above (2.14). Under the conditions of Theorem 2.3, The angle between $\mathbf{v}_k$, $k \in \{r_{i-1} + 1, ..., r_i\}$ and the subspace spanned by $\{\xi_j, j = r_{i-1} + 1, ..., r_i\}$ tends to 0 in probability. In other words, we have*

$$\mathbf{v}_k^{\mathsf{T}} \Big(\sum_{j=r_{i-1}+1}^{r_i} \xi_j \xi_j^{\mathsf{T}} \Big) \mathbf{v}_k \xrightarrow{i.p.} 1, \ \ k \in \{r_{i-1} + 1, ..., r_i\}.$$

Corollary 3 shows that the sample eigenvectors $\{\xi_j, j = r_{i-1} + 1, ..., r_j\}$ are close to the space spaned by $\{\mathbf{v}_j, j = r_{i-1} + 1, ..., r_j\}$.

5. Centralized sample covariance matrices. So far we have focused on the non-centralized sample covariance matrix S_n. We now turn to its centralized version

$$\tilde{\mathbf{S}}_n = \frac{1}{n}\sum_{i=1}^n \mathbf{\Gamma}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\intercal\mathbf{\Gamma}^\intercal = \mathbf{\Gamma}\mathbf{X}(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^\intercal)\mathbf{X}^\intercal\mathbf{\Gamma}^\intercal,$$

where $\mathbf{1}$ is the $n \times 1$ vector with all elements being 1. Denote $(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^\intercal)$ by $\mathbf{\Upsilon}$. First we have the following Lemma.

LEMMA 1. *Under the conditions of Theorem 1, we have*

$$(5.1)\quad \frac{\sqrt{n}}{\tilde{\sigma}_1}\left(\mathbf{w}_1^\intercal\mathbf{X}\mathbf{\Upsilon}(n\mathbf{I} - \mathbf{\Upsilon}\mathbf{X}^\intercal\frac{\mathbf{\Sigma}_1}{\theta}\mathbf{X}\mathbf{\Upsilon})^{-1}\mathbf{X}^\intercal\mathbf{w}_1 + \tilde{m}_\theta(1)\right) \xrightarrow{D} N(0,1)$$

and

$$(5.2)\quad \frac{\sqrt{n}}{\tilde{\sigma}_{12}}\mathbf{w}_1^\intercal\mathbf{X}\mathbf{\Upsilon}(n\mathbf{I} - \mathbf{\Upsilon}\mathbf{X}^\intercal\frac{\mathbf{\Sigma}_1}{\theta}\mathbf{X}\mathbf{\Upsilon})^{-1}\mathbf{\Upsilon}\mathbf{X}^\intercal\mathbf{w}_2 \xrightarrow{D} N(0,1)$$

where $\tilde{\sigma}_1^2 = \sum_{j=1}^{p+l}[(\gamma_{4j} - 3)\mathbf{w}_{1j}^4] + 2$, $\tilde{\sigma}_{12}^2 = \sum_{s=1}^{p+l}[(\gamma_{4s} - 3)\mathbf{w}_{1s}^2\mathbf{w}_{2s}^2] + 1$ *and* $\mathbf{w}_{ij}$ *is the j-th element of* $\mathbf{w}_i$, $i = 1, 2$.

By Lemma 1 and checking carefully the proofs of the main results, it can be seen that all arguments remain valid if $\mathbf{X}$ is replaced by $\mathbf{X}\mathbf{\Upsilon}$ (note that $\mathbf{\Upsilon}^2 = \mathbf{\Upsilon}$). So Theorem 2.1–Corollary 3 hold for $\frac{1}{n}\mathbf{\Gamma}\mathbf{X}\mathbf{\Upsilon}\mathbf{X}^\intercal\mathbf{\Gamma}^\intercal$ as well.

6. Numerical Results . In this section we illustrate some of the theoretical results obtained earlier through numerical experiments. We first use simulation to confirm that the asymptotic behavior of the spiked eigenvalues is indeed affected by the population eigenvectors.

Let $K = 2$ and $\Lambda_P = \mathsf{diag}(\mu_3, ..., \mu_p)$. Suppose that $\{\mu_i, \ i = 3, ..., p\}$ are i.i.d. copies of the uniform random variable $U(1, 2)$. Define $\mathbf{v}_1 = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})^\intercal$, $\mathbf{v}_2 = (\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}})^\intercal$, $\check{\mathbf{V}} = (\mathbf{v}_1, \mathbf{v}_2)$ and $\Lambda_S = \mathsf{diag}(800, 200)$. We now define two different population matrices

$$\mathbf{\Sigma}_2 = \begin{pmatrix} \mathbf{\Lambda}_S & 0 \\ 0 & \mathbf{\Lambda}_P \end{pmatrix}, \qquad \mathbf{\Sigma}_3 = \begin{pmatrix} \check{\mathbf{V}}\mathbf{\Lambda}_S\check{\mathbf{V}}^\intercal & 0 \\ 0 & \mathbf{\Lambda}_P \end{pmatrix}.$$

Then the eigenvalues of $\mathbf{\Sigma}_2$ and $\mathbf{\Sigma}_3$ are the same but the eigenvectors corresponding to the first two largest eigenvalues are different. Consider the case

$p = n$ and $\mathbf{X} = (x_{ij})$ are i.i.d. $U(-\sqrt{3}, \sqrt{3})$. Denote by $\check{\lambda}_1$ and $\breve{\lambda}_1$ respectively the largest eigenvalues of the sample covariance matrices $\frac{1}{n}\boldsymbol{\Sigma}_2^{1/2}\mathbf{X}\mathbf{X}^\intercal\boldsymbol{\Sigma}_2^{1/2}$ and $\frac{1}{n}\boldsymbol{\Sigma}_3^{1/2}\mathbf{X}\mathbf{X}^\intercal\boldsymbol{\Sigma}_3^{1/2}$. Table 1 reports the sample variance of the rescaled eigenvalues $\frac{\sqrt{n}\check{\lambda}_1}{800}$ and $\frac{\sqrt{n}\breve{\lambda}_1}{800}$. It can be seen that the behavior of the spiked sample eigenvalues is indeed affected by the population eigenvectors.

TABLE 1

The variances of the rescaled largest eigenvalues

p	200	400	600	800	1000
$\boldsymbol{\Sigma}_2$	0.8111	0.7965	0.8287	0.7574	0.7874
$\boldsymbol{\Sigma}_3$	1.2507	1.4051	1.2800	1.5012	1.3911

We now consider estimating the number of factors under the factor model (3.7):

$$\mathbf{Y} = \Lambda_1\mathbf{F} + \mathbf{T}\mathbf{Z}.$$

In the simulation, the entries of $\mathbf{F}$ and $\mathbf{Z}$ follow the standard Gaussian distribution. Consider two choices: $\mathbf{T} = \mathbf{T}_1$ or $\mathbf{T}_2$, where $\mathbf{T}_1 = \mathbf{I}$, $\mathbf{T}_2 = \mathsf{diag}(\underbrace{1, 1, ..., 1}_{p/2}, \underbrace{\frac{1}{\sqrt{2}}, ..., \frac{1}{\sqrt{2}}}_{p/2})$. Let Λ_1 be a $p \times K$ matrix with nonzero entries being $(\Lambda_{11}, ..., \Lambda_{KK}) = (\sqrt{b_1^2 - 1}, ..., \sqrt{b_K^2 - 1})$ where $K = 5\lceil n^{1/7}\rceil + 1$, and $(b_1, .., b_K) = \sqrt{(6, ..., 6 + K - 1) * r} + 1$, $0 \le r \le 1$.

We compare the accuracy of three methods for estimating the number of factors K: our procedure proposed in Section 3, the method introduced in [34] , and the approach given in [9], which are denoted by CHP, Ons, and BYK, respectively. Here we omit the simulation results of BIC used in [9] for reasons of space. The initial value of $\hat{p}_0$ is given in (3.5) and note that $K \ll n^{1/6}$. However this requirement may be violated in practice when the sample size n is not sufficiently large. For example when $n = 100$ in our simulation setting $n^{1/6}$ is as small as three while $K = 5\lceil n^{1/7}\rceil + 1 = 11$. Therefore one has to replace the initial value of $\hat{p}_0$ by $\lambda_{c_1\lceil n^{1/6}\rceil} + \log n \times n^{-5/9}$ where c_1 is a constant such that $K = 5\lceil n^{1/7}\rceil + 1 \ll c_1\lceil n^{1/6}\rceil$. Here we set $c_1 = 15$ according to our extensive simulations in order to reduce the number of updating iteration (such a replacement does not change the conclusions of the theoretical results developed in Section 3). In contrast to the size of n, such an initial value of $\hat{p}_0$ is essentially a conservative choice. One can see that $\lambda_{15\lceil n^{1/6}\rceil}$ is a nonspiked eigenvalue. The approach in Section IV of [34] uses an iteration approach to estimate K, which also requires the rough information of the number of nonspiked eigenvalues. In addition, we also set $15\lceil n^{1/6}\rceil$ as the initial value for the algorithm in [34] when running the

algorithm there. [9] uses AIC based on sample eigenvalues to estimate K.

