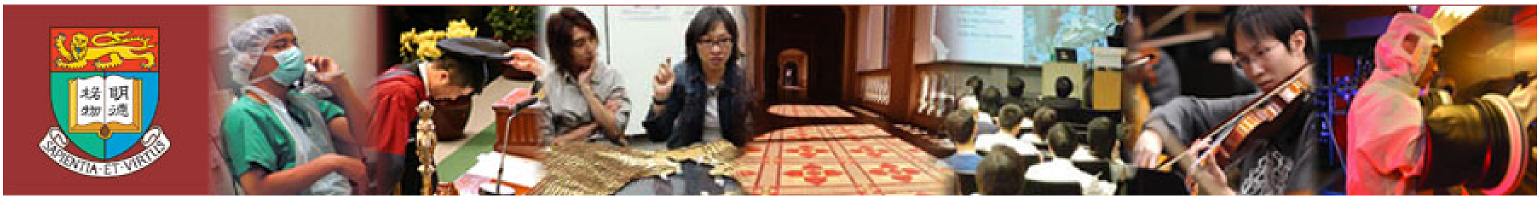

On estimation of the noise variance in high-dimensional probabilistic principal component analysis

Damien Passemier, Zhaoyuan Li and Jianfeng Yao

Department of Statistics and Actuarial Science

The University of Hong Kong

Abstract

In this paper, we develop new statistical theory for probabilistic principal component analysis models in high dimensions. The focus is the estimation of the noise variance, which is an important and unresolved issue when the number of variables is large in comparison with the sample size. We first unveil the reasons of an observed downward bias of the maximum likelihood estimator of the noise variance when the data dimension is high. We then propose a bias-corrected estimator using random matrix theory and establish its asymptotic normality. The superiority of the new and bias-corrected estimator over existing alternatives is checked by Monte-Carlo experiments with various combinations of (p, n) (dimension and sample size). Next, we construct a new criterion based on the bias-corrected estimator to determine the number of the principal components, and a consistent estimator is obtained. Its good performance is confirmed by simulation study and real data analysis. The bias-corrected estimator is also used to derive new asymptotic for the related goodness-of-fit statistic under the high-dimensional scheme.

Keywords. Probabilistic principal component analysis, high-dimensional data, noise variance estimator, number of principal components, random matrix theory, goodness-of-fit.

1 Introduction

Principal component analysis (PCA) is a very popular technique in multivariate analysis for dimensionality reduction and feature extraction. Due to dramatic development in data-collection technology, high-dimensional data are nowadays common in many fields. Natural high-dimensional data, such as images, signal processing, documents and biological data often reside in a low-dimensional subspace or low-dimensional manifold (Ding et al., 2011). In financial econometrics, it is commonly believed that the variations in a large number of economic variables can be modeled by a small number of reference variables (Forni et al., 2000; Bai and Ng, 2002; Bai, 2003). Consequently, PCA is a recommended tool for analysis of such high-dimensional data.

There is an underlying probabilistic model behind PCA, called probabilistic principal component analysis (PPCA), defined as follows. The observation vectors $\{\mathbf{x}_i\}_{1\leq i\leq n}$ are p-dimensional and satisfy the equation

$$\mathbf{x}_i = \mathbf{\Lambda}\mathbf{f}_i + \mathbf{e}_i + \boldsymbol{\mu}\ , \quad i = 1, \ldots, n. \tag{1}$$

Here, $\mathbf{f}_i$ is a m-dimensional *principal components* with $m \ll p$, $\mathbf{\Lambda}$ is a $p \times m$ matrix of *loadings*, and $\boldsymbol{\mu}$ represents the general mean and $\{\mathbf{e}_i\}_{1\leq i\leq n}$ are a sequence of independent errors with covariance matrix $\mathbf{\Psi} = \sigma^2\mathbf{I}_p$. The parameter σ^2 is the noise variance we are interested in. None of the quantities in the right-hand side of (1) is known or observed (except their sum $\mathbf{x}_i$).

To ensure the identification of the model, constraints have to be introduced on the parameters. There are several possibilities for the choice of such constraints, see Table 1 in Bai and Li (2012). A traditional choice is the following (Anderson, 2003, Chapter 14):

- $\mathbb{E}\mathbf{f}_i = \mathbf{0}$ and $\mathbb{E}\mathbf{f}_i\mathbf{f}_i' = \mathbf{I}$;
- The matrix $\mathbf{\Gamma} := \mathbf{\Lambda}'\mathbf{\Lambda}$ is diagonal with distinct diagonal elements.

Therefore, the population covariance matrix of $\{\mathbf{x}_i\}_{1\leq i\leq n}$ is $\mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Lambda}' + \sigma^2\mathbf{I}$. Finding a reliable estimator of the noise variance σ^2 is a nontrivial issue for high-dimensional data which we now pursue.

The PPCA model (1) can be viewed as a special instance of the approximate factor model (Chamberlain and Rothschild, 1983) where the noise covariance $\mathbf{\Psi}$ can be a general diagonal matrix (the model is also called a strict factor model in statistical literature, see Anderson, 2003). For related recent papers on inference of large approximate (or dynamic) factor models, we refer to Bai (2003), Forni et al. (2000) and Doz et al. (2012).

Let $\bar{\mathbf{x}}$ be the sample mean and define the sample covariance matrix

$$\mathbf{S}_n = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'. \tag{2}$$

Let $\lambda_{n,1} \geq \lambda_{n,2} \geq \cdots \geq \lambda_{n,p}$ be the eigenvalues of $\mathbf{S}_n$. Under the normality assumption on both $\{\mathbf{f}_i\}$ and $\{\mathbf{e}_i\}$, the maximum likelihood estimator of σ^2 is

$$\widehat{\sigma}^2 \quad = \quad \frac{1}{p-m}\sum_{i=m+1}^{p}\lambda_{n,i}. \tag{3}$$

In the classic setting where the dimension p is relatively small compared to the sample size n (low-dimensional setting), the consistency of $\widehat{\sigma}^2$ is established in Anderson and Rubin (1956). Moreover, it is asymptotically normal with the standard $\sqrt{n}$-convergence rate: as $n \to \infty$,

$$\sqrt{n}(\widehat{\sigma}^2 - \sigma^2) \xrightarrow{\mathcal{D}} \mathcal{N}(0, s^2), \quad s^2 = \frac{2\sigma^4}{p-m}. \tag{4}$$

Actually, Anderson and Amemiya (1988) provides a general CLT in an approximate factor model that encompasses the present PPCA model. For the reader's convenience, we provide in Supplement a detailed deviation of (4) from this general CLT.

The situation is, however radically different when p is large compared to the sample size n. Recent advances in high-dimensional statistics indicate that in such high-dimensional situation, the above asymptotic result is no more valid and indeed, it has been reported in the literature that $\widehat{\sigma}^2$ seriously underestimates the true noise variance σ^2, see Kritchman and Nadler (2008). However, the exact form of this bias has not been determined (to our best knowledge). Basically, when p becomes large, the sample principal eigenvalues and principal components are no longer consistent estimates of their population counterparts (Baik and Silverstein, 2006; Johnstone and Lu, 2009; Kritchman and Nadler, 2008). Many estimation methods developed in low-dimensional setting have been shown to perform poorly even for moderately large p and n (Cragg and Donald, 1997).

As all meaningful inference procedures in the model will unavoidably use some estimator of the noise variance σ^2, such a severe bias needs to be corrected for high-dimensional data. There are several estimators proposed to deal with the high-dimensional situation. Kritchman and Nadler (2008) proposes an estimator by solving a system of implicit equations; Ulfarsson and Solo (2008) introduces an estimator using the median of the sample eigenvalues $\{\lambda_{n,i}\}$; and Johnstone and Lu (2009) uses the median of the sample variances. However, these estimators are assessed by Monte-Carlo experiments only and their theoretical properties have not been investigated.

The main aim of this paper is to provide a new estimator of the noise variance for which a rigorous asymptotic theory can be established in the high-dimensional setting. First, by using recent advances in random matrix theory, we found a CLT for the m.l.e. $\widehat{\sigma}^2$ in the high-dimensional setting. Next, using this identification, we propose a new estimator $\widehat{\sigma}^2_*$ for the noise variance by correcting this bias. The asymptotic normality of the new estimator is thus established with explicit asymptotic mean and variance.

Although the asymptotic Gaussian distribution of the new estimator $\widehat{\sigma}^2_*$ is established under the high-dimensional setting $p \to \infty$, $n \to \infty$ and $p/n \to c > 0$, if we set $c = 0$, i.e. the dimension p is infinitely smaller than n, this Gaussian limit coincides with the classical low-dimensional limit given in (4). In this sense, the new asymptotic theory extends in a continuous manner the classical low-dimensional result to the high-dimensional situation. Finite sample properties of the new estimator $\widehat{\sigma}^2_*$ have been checked via Monte-Carlo experiments in comparison with the above-mentioned four existing estimators. In terms of mean squared errors and in all the tested scenarios, $\widehat{\sigma}^2_*$ outperforms very significantly three of them, and is slightly preferable than the other one, see Table 2.

In order to demonstrate further potential benefits of the new estimator $\widehat{\sigma}^2_*$, we consider an important inference problem in PPCA, namely, the determination of the number of principal components (PCs). Bai and Ng (2002) developed six criteria with penalty on both p and n to identify the number of factors in the approximate factor model. The approximate factor model allows the components of the errors $\{\mathbf{e}_i\}$ be correlated. PPCA can be considered as a simplified instance of this model and indeed, Bai and Ng (2002) also applied their criteria to PPCA. It is worth noticing that the determination of the number m of PCs and the estimation of the noise variance σ^2 are inter-related and Bai and Ng's criteria provide a consistent and joint inference on (m, σ^2) in the high-dimensional context. However, this consistency is obtained under the assumption that the variances (or the strengths) of the PCs grow up to infinity with the dimension (see Assumption B of their paper), while in our context these variances could be bounded. Therefore, we propose a modified estimator of both (m, σ^2) by implementing our new estimator $\widehat{\sigma}_*^2$ in the criteria of Bai and Ng (2002). Furthermore, in order to deal with possibly bounded variances of PCs, a new penalty function is found based on our new estimator $\widehat{\sigma}_*^2$. The resulting procedure provides a consistent joint estimator of (m, σ^2). Moreover, as predicted by our theory, this new procedure has a better performance than the original Bai and Ng's procedures in our context with possibly bounded PC variances.

As a final application of the new estimator $\widehat{\sigma}_*^2$, we consider the goodness-of-fit test for the PPCA model. The likelihood ratio test statistic as well as their classical (low-dimensional) chi-squared asymptotic theory are well-known since the work of Amemiya and Anderson (1990). These results are again challenged by high-dimensional data and the classical chi-squared limit is no more valid. We propose a correction to this goodness-of-fit test statistic involving our new estimator $\widehat{\sigma}_*^2$ to cope with the high-dimensional effects and establish its asymptotic normality.