Different combinations of n and p are considered. The following tables report the proportion of times the number of factors is correctly identified, i.e. $\hat{K} = K$, where for each (n, p) we repeat 500 times. Different choices of r (ranging from 0.3 to 1) are also considered. From Tables 2 and 3, one can see that the accuracy of our approach increases as (n, p) become larger. Our approach works better in comparison to [34]. This is likely due to the fact that our method allows the number of factors K to be increasing with n, while [34] requires K to be fixed. Tables 2 and 3 show that the method based on the AIC criterion and our procedure have similar performance. But as mentioned earlier in Remark 2, the model in [9] only allows that $\mu_{K+1} = \ldots = \mu_p = 1$, which is a special case of what we consider in the present paper. Indeed, Table 4 also confirms that for the non-identity matrix $\mathbf{T}_2$, the method based on the AIC criterion performs much worse than our approach. Therefore, our procedure is preferred for the case where $\mu_{K+1}, \ldots, \mu_p$ are unknown.

7. Proofs. In this section, we prove only one of the main results, Theorem 2.4. The proof of Theorem 2.2 is involved and is given in the supplement [16]. The proofs of the other results and additional technical lemmas are also provided in the supplement [16].

7.1. *Proof of Theorem 2.4.* The main idea of this proof is to first express $\mathbf{w}_1^\mathsf{T}\mathbf{X}(n\mathbf{I} - \mathbf{X}^\mathsf{T}\frac{\mathbf{\Sigma}_1}{\theta}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{w}_1$ as a sum of martingale differences and then apply the central limit theorem for the martingale difference.

We below consider the case $p \gtrsim n$ and prove (2.15) only because the case $\frac{p}{n} \to 0$ and (2.16) can be proved similarly. First of all, we need to do truncation and centralization on $\mathbf{x}_{ij}$ as in the first paragraph of Section 12 in the supplement [16]. In fact, by (12.2)-(12.6) in [16], we conclude that the truncation and centralization do not affect the CLT. i.e. we can get the following inequality similar to (12.7) in [16]

$$\mathbf{w}_1^\mathsf{T}\mathbf{X}(n\mathbf{I} - \mathbf{X}^\mathsf{T}\frac{\mathbf{\Sigma}_1}{\theta}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{w}_1 = \mathbf{w}_1^\mathsf{T}\tilde{\mathbf{X}}(n\mathbf{I} - \tilde{\mathbf{X}}^\mathsf{T}\frac{\mathbf{\Sigma}_1}{\theta}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^\mathsf{T}\mathbf{w}_1 + o_p(\frac{1}{\sqrt{n}}),$$

where $\tilde{\mathbf{X}}$ is the truncated and centralized version of $\mathbf{X}$. The argument is standard and we omit the details here. Therefore, for simplicity we below assume that

$$E\mathbf{x}_{ij} = 0, \quad |\mathbf{x}_{ij}| \leq \delta_n \sqrt[4]{np}.$$

CLT of the random part. Define the following events

$$F_d = \{\|\frac{1}{n}\mathbf{X}^\mathsf{T}\mathbf{\Sigma}_1\mathbf{X}\| \leq 4\|\mathbf{\Sigma}_1\|(1+\frac{p}{n})\}, \ F_d^{(k)} = \{\|\frac{1}{n}\mathbf{X}_k^\mathsf{T}\mathbf{\Sigma}_1\mathbf{X}_k\| \leq 4\|\mathbf{\Sigma}_1\|(1+\frac{p}{n})\}, k = 1, \ldots, n,$$

TABLE 2

Ratio of Identifying The Correct Number of Factors with $\mathbf{T}_1$

$r\backslash(n,p)$	(50,50)			(50,100)			(50,150)		
	CHP	Ons	BYK	CHP	Ons	BYK	CHP	Ons	BYK
0.3	0.608	0.000	0.610	0.192	0.000	0.330	0.068	0.002	0.122
0.4	0.816	0.020	0.706	0.442	0.000	0.618	0.184	0.000	0.368
0.5	0.904	0.008	0.662	0.676	0.012	0.788	0.450	0.002	0.606
0.6	0.892	0.044	0.612	0.832	0.012	0.880	0.638	0.006	0.800
0.7	0.906	0.040	0.636	0.880	0.014	0.870	0.756	0.002	0.866
0.8	0.914	0.040	0.638	0.918	0.022	0.886	0.868	0.010	0.880
0.9	0.908	0.030	0.648	0.948	0.022	0.866	0.916	0.016	0.910
1.0	0.914	0.042	0.616	0.946	0.014	0.872	0.912	0.020	0.896

TABLE 3

Ratio of Identifying The Correct Number of Factors with $\mathbf{T}_1$

$r\backslash(n,p)$	(100,100)			(100,200)			(100,300)		
	CHP	Ons	BYK	CHP	Ons	BYK	CHP	Ons	BYK
0.3	0.954	0.130	0.974	0.772	0.056	0.854	0.392	0.006	0.482
0.4	0.980	0.234	0.982	0.942	0.088	0.984	0.782	0.005	0.908
0.5	0.956	0.272	0.974	0.964	0.148	0.990	0.938	0.068	0.976
0.6	0.972	0.330	0.976	0.980	0.162	0.994	0.966	0.124	0.990
0.7	0.970	0.396	0.974	0.978	0.234	0.986	0.972	0.158	0.996
0.8	0.954	0.412	0.974	0.972	0.272	0.998	0.984	0.178	0.980
0.9	0.954	0.446	0.980	0.970	0.316	0.986	0.980	0.240	0.984
1.0	0.950	0.444	0.972	0.958	0.326	0.984	0.982	0.290	0.988

TABLE 4

Ratio of Identifying The Correct Number of Factors with $\mathbf{T}_2$

$r\backslash(n,p)$	(100,100)			(100,200)			(100,300)		
	CHP	Ons	BYK	CHP	Ons	BYK	CHP	Ons	BYK
0.3	0.946	0.264	0.490	0.938	0.088	0.658	0.792	0.042	0.716
0.4	0.928	0.296	0.454	0.974	0.178	0.624	0.968	0.070	0.710
0.5	0.944	0.360	0.424	0.968	0.236	0.682	0.986	0.148	0.704
0.6	0.926	0.400	0.440	0.966	0.276	0.672	0.978	0.206	0.654
0.7	0.926	0.466	0.434	0.970	0.336	0.662	0.972	0.262	0.670
0.8	0.918	0.512	0.450	0.978	0.390	0.650	0.986	0.270	0.660
0.9	0.928	0.510	0.434	0.978	0.402	0.608	0.980	0.310	0.670
1.0	0.930	0.544	0.410	0.980	0.418	0.614	0.976	0.386	0.658

where $\mathbf{X}_k = \mathbf{X} - \mathbf{x}_k \mathbf{e}_k^\intercal$, $\mathbf{x}_k$ is the k-th column of $\mathbf{X}$ and $\mathbf{e}_k = (0, .., 0, 1, 0, ..., 0)^\intercal$ is a M-dimensional vector with only k-th element being 1. By Theorem 2 of [18], we have

$$(7.1) \qquad I(F_d) = 1 \quad \text{and} \quad I(F_d^{(k)}) = 1, \; k = 1, ..., n$$

with high probability.