The remaining sections are organized as follows. In Section 2, we present the main results of the paper. The new estimator $\widehat{\sigma}_*^2$ of the noise variance is proposed first, then a new joint estimator of the number of PCs and the noise variance is constructed. In Section 3, we develop the corrected likelihood ratio test for the goodness-of-fit of a PPCA model in the high-dimensional framework using the new estimator $\widehat{\sigma}_*^2$. Section 4 concludes. The most important technical proofs are gathered in Appendix while the remaining ones are relegated to the supplementary report. This report contains also many numerical results and additional applications. Lastly, all the codes permitting the reproduction of the results in Tables 1-7 of the paper and some of the data sets used in the paper are available at `http://web.hku.hk/~jeffyao/papersInfo.html`

2 Main results

The PPCA model (1) is a spiked population model (Johnstone, 2001) since the eigenvalues of the population covariance matrix $\mathbf{\Sigma}$ are

$$\begin{aligned} \operatorname{spec}(\mathbf{\Sigma}) &= (\alpha_1, \dots, \alpha_m, \underbrace{0, \dots, 0}_{p-m}) + \sigma^2(\underbrace{1, \dots, 1}_{p}) \\ &= \sigma^2(\alpha_1^*, \dots, \alpha_m^*, \underbrace{1, \cdots, 1}_{p-m}), \end{aligned} \tag{5}$$

where $\{\alpha_i\}$ are m non-null eigenvalues of $\mathbf{\Lambda\Lambda}'$ and the notation $\alpha_i^* = \alpha_i/\sigma^2 + 1$ is used. To develop a meaningful asymptotic theory in the high-dimensional context, we assume that p and n are related so that when $n \to \infty$, $c_n = p/(n-1) \to c > 0$, that is, p can be large compared to the sample size n and for the asymptotic theory, p and n tend to infinity proportionally. Let

$$\phi(\alpha) = \alpha + \frac{c\alpha}{\alpha - 1}\ , \quad \alpha \neq 1\ .$$

Following Baik and Silverstein (2006), assumed that $\alpha_1^* \geq \cdots \geq \alpha_m^* > 1 + \sqrt{c}$, i.e all the eigenvalues α_i are greater than $\sigma^2\sqrt{c}$. It is then known that, for the spiked sample eigenvalues $\{\lambda_{n,i}\}_{1 \leq i \leq m}$ of $\mathbf{S}_n$, almost surely,

$$\lambda_{n,i} \to \sigma^2\phi(\alpha_i^*) = \psi(\sigma^2, c, \alpha_i) = \alpha_i + \sigma^2 + \sigma^2 c\left(1 + \frac{\sigma^2}{\alpha_i}\right). \tag{6}$$

Moreover, the remaining sample eigenvalues $\{\lambda_{n,i}\}_{m < i \leq p}$, called *noise eigenvalues*, will converge to a continuous distribution with support interval $[a(c), b(c)]$ where $a(c) = \sigma^2(1-\sqrt{c})^2$ and $b(c) = \sigma^2(1+\sqrt{c})^2$. In particular, for all $1 \leq j \leq L$ with a prefixed range L and almost surely, $\lambda_{n,m+j} \to b(c)$. It is worth noticing that, if $c \to 0$, we recover the low-dimensional limits $\lambda_{n,i} \to \alpha_i + \sigma^2$ (population spike eigenvalues) and $\lambda_{n,i} \to \sigma^2$ (population noise eigenvalues) discussed earlier. In addition, CLT for the spiked eigenvalues is established in Bai and Yao (2008): $\sqrt{n}(\lambda_{n,i} - \sigma^2\phi(\alpha_i^*))$ is asymptotically Gaussian.

As explained in Introduction, when the dimension p is large compared to the sample size n, the m.l.e. $\widehat{\sigma}^2$ has a negative bias. In order to identify this bias, we first establish a CLT for $\widehat{\sigma}^2$ under the high-dimensional scheme.

Theorem 1. *Consider the PPCA model (1) with population covariance matrix* $\mathbf{\Sigma} = \mathbf{\Lambda\Lambda}' + \sigma^2\mathbf{I}_p$ *where both the principal components and the noise are Gaussian. Assume that* $p \to \infty$, $n \to \infty$ *and* $c_n = p/(n-1) \to c > 0$, *and the non-null eigenvalues of* $\mathbf{\Lambda\Lambda}'$ $\{\alpha_i\}$ *satisfy* $\alpha_i \geq \sigma^2\sqrt{c}$ $(1 \leq i \leq m)$. *Then, we have*

$$\frac{(p-m)}{\sigma^2\sqrt{2c}}(\widehat{\sigma}^2 - \sigma^2) + b(\sigma^2) \xrightarrow{\mathcal{D}} \mathcal{N}(0,1),$$

where $b(\sigma^2) = \sqrt{\frac{c}{2}}\left(m + \sigma^2\sum_{i=1}^m \frac{1}{\alpha_i}\right)$.

The proof is given in Appendix. Therefore, for high-dimensional data, the m.l.e. $\widehat{\sigma}^2$ has an asymptotic bias $-b(\sigma^2)$ (after normalization). This bias is a complex function of the noise variance and the m non-null eigenvalues of the loading matrix $\mathbf{\Lambda\Lambda}'$. The above CLT is still valid if $\tilde{c}_n = (p-m)/n$ is substituted for c. Now if indeed $p \ll n$, i.e. the dimension p is infinitely smaller than the sample size n, so that $\tilde{c}_n \simeq 0$ and $b(\sigma^2) \simeq 0$, and hence

$$\frac{(p-m)}{\sigma^2\sqrt{2c}}(\widehat{\sigma}^2 - \sigma^2) + b(\sigma^2) \simeq \frac{\sqrt{p-m}}{\sigma^2\sqrt{2}}\sqrt{n}(\widehat{\sigma}^2 - \sigma^2) \xrightarrow{\mathcal{D}} \mathcal{N}(0,1)\ .$$

This is the CLT (4) for $\widehat{\sigma}^2$ known under the classical low-dimensional scheme. In a sense, Theorem 1 constitutes a natural and continuous extension of the classical CLT to the high-dimensional context.

Theorem 1 recommends to correct the negative bias of $\widehat{\sigma}^2$. As the bias depends on σ^2 which we want to estimate, a natural correction is to use the plug-in estimator

$$\widehat{\sigma}^2_* = \widehat{\sigma}^2 + \frac{b(\widehat{\sigma}^2)}{p-m}\widehat{\sigma}^2\sqrt{2c_n}. \tag{7}$$

This estimator will be hereafter referred as the *bias-corrected estimator*. The following CLT is an direct consequence of Theorem 1.

Theorem 2. *We assume the same conditions as in Theorem 1. Then, we have*

$$\frac{p-m}{\sigma^2\sqrt{2c_n}}\left(\widehat{\sigma}_*^2-\sigma^2\right)\xrightarrow{\mathcal{D}}\mathcal{N}(0,1)\ .$$

The proof is given in Appendix. Compared to the m.l.e. $\widehat{\sigma}^2$ in Theorem 1, the bias-corrected estimator $\widehat{\sigma}_*^2$ has no more a bias after normalization by $\frac{p-m}{\sigma^2\sqrt{2c_n}}$, and it should be a much better estimator than $\widehat{\sigma}^2$.

For the implementation of $\widehat{\sigma}_*^2$ in practice, we need the value of $b(\widehat{\sigma}^2)$ which depends on the unknown spike values $\{\alpha_i\}$. It is remarked that only consistent estimate of $b(\widehat{\sigma}^2)$ is needed here and this is achieved by substituting some consistent estimates $\widehat{\alpha}_i$ for α_i in $b(\widehat{\sigma}^2)$. This is done as follows. Following Theorem 1, $\widehat{\sigma}^2\xrightarrow{P}\sigma^2$. Then using the function ψ in (6), and by solving in α_i the equation $\lambda_{n,i}=\psi(\widehat{\sigma}^2,p/n,\alpha_i)$, we find an estimator $\widehat{\alpha}_i$ for α_i. Since $p/n\to c$, $\widehat{\sigma}^2\xrightarrow{P}\sigma^2$ and ψ is known to be invertible, we deduce easily that $\widehat{\alpha}_i\xrightarrow{P}\alpha_i$. This procedure will be used for real data analysis in Section 2.4.

2.1 Monte-Carlo experiments

We first check by simulation the effect of bias-correction obtained in $\widehat{\sigma}_*^2$ and its asymptotic normality. Independent Gaussian samples of size n are considered in three different settings:

- Model 1: $\text{spec}(\mathbf{\Sigma})=(25,16,9,0,\ldots,0)+\sigma^2(1,\ldots,1)$, $\sigma^2=4$, $c=1$;
- Model 2: $\text{spec}(\mathbf{\Sigma})=(4,3,0,\ldots,0)+\sigma^2(1,\ldots,1)$, $\sigma^2=2$, $c=0.2$;
- Model 3: $\text{spec}(\mathbf{\Sigma})=(12,10,8,8,0,\ldots,0)+\sigma^2(1,\ldots,1)$, $\sigma^2=3$, $c=1.5$.

Table 1: Comparison between the empirical and the theoretical bias.

Settings			Empirical bias	Theoretical bias	|Difference|
Mod.	p	n			
1	100	100	-0.1556	-0.1589	0.0024
	400	400	-0.0391	-0.0388	0.0003
	800	800	-0.0197	-0.0193	0.0003
2	20	100	-0.0625	-0.0704	0.0052
	80	400	-0.0166	-0.0162	0.0027
	200	1000	-0.0064	-0.0064	0.0011
3	150	100	-0.1609	-0.1634	0.0025
	600	400	-0.0401	-0.0400	0.0001
	1500	1000	-0.0161	-0.0159	0.0002

In Table 1, we compare the empirical bias of $\widehat{\sigma}^2$ (i.e. the empirical mean of $\widehat{\sigma}^2 - \sigma^2 = \frac{1}{p-m}\sum_{i=m+1}^{p} \lambda_{n,i} - \sigma^2$) over 1000 replications with the theoretical one $-\sigma^2\sqrt{2c}b(\sigma^2)/(p-m)$. In all the three models, the empirical and theoretical bias are close each other. As expected, their difference vanishes when p and n increase. The table also shows that this bias is quite significant even for large dimension and sample size such as $(p, n) = (1500, 1000)$. In addition, we have drawn the histograms from 1000 replications of $(p-m)(\sigma^2\sqrt{2c_n})^{-1}(\widehat{\sigma}^2 - \sigma^2) + b(\sigma^2)$ of the three models above, with sample size $n = 100$ and dimensions $p = c \times n$ and they match very well the density of the standard Gaussian distribution (see the supplementary report).

Next, we compare our bias-corrected estimator $\widehat{\sigma}^2_*$ to the m.l.e. $\widehat{\sigma}^2$ and other three existing estimators in the literature. For the reader's convenience, we recall their definitions:

1. The estimator $\widehat{\sigma}^2_{\text{KN}}$ of Kritchman and Nadler (2008): it is defined as the solution of the following non-linear system of $m+1$ equations involving the $m+1$ unknowns $\widehat{\rho}_1, \ldots, \widehat{\rho}_m$ and $\widehat{\sigma}^2_{\text{KN}}$:

$$\widehat{\sigma}^2_{\text{KN}} - \frac{1}{p-m}\left[\sum_{j=m+1}^{p} \lambda_{n,j} + \sum_{j=1}^{m}(\lambda_{n,j} - \widehat{\rho}_j)\right] = 0, \quad \text{and}$$

$$\widehat{\rho}_j^2 - \widehat{\rho}_j\left(\lambda_{n,j} + \widehat{\sigma}^2_{\text{KN}} - \widehat{\sigma}^2_{\text{KN}}\frac{p-m}{n}\right) + \lambda_{n,j}\widehat{\sigma}^2_{\text{KN}} = 0, \quad j = 1, \ldots, m.$$

   We use the code available on the authors' web-page to carry out the simulation. Notice that $\widehat{\sigma}^2_{\text{KN}}$ is only implicitly defined and no precise asymptotic analysis has been provided in Kritchman and Nadler (2008) for $\widehat{\sigma}^2_{\text{KN}}$. We mention one common feature shared by $\widehat{\sigma}^2_{\text{KN}}$ and $\widehat{\sigma}^2_{*}$: both methods use the same relationship between the population spike eigenvalues and their asymptotic limits which are equation (6) in our paper and equation (16) in Kritchman and Nadler (2008) (this leads to the second equation in the system above satisfied by $\widehat{\sigma}^2_{\text{KN}}$).

2. The estimator $\widehat{\sigma}^2_{\text{US}}$ of Ulfarsson and Solo (2008): it is defined as the ratio

$$\widehat{\sigma}^2_{\text{US}} = \frac{\text{median}(\lambda_{n,m+1}, \ldots, \lambda_{n,p})}{m_{p/n,1}},$$

   where $m_{\alpha,1}$ is the median of the Marčenko-Pastur distribution $F_{\alpha,1}$.

3. The estimator $\widehat{\sigma}^2_{\text{median}}$ of Johnstone and Lu (2009): it is defined as the median of the p sample variances (the data $\{x_{ij}\}$ are assumed centered)

$$\widehat{\sigma}^2_{\text{median}} = \text{median}\left(\frac{1}{n}\sum_{i=1}^{n} x_{ij}^2, \quad 1 \le j \le p\right).$$

Table 2 presents the ratios of the empirical MSEs of these estimators over the empirical MSE of the bias-corrected estimator $\widehat{\sigma}^2_{*}$. The performance of $\widehat{\sigma}^2_{*}$ and $\widehat{\sigma}^2_{\text{KN}}$ are similar but $\widehat{\sigma}^2_{*}$ is slightly better. The estimator $\widehat{\sigma}^2_{\text{median}}$ is better than $\widehat{\sigma}^2_{\text{US}}$ and

Table 2: Comparison between four existing estimators and the proposed $\widehat{\sigma}_*^2$ in terms of ratios of MSEs: $\frac{\mathrm{MSE}(\widehat{\sigma}^2)}{\mathrm{MSE}\left(\widehat{\sigma}_*^2\right)}$, $\frac{\mathrm{MSE}\left(\widehat{\sigma}_{\mathrm{KN}}^2\right)}{\mathrm{MSE}\left(\widehat{\sigma}_*^2\right)}$, $\frac{\mathrm{MSE}\left(\widehat{\sigma}_{\mathrm{US}}^2\right)}{\mathrm{MSE}\left(\widehat{\sigma}_*^2\right)}$ and $\frac{\mathrm{MSE}\left(\widehat{\sigma}_{\mathrm{median}}^2\right)}{\mathrm{MSE}\left(\widehat{\sigma}_*^2\right)}$.

Settings				$\widehat{\sigma}^2$	$\widehat{\sigma}_{KN}^2$	$\widehat{\sigma}_{US}^2$	$\widehat{\sigma}_{\mathrm{median}}^2$
Mod.	p	n	σ^2				
1	100	100	4	7.8232	1.0130	14.6394	1.5085
	400	400		8.5905	0.9980	25.5941	1.6429
	800	800		8.1162	1.0019	39.9444	1.6639
2	20	100	2	1.7045	1.0220	2.4980	1.5926
	80	400		2.0406	1.0045	3.8686	1.5433
	200	1000		1.9729	1.0011	3.8731	1.5427
3	150	100	3	19.2114	1.2292	41.7319	1.4274
	600	400		20.8471	0.9958	48.3130	1.6096
	1500	1000		21.6207	1.0001	51.9302	1.8071

the m.l.e. $\widehat{\sigma}^2$. But $\widehat{\sigma}_{\mathrm{median}}^2$ and $\widehat{\sigma}_{\mathrm{US}}^2$ performs poorly compared to $\widehat{\sigma}_*^2$ and $\widehat{\sigma}_{\mathrm{KN}}^2$. The reader is, however reminded that the theoretic properties of $\widehat{\sigma}_{\mathrm{KN}}^2$, $\widehat{\sigma}_{\mathrm{US}}^2$ and $\widehat{\sigma}_{\mathrm{median}}^2$ are unknown and so far they have been checked via simulations only. A careful look at the defining formula of both $\widehat{\sigma}_{\mathrm{US}}^2$ and $\widehat{\sigma}_{\mathrm{median}}^2$ reveals that these estimators are close to $\widehat{\sigma}^2$, all of them being close to average or median of sample eigenvalues. Therefore they might have a similar performance as $\widehat{\sigma}^2$.

2.2 Extension to non-Gaussian data

In this section, we provide some extension of the main results to cover non-Gaussian data. Following a common approach in high-dimensional statistics (Bai and Saranadasa, 1996), we assume $\mathbf{x}_i$ can be generated as

$$\mathbf{x}_i = \mathbf{A}\mathbf{y}_i + \boldsymbol{\mu}, \tag{8}$$

where $\mathbf{A} = \mathbf{\Sigma}^{1/2}$ and $\mathbf{y}_i = \{y_{ij}\}_{1\le j\le p}$ has p i.i.d and standardized components. We set $\gamma = E|y_{11}|^4 - 1$ ($\gamma = 2$ under normal assumption).

Theorem 3. *Consider the PPCA model (1) where the observation vectors $\{\mathbf{x}_i\}_{1\le i\le n}$ are generated as in (8). Assume that $p \to \infty, n \to \infty$ and $c_n = p/(n-1) \to c > 0$. Then, we have*

$$\frac{(p-m)}{\sigma^2\sqrt{\gamma c}}(\widehat{\sigma}^2 - \sigma^2) + b(\sigma^2) \xrightarrow{\mathcal{D}} \mathcal{N}(0,1).$$

The proof is given in Appendix. Similarly to the the plug-in estimator $\widehat{\sigma}^2_*$ in (7) for Gaussian data, we define a new plug-in estimator

$$\widehat{\sigma}^2_{*0} = \widehat{\sigma}^2 + \frac{b(\widehat{\sigma}^2)}{p-m}\widehat{\sigma}^2\sqrt{\gamma c_n}. \tag{9}$$

Notice that when $\gamma = 2$ (Gaussian case), $\widehat{\sigma}^2_{*0}$ coincides with $\widehat{\sigma}^2_*$.

Theorem 4. *We assume the same conditions as in Theorem 3. Then, we have*

$$\frac{p-m}{\sigma^2\sqrt{\gamma c_n}}\left(\widehat{\sigma}^2_{*0} - \sigma^2\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0,1).$$

The proof of this theorem is omitted because it is the same for Theorem 2. For the implementation of $\widehat{\sigma}^2_{*0}$, there is however a new problem to solve with non-Gaussian data, namely the parameter γ in (9) needs to be first estimated. Again we use the random matrix theory to resolve this issue.

Proposition 1. *Under the same conditions in Theorem 3, as $p, n \to \infty$,*

$$\sum_{i=1}^{p}\lambda^2_{n,i} - p\left(\beta_2 + c_n\beta_1^2\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(c\sigma^4(\gamma-1), v\right),$$

where $v > 0$ denotes an (computable) asymptotic variance, and

$$\beta_1 = \sigma^2 + \frac{1}{p}\sum_{j=1}^{m}\alpha_j, \quad \beta_2 = \sigma^4 + \frac{1}{p}\sum_{j=1}^{m}\alpha_j^2 + \frac{2}{p}\sigma^2\sum_{j=1}^{m}\alpha_j.$$

This result does not provide directly an consistent estimator of γ. Therefore, bootstrap method is applied to the sample eigenvalues $\{\lambda_{n,i}\}_{1\leq i\leq p}$ to get say B bootstrapped sample $\{\lambda^*_{n,i}\}_{1\leq i\leq p}$. We find the bootstrap mean w^* of $w = \left(\sum_{i=1}^{p} \lambda^2_{n,i} - p(\beta_2 + c_n\beta_1^2)\right)/c\sigma^4$ (here σ^2 is approximated by the m.l.e. $\widehat{\sigma}^2$), and finally by letting $w^* = \widehat{\gamma}^* - 1$ we find a bootstrap estimator $\widehat{\gamma}^*$ of the unknown γ. Plugging $\widehat{\gamma}^*$ in (9), the final bias-corrected estimator we propose is

$$\widehat{\sigma}^2_{**} = \widehat{\sigma}^2 + \frac{b(\widehat{\sigma}^2)}{p-m}\widehat{\sigma}^2\sqrt{\widehat{\gamma}^* c_n}. \tag{10}$$

We conclude the section by some simulation experiments to check the performance of $\widehat{\sigma}^2_{**}$ for non-Gaussian data. We start with the same setting of covariance matrix of Table 1 and use independent gamma and continuous uniform distributed variables for the components of $\mathbf{x}_i$'s. The gamma distributed data are drawn from $Gamma(k, \theta)$ with fixed shape parameter $k = 2$ and accordingly determined scale parameter θ; the uniform distributed data are drawn from $U(a, b)$ with fixed $a = 0$ and accordingly determined b. The bootstrap sample are repeated $B = 500$ times. The simulation results are summarized in Table 3. The estimator $\widehat{\sigma}^2_{**}$ has a very good performance for both tested non-Gaussian distributions. The MAD and MSE decrease when p and n increase in all model settings.

2.3 Determination of the number of PCs

So far we have assumed that the number m of PCs is known. It is however desirable to estimate m directly from the data. In the literature, consistent estimators of m have been proposed in the high-dimensional context, e.g. in Kritchman and Nadler (2008), Ulfarsson and Solo (2008), Onatski (2009) and Passemier and Yao (2012). As a benchmark work, Bai and Ng (2002) proposes six criteria to determine the number of PCs (or factors) under the framework of large cross-sections (N)

Table 3: Empirical mean, MAD and MSE of $\widehat{\sigma}^2_{**}$ for gamma and uniform samples.

Settings				Gamma			Continuous uniform		
Mod.	p	n	σ^2	$\widehat{\sigma}^2_{**}$	MAD	MSE	$\widehat{\sigma}^2_{**}$	MAD	MSE
1	100	100	4	4.0807	0.1697	0.0411	4.0150	0.1052	0.0161
	400	400		4.0561	0.0571	0.0040	4.0424	0.0429	0.0022
	800	800		4.0304	0.0306	0.0011	4.0236	0.0237	0.0006
2	20	100	2	1.9570	0.1150	0.0194	1.9175	0.0849	0.0089
	80	400		1.9978	0.0241	0.0009	1.9858	0.0166	0.0004
	200	1000		2.0016	0.0090	0.0001	1.9968	0.0054	$< 10^{-4}$
3	150	100	3	3.3637	0.3637	0.1415	3.3336	0.3336	0.1143
	600	400		3.1060	0.1060	0.0116	3.0957	0.0957	0.0093
	1500	1000		3.0434	0.0434	0.0019	3.0391	0.0391	0.0015

and large time dimensions (T). These criteria are popular and widely used in factor modeling literature: for example, they are one of the starting-blocks of the newly proposed POET-estimator in Fan et al. (2013). Notice that the dimension-sample-size pair is denoted here as (N, T) instead of (p, n). Three of these criteria applicable to PPCA models are

$$PC_j(m) = V(m, \widehat{F}^m) + m\widehat{\sigma}^2_{\mathrm{BN}} g_j(N, T), \quad j \in \{1, 2, 3\}, \tag{11}$$

where $\widehat{\sigma}^2_{\mathrm{BN}}$ is a consistent estimate of $(NT)^{-1}\sum_{i=1}^{N}\sum_{j=1}^{T} E(e_{ij})^2$, $V(m, \widehat{F}^m) = (NT)^{-1}\sum_{i=1}^{N} \hat{\mathbf{e}}_i'\hat{\mathbf{e}}_i$, and $g_j(N, T)$ denote the penalty functions

$$g_1(N, T) = \frac{N+T}{NT}\ln\frac{NT}{N+T}, \quad g_2(N, T) = \frac{N+T}{NT}\ln\widetilde{N}, \quad g_3(N, T) = \frac{\ln\widetilde{N}}{\widetilde{N}},$$

with $\widetilde{N} = \min\{N, T\}$. The corresponding estimators of the number of PCs are $\widehat{m}_j = \arg\min_{0\le m\le m_0} PC_j(m)$, $j \in \{1, 2, 3\}$, where m_0 is a predetermined maximum value of m. In applications, $\widehat{\sigma}^2_{\mathrm{BN}}$ is replaced by $V(m_0, \widehat{F}^{m_0})$. The calculations of $\widehat{\sigma}^2_{\mathrm{BN}}$ and $V(m, \widehat{F}^m)$ have no explicit formula and are based on the estimation of the residuals $\{\hat{\mathbf{e}}_i\}$. It is worth mentioning that $V(m, \widehat{F}^m)$ and $\widehat{\sigma}^2_{\mathrm{BN}}$ are indeed the estimates of the noise variance if the underlying model is the PPCA model.