We define $\frac{\mathbf{\Sigma}_1}{\theta} = \tilde{\mathbf{\Sigma}}_1$, $\mathbf{A} = \mathbf{I} - \frac{1}{n}\mathbf{X}^\intercal\tilde{\mathbf{\Sigma}}_1\mathbf{X}$, $\mathbf{A}_k = \mathbf{I} - \frac{1}{n}\mathbf{X}_k^\intercal\tilde{\mathbf{\Sigma}}_1\mathbf{X}_k$ and $\mathbf{A}_{(k)} = \mathbf{A}_k - \frac{1}{n}\mathbf{X}_k^\intercal\tilde{\mathbf{\Sigma}}_1\mathbf{x}_k\mathbf{e}_k^\intercal$. Then $\mathbf{A} = \mathbf{A}_k - \frac{1}{n}(\mathbf{e}_k\mathbf{x}_k^\intercal\tilde{\mathbf{\Sigma}}_1\mathbf{X}_k + \mathbf{X}_k^\intercal\tilde{\mathbf{\Sigma}}_1\mathbf{x}_k\mathbf{e}_k^\intercal + \mathbf{e}_k\mathbf{x}_k^\intercal\tilde{\mathbf{\Sigma}}_1\mathbf{x}_k\mathbf{e}_k^\intercal)$. Therefore,

$$(7.2) \qquad \mathbf{w}_1^\intercal\mathbf{X}(n\mathbf{I} - \mathbf{X}^\intercal\frac{\mathbf{\Sigma}_1}{\theta}\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{w}_1 = \frac{1}{n}\mathbf{w}_1^\intercal\mathbf{X}\mathbf{A}^{-1}\mathbf{X}^\intercal\mathbf{w}_1.$$

By the definition of $\mathbf{X}_k$ and $\mathbf{A}_k$, we observe that the k-th row and k-th colomn of $\mathbf{A}_k$ are 0 except for the diagonal entry. Hence it is not hard to conclude the following important facts

$$(7.3) \qquad \mathbf{e}_k^\intercal\mathbf{A}_k^{-1}\mathbf{e}_k = 1,$$

$$(7.4) \qquad \mathbf{e}_i^\intercal\mathbf{A}_k^{-1}\mathbf{e}_k = 0, \;\; i \neq k$$

and

$$(7.5) \qquad \mathbf{X}_k\mathbf{A}_k^{-1}\mathbf{e}_k = \mathbf{X}_k\mathbf{e}_k = 0.$$

In the sequel, we prove the central limit theorem for $\frac{1}{n}\mathbf{w}_1^\intercal\mathbf{X}\mathbf{A}^{-1}\mathbf{X}^\intercal\mathbf{w}_1 I(F_d)$ instead of $\frac{1}{n}\mathbf{w}_1^\intercal\mathbf{X}\mathbf{A}^{-1}\mathbf{X}^\intercal\mathbf{w}_1$. In fact, it follows from (7.1) that $I(F_d) = 1$ with high probability. Therefore $\frac{1}{n}\mathbf{w}_1^\intercal\mathbf{X}\mathbf{A}^{-1}\mathbf{X}^\intercal\mathbf{w}_1$ and $\frac{1}{n}\mathbf{w}_1^\intercal\mathbf{X}\mathbf{A}^{-1}\mathbf{X}^\intercal\mathbf{w}_1 I(F_d)$ have the same central limit theorem. Let $\mathbb{E}_k = \mathbb{E}(.|\mathbf{x}_1, ..., \mathbf{x}_k)$, $\mathbb{E} = \mathbb{E}(.)$ and write

$$(7.6) \mathbf{w}_1^\intercal\mathbf{X}\mathbf{A}^{-1}\mathbf{X}^\intercal\mathbf{w}_1 I(F_d) - \mathbb{E}\mathbf{w}_1^\intercal\mathbf{X}\mathbf{A}^{-1}\mathbf{X}^\intercal\mathbf{w}_1 I(F_d)$$

$$= \sum_{k=1}^{n}(\mathbb{E}_k - \mathbb{E}_{k-1})\mathbf{w}_1^\intercal\mathbf{X}\mathbf{A}^{-1}\mathbf{X}^\intercal\mathbf{w}_1 I(F_d)$$

$$= \sum_{k=1}^{n}(\mathbb{E}_k - \mathbb{E}_{k-1})\big(\mathbf{w}_1^\intercal\mathbf{X}\mathbf{A}^{-1}\mathbf{X}^\intercal\mathbf{w}_1 I(F_d) - \mathbf{w}_1^\intercal\mathbf{X}_k\mathbf{A}_k^{-1}\mathbf{X}_k^\intercal\mathbf{w}_1 I(F_d^{(k)})\big)$$

$$= \sum_{k=1}^{n}(\mathbb{E}_k - \mathbb{E}_{k-1})\big(\mathbf{w}_1^\intercal\mathbf{X}\mathbf{A}^{-1}\mathbf{X}^\intercal\mathbf{w}_1 - \mathbf{w}_1^\intercal\mathbf{X}_k\mathbf{A}_k^{-1}\mathbf{X}_k^\intercal\mathbf{w}_1\big)I(F_d) + o_p(n^{-2})$$

$$= \sum_{k=1}^{n}(\mathbb{E}_k - \mathbb{E}_{k-1})(I_1 + 2I_2 + I_3 - \mathbf{w}_1^\intercal\mathbf{X}_k\mathbf{A}_k^{-1}\mathbf{X}_k^\intercal\mathbf{w}_1)I(F_d) + o_p(n^{-2}),$$

where the third equality follows from (7.1), $I_1 = (\mathbf{w}_1^\intercal \mathbf{x}_k)^2 \mathbf{e}_k^\intercal \mathbf{A}^{-1} \mathbf{e}_k$, $I_2 = \sum_{i \neq k} \mathbf{w}_1^\intercal \mathbf{x}_k \mathbf{w}_1^\intercal \mathbf{x}_i \mathbf{e}_i^\intercal \mathbf{A}^{-1} \mathbf{e}_k$, and $I_3 = \sum_{i,j \neq k} \mathbf{w}_1^\intercal \mathbf{x}_i \mathbf{w}_1^\intercal \mathbf{x}_j \mathbf{e}_i^\intercal \mathbf{A}^{-1} \mathbf{e}_j$. We define

$$(7.7) \qquad a_k = 1 - \frac{1}{n}(\mathbf{x}_k^\intercal \tilde{\Sigma}_1 \mathbf{X}_k \mathbf{A}_{(k)}^{-1} \mathbf{e}_k + \mathbf{x}_k^\intercal \tilde{\Sigma}_1 \mathbf{x}_k \mathbf{e}_k^\intercal \mathbf{A}_{(k)}^{-1} \mathbf{e}_k)$$

and

$$(7.8) \qquad b_k = 1 - \frac{1}{n}\mathbf{e}_k^\intercal \mathbf{A}_k^{-1} \mathbf{X}_k^\intercal \tilde{\Sigma}_1 \mathbf{x}_k.$$

We next simplify the formula. Noting that $\mathbf{w}_1^\intercal \mathbf{X} = \mathbf{w}_1^\intercal \mathbf{X}_k + \mathbf{w}_1^\intercal \mathbf{x}_k \mathbf{e}_k^\intercal$, by the formulas

$$(7.9) \qquad \mathbf{W}^{-1} = \mathbf{Q}^{-1} - \frac{\mathbf{Q}^{-1}(\mathbf{W} - \mathbf{Q})\mathbf{Q}^{-1}}{1 + tr(\mathbf{Q}^{-1}(\mathbf{W} - \mathbf{Q}))}$$

and

$$(7.10) \qquad (\mathbf{Q} + \sum_{j=1}^m ab_j^\intercal)^{-1}a = \frac{\mathbf{Q}^{-1}a}{1 + \sum_{j=1}^m b_j^\intercal \mathbf{Q}^{-1}a},$$

we have

$$(7.11)$$

$$
\begin{aligned}
\mathbf{A}^{-1} &= \mathbf{A}_{(k)}^{-1} + \frac{\mathbf{A}_{(k)}^{-1}(\mathbf{e}_k \mathbf{x}_k^\intercal \tilde{\Sigma}_1 \mathbf{X}_k + \mathbf{e}_k \mathbf{x}_k^\intercal \tilde{\Sigma}_1 \mathbf{x}_k \mathbf{e}_k^\intercal)\mathbf{A}_{(k)}^{-1}}{na_k} \\
&= \mathbf{A}_k^{-1} + \frac{\mathbf{A}_k^{-1}\mathbf{X}_k^\intercal \tilde{\Sigma}_1 \mathbf{x}_k \mathbf{e}_k^\intercal \mathbf{A}_k^{-1}}{nb_k} + \frac{\mathbf{A}_{(k)}^{-1}(\mathbf{e}_k \mathbf{x}_k^\intercal \tilde{\Sigma}_1 \mathbf{X}_k + \mathbf{e}_k \mathbf{x}_k^\intercal \tilde{\Sigma}_1 \mathbf{x}_k \mathbf{e}_k^\intercal)\mathbf{A}_{(k)}^{-1}}{na_k}
\end{aligned}
$$

and

$$
\begin{aligned}
(7.12)\, I_1 &= (\mathbf{w}_1^\intercal \mathbf{x}_k)^2 \mathbf{e}_k^\intercal \mathbf{A}^{-1} \mathbf{e}_k = \frac{(\mathbf{w}_1 \mathbf{x}_k)^2 \mathbf{e}_k^\intercal \mathbf{A}_{(k)}^{-1} \mathbf{e}_k}{a_k} \\
&= \frac{(\mathbf{w}_1 \mathbf{x}_k)^2 \mathbf{e}_k^\intercal \mathbf{A}_k^{-1} \mathbf{e}_k}{a_k(1 - \frac{1}{n}\mathbf{e}_k^\intercal \mathbf{A}_k^{-1} \mathbf{X}_k^\intercal \tilde{\Sigma}_1 \mathbf{x}_k)} = \frac{(\mathbf{w}_1 \mathbf{x}_k)^2 \mathbf{e}_k^\intercal \mathbf{A}_k^{-1} \mathbf{e}_k}{a_k b_k} = \frac{(\mathbf{w}_1 \mathbf{x}_k)^2}{a_k b_k}.
\end{aligned}
$$

Moreover, it follows from (7.3), (7.4) and (7.9) that

$$(7.13) \qquad b_k = 1 - \frac{1}{n}\mathbf{e}_k^\intercal \mathbf{A}_k^{-1} \mathbf{X}_k^\intercal \tilde{\Sigma}_1 \mathbf{x}_k = 1$$

and

$$
\begin{aligned}
(7.14)\, a_k &= 1 - \frac{1}{n}\mathbf{x}_k^\intercal \tilde{\Sigma}_1 \mathbf{X}_k \mathbf{A}_{(k)}^{-1} \mathbf{e}_k = 1 - \frac{1}{n^2}\mathbf{e}_k^\intercal \mathbf{A}_k^{-1} \mathbf{e}_k \mathbf{x}_k^\intercal \tilde{\Sigma}_1 \mathbf{X}_k \mathbf{A}_k^{-1} \mathbf{X}_k^\intercal \tilde{\Sigma}_1 \mathbf{x}_k \\
&= 1 - \frac{1}{n^2}\mathbf{x}_k^\intercal \tilde{\Sigma}_1 \mathbf{X}_k \mathbf{A}_k^{-1} \mathbf{X}_k^\intercal \tilde{\Sigma}_1 \mathbf{x}_k.
\end{aligned}
$$

By the Cauchy interlacing property we know

$$(7.15)\frac{1}{n^2}\mathbf{x}_k^\mathsf{T}\tilde{\Sigma}_1\mathbf{X}_k\mathbf{A}_k^{-1}\mathbf{X}_k^\mathsf{T}\tilde{\Sigma}_1\mathbf{x}_kI(F_d) \leq \frac{1}{n^2}\mathbf{x}_k^\mathsf{T}\tilde{\Sigma}_1\mathbf{x}_k\|\tilde{\Sigma}_1^{1/2}\mathbf{X}_k\mathbf{A}_k^{-1}\mathbf{X}_k^\mathsf{T}\tilde{\Sigma}_1^{1/2}\|I(F_d)$$

$$= \frac{1}{n^2}\mathbf{x}_k^\mathsf{T}\tilde{\Sigma}_1\mathbf{x}_k\|\mathbf{A}_k^{-1}\mathbf{X}_k^\mathsf{T}\tilde{\Sigma}_1\mathbf{X}_k\|I(F_d) \leq \frac{1}{n^2}\mathbf{x}_k^\mathsf{T}\tilde{\Sigma}_1\mathbf{x}_k\|\mathbf{A}_k^{-1}\|\|\mathbf{X}_k^\mathsf{T}\tilde{\Sigma}_1\mathbf{X}_k\|I(F_d)$$

$$\leq 2(\frac{p}{n\theta})^2.$$

This implies that

$$(7.16) \qquad\qquad a_kI(F_d) = 1 + O((\frac{p}{n\theta})^2).$$

As for the term $i \neq k$, by (7.4), (7.5), (7.9) and (7.10) we have
(7.17)

$$\mathbf{A}^{-1}\mathbf{e}_k = \frac{\mathbf{A}_{(k)}^{-1}\mathbf{e}_k}{a_k} = \frac{\mathbf{A}_k^{-1}\mathbf{e}_k}{a_k} + \frac{\mathbf{A}_k^{-1}\mathbf{X}_k^\mathsf{T}\tilde{\Sigma}_1\mathbf{x}_k}{a_kb_k} = \frac{\mathbf{A}_k^{-1}\mathbf{e}_k}{a_k} + \frac{\mathbf{A}_k^{-1}\mathbf{X}_k^\mathsf{T}\tilde{\Sigma}_1\mathbf{x}_k}{a_k}.$$

We then conclude that

$$(7.18) \quad I_2 = \sum_{i\neq k}\mathbf{w}_1^\mathsf{T}\mathbf{x}_k\mathbf{w}_1^\mathsf{T}\mathbf{x}_i\mathbf{e}_i^\mathsf{T}\mathbf{A}^{-1}\mathbf{e}_k = \frac{\mathbf{w}_1^\mathsf{T}\mathbf{X}_k\mathbf{A}_k^{-1}\mathbf{X}_k^\mathsf{T}\tilde{\Sigma}_1\mathbf{x}_k\mathbf{x}_k^\mathsf{T}\mathbf{w}_1}{na_k}.$$

It follows from (7.4), (7.5) and (7.11) that for $i, j \neq k$

$$(7.19)$$

$$I_3 = \sum_{i,j\neq k}\mathbf{w}_1^\mathsf{T}\mathbf{x}_i\mathbf{w}_1^\mathsf{T}\mathbf{x}_j\mathbf{e}_i^\mathsf{T}\mathbf{A}^{-1}\mathbf{e}_j$$

$$= \sum_{i,j\neq k}\mathbf{w}_1^\mathsf{T}\mathbf{x}_i\mathbf{w}_1^\mathsf{T}\mathbf{x}_j\mathbf{e}_i^\mathsf{T}\mathbf{A}_k^{-1}\mathbf{e}_j + \sum_{i,j\neq k}\mathbf{w}_1^\mathsf{T}\mathbf{x}_i\mathbf{w}_1^\mathsf{T}\mathbf{x}_j\mathbf{e}_i^\mathsf{T}\frac{\mathbf{A}_{(k)}^{-1}(\mathbf{e}_k\mathbf{x}_k^\mathsf{T}\tilde{\Sigma}_1\mathbf{x}_k\mathbf{e}_k^\mathsf{T} + \mathbf{e}_k\mathbf{x}_k^\mathsf{T}\tilde{\Sigma}_1\mathbf{X}_k)\mathbf{A}_{(k)}^{-1}}{na_k}\mathbf{e}_j$$

$$= \mathbf{w}_1^\mathsf{T}\mathbf{X}_k\mathbf{A}_k^{-1}\mathbf{X}_k^\mathsf{T}\mathbf{w}_1 + \frac{\mathbf{w}_1^\mathsf{T}\mathbf{X}_k\mathbf{A}_{(k)}^{-1}(\mathbf{e}_k\mathbf{x}_k^\mathsf{T}\tilde{\Sigma}_1\mathbf{x}_k\mathbf{e}_k^\mathsf{T} + \mathbf{e}_k\mathbf{x}_k^\mathsf{T}\tilde{\Sigma}_1\mathbf{X}_k)\mathbf{A}_{(k)}^{-1}\mathbf{X}_k^\mathsf{T}\mathbf{w}_1}{na_k}.$$

Consider $(\mathbb{E}_k - \mathbb{E}_{k-1})(I_3 - \mathbf{w}_1^\mathsf{T}\mathbf{X}_k\mathbf{A}_k^{-1}\mathbf{X}_k^\mathsf{T}\mathbf{w}_1)I(F_d)$ next.
We claim that

$$(7.20) \qquad \frac{\mathbf{w}_1^\mathsf{T}\mathbf{X}_k\mathbf{A}_{(k)}^{-1}(\mathbf{e}_k\mathbf{x}_k^\mathsf{T}\tilde{\Sigma}_1\mathbf{x}_k\mathbf{e}_k^\mathsf{T} + \mathbf{e}_k\mathbf{x}_k^\mathsf{T}\tilde{\Sigma}_1\mathbf{X}_k)\mathbf{A}_{(k)}^{-1}\mathbf{X}_k^\mathsf{T}\mathbf{w}_1}{na_k}$$

is negligible. Let $\mathbf{B}_k = \tilde{\Sigma}_1\mathbf{X}_k\mathbf{A}_k^{-1}\mathbf{X}_k^\mathsf{T}\mathbf{w}_1\mathbf{w}_1^\mathsf{T}\mathbf{X}_k\mathbf{A}_k^{-1}\mathbf{X}_k^\mathsf{T}\tilde{\Sigma}_1$. Indeed, by (7.9) and (7.3)-(7.5) we have $\mathbf{A}_{(k)}^{-1} = \mathbf{A}_k^{-1} + \frac{1}{n}\mathbf{A}_k^{-1}\mathbf{X}_k^\mathsf{T}\tilde{\Sigma}_1\mathbf{x}_k\mathbf{e}_k^\mathsf{T}\mathbf{A}_k^{-1}$. This, together with (7.3), (7.4) and (7.5) implies that

$$(7.20) \quad = \quad \frac{\mathbf{w}_1^\mathsf{T}\mathbf{X}_k\mathbf{A}_k^{-1}\mathbf{X}_k^\mathsf{T}\tilde{\Sigma}_1\mathbf{x}_k\mathbf{e}_k^\mathsf{T}\mathbf{A}_k^{-1}\mathbf{e}_k\mathbf{x}_k^\mathsf{T}\tilde{\Sigma}_1\mathbf{X}_k\mathbf{A}_k^{-1}\mathbf{X}_k^\mathsf{T}\mathbf{w}_1}{n^2a_k} = \frac{\mathbf{x}_k^\mathsf{T}\mathbf{B}_k\mathbf{x}_k}{n^2a_k}.$$