To start with and in order to assess the quality of our bias-corrected estimator $\widehat{\sigma}^2_*$ in the current context, we substitute $\widehat{\sigma}^2_*$ for empirical $V(m, \widehat{F}^m)$ and $\widehat{\sigma}^2_{\mathrm{BN}}$ in the criteria PC_j's. Notice that the noise variance estimator $\widehat{\sigma}^2_*$ depends on the supposed number m of PCs, and we let $\widehat{\sigma}^2_*(m) = \widehat{\sigma}^2_*$ to denote explicitly this dependency. The modified criteria and estimators using $\widehat{\sigma}^2_*(m)$ are thus

$$PC^*_j(m) = \widehat{\sigma}^2_*(m) + m\widehat{\sigma}^2_*(m_0)g_j(N,T),$$

and

$$\widehat{m}^*_j = \arg\min_{0\le m\le m_0} PC^*_j(m), \quad j \in \{1,2,3\}, \tag{12}$$

respectively. These modified criteria PC^*_j's will be compared below to their original counterparts by simulation.

Under appropriate conditions, Bai and Ng (2002) established the consistency of the criteria PC_j when both N and T grow to infinity. A careful examination of their method reveals that the modified criteria PC^*_j are also consistent under the same conditions (this thus means that potential differences between the two families of criteria are of higher asymptotic order). There is however a main issue here: such consistency results require that the variances of the PCs (or their strengths) grow to infinity with the dimension N, see Assumption B of their paper. This *pervasiveness* assumption is not satisfied in our context where these variances can be weaker and remain bounded. Consequently, the proof of Bai and Ng (2002) does not apply here and we are forced to seek for a new asymptotic result. We thus introduce a new penalty function

$$g(N,T) = \frac{(c + 2\sqrt{c})(1 + T/N^{1+\delta})}{N}, \tag{13}$$

and define a new criterion

$$PC^*(m) = \widehat{\sigma}_*^2(m) + m\widehat{\sigma}_*^2(m_0)g(N,T). \tag{14}$$

Here $\delta > 0$ is a small pre-fixed constant. As another main contribution of the paper, we establish the consistency of the corresponding estimator

$$\widehat{m}^* = \arg\min_{0\leq m\leq m_0} PC^*(m). \tag{15}$$

Theorem 5. *We assume the same conditions as in Theorem 1: in particular $N, T \to \infty$ and $c_n = N/(T-1) \to c > 0$. With the condition $\alpha_i > \sigma^2\sqrt{c}$, we have $\lim_{N,T\to\infty} Prob(\widehat{m}^* = \bar{m}) = 1$ where $\bar{m}$ is the true number of PCs.*

The proof is given in Appendix. Simulation experiments are conducted to show the performance of the new estimator $\widehat{m}^*$ in comparison with both the modified estimators $\widehat{m}_j^*$'s and the original $\widehat{m}_j$'s. As in Bai and Ng (2002), the data are generated from the model:

$$X_{it} = \sum_{j=1}^{m} \lambda_{ij}F_{tj} + \sqrt{\theta}e_{it},$$

where the PCs, the loadings and the errors (e_{it}) are $N(0,1)$ variates, the common component of X_{it} has variance m and the idiosyncratic component has variance θ. The noise variance is $\sigma^2 = \theta$ and $\mathbf{\Lambda} = (\lambda_{ij})$. Typically, a PC corresponding to α_j is detectable when $\alpha_j \geq \sqrt{\frac{N}{T}}\theta$, see (6). We conduct extensive simulation by reproducing the configuration of N and T used in Bai and Ng (2002). In all the experiments, the same value of $\delta = 0.05$ is used in (13).

Tables 4 and 5 report the empirical means of the estimator of the number of PCs over 1000 replications, for $m = 1$ and 5 respectively, with standard errors in parentheses. When a standard error is actually zero, no standard error is thus indicated. For all cases, the predetermined maximum number m_0 of PCs is set

Table 4: Comparison between PC^*, PC_j^* and PC_j for $m = 1, \theta = 1$.

N	T	PC^*	PC_1^*	PC_2^*	PC_3^*	PC_1	PC_2	PC_3
100	40	1.00	1.00	1.00	1.00	1.17(0.37)	1.01(0.10)	3.78(0.75)
100	60	1.00	1.00	1.00	1.00	1.00	1.00	3.63(0.76)
200	60	1.00	1.00	1.00	1.00	1.00	1.00	1.00
500	60	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1000	60	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2000	60	1.00	1.00	1.00	1.00	1.00	1.00	1.00
100	100	1.00	1.00	1.00	1.00	1.00	1.00	5.36(0.80)
40	100	1.00	1.00	1.00	1.00	1.79(0.72)	1.19(0.40)	4.91(0.90)
60	100	1.00	1.00	1.00	1.00	1.01(0.08)	1.00	4.30(0.85)
60	200	1.00	1.00	1.00	1.00	1.00	1.00	1.02(0.16)
10	50	7.19(1.92)	8.00	8.00	8.00	8.00	8.00	8.00
10	100	4.95(3.13)	8.00	8.00	8.00	8.00	8.00	8.00
20	100	1.00(0.03)	1.01(0.15)	1.01(0.12)	1.08(0.53)	6.96(0.88)	6.35(0.98)	7.84(0.40)
100	10	2.10(2.52)	1.08(0.73)	1.03(0.50)	1.15(1.01)	8.00	8.00	8.00
100	20	1.00	1.00(0.03)	1.00(0.03)	1.00(0.03)	5.88(0.76)	5.12(0.77)	7.35(0.63)

to 8. When the true number of PCs is 1 (Table 4), the new criterion PC^* and the modified criteria PC_j^* can correctly detect the number almost surely and the corresponding standard errors are all zeros. In comparison, there are 11 cases where the original criteria PC_j lose efficiency in finding the true number of PCs with a non-zero standard error. In the small dimensions situations (last five rows), all the modified PC_j^* and the original PC_j fail when the value of N is 10: they all report the maximum value m_0. But the new criterion PC^* outperforms the others in all cases in terms of mean and standard error. Meanwhile, the modified criteria PC_j^* globally perform better than the original PC_j's and this establishes the superiority of the bias-corrected estimator $\widehat{\sigma}_*^2$. In Table 5, the common component has variance 5 and the idiosyncratic component has a smaller variance 3, and the situation is a bit more difficult. We can however draw the same conclusion that the new criterion PC^* outperforms the other criteria in all tested cases, and the modified criteria PC_j^* have an overall better performance than the original PC_j's. In both Tables 4 and 5, only a part of the tested combinations of N and T is reported and the

Table 5: Comparison between PC^*, PC_j^* and PC_j for $m = 5, \theta = 3$.

N	T	PC^*	PC_1^*	PC_2^*	PC_3^*	PC_1	PC_2	PC_3
100	40	5.00	4.91(0.30)	4.81(0.41)	4.99(0.11)	5.00(0.03)	5.00	5.59(0.57)
100	60	5.00	5.00(0.04)	4.99(0.11)	5.00	5.00	5.00	5.58(0.57)
200	60	5.00	5.00	5.00	5.00	5.00	5.00	5.00
500	60	5.00	5.00	5.00	5.00	5.00	5.00	5.00
1000	60	5.00	5.00	5.00	5.00	5.00	5.00	5.00
2000	60	5.00	5.00	5.00	5.00	5.00	5.00	5.00
100	100	5.00	5.00	5.00	5.00	5.00	5.00	6.84(0.65)
40	100	4.98(0.14)	4.97(0.17)	4.92(0.27)	5.00(0.04)	5.02(0.12)	5.00	6.22(0.66)
60	100	5.00	5.00(0.04)	4.99(0.08)	5.00	5.00	5.00	6.03(0.64)
60	200	5.00	5.00	5.00	5.00	5.00	5.00	6.03(0.03)
10	50	7.47(1.00)	8.00	8.00	8.00	8.00	8.00	8.00
10	100	5.77(1.53)	8.00	8.00	8.00	8.00	8.00	8.00
20	100	3.74(0.84)	4.74(0.51)	4.62(0.57)	4.92(0.45)	7.11(0.63)	6.65(0.64)	7.85(0.37)
100	10	6.66(1.97)	4.59(1.99)	4.35(1.91)	4.88(2.09)	8.00	8.00	8.00
100	20	4.68(0.51)	3.86(0.79)	3.69(0.81)	4.13(0.73)	6.74(0.63)	6.19(0.62)	7.77(0.43)

other combinations where all criteria detect the right number m with zero error are omitted. Additional simulation results and tables are in the supplementary report. In conclusion, the proposed criterion has the best performance in determining the number of PCs, and the modified criteria perform better by using the bias-corrected estimator proposed in this paper for PPCA model.

2.4 Real data

Though the new and modified estimators seems to perform better than the original ones in the simulation experiments, we now compare them on two real data sets. The first data set contains stock returns. Following Bai and Ng (2002), we extract data from the CRSP US Stock Database using the monthly returns for all common stocks listed in NYSE, Amex, and NASDAQ over twenty years (January 1991 to December 2010). Stocks that do not trade for cumulative two years during the period are deleted. The final data set includes 1913 stocks with 240 monthly

Table 6: Comparison between the modified and the original criteria.

	PC^*	PC_1^*	PC_2^*	PC_3^*	PC_1	PC_2	PC_3
Data 1 ($m_0 = 15$)	2	2	2	2	4	4	6
Data 2 ($m_0 = 20$)	13	18	18	19	20	20	20

returns for each of them ($T = 240, N = 1913$). Notice that the data set does not match exactly the one used in Bai and Ng (2002) as they selected 4883 firms for a shorter period January 1994 to December 1998; however this selected data set is not publicly available. The second one is the fMRI data set. This data set is freely available on the web-site *http://afni.nimh.nih.gov/afni/*. A human brain was scanned when the person performed finger-thumb opposition. There are $T = 124$ observations on 21 brain slices. We pick out one brain slice and only keep the variables (pixels) that significantly corresponded to brain tissue, so that, $N = 1126$ variables are selected. We transform both data respectively so that each series is mean zero. The results of rank estimates of the new and modified criteria on these two data sets are shown in Table 6. The original criteria PC_j display a significant variation for the first data set and fail for the second one by only reporting the maximum value $m_0 = 20$. In contrary, the new criterion PC^* and the modified criteria PC_j^*'s with the proposed variance estimator $\widehat{\sigma}_*^2$ seem mutually consistent by giving very close if not identical rank estimates. In particular, the original criteria PC_j have a significant over-estimation effect and this is much reduced and stabilized either with a more accurate estimation of the noise variance by $\widehat{\sigma}_*^2$ or with the new penalty function $g(N, T)$ in (13).