It follows from (7.19) and (7.3)-(7.5) that

$$\sum_{k=1}^{n}(\mathbb{E}_k-\mathbb{E}_{k-1})(I_3-\mathbf{w}_1^\intercal\mathbf{X}_k\mathbf{A}_k^{-1}\mathbf{X}_k^\intercal\mathbf{w}_1)I(F_d) = \sum_{k=1}^{n}(\mathbb{E}_k-\mathbb{E}_{k-1})\frac{\mathbf{x}_k^\intercal\mathbf{B}_k\mathbf{x}_k}{n^2a_k}I(F_d^{(k)})+o_p(n^{-2}).$$

Considering the second moment of the above equation, by Lemma 8.10 of [7] we have

$$(7.21)\quad \sum_{k=1}^{n}\mathbb{E}|(\mathbb{E}_k-\mathbb{E}_{k-1})\frac{\mathbf{x}_k^\intercal\mathbf{B}_k\mathbf{x}_k}{n^2a_k}|^2I(F_d^{(k)}) \leq \frac{4}{n^4}\sum_{k=1}^{n}\mathbb{E}|\mathbf{x}_k^\intercal\mathbf{B}_k\mathbf{x}_k|^2I(F_d^{(k)})$$

$$\leq \frac{8}{n^4}\sum_{k=1}^{n}\mathbb{E}|\mathbf{x}_k^\intercal\mathbf{B}_k\mathbf{x}_k - tr\mathbf{B}_k|^2I(F_d^{(k)}) + \frac{8}{n^4}\sum_{k=1}^{n}\mathbb{E}|tr\mathbf{B}_k|^2I(F_d^{(k)})$$

$$\leq \frac{Cp^2}{n\theta^2} \ll N,$$

where we used the inequality

$$tr\mathbf{B}_k \leq \mathbf{X}_k\mathbf{A}_k^{-1}\mathbf{X}_k^\intercal\tilde{\Sigma}_1^2\mathbf{X}_k\mathbf{A}_k^{-1}\mathbf{X}_k^\intercal I(F_d^{(k)}) = O(\frac{p^2}{\theta^2}).$$

We conclude that

$$\frac{1}{n}\sum_{k=1}^{n}(\mathbb{E}_k - \mathbb{E}_{k-1})(I_3 - \mathbf{w}_1^\intercal\mathbf{X}_k\mathbf{A}_k^{-1}\mathbf{X}_k^\intercal\mathbf{w}_1)I(F_d) = o_p(\frac{1}{\sqrt{n}}),$$

which is negligible.

Next we consider I_1 and I_2. It follows from (7.12) and (7.18) that

$$(7.22)\quad \frac{1}{\sqrt{n}}\sum_{k=1}^{n}(\mathbb{E}_k - \mathbb{E}_{k-1})(I_1 + 2I_2)I(F_d)$$

$$= \frac{2}{\sqrt{n}}\sum_{k=1}^{n}(\mathbb{E}_k - \mathbb{E}_{k-1})\Big(\frac{(\mathbf{w}_1\mathbf{x}_k)^2}{2a_k} + \frac{\mathbf{w}_1^\intercal\mathbf{X}_k\mathbf{A}_k^{-1}\mathbf{X}_k^\intercal\tilde{\Sigma}_1\mathbf{x}_k\mathbf{x}_k^\intercal\mathbf{w}_1}{na_k}\Big)I(F_d).$$

We claim that the second term of (7.22) is negligible. Actually, similar to (7.21), it is easy to show that

$$\sum_{k=1}^{n}(\mathbb{E}_k - \mathbb{E}_{k-1})\frac{\mathbf{w}_1^\intercal\mathbf{X}_k\mathbf{A}_k^{-1}\mathbf{X}_k^\intercal\tilde{\Sigma}_1\mathbf{x}_k\mathbf{x}_k^\intercal\mathbf{w}_1}{na_k}I(F_d) = o_p(\sqrt{n})$$

Therefore, the leading term of (7.22) is

$$\frac{1}{\sqrt{n}}\sum_{k=1}^{n}(\mathbb{E}_k - \mathbb{E}_{k-1})\frac{(\mathbf{w}_1^\intercal\mathbf{x}_k)^2}{a_k}I(F_d)$$

$$= \frac{1}{\sqrt{n}}\sum_{k=1}^{n}(\mathbb{E}_k - \mathbb{E}_{k-1})\frac{(1-a_k)(\mathbf{w}_1^\intercal\mathbf{x}_k)^2}{a_k}I(F_d) + \frac{1}{\sqrt{n}}\sum_{k=1}^{n}(\mathbb{E}_k - \mathbb{E}_{k-1})(\mathbf{w}_1^\intercal\mathbf{x}_k)^2I(F_d).$$

Similar to (7.21), by (7.16) we can show that

$$\frac{1}{\sqrt{n}} \sum_{k=1}^{n} (\mathbb{E}_k - \mathbb{E}_{k-1}) \frac{(1 - a_k)(\mathbf{w}_1^\intercal \mathbf{x}_k)^2}{a_k} I(F_d) = o_p(1).$$

It suffices to show CLT for

$$(7.23) \quad \frac{1}{\sqrt{n}} \sum_{k=1}^{n} (\mathbb{E}_k - \mathbb{E}_{k-1})(\mathbf{w}_1^\intercal \mathbf{x}_k)^2 = \frac{1}{\sqrt{n}} \sum_{k=1}^{n} \left[(\mathbf{w}_1^\intercal \mathbf{x}_k)^2 - 1 \right].$$

By the CLT for the sum of i.i.d. variables, we conclude that

$$\frac{1}{\sqrt{n}\sigma} \sum_{k=1}^{n} (\mathbb{E}_k - \mathbb{E}_{k-1})(\mathbf{w}_1^\intercal \mathbf{x}_k)^2 \xrightarrow{D} N(0, \sigma^2),$$

where

$$(7.24) \sigma^2 = \frac{1}{n} \mathbb{E} \left[(\mathbf{w}_1^\intercal \mathbf{x}_k)^2 - 1 \right]^2 = \frac{\sum_{i=1}^{p+l} \gamma_{4i} \mathbf{w}_{1i}^4 + 3 \sum_{i \neq j}^{p+l} \mathbf{w}_{1i}^2 \mathbf{w}_{1j}^2 - 1}{n}$$

$$= \sum_{i=1}^{p+l} (\gamma_{4i} - 3) \mathbf{w}_{1i}^4 + 2.$$

7.2. *Calculation of the Mean.* This section is to calculate the expectation of $\frac{1}{n} \mathbf{w}_1^\intercal \mathbf{X} \mathbf{A}^{-1} \mathbf{X}^\intercal \mathbf{w}_1 I(F_d)$. The strategy is to prove that

$$(7.25) \quad \sqrt{n} \mathbb{E} \left[\frac{1}{n} \mathbf{w}_1^\intercal \mathbf{X}^0 \mathbf{A}^{-1} (\mathbf{X}^0)^\intercal \mathbf{w}_1 I(F_d) + \tilde{m}_\theta(1) \right] \to 0,$$

and

$$(7.26) \frac{1}{\sqrt{n}} \mathbb{E} \left[\mathbf{w}_1^\intercal \mathbf{X} \mathbf{A}^{-1} \mathbf{X}^\intercal \mathbf{w}_1 I(F_d) - \mathbf{w}_1^\intercal \mathbf{X}^0 \mathbf{A}^{-1} (\mathbf{X}^0)^\intercal \mathbf{w}_1 I(F_d) \right] \to 0,$$

where $\mathbf{X}^0 = (\mathbf{x}_1^0, ..., \mathbf{x}_n^0)$ is $(p + l) \times n$ matrix with i.i.d. standard Gaussian random variables. As before, we omit $I(F_d)$ in the following proof.