3 Application to the goodness-of-fit test of a PPCA model

As an additional application of the bias-corrected estimator $\widehat{\sigma}^2_*$, we consider the following goodness-of-fit test for the PPCA model (1). The null hypothesis is

$$\mathcal{H}_0: \ \mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Lambda}' + \sigma^2 \mathbf{I}_p,$$

where the number of PCs m is specified. Following Anderson and Rubin (1956), the likelihood ratio test (LRT) statistic is

$$T_n = -nL^*, \text{ with } L^* = \sum_{j=m+1}^{p} \log \frac{\lambda_{n,j}}{\widehat{\sigma}^2},$$

and $\widehat{\sigma}^2$ is the m.l.e. (3) of the variance. Keeping p fixed while letting $n \to \infty$, the classical low-dimensional theory states that T_n converges to χ^2_q, where $q = p(p+1)/2 + m(m-1)/2 - pm - 1$. However, this classical approximation is again useless in large-dimensional situation. Indeed, this criterion leads to a high false-positive rate (see Table 7).

In a way similar to Section 2, we now construct a corrected version of T_n using the calculus done in Bai et al. (2009) and Zheng (2012). As we consider the logarithm of the eigenvalues of the sample covariance matrix, we will assume in the sequel that $p < n$ and $c < 1$ to avoid null eigenvalues.

Theorem 6. *Assume the same conditions as in Theorem 1 and in addition $c < 1$. Then, we have*

$$v(c)^{-\frac{1}{2}} \left\{ L^* - m(c) - ph(c_n) + \eta + (p-m)\log(\beta) \right\} \ \xrightarrow{\mathcal{D}} \ \mathcal{N}(0,1),$$

where

$$m(c) = \frac{\log(1-c)}{2}, \ h(c_n) = \frac{c_n - 1}{c_n}\log(1-c_n) - 1, \ \eta = \sum_{i=1}^{m}\log(1 + c\sigma^2\alpha_i^{-1}),$$
$$\beta = 1 - \frac{c}{p-m}(m + \sigma^2\sum_{i=1}^{m}\alpha_i^{-1}), \ v(c) = -2\log(1-c) + \frac{2c}{\beta}\left(\frac{1}{\beta} - 2\right) .$$

The proof is given in the supplementary report. The above statistic depends on the unknown variance σ^2 and the spike eigenvalues $\{\alpha_i\}$. First of all, as explained in Section 2, consistent estimates of $\{\alpha_i\}$ are available. By using these estimates and substituting bias-corrected estimate $\widehat{\sigma}^2_*$ for σ^2, we obtain consistent estimates $\widehat{v}(c_n)$, $\widehat{\eta}$ and $\widehat{\beta}$ of $v(c)$, η and β, respectively. Therefore, to test $\mathcal{H}_0$, it is natural to use the statistic

$$\Delta_n := \widehat{v}(c_n)^{-\frac{1}{2}}\left(L^* - m(c_n) - ph(c_n) + \widehat{\eta} + (p-m)\log(\widehat{\beta})\right) .$$

Since Δ_n is asymptotically standard normal, the critical region $\{\Delta_n > q_\alpha\}$ where q_α is the αth upper quantile of the standard normal, will have an asymptotic size α. This test is referred as the corrected likelihood ratio test (CLRT).

Next, we present some simulation experiments to compare the classical likelihood ratio test and the corrected likelihood ratio test. We consider again Models 1 and 2 described in Section 2, and a new one (Model 4):

- Model 1: $\text{spec}(\mathbf{\Sigma}) = (25, 16, 9, 0, \ldots, 0) + \sigma^2(1, \ldots, 1)$, $\sigma^2 = 4$, $c = 0.9$;
- Model 2: $\text{spec}(\mathbf{\Sigma}) = (4, 3, 0, \ldots, 0) + \sigma^2(1, \ldots, 1)$, $\sigma^2 = 2$, $c = 0.2$;
- Model 4: $\text{spec}(\mathbf{\Sigma}) = (8, 7, 0, \ldots, 0) + \sigma^2(1, \ldots, 1)$, $\sigma^2 = 1$, varying c.

Table 7 presents the empirical sizes of the LRT and the CLRT. For the LRT, we use the correction proposed by Bartlett (1950), that is replacing $T_n = -nL^*$

Table 7: Comparison of the empirical size of LRT and CLRT in various settings.

Settings			Empirical size of CLRT	Empirical size of LRT
Mod.	p	n		
1	90	100	0.0522	0.9997
	180	200	0.0515	1.0000
	720	800	0.0483	1.0000
2	20	100	0.0375	0.0321
	80	400	0.0440	0.0368
	200	1000	0.0481	0.0514
4	5	500	0.0122	0.0475
	10	500	0.0217	0.0482
	50	500	0.0421	0.0419
	100	500	0.0438	0.0424
	200	500	0.0498	0.2216
	250	500	0.0501	0.7416
	300	500	0.0461	0.9991

by $\tilde{T}_n = -(n - (2p + 11)/6 - 2m/3)L^*$. The computations are done under 10000 independent replications and the nominal test level is 0.05. The empirical sizes of the CLRT are very close to the nominal one, except when the ratio p/n is very small (less than 0.1). On the contrary, the empirical sizes of the classical LRT are much higher than the nominal level especially when c is not too small, and the test will always reject the null hypothesis when p becomes large. In particular when $p/n \geq \frac{1}{2}$, the LRT test tends to reject automatically the null.

4 Conclusions

In this paper, we propose a bias-corrected estimator of the noise variance for PPCA model in the high-dimensional framework. The main appeal of our estimator is that it is developed under the assumption that $p/n \to c > 0$ as $p, n \to \infty$ and is thus appropriate for a wide range of large-dimensional data sets. Extensive Monte-Carlo experiments demonstrated the superiority of the proposed estimator over several existing estimators (however no theoretical justification has been proposed in the literature for these estimators). In addition, by implementing the proposed estimator of the noise variance within the well-known determination algorithms for the number of principal components proposed by Bai and Ng (2002), we construct a new joint consistent estimator of the pair (m, σ^2) with a new penalty function to cope with non pervasive PCs. In an additional application of our methodology, we develop an asymptotic theory of the goodness-of-fit test for high-dimensional PPCA model. The overall message from the paper is that in a high-dimensional PPCA model, when an estimator of the noise variance σ^2 is needed, the bias-corrected estimator $\widehat{\sigma}_*^2$ from the paper should be recommended.

To conclude, we like to mention an important question that requires further investigation, namely the impact of the size of m and of the PC eigenvalues on the methodology developed in this paper. In the numerical simulations, m is typically small compared to $\min\{p, n\}$. Can one expect different behavior if (a) some of the PC eigenvalues are fairly big compared to σ^2, say of the order $O(p)$, as is often the case in some econometric problems (**?**) or, (b) when m is relatively big, e.g. increasing with p, while many of the PC eigenvalues are small, possibly below the size where the phase transition of eigenvalues take place? Unfortunately, these questions go much beyond the scope the existing literature including this paper: for instance we are not aware of any work capable of integrating both large and small

PC eigenvalues. Similarly, the treatment of varying number m of PC components will require the development of new mathematical techniques. Needless to say, these questions are of fundamental importance and worth much research effort in the future.

References

Y. Amemiya and T. W. Anderson. Asymptotic chi-square tests for a large class of factor analysis models. *Ann. Statist.*, 18(3):1453–1463, 1990.

T. W. Anderson. *An introduction to multivariate statistical analysis.* Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, third edition, 2003.

T. W. Anderson and Y. Amemiya. The asymptotic normal distribution of estimators in factor analysis under general conditions. *Ann. Statist.*, 16(2):759–771, 1988.

T. W. Anderson and H. Rubin. Statistical inference in factor analysis. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, 1954–1955, vol. V*, pages 111–150, Berkeley and Los Angeles, 1956. University of California Press.

J. Bai. Inference theory for factor models of large dimensions. *Econometrica*, 71 (1):135–171, 2003.

J. Bai and K. Li. Statistical analysis of factor models of high dimension. *Ann. Statist.*, 40(1):436–465, 2012.

J. Bai and S. Ng. Determining the number of factors in approximate factor models. *Econometrica*, 70(1):191–221, 2002.

Z. Bai and H. Saranadasa. Effect of high dimension: by an example of a two sample problem. *Statist. Sinica*, 6(2):311–329, 1996.

Z. Bai and J. Yao. Central limit theorems for eigenvalues in a spiked population model. *Ann. Inst. Henri Poincaré Probab. Stat.*, 44(3):447–474, 2008.

Z. Bai, D. Jiang, J. Yao, and S. Zheng. Corrections to LRT on large-dimensional covariance matrix by RMT. *Ann. Statist.*, 37(6B):3822–3840, 2009.

Z. Bai, J. Chen, and J. Yao. On estimation of the population spectral distribution from a high-dimensional sample covariance matrix. *Aust. N. Z. J. Stat.*, 52(4): 423–437, 2010.

J. Baik and J. W. Silverstein. Eigenvalues of large sample covariance matrices of spiked population models. *J. Multivariate Anal.*, 97(6):1382–1408, 2006.

M. S. Bartlett. Test of significance in factor analysis. *Brit. Jour. Psych.*, 3:97–104, 1950.

G. Chamberlain and M. Rothschild. Arbitrage, factor structure, and mean-variance analysis on large asset markets. *Econometrica*, 51(5):1281–1304, 1983.

J. G. Cragg and S. G. Donald. Inferring the rank of a matrix. *Journal of econometrics*, 76(1):223–250, 1997.

X. Ding, L. He, and L. Carin. Bayesian robust principal component analysis. *Image Processing, IEEE Transactions on*, 20(12):3419–3430, 2011.

C. Doz, D. Giannone, and L. Reichlin. A quasi–maximum likelihood approach for large, approximate dynamic factor models. *Rev. Econ. Stat.*, 94(4):1014–1024, 2012.

J. Fan, Y. Liao, and M. Mincheva. Large covariance estimation by thresholding principal orthogonal complements. *J. R. Statist. Soc. B*, 75(4):603–680, 2013.

M. Forni, M. Hallin, M. Lippi, and L. Reichlin. *Reference cycles: the NBER methodology revisited*. Number 2400. Citeseer, 2000.

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

I. M Johnstone and A. Y. Lu. On consistency and sparsity for principal components analysis in high dimensions. *JASA*, 104(486), 2009.

S. Kritchman and B. Nadler. Determining the number of components in a factor model from limited noisy data. *Chem. Int. Lab. Syst.*, 94(10):19–32, 2008.

A. Onatski. Testing hypotheses about the number of factors in large factor models. *Econometrica*, 77(5):1447–1479, 2009.

D. Passemier and J. Yao. On determining the number of spikes in a high-dimensional spiked population model. *Random Matrices: Theory and Applications*, 1(1):1150002, 2012.

M. O. Ulfarsson and V. Solo. Dimension estimation in noisy PCA with SURE and random matrix theory. *IEEE Trans. Signal Process.*, 56(12):5804–5816, 2008.

Q. Wang and J. Yao. On the sphericity test with large-dimensional observations. *Electron. J. Stat.*, 7:2164–2192, 2013.

S. Zheng. Central limit theorems for linear spectral statistics of large dimensional f-matrices. *Ann. Inst. Henri Poincaré Probab. Stat.*, 48(2):444–476, 2012.

Appendix

Proof of Theorem 1. We have $(p-m)\widehat{\sigma}^2 = \sum_{i=1}^{p} \lambda_{n,i} - \sum_{i=1}^{m} \lambda_{n,i}$. By (6),

$$\sum_{i=1}^{m} \lambda_{n,i} \quad \longrightarrow \quad \sum_{i=1}^{m} \left(\alpha_i + \frac{c\sigma^4}{\alpha_i} \right) + \sigma^2 m(1+c) \text{ a.s.} \tag{16}$$

For the first term, we have

$$\begin{aligned} \sum_{i=1}^{p} \lambda_i &= p \int x dF_n(x) \\ &= p \int x \, \mathrm{d}(F_n - F_{c_n,H_n})(x) + p \int x \, \mathrm{d}F_{c_n,H_n}(x) \\ &= G_n(x) + p \int x \, \mathrm{d}F_{c_n,H_n}(x). \end{aligned}$$

By Proposition 1 in the supplement report, the first term is asymptotically normal

$$G_n(x) = \sum_{i=1}^{p} \lambda_{n,i} - p \int x \, \mathrm{d}F_{c_n,H_n}(x) \xrightarrow{\mathcal{D}} \mathcal{N}\left(m(x), v(x)\right),$$

with asymptotic mean

$$m(x) = 0 \tag{17}$$

and asymptotic variance

$$v(x) = 2c\sigma^4. \tag{18}$$

Furthermore, by Lemma 1 of Bai et al. (2010),

$$\int x \, \mathrm{d}F_{c_n,H_n}(x) \quad = \quad \int t \, \mathrm{d}H_n(t) = \sigma^2 + \frac{1}{p} \sum_{i=1}^{m} \alpha_i.$$

So we have

$$\sum_{i=1}^{p} \lambda_{n,i} - p\sigma^2 - \sum_{i=1}^{m} \alpha_i \quad \xrightarrow{\mathcal{D}} \quad \mathcal{N}(0, 2c\sigma^4). \tag{19}$$

By (16) and (19) and using Slutsky's lemma, we obtain

$$(p-m)(\widehat{\sigma}^2-\sigma^2)+c\sigma^2\left(m+\sigma^2\sum_{i=1}^{m}\frac{1}{\alpha_i}\right)\xrightarrow{\mathcal{D}}\mathcal{N}(0,2c\sigma^4).$$

Proof of Theorem 2. We have

$$\begin{aligned}\frac{p-m}{\sigma^2\sqrt{2c_n}}\left(\widehat{\sigma}^2_*-\sigma^2\right) &= \frac{p-m}{\sigma^2\sqrt{2c_n}}\left(\widehat{\sigma}^2-\sigma^2\right)+b\left(\widehat{\sigma}^2\right)\frac{\widehat{\sigma}^2}{\sigma^2}\\ &= \left\{\frac{p-m}{\sigma^2\sqrt{2c_n}}\left(\widehat{\sigma}^2-\sigma^2\right)+b(\sigma^2)\right\}+\frac{1}{\sigma^2}\left\{b\left(\widehat{\sigma}^2\right)\widehat{\sigma}^2-b(\sigma^2)\sigma^2\right\}.\end{aligned}$$

Since $\widehat{\sigma}^2\xrightarrow{\mathcal{P}}\sigma^2$, by continuity, the second expression tends to 0 in probability and the conclusion follows from Theorem 1.

Proof of Theorem 3. By Lemma 2.2 of Wang and Yao (2013), we have

$$G_n(x)=\sum_{i=1}^{p}\lambda_{n,i}-p\int x\,\mathrm{d}F_{c_n,H_n}(x)\xrightarrow{\mathcal{D}}\mathcal{N}(m(x),v(x)).$$

The asymptotic mean does not change for non-Gaussian data, $m(x)=0$, but the asymptotic variance is $v(x)=c\gamma\sigma^4$.

Proof of Theorem 5 We prove that $\lim_{N,T\to\infty}P[PC^*(m)<PC^*(\bar{m})]=0$ for all $m\neq\bar{m}$ and $m\leq m_0$. Notice that by definition,

$$\begin{aligned}&PC^*(m)-PC^*(\bar{m})<0\\ \Leftrightarrow\ &\widehat{\sigma}^2_*(m)-\widehat{\sigma}^2_*(\bar{m})<(\bar{m}-m)\widehat{\sigma}^2_*(m_0)g(N,T)\\ \Leftrightarrow\ &\widehat{\sigma}^2_*(\bar{m})-\widehat{\sigma}^2_*(m)>(m-\bar{m})\widehat{\sigma}^2_*(m_0)g(N,T).\end{aligned}$$

Consider first $m<\bar{m}$. We have by (7)

$$\widehat{\sigma}^2_*(m)-\widehat{\sigma}^2_*(\bar{m})=\{\widehat{\sigma}^2(m)-\widehat{\sigma}^2(\bar{m})\}\{1+o_p(1)\}.$$

Moreover,

$$
\begin{aligned}
(N-m)\{\widehat{\sigma}^2(m)-\widehat{\sigma}^2(\bar{m})\} &= \sum_{m<i\le\bar{m}}\lambda_i-(\bar{m}-m)\widehat{\sigma}^2(\bar{m}) \\
&\ge (\bar{m}-m)\{\lambda_{\bar{m}}-\widehat{\sigma}^2(\bar{m})\}.
\end{aligned}
$$

Since $\lambda_{\bar{m}}\to\sigma^2\left[\frac{\alpha_{\bar{m}}}{\sigma^2}+1+c\left(1+\frac{\sigma^2}{\alpha_{\bar{m}}}\right)\right]$ and $\widehat{\sigma}^2(\bar{m})\to\sigma^2$ (in probability), the lower bound above converges to $(\bar{m}-m)\sigma^2\{\alpha_{\bar{m}}/\sigma^2+c(1+\sigma^2/\alpha_{\bar{m}})\}$ which is positive. The conclusion $P[PC^*(m)<PC^*(\bar{m})]\to 0$ will follow if the penalty satisfies

$$
(N-m)g(N,T)<\frac{\sigma^2}{\widehat{\sigma}^2_*(m_0)}\left[\frac{\alpha_{\bar{m}}}{\sigma^2}+c\left(1+\frac{\sigma^2}{\alpha_{\bar{m}}}\right)\right], \tag{20}
$$

for large N,T. Notice that by assumption $\alpha_{\bar{m}}/\sigma^2>\sqrt{c}$ which implies that $\left[\frac{\alpha_{\bar{m}}}{\sigma^2}+1+c\left(1+\frac{\sigma^2}{\alpha_{\bar{m}}}\right)\right]>c+2\sqrt{c}$. On the other hand, we have $\sigma^2/\widehat{\sigma}^2_*(m_0)=1+\beta/T+o_p(1/T)$ where β is some constant (depending on m_0, c and σ^2). So with the $g(N,T)$ in (13), we have $(N-m)g(N,T)=(c+2\sqrt{c})(1+\frac{T}{N^{1+\delta}})\ll(c+2\sqrt{c})(1+\beta/T+o_p(1/T))$ and the conclusion follows. Next, consider the case where $m>\bar{m}$. We have

$$
\begin{aligned}
(N-m)\{\widehat{\sigma}^2(\bar{m})-\widehat{\sigma}^2(m)\} &= \sum_{\bar{m}<i\le m}\lambda_i-(m-\bar{m})\widehat{\sigma}^2(\bar{m}) \\
&\overset{i.p.}{\to} (m-\bar{m})\sigma^2(c+2\sqrt{c}),
\end{aligned}
$$

due to $\lambda_i\overset{i.p.}{\to}\sigma^2(1+\sqrt{c})^2$ for $\bar{m}<i\le m$. Notice that $\widehat{\sigma}^2_*(m_0)\overset{i.p.}{\to}\sigma^2$ and with the $g(N,T)$ in (13), we have

$$
\liminf_{N,T\to\infty}(N-m)g(N,T)\ge c+2\sqrt{c}.
$$

The conclusion follows.

On estimation of the noise variance in high-dimensional probabilistic principal component analysis

(Supplementary report)

Damien Passemier, Zhaoyuan Li and Jianfeng Yao
Department of Statistics and Actuarial Science
The University of Hong Kong

1 Figure for Section 2.1

Figure 1 presents the histograms from 1000 replications of

$$\frac{(p-m)}{\sigma^2\sqrt{2c_n}}(\widehat{\sigma}^2-\sigma^2)+b(\sigma^2)$$

of the three models in Section 2.1, with sample size $n = 100$ and dimensions $p = c \times n$, compared to the density of the standard Gaussian distribution. The sampling distribution is almost normal.

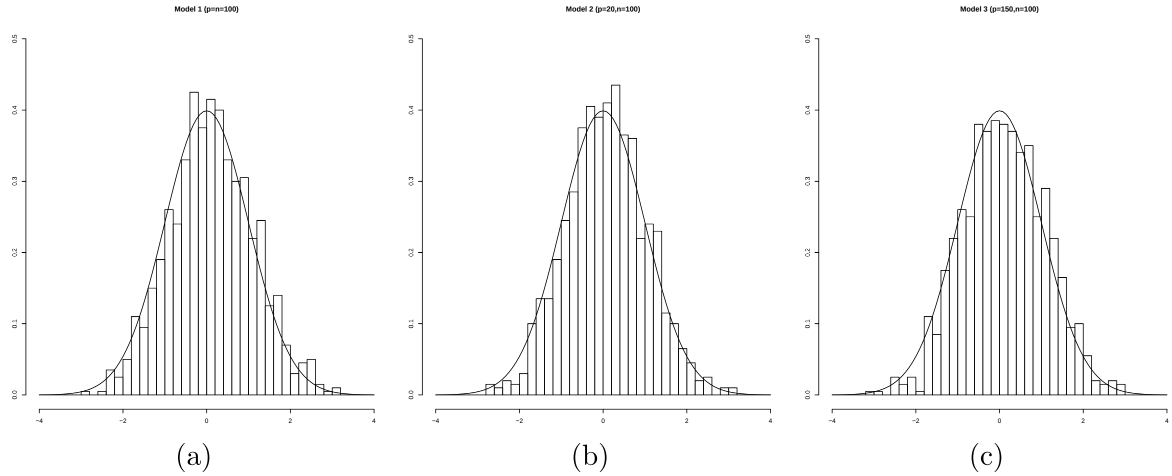

Figure 1. Histogram of $\frac{(p-m)}{\sigma^2\sqrt{2c}}(\widehat{\sigma}^2-\sigma^2)+b(\sigma^2)$ compared with the density of a standard Gaussian distribution.

2 More Monte-Carlo experiments for Section 2.3

Tables 1 and 2 report the empirical means of the estimator of the number of PCs over 1000 replications, for $m = 3$ and 5 respectively, with standard errors in parentheses. When a standard error is actually zero, no standard error is thus indicated. For all cases, the predetermined maximum number m_0 of PCs is set to 8.