We prove (7.26) first by the Lindeberg's strategy. Define

$$\mathbf{Z}_k^1 = \sum_{i=1}^{k} \mathbf{x}_i \mathbf{e}_i^\intercal + \sum_{i=k+1}^{n} \mathbf{x}_i^0 \mathbf{e}_i^\intercal, \ \mathbf{Z}_k^0 = \sum_{i=1}^{k-1} \mathbf{x}_i \mathbf{e}_i^\intercal + \sum_{i=k}^{n} \mathbf{x}_i^0 \mathbf{e}_i^\intercal,$$

$$\mathbf{Z}_k = \sum_{i=1}^{k-1} \mathbf{x}_i \mathbf{e}_i^\intercal + \sum_{i=k+1}^{N} \mathbf{x}_i^0 \mathbf{e}_i^\intercal, \ \hat{\mathbf{A}}_k^1 = \mathbf{I} - \frac{1}{n} (\mathbf{Z}_k^1)^\intercal \tilde{\Sigma}_1 \mathbf{Z}_k^1,$$

$$\hat{\mathbf{A}}_k^0 = \mathbf{I} - \frac{1}{n} (\mathbf{Z}_k^0)^\intercal \tilde{\Sigma}_1 \mathbf{Z}_k^0 \quad \text{and} \quad \hat{\mathbf{A}}_k = \mathbf{I} - \frac{1}{n} \mathbf{Z}_k^\intercal \tilde{\Sigma}_1 \mathbf{Z}_k.$$

Then we have $\mathbf{X} = \mathbf{Z}_N^1$, $\mathbf{X}^0 = \mathbf{Z}_1^0$, $\mathbf{Z}_{k+1}^0 = \mathbf{Z}_k^1$. It follows that

$$(7.27) \quad \frac{1}{\sqrt{n}}\mathbb{E}\left[\mathbf{w}_1^\intercal\mathbf{X}\mathbf{A}^{-1}\mathbf{X}^\intercal\mathbf{w}_1 - \mathbf{w}_1^\intercal\mathbf{X}^0\mathbf{A}^{-1}(\mathbf{X}^0)^\intercal\mathbf{w}_1\right]$$

$$= \frac{1}{\sqrt{n}}\sum_{k=1}^n\mathbb{E}\left[\mathbf{w}_1^\intercal\mathbf{Z}_k^1(\hat{\mathbf{A}}_k^1)^{-1}(\mathbf{Z}_k^1)^\intercal\mathbf{w}_1 - \mathbf{w}_1^\intercal\mathbf{Z}_k^0(\hat{\mathbf{A}}_k^0)^{-1}(\mathbf{Z}_k^0)^\intercal\mathbf{w}_1\right]$$

$$= \frac{1}{\sqrt{n}}\sum_{k=1}^n\mathbb{E}\left[\mathbf{w}_1^\intercal\mathbf{Z}_k^1(\hat{\mathbf{A}}_k^1)^{-1}(\mathbf{Z}_k^1)^\intercal\mathbf{w}_1 - \mathbf{w}_1^\intercal\mathbf{Z}_k\hat{\mathbf{A}}_k^{-1}\mathbf{Z}_k^\intercal\mathbf{w}_1\right]$$

$$+ \frac{1}{\sqrt{n}}\sum_{k=1}^n\mathbb{E}\left[\mathbf{w}_1^\intercal\mathbf{Z}_k\hat{\mathbf{A}}_k^{-1}\mathbf{Z}_k^\intercal\mathbf{w}_1 - \mathbf{w}_1^\intercal\mathbf{Z}_k^0(\hat{\mathbf{A}}_k^0)^{-1}(\mathbf{Z}_k^0)^\intercal\mathbf{w}_1\right].$$

For any k, similar to the expansions from (7.11)-(7.20), we can get

$$(7.28) \quad \mathbb{E}\left[\mathbf{w}_1^\intercal\mathbf{Z}_k^1(\hat{\mathbf{A}}_k^1)^{-1}(\mathbf{Z}_k^1)^\intercal\mathbf{w}_1 - \mathbf{w}_1^\intercal\mathbf{Z}_k\hat{\mathbf{A}}_k^{-1}\mathbf{Z}_k^\intercal\mathbf{w}_1\right]$$

$$= \mathbb{E}\left[\frac{(\mathbf{w}_1\mathbf{x}_k)^2}{\hat{a}_k} + \frac{2\mathbf{w}_1^\intercal\mathbf{Z}_k\hat{\mathbf{A}}_k^{-1}\mathbf{Z}_k^\intercal\tilde{\Sigma}_1\mathbf{x}_k\mathbf{x}_k^\intercal\mathbf{w}_1}{n\hat{a}_k} + \frac{\mathbf{x}_k^\intercal\hat{\mathbf{B}}_k\mathbf{x}_k}{n^2\hat{a}_k}\right],$$

where $\hat{\mathbf{B}}_k = \tilde{\Sigma}_1\mathbf{Z}_k\hat{\mathbf{A}}_k^{-1}\mathbf{Z}_k^\intercal\mathbf{w}_1\mathbf{w}_1^\intercal\mathbf{Z}_k\hat{\mathbf{A}}_k^{-1}\mathbf{Z}_k^\intercal\tilde{\Sigma}_1$ and $\hat{a}_k = 1 - \frac{1}{n^2}\mathbf{x}_k^\intercal\tilde{\Sigma}_1\mathbf{Z}_k\hat{\mathbf{A}}_k^{-1}\mathbf{Z}_k^\intercal\tilde{\Sigma}_1\mathbf{x}_k$. Let $\bar{a}_k = 1 - \frac{1}{n^2}tr\tilde{\Sigma}_1\mathbf{Z}_k\hat{\mathbf{A}}_k^{-1}\mathbf{Z}_k^\intercal\tilde{\Sigma}_1$, $\tau_k = \hat{a}_k - \bar{a}_k$. Then we have

$$(7.29) \quad \frac{1}{\hat{a}_k} = \frac{1}{\bar{a}_k} - \frac{\tau_k}{\hat{a}_k\bar{a}_k}.$$

By Lemma 8.10 of [7], we conclude that

$$(7.30) \quad \mathbb{E}|\tau_k|^2 = \mathbb{E}|\frac{1}{n^2}\mathbf{x}_k^\intercal\tilde{\Sigma}_1\mathbf{Z}_k\hat{\mathbf{A}}_k^{-1}\mathbf{Z}_k^\intercal\tilde{\Sigma}_1\mathbf{x}_k - \frac{1}{n^2}tr\tilde{\Sigma}_1\mathbf{Z}_k\hat{\mathbf{A}}_k^{-1}\mathbf{Z}_k^\intercal\tilde{\Sigma}_1|^2$$

$$\leq \frac{C}{n^4}tr(\tilde{\Sigma}_1\mathbf{Z}_k\hat{\mathbf{A}}_k^{-1}\mathbf{Z}_k^\intercal\tilde{\Sigma}_1)^2 = O(\frac{d^2}{p}).$$

Consider the first term at the right hand side of (7.28). It follows from (7.29), (7.30) and Holder's inequality that

$$(7.31)$$

$$|\mathbb{E}(\frac{(\mathbf{w}_1\mathbf{x}_k)^2}{\hat{a}_k} - \frac{(\mathbf{w}_1\mathbf{x}_k)^2}{\bar{a}_k})| = |\mathbb{E}\frac{(\mathbf{w}_1\mathbf{x}_k)^2\tau_k}{\hat{a}_k\bar{a}_k}| \leq C\sqrt{\mathbb{E}(\mathbf{w}_1\mathbf{x}_k)^4}\sqrt{\mathbb{E}\tau_k^2} = O(\frac{d}{\sqrt{p}}).$$

Thus we conclude that

$$\mathbb{E}\frac{(\mathbf{w}_1\mathbf{x}_k)^2}{\hat{a}_k} = \mathbb{E}\frac{(\mathbf{w}_1\mathbf{x}_k)^2}{\bar{a}_k} + O(\frac{d}{\sqrt{p}}) = \mathbb{E}\frac{1}{\bar{a}_k} + o(\frac{1}{\sqrt{n}}).$$

Moreover, a similar approach can be applied to the other terms at the right hand side of (7.28) and thus we have

$$(7.32) \quad \frac{1}{\sqrt{n}} \sum_{k=1}^{n} \mathbb{E} \left[\mathbf{w}_1^\intercal \mathbf{Z}_k^1 (\hat{\mathbf{A}}_k^1)^{-1} (\mathbf{Z}_k^1)^\intercal \mathbf{w}_1 - \mathbf{w}_1^\intercal \mathbf{Z}_k \hat{\mathbf{A}}_k^{-1} \mathbf{Z}_k^\intercal \mathbf{w}_1 \right]$$

$$= \frac{1}{\sqrt{n}} \sum_{k=1}^{n} \mathbb{E} \left[\frac{1}{\bar{a}_k} + \frac{2 \mathbf{w}_1^\intercal \mathbf{Z}_k \hat{\mathbf{A}}_k^{-1} \mathbf{Z}_k^\intercal \tilde{\Sigma}_1 \mathbf{w}_1}{n \bar{a}_k} + \frac{tr \hat{\mathbf{B}}_k}{n^2 \bar{a}_k} \right] + o(1).$$

By the same arguments above, we can also get

$$(7.33) \quad \frac{1}{\sqrt{n}} \sum_{k=1}^{n} \mathbb{E} \left[\mathbf{w}_1^\intercal \mathbf{Z}_k \hat{\mathbf{A}}_k^{-1} \mathbf{Z}_k^\intercal \mathbf{w}_1 - \mathbf{w}_1^\intercal \mathbf{Z}_k^0 (\hat{\mathbf{A}}_k^0)^{-1} (\mathbf{Z}_k^0)^\intercal \mathbf{w}_1 \right]$$

$$= -\frac{1}{\sqrt{n}} \sum_{k=1}^{n} \mathbb{E} \left[\frac{1}{\bar{a}_k} + \frac{2 \mathbf{w}_1^\intercal \mathbf{Z}_k \hat{\mathbf{A}}_k^{-1} \mathbf{Z}_k^\intercal \tilde{\Sigma}_1 \mathbf{w}_1}{n \bar{a}_k} + \frac{tr \hat{\mathbf{B}}_k}{n^2 \bar{a}_k} \right] + o(1).$$

Combining (7.27), (7.32) and (7.33), the equation (7.26) holds.