3 Application to the SURE criterion

Ulfarsson and Solo (2008) proposes to use the SURE criteria to choose the number of PCs. This criterion uses the noise variance estimator $\widehat{\sigma}^2_{US}$ defined in Section 2. It aims at minimizing the Euclidean distance between the underlying estimator of the population mean $\boldsymbol{\mu}$ and its true value. The proposed SURE criterion for m number of PCs (to be minimized) is

$$\begin{aligned} R_m &= (p-m)\widehat{\sigma}^2_{US} + \widehat{\sigma}^4_{US}\sum_{j=1}^{m}\frac{1}{\lambda_j} + 2\widehat{\sigma}^2_{US}(1-1/n)m \\ &\quad -2\widehat{\sigma}^4_{US}(1-1/n)\sum_{j=1}^{m}\frac{1}{\lambda_j} + \frac{4(1-1/n)\widehat{\sigma}^4_{US}}{n}\sum_{j=1}^{m}\frac{1}{\lambda_j} + C_m, \end{aligned} \tag{1}$$

where

$$\begin{aligned} C_m &= \frac{4(1-1/n)\widehat{\sigma}^2_{US}}{n}\sum_{j=1}^{m}\sum_{i=m+1}^{p}\frac{\lambda_j-\widehat{\sigma}^2_{US}}{\lambda_j-\lambda_i} + \frac{2(1-1/n)\widehat{\sigma}^2_{US}}{n}m(m-1) \\ &\quad -\frac{2(1-1/n)\widehat{\sigma}^2_{US}}{n}(p-1)\sum_{j=1}^{m}\left(1-\frac{\widehat{\sigma}^2_{US}}{\lambda_j}\right). \end{aligned}$$

Recall that $\widehat{\sigma}^2_{US}$ is also related to m. From Section 2, we have known that $\widehat{\sigma}^2_{US}$ is not as good as our bias-corrected estimator. To examine further this difference, we replace $\widehat{\sigma}^2_{US}$ with $\widehat{\sigma}^2_*$ in (11), referred then as SURE*, to see whether the performance of SURE can be improved.

Then simulation experiments are conducted to check the performance of SURE*. The setup follows the paper Ulfarsson and Solo (2008) and the data are simulated according to (1) with the parameters $p = 64, p/n = [2/3, 1/2, 2/5], m = [5, 10, 15, 20]$ and $\sigma^2 = 1$. The loading matrix is set to $\boldsymbol{\Lambda} = \mathbf{F}\mathbf{D}^{1/2}$, where $\mathbf{F}$ is constructed by generating a $p \times m$ matrix of Gaussian random variables and then orthogonalizing the resulting matrix, and $\mathbf{D} = \text{diag}\big((m+1)^2, m^2, \ldots, 3^2, \lambda_m\big), \lambda_m = 1.5$. All simulations were repeated 1500 times. Table 3 presents the percentage of correct selection of number of PCs for SURE and SURE*, and the results of SURE are from Table II of Ulfarsson and Solo (2008). SURE* largely outperforms SURE in all of the tested cases, most of times by a wide margin. All the percentages of correct selection of SURE* are larger than 90% and in 4 out of 12 cases, the

Table 1: Comparison between $PC_p^* s$ and $PC_p s$ in terms of the mean estimation numbers of PCs for $m = 3, \theta = 3$.

N	T	PC_{p1}^*	PC_{p2}^*	PC_{p3}^*	PC_{p1}	PC_{p2}	PC_{p3}
100	40	2.98(0.15)	2.95(0.22)	3.00(0.06)	3.00	3.00	3.90
100	60	3.00(0.03)	3.00(0.04)	3.00	3.01(0.08)	3.00	4.37(0.64)
200	60	3.00	3.00	3.00	3.00	3.00	4.18(0.63)
500	60	3.00	3.00	3.00	3.00	3.00	3.00
1000	60	3.00	3.00	3.00	3.00	3.00	3.00
2000	60	3.00	3.00	3.00	3.00	3.00	3.00
100	100	3.00	3.00	3.00	3.00	3.00	5.62(0.72)
200	100	3.00	3.00	3.00	3.00	3.00	3.00
500	100	3.00	3.00	3.00	3.00	3.00	3.00
1000	100	3.00	3.00	3.00	3.00	3.00	3.00
2000	60	3.00	3.00	3.00	3.00	3.00	3.00
40	100	2.99(0.10)	2.98(0.14)	3.00	3.07(0.26)	3.01(0.07)	5.04(0.72)
60	100	3.00	3.00(0.03)	3.00	3.00	3.00	4.65(0.69)
60	200	3.00	3.00	3.00	3.00	3.00	3.00
60	500	3.00	3.00	3.00	3.00	3.00	3.00
60	1000	3.00	3.00	3.00	3.00	3.00	3.00
60	2000	3.00	3.00	3.00	3.00	3.00	3.00
4000	60	3.00	3.00	3.00	3.00	3.00	3.00
4000	100	3.00	3.00	3.00	3.00	3.00	3.00
8000	60	3.00	3.00	3.00	3.00	3.00	3.00
8000	100	3.00	3.00	3.00	3.00	3.00	3.00
60	4000	3.00	3.00	3.00	3.00	3.00	3.00
100	4000	3.00	3.00	3.00	3.00	3.00	3.00
60	8000	3.00	3.00	3.00	3.00	3.00	3.00
100	8000	3.00	3.00	3.00	3.00	3.00	3.00
10	50	8.00	8.00	8.00	8.00	8.00	8.00
10	100	8.00	8.00	8.00	8.00	8.00	8.00
20	100	2.89(0.32)	2.85(0.37)	2.95(0.27)	6.55(0.74)	5.96(0.77)	7.62(0.55)
100	10	2.57(1.35)	2.43(1.19)	2.77(1.54)	8.00	8.00	8.00
100	20	2.46(0.63)	2.37(0.65)	2.65(0.52)	6.15(0.69)	5.46(0.68)	7.49(0.59)

Table 2: Comparison between PC_p^*s and PC_ps in terms of the mean estimation numbers of PCs for $m = 5, \theta = 5$.

N	T	PC_{p1}^*	PC_{p2}^*	PC_{p3}^*	PC_{p1}	PC_{p2}	PC_{p3}
100	40	3.83(0.77)	3.49(0.77)	4.51(0.58)	5.00(0.07)	4.98(0.15)	5.36(0.51)
100	60	4.66(0.50)	4.36(0.61)	4.98(0.13)	5.00(0.03)	5.00(0.06)	5.27(0.45)
200	60	4.95(0.22)	4.90(0.30)	4.99(0.08)	5.00	5.00	5.00
500	60	5.00(0.04)	5.00(0.07)	5.00(0.03)	5.00	5.00	5.00
1000	60	5.00(0.04)	5.00(0.04)	5.00	5.00	5.00	5.00
2000	60	5.00(0.03)	5.00(0.03)	5.00(0.03)	5.00	5.00	5.00
100	100	4.(0.12)	4.90(0.30)	5.00	5.00	5.00	6.18(0.63)
200	100	5.00	5.00	5.00	5.00	5.00	5.00
500	100	5.00	5.00	5.00	5.00	5.00	5.00
1000	100	5.00	5.00	5.00	5.00	5.00	5.00
2000	60	5.00	5.00	5.00	5.00	5.00	5.00
40	100	4.25(0.68)	3.92(0.75)	4.77(0.44)	4.98(0.04)	5.66(0.14)	5.66(0.57)
60	100	4.76(0.44)	4.47(0.60)	4.76(0.10)	5.00(0.03)	4.99(0.08)	5.46(0.56)
60	200	4.97(0.17)	4.94(0.24)	5.00	5.00	5.00	5.00
60	500	5.00(0.05)	5.00(0.06)	5.00(0.04)	5.00	5.00	5.00
60	1000	5.00(0.03)	5.00(0.03)	5.00	5.00	5.00	5.00
60	2000	5.00	5.00	5.00	5.00	5.00	5.00
4000	60	5.00	5.00	5.00	5.00	5.00	5.00
4000	100	5.00	5.00	5.00	5.00	5.00	5.00
8000	60	5.00	5.00	5.00	5.00	5.00	5.00
8000	100	5.00	5.00	5.00	5.00	5.00	5.00
60	4000	5.00	5.00	5.00	5.00	5.00	5.000
100	4000	5.00	5.00	5.00	5.00	5.00	5.00
60	8000	5.00	5.00	5.00	5.00	5.00	5.00
100	8000	5.00	5.00	5.00	5.00	5.00	5.00
10	50	8.00	8.00	8.00	8.00	8.00	8.00
10	100	8.00	8.00	8.00	8.00	8.00	8.00
20	100	3.64(0.91)	3.38(0.94)	4.08(0.79)	6.65(0.64)	6.12(0.64)	7.63(0.51)
100	10	3.10(2.01)	2.83(1.86)	3.53(2.27)	8.00	8.00	8.00
100	20	2.18(0.92)	1.93(0.92)	2.65(0.0.90)	6.56(0.62)	5.97(0.62)	7.66(0.50)

Table 3: Comparison between SURE and SURE* in terms of percentage of correct selection of PCs.

m		$p/n = 2/3$	$p/n = 1/2$	$p/n = 2/5$
5	SURE*	1.000	1.000	1.000
	SURE	0.408	0.621	0.807
10	SURE*	0.992	1.000	0.998
	SURE	0.512	0.739	0.858
15	SURE*	0.920	0.978	0.989
	SURE	0.598	0.783	0.911
20	SURE*	0.909	0.966	0.990
	SURE	0.617	0.810	0.899

detection rate is 100%. Therefore, by implementing our bias-corrected estimator of the noise variance instead of the one provided by its authors, the SURE criterion has a much better performance.

4 Proofs

Before giving the proofs, we first recall some important results from the random matrix theory which laid the foundation for the proofs of the main results of the paper.

4.1 Useful results from random matrix theory

Random matrix theory has become a powerful tool to address new inference problems in high-dimensional scheme. For general background and references, we refer to review papers Johnstone (2007) and Johnstone and Titterington (2009).

Let H be a probability measure on $\mathbb{R}^+$ and $c > 0$ a constant. We define the map

$$g(s) = g_{c,H}(s) = \frac{1}{s} + c \int \frac{t}{1+ts} \,\mathrm{d}H(t) \tag{2}$$

in the set $\mathbb{C}^+ = \{z \in \mathbb{C} : \Im z > 0\}$. The map g is a one-to-one mapping from $\mathbb{C}^+$ onto itself (see Bai and Silverstein, 2010, Chapter 6), and the inverse map $m = g^{-1}$ satisfies all the requirements of the Stieltjes transform of a probability measure on $[0, \infty)$. We call this measure $\underline{F}_{c,H}$. Next, a companion measure $F_{c,H}$ is introduced by the equation $cF_{c,H} =$

$(c-1)\,\delta_0 + \underline{F}_{c,H}$ (note that in this equation, measures can be signed). The measure $F_{c,H}$ is referred as the generalized Marčenko-Pastur distribution with index (c, H).

Let $F_n = \frac{1}{p}\sum_{i=1}^{p} \delta_{\lambda_{n,i}}$ be the empirical spectral distribution (ESD) of the sample covariance matrix $\mathbf{S}_n$ defined in (2) in main paper with the $\{\lambda_{n,i}\}$ denoting its eigenvalues. Then, it is well-known that under suitable moment conditions, F_n converges to the Marčenko-Pastur distribution of index (c, δ_{σ^2}), simply denoted as F_{c,σ^2}, with the following density function

$$p_{c,\sigma^2}(x) = \begin{cases} \frac{1}{2\pi x c\sigma^2}\sqrt{\{b(c)-x\}\{x-a(c)\}}\ , & a(c) \le x \le b(c)\ , \\ 0\ , & \text{otherwise.} \end{cases}$$

The distribution has an additional mass $(1-1/c)$ at the origin if $c > 1$.