We next prove (7.25). To simplify notation, we use $\mathbf{X}$ for $\mathbf{X}^0$ and hence assume that $\mathbf{X}$ follows standard normal distribution. By $\mathbf{w}_1^\intercal \mathbf{U}_2^\intercal = 0$, we conclude that $\mathbf{w}_1^\intercal \mathbf{X}$ is independent of $\mathbf{A}$ and hence $\frac{1}{n} \mathbb{E} \mathbf{w}_1^\intercal \mathbf{X} \mathbf{A}^{-1} \mathbf{X}^\intercal \mathbf{w}_1 = \frac{1}{n} \mathbb{E} tr \mathbf{A}^{-1}$. By (6.2.4) of [7](or Lemma 3.1 of [11]), we have

$$\frac{1}{n} \mathbb{E} tr \mathbf{A}^{-1} = \mathbb{E} \frac{1}{1 + \mathbf{r}_1^\intercal \underline{\mathbf{A}}_1^{-1} \mathbf{r}_1},$$

where we denote $\underline{\mathbf{A}} = \tilde{\Sigma}_1^{1/2} \mathbf{X} \mathbf{X}^\intercal \tilde{\Sigma}_1^{1/2} - \mathbf{I}$, $\mathbf{r}_i = \frac{1}{\sqrt{N}} \tilde{\Sigma}_1^{1/2} \mathbf{x}_i$ and $\underline{\mathbf{A}}_j = \sum_{i \neq j} \mathbf{r}_i \mathbf{r}_i^\intercal - \mathbf{I}$. By Lemma 8.10 of [7], we have

$$(7.34) \quad \mathbb{E} | \mathbf{r}_1^\intercal \underline{\mathbf{A}}_1^{-1} \mathbf{r}_1 - \frac{1}{\theta N} tr \underline{\mathbf{A}}_1^{-1} \Sigma_1 | \leq \frac{C}{n^2} tr \tilde{\Sigma}_1^2 = o(M^{-1}),$$

which concludes that $\mathbb{E} \frac{1}{1 + \mathbf{r}_1^\intercal \underline{\mathbf{A}}_1^{-1} \mathbf{r}_1} = \mathbb{E} \frac{1}{1 + \frac{1}{\theta N} tr \underline{\mathbf{A}}_1^{-1} \Sigma_1} + o(n^{-1/2})$. Moreover,

$$\mathbb{E} | \frac{1}{1 + \frac{1}{\theta N} tr \underline{\mathbf{A}}_1^{-1} \Sigma_1} - \frac{1}{1 + \frac{1}{\theta N} \mathbb{E} tr \underline{\mathbf{A}}_1^{-1} \Sigma_1} |^2 \leq \frac{C}{n^2} \mathbb{E} | tr \underline{\mathbf{A}}_1^{-1} \Sigma_1 - \mathbb{E} tr \underline{\mathbf{A}}_1^{-1} \Sigma_1 |^2$$

$$(7.35) \qquad\qquad\qquad\qquad\qquad \leq \frac{C}{n} \mathbb{E} | \beta_{12} \mathbf{r}_2^\intercal \underline{\mathbf{A}}_{12}^{-2} \mathbf{r}_2 |^2 = o(n^{-1}).$$

Hence $\mathbb{E} \frac{1}{1 + \frac{1}{\theta N} tr \underline{\mathbf{A}}_1^{-1} \Sigma_1} = \frac{1}{1 + \frac{1}{\theta N} \mathbb{E} tr \underline{\mathbf{A}}_1^{-1} \Sigma_1} + o(n^{-1/2})$. Define $\beta_i = \frac{1}{1 + \mathbf{r}_i^\intercal \underline{\mathbf{A}}_i^{-1} \mathbf{r}_i}$, $b_i = \frac{1}{1 + \frac{1}{n \theta} \mathbb{E} tr \Sigma_1 \underline{\mathbf{A}}_i^{-1}}$, and $\alpha_i = \mathbf{r}_i^\intercal \underline{\mathbf{A}}_i^{-1} \mathbf{r}_i - \frac{1}{n \theta} tr \Sigma_1 \underline{\mathbf{A}}_i^{-1}$. By the equality that

$$\underline{\mathbf{A}}_1 + \mathbf{I} - b(\theta) \tilde{\Sigma}_1 = \sum_{i \neq 1} \mathbf{r}_i \mathbf{r}_i^\intercal - b(\theta) \tilde{\Sigma}_1,$$

we have

(7.36) $\underline{\mathbf{A}}_1^{-1} = -(\mathbf{I} - b_1(\theta)\tilde{\Sigma}_1)^{-1} + b_1(z)A(\theta) + B(\theta) + C(\theta),$

where

$$A(\theta) = \sum_{i \neq 1}(\mathbf{I} - b_1(\theta)\tilde{\Sigma}_1)^{-1}(\mathbf{r}_i\mathbf{r}_i^\mathsf{T} - \frac{1}{n\theta}\Sigma_1)\underline{\mathbf{A}}_i^{-1},$$

$$B(\theta) = \sum_{i \neq 1}(\beta_i - b_1)(\mathbf{I} - b_1(\theta)\tilde{\Sigma}_1)^{-1}\mathbf{r}_i\mathbf{r}_i^\mathsf{T}\underline{\mathbf{A}}_i^{-1},$$

$$C(\theta) = n^{-1}b_1(\mathbf{I} - b_1(\theta)\Sigma_1)^{-1}\tilde{\Sigma}_1\sum_{i \neq 1}(\underline{\mathbf{A}}_1^{-1} - \underline{\mathbf{A}}_{1i}^{-1}).$$

For $A(\theta)$, similar to (7.34) we have
(7.37)
$$\frac{1}{n}\mathbb{E}|trA(\theta)\tilde{\Sigma}_1| \leq \frac{1}{n}\sum_{i \neq 2}\mathbb{E}|\mathbf{r}_i^\mathsf{T}\underline{\mathbf{A}}_i^{-1}\tilde{\Sigma}_1(\mathbf{I}-b_1(\theta)\tilde{\Sigma}_1)^{-1}\mathbf{r}_i - \frac{1}{n\theta}tr(\Sigma_1\underline{\mathbf{A}}_i^{-1}\tilde{\Sigma}_1(\mathbf{I}-b_1(\theta)\tilde{\Sigma}_1)^{-1})| = o(M^{-1}).$$

Similar to the previous inequalities (7.34)-(7.35) or as in Chapter 9 of [7], we can also show that $B(\theta)$ and $C(\theta)$ are negligible. Hence we get

(7.38) $\frac{1}{n}\mathbb{E}tr\underline{\mathbf{A}}_1^{-1}\tilde{\Sigma}_1 = -\frac{1}{n}tr(\mathbf{I} - b_1(\theta)\tilde{\Sigma}_1)^{-1}\tilde{\Sigma}_1 + o(n^{-1/2}),$

which implies that

(7.39) $\frac{1}{n}\mathbb{E}tr\mathbf{A}^{-1} = \dfrac{1}{1 - \frac{1}{n}tr(\mathbf{I} - \frac{1}{n}(\mathbb{E}tr\mathbf{A}^{-1})\tilde{\Sigma}_1)^{-1}\tilde{\Sigma}_1} + o(n^{-1/2}),$

By the Steiltjes transform of the limit of the ESD of any sample covariance matrix, there exists only one $\tilde{m}_\theta(z)$ such that (One can also refer to (1.6) of [11] or (6.12)-(6.15) of [7])

(7.40) $\tilde{m}_\theta(z) = -\dfrac{1}{z - \frac{1}{n}tr(\mathbf{I} + \tilde{m}_\theta(z)\tilde{\Sigma}_1)^{-1}\tilde{\Sigma}_1}, \quad z \in \mathbb{C}^+.$

Consider the difference between (7.39)-(7.40) and denote $\delta = \frac{1}{n}\mathbb{E}tr\mathbf{A}^{-1} + \tilde{m}_\theta(1)$. It is easy to conclude that

$$\delta(1 + \dfrac{\frac{1}{n}tr\left[(\mathbf{I} - \frac{1}{n}(\mathbb{E}tr\mathbf{A}^{-1})\tilde{\Sigma}_1)^{-1}\tilde{\Sigma}_1(\mathbf{I} + \tilde{m}_\theta(1)\tilde{\Sigma}_1)^{-1}\tilde{\Sigma}_1\right]}{(1 - \frac{1}{n}tr(\mathbf{I} - \frac{1}{n}(\mathbb{E}tr\mathbf{A}^{-1})\tilde{\Sigma}_1)^{-1}\tilde{\Sigma}_1)(1 - \frac{1}{n}tr(\mathbf{I} + \tilde{m}_\theta(1)\tilde{\Sigma}_1)^{-1}\tilde{\Sigma}_1)}) = o(n^{-1/2}).$$

Together with the fact that $\|\tilde{\Sigma}_1\| = O(\theta^{-1})$, it follows that $\delta = o(1/\sqrt{n})$. Therefore, we have shown that

(7.41) $\sqrt{n}(\frac{1}{n}\mathbb{E}tr\mathbf{A}^{-1} + \tilde{m}_\theta(1)) \to 0. \quad \square$

References.

[1] Anderson, T. W. (1984). *An Introduction to Multivariate Statistical Analysis*, 2nd ed. Wiley, New York.

[2] Baik, J and Silverstein, J. W. (2006). Eigenvalues of large sample covariance matrices of spiked population models. *Journal of Multivariate Analysis* **97**, 1382-1408.

[3] Baik, J, Arous, G. B., and Péché, S. (2005). Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *The Annals of Probability* **33**, 1643-1697.

[4] Bai, J. and Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica* **70**, 191-221.

[5] Bai, J. (2003). Inferential Theory for Factor Models of Large Dimensionss. *Econometrica* **71(1)**, 135-171.

[6] Bai, J. and Ng, S. (2008). Large Dimensional Factor Analysis. *Foundations and Trends in Econometrics* **3(2)**, 89-163.

[7] Bai, Z. D. and Silverstein, J. W. (2006). *Spectral Analysis of Large Dimensional Random Matrices*, 1st ed. Springer, New York.

[8] Bai, Z. D. and Yao, J. F., (2008). Central limit theorems for eigenvalues in a spiked population model. *Annales de Institut Henri Poincaré - Probabilités et Statistiques* **44**, 447-474.

[9] Bai, Z. D., Yasunori, F., and Kwok, P. C., (2017). High-dimensional consistency of AIC and BIC for estimating the number of significant components in principal component analysis. *The Annals of Statistics*, to appear.

[10] Bao, Z. G., Pan, G. M., and Zhou, W. (2015). Universality for the largest eigenvalue of sample covariance matrices with general population. *The Annals of Statistics* **43**, 382-421.

[11] Bao, Z. G., Pan, G. M. and Zhou, W. (2014). Local density of the spectrum on the edge for sample covariance matrices with general population. *Preprint.* Available at `http://www.ntu.edu.sg/home/gmpan/publications.html`.

[12] Birnbaum, A., Johnstone, I. M., Nadler, B., and Paul, D. (2013). Minimax bounds for sparse PCA with noisy high-dimensional data. *The Annals of Statistics* **41**, 1055–1084.

[13] Bloemendal, A., Knowles, A., Yau, H. Z., and Yin, J. (2015). On the principal components of sample covariance matrices. *Preprint.* http://arxiv.org/pdf/1404.0788v3.pdf.

[14] Cai, T. T., Ma, Z., and Wu, Y. (2013). Sparse PCA: Optimal rates and adaptive estimation. *The Annals of Statistics* **41**, 3074-3110.

[15] Cai, T. T., Ma, Z., and Wu, Y. (2015). Optimal estimation and rank detection for sparse spiked covariance matrices. *Probability Theory and Related Fields* **161**, 781-815.

[16] Cai, T. T., Han, X., and Pan, G. M. (2017). Supplement to "Limiting Laws for Divergent Spiked Eigenvalues and Largest Non–spiked Eigenvalue of Sample Covariance Matrices".

[17] Chamberlain, G. and Rothschild, M. (1983). Arbitrage, factor structure, and mean-variance analysis on large asset markets. *Econometrica* **51** 1281–1304.

[18] Chen, B. B. and Pan, G. M. (2012). Convergence of the largest eigenvalue of normalized sample covariance matrices when p and n both tend to infinity with their ratio converging to zero. *Bernoulli* **18**, 1405–1420.

[19] Davies, E. B. (1995). The functional calculus. *Journal of the London Mathematical Society* **52**, 166-176.

[20] El Karoui, N. (2007). Tracy-Widom limit for the largest eigenvalue of a large class of complex sample covariance matrices. *The Annals of Probability* **35**, 663-714.

[21] Fan J., Liao Y., and Mincheva M. (2013). Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society: Series B*

(Statistical Methodology) **75**, 603-680.

[22] Fujikoshi, Y., Yamada, T., Watanabe, D. and Sugiyama, T. (2007). Asymptotic distribution of LR statistic for equality of the smallest eigenvalues in high-dimensional principal component analysis. *Journal of Multivariate Analysis* **98**, 2002-2008

[23] Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning 2nd edition*. Springer, New York.

[24] Hoyle, D.C. and Rattray, M. (2004). Principal-component-analysis eigenvalue spectra from data with symmetry-breaking structure. *Physics Review E* **69**, 026124.

[25] Knowles, A. and Yin, J. (2013). The isotropic semicircle law and deformation of Wigner matrices. *Communications on Pure and Applied Mathematics* **66**, 1663-1749.

[26] Knowles, A. and Yin, J. (2015). Anisotropic local laws for random matrices. *arXiv:1410.3516v3.*

[27] Lee, S., Zou, F., and Wright, F. A., (2014). Convergence of sample eigenvalues, eigenvectors, and principal component scores for ultra-high dimensional data. *Biometrika* **101**, 484-490.

[28] Jiang, T. (2005). Maxima of entries of Haar distributed matrices. *Probability Theory and Related Fields* **131(1)**, 121-144.

[29] Johnstone. I. M. (2001). On the distribution of the largest eigenvalue in principal component analysis. *The Annals of Statistics* **29**, 295-327.

[30] Johnstone, I. M. and Lu, A. Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association* **104**, 682-693.

[31] Jung, S. and Marron, J. S. (2009). PCA consistency in High dimension, low sample size context. *The Annals of Statistics* **37**, 4104-4130.

[32] Nadler, B. (2008). Finite sample approximation results for principal component analysis: A matrix perturbation approach. *The Annals of Statistics* **36** 2791–2817.

[33] Onatski, A. (2009). Testing hypotheses about the number of factors in large factor models. *Econometrica* **77**, 1447-1479.

[34] Onatski, A. (2010). Determining the number of factors from empirical distribution of eigenvalues. *The Review of Economics and Statistics* **92(4)**, 1004-1016.

[35] Onatski, A. (2012). Asymptotics of the principal components estimator of large factor models with weakly influential factors. *Journal of Econometrics* **168(2)**, 244-258.

[36] Paul, D. (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica* **17**, 1617-1642.

[37] Schott, J. R. (2006). A high-dimensional test for the equality of the smallest eigenvalues of a covariance matrix. *Journal of Multivariate Analysis* **97**, 827-843.

[38] Shen, D., Shen, H., Zhu, H. and Marron, J. S. (2013). Surprising asymptotic conical structure in critical sample eigen-directions. http://arxiv.org/abs/1303.6171.

[39] Shi, D. (2013). Asymptotic joint distribution of extreme sample eigenvalues and eigenvectors in the spiked population model. http://arXiv:1304.6113.

[40] Stock, J. and Watson, M. (2002). Forecasting using principal components from a large number of predictors. *Journal of the American statistical association* **97(460)**, 1167-1179.

[41] Wang, W. C. and Fan, J. (2016). Asymptotics of Empirical Eigen-structure for High Dimensional Spiked Covariance. *The Annals of Statistics*, to appear.

[42] Yataa, K and Aoshima, M. (2012). Effective PCA for high-dimension, low-sample-size data with noise reduction via geometric representations. *Journal of Multivariate Analysis*, **105**, 193-215 .

T. Tony Cai   Xiao Han and Guangming Pan
3730 Walnut St./Jon M. Huntsman Hall   School of Physical & Mathematical Sciences
Department of Statistics   Nanyang Technological University
The Wharton School   Nanyang Link, Singapore 637371
University of Pennsylvania
Philadelphia, PA 19104
E-mail: tcai@wharton.upenn.edu
        xhan011@e.ntu.edu.sg
        gmpan@ntu.edu.sg