The ESD H_n of $\mathbf{\Sigma}$ is

$$H_n = \frac{p-m}{p}\delta_{\sigma^2} + \frac{1}{p}\sum_{i=1}^{m}\delta_{\alpha_i+\sigma^2}\ ,$$

and $H_n \to \delta_{\sigma^2}$. Define the normalized empirical process

$$G_n(f) = p\int_{\mathbb{R}} f(x)[F_n - F_{c_n,H_n}](\mathrm{d}x),\ f \in \mathcal{A},$$

where $\mathcal{A}$ is the set of analytic functions $f : \mathcal{U} \to \mathbb{C}$, with $\mathcal{U}$ an open set of $\mathbb{C}$ such that $[1_{(0,1)}(c)a(c), b(c)] \subset \mathcal{U}$. We will need the following CLT which is a combination of Theorem 1.1 of Bai and Silverstein (2004) and a recent addition proposed in Zheng et al. (2015).

Proposition 1. *We assume the same conditions as in Theorem 1. Then, for any functions $f_1, \ldots, f_k \in \mathcal{A}$, the random vector $(G_n(f_1), \ldots, G_n(f_k))$ converges to a k-dimensional Gaussian vector with mean vector*

$$m(f_j) = \frac{f_j(a(c)) + f_j(b(c))}{4} - \frac{1}{2\pi}\int_{a(c)}^{b(c)} \frac{f_j(x)}{\sqrt{4c\sigma^4 - (x-\sigma^2-c\sigma^2)^2}}\,\mathrm{d}x,\ j = 1, \ldots, k,$$

and covariance function

$$v(f_j, f_l) \quad = \quad -\frac{1}{2\pi^2}\oint_{\mathcal{C}_1}\oint_{\mathcal{C}_2} \frac{f_j(z_1)f_l(z_2)}{(\underline{m}(z_1)-\underline{m}(z_2))^2}\,\mathrm{d}\underline{m}(z_1)\mathrm{d}\underline{m}(z_2),\ j, l = 1, \ldots, k, \qquad (3)$$

where $\underline{m}(z)$ is the Stieltjes transform of $\underline{F}_{c,\sigma^2} = (1-c)\delta_0 + cF_{c,\sigma^2}$. The contours $\mathcal{C}_1$ and $\mathcal{C}_2$ are non overlapping and both contain the support of F_{c,σ^2}.

An important and subtle point here is that the centering term in $G_n(f)$ in the above CLT is defined with respect to the Marcčenko-Pastur distribution F_{c_n,H_n} with "current" index

(c_n, H_n) instead of the limiting distribution F_{c,σ^2} with index (c, σ^2). In contrast, the limiting mean function $m(f_j)$ and covariance function $v(f_j, f_l)$ depend on the limiting distribution F_{c,σ^2} only.

4.2 Proof of Proposition 1 in main paper

By Lemma 2.2 of Wang and Yao (2013),

$$\sum_{i=1}^{p} \lambda_i^2 - p \int x^2 dF_{c_n,H_n}(x) \xrightarrow{\mathcal{D}} \mathcal{N}(m(x^2), v),$$

with $m(x^2) = c\sigma^4(\gamma - 1)$ and some computable $v > 0$. Furthermore, by Lemma 1 of Bai et al. (2010),

$$\int x^2 dF_{c_n,H_n}(x) = \beta_2 + \frac{p}{n}\beta_1^2,$$

where

$$\beta_1 = \sigma^2 + \frac{1}{p}\sum_{j=1}^{m} \alpha_j, \text{ and } \beta_2 = \sigma^4 + \frac{1}{p}\sum_{j=1}^{m} \alpha_j^2 + \frac{2}{p}\sigma^2 \sum_{j=1}^{m} \alpha_j.$$

The conclusion follows.

4.3 Proof of Theorem 6 in main paper

We have

$$\begin{aligned} L^* &= \sum_{i=m+1}^{p} \log \frac{\lambda_{n,i}}{\widehat{\sigma}^2} \\ &= \sum_{i=m+1}^{p} \log \frac{\lambda_{n,i}}{\sigma^2} - \sum_{i=m+1}^{p} \log \frac{\widehat{\sigma}^2}{\sigma^2} \\ &= \sum_{i=m+1}^{p} \log \frac{\lambda_{n,i}}{\sigma^2} - (p-m)\log\left(\frac{1}{p-m}\sum_{i=m+1}^{p} \frac{\lambda_{n,i}}{\sigma^2}\right) \\ &= L_1 - (p-m)\log\left(\frac{L_2}{p-m}\right), \end{aligned}$$

where we have defined a two-dimensional vector $(L_1, L_2) = (\sum_{i=m+1}^{p} \log \frac{\lambda_{n,i}}{\sigma^2}, \sum_{i=m+1}^{p} \frac{\lambda_{n,i}}{\sigma^2})$.

CLT when $\sigma^2 = 1$. To start with, we consider the case $\sigma^2 = 1$. We have

$$\begin{aligned} L_1 &= p\int \log(x)\,\mathrm{d}F_n(x) - \sum_{i=1}^{m}\log\lambda_{n,i} \\ &= p\int \log(x)\,\mathrm{d}(F_n - F_{c_n,H_n})(x) + p\int \log(x)\,\mathrm{d}F_{c_n,H_n}(x) - \sum_{i=1}^{m}\log\lambda_{n,i}. \end{aligned}$$

Similarly, we have

$$L_2 = p\int x\,\mathrm{d}(F_n - F_{c_n,H_n})(x) + p\int x\,\mathrm{d}F_{c_n,H_n}(x) - \sum_{i=1}^{m}\lambda_{n,i}.$$

By Proposition 1, we find that

$$p\begin{pmatrix} \int \log(x)\,\mathrm{d}(F_n - F_{c_n,H_n})(x) \\ \int x\,\mathrm{d}(F_n - F_{c_n,H_n})(x) \end{pmatrix} \xrightarrow{\mathcal{D}} \mathcal{N}\left(\begin{pmatrix} m_1(c) \\ m_2(c) \end{pmatrix}, \begin{pmatrix} v_1(c) & v_{1,2}(c) \\ v_{1,2}(c) & v_2(c) \end{pmatrix}\right) \tag{4}$$

with $m_2(c) = 0$ and $v_2(c) = 2c$ and

$$m_1(c) = \frac{\log(1-c)}{2}, \tag{5}$$

$$v_1(c) = -2\log(1-c), \tag{6}$$

$$v_{1,2}(c) = 2c. \tag{7}$$

Formulae of m_2 and v_2 have been established in the proof of Theorem 1 and the others are derived in next subsection.

In Theorem 1, with $\sigma^2 = 1$, we found that

$$\int x\,\mathrm{d}F_{c_n,H_n}(x) = 1 + \frac{1}{p}\sum_{i=1}^{m}\alpha_i,$$

and

$$\sum_{i=1}^{m}\lambda_{n,i} \xrightarrow{\text{a.s.}} \sum_{i=1}^{m}\left(\alpha_i + \frac{c}{\alpha_i}\right) + m(1+c).$$

For the last term of L_1, by (6) in main paper, we have

$$\log\lambda_{n,i} \longrightarrow \log(\phi(\alpha_i+1)) = \log\left((\alpha_i+1)(1+c\alpha_i^{-1})\right) \text{ a.s.}$$

Furthermore, by Wang et al. (2014), we have

$$\int \log(x)\,\mathrm{d}F_{c_n,H_n}(x) = \frac{1}{p}\sum_{i=1}^{m}\log(\alpha_i+1) + h(c_n) + o\left(\frac{1}{p}\right),$$

where

$$h(c_n) \quad = \quad \int \log(x)dF_{c_n,\delta_1}(x) = \frac{c_n-1}{c_n}\log(1-c_n) - 1. \tag{8}$$

can be calculated using the density of the Marčenko-Pastur law (see 4.1). Summarising, we have obtained that

$$L_1 - m_1(c) - ph(c_n) + \eta(c,\alpha) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, v_1(c)\right),$$

where $h(c_n) = \frac{c_n-1}{c_n}\log(1-c_n)-1$ and $\eta(c,\alpha) = \sum_{i=1}^{m}\log(1+c\sigma^2\alpha_i^{-1})$. Similarly, we have

$$L_2 - (p-m) + \rho(c,\alpha) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, v_2(c)\right),$$

where $\rho(c,\alpha) = c(m + \sum_{i=1}^{m}\alpha_i^{-1})$.

Using (4) and Slutsky's lemma,

$$\begin{pmatrix} L_1 \\ L_2 \end{pmatrix} \xrightarrow{\mathcal{D}} \mathcal{N}\left(\begin{pmatrix} m_1(c) + ph(c_n) - \eta(c,\alpha) \\ p - m - \rho(c,\alpha) \end{pmatrix}, \begin{pmatrix} v_1(c) & v_{1,2}(c) \\ v_{1,2}(c_n) & v_2(c_n) \end{pmatrix}\right),$$

with $h(c_n) = \frac{c_n-1}{c_n}\log(1-c_n)-1$, $\eta(c,\alpha) = \sum_{i=1}^{m}\log(1+c\sigma^2\alpha_i^{-1})$ and $\rho(c,\alpha) = c(m + \sum_{i=1}^{m}\alpha_i^{-1})$.

CLT with general σ^2. When $\sigma^2 = 1$,

$$\mathrm{spec}(\mathbf{\Sigma}) = (\alpha_1+1,\ldots,\alpha_m+1,1,\ldots,1),$$

whereas in the general case

$$\begin{aligned} \mathrm{spec}(\mathbf{\Sigma}) &= (\alpha_1+\sigma^2,\ldots,\alpha_m+\sigma^2,\sigma^2,\ldots,\sigma^2) \\ &= \sigma^2\left(\frac{\alpha_1}{\sigma^2}+1,\ldots,\frac{\alpha_m}{\sigma^2}+1,\ldots,1\right). \end{aligned}$$

Thus, if we consider λ_i/σ^2, we will find the same CLT by replacing the $(\alpha_i)_{1\le i\le m}$ by α_i/σ^2. Furthermore, we divide L_2 by $p-m$ to find

$$\begin{pmatrix} L_1 \\ \frac{L_2}{p-m} \end{pmatrix} \xrightarrow{\mathcal{D}} \mathcal{N}\left(\begin{pmatrix} m_1(c) + ph(c_n) - \eta(c,\alpha/\sigma^2) \\ 1 - \frac{\rho(c,\alpha/\sigma^2)}{p-m} \end{pmatrix}, \begin{pmatrix} \frac{2c}{(p-m)^2} & \frac{2c}{p-m} \\ \frac{2c}{p-m} & -2\log(1-c) \end{pmatrix}\right), \tag{9}$$

with $\eta(c, \alpha/\sigma^2) = \sum_{i=1}^{m} \log(1 + c\sigma^2 \alpha_i^{-1})$, $\rho(c, \alpha/\sigma^2) = c(m + \sigma^2 \sum_{i=1}^{m} \alpha_i^{-1})$ and $h(c_n) = \frac{c_n - 1}{c_n} \log(1 - c_n) - 1$.

Asymptotic distribution of L^*. We have $L^* = g(L_1, L_2/(p-m))$, with $g(x,y) = x - (p-m)\log(y)$. We will apply the multivariate delta-method on (9) with the function g. We have $\bigtriangledown g(x,y) = \left(1, -\frac{p-m}{y}\right)$ and

$$L^* \xrightarrow{\mathcal{D}} \mathcal{N}(\beta_1 - (p-m)\log(\beta_2), \bigtriangledown g(\beta_1, \beta_2) \ \mathrm{cov}(L_1, L_2/(p-m)) \bigtriangledown g(\beta_1, \beta_2)'),$$

with $\beta_1 = m_1(c) + ph(c_n) - \eta(c, \alpha/\sigma^2)$ and $\beta_2 = 1 - \frac{\rho(c, \alpha/\sigma^2)}{p-m}$. After some standard calculation, we finally find

$$L^* \xrightarrow{\mathcal{D}} \mathcal{N}\left(m_1(c) + ph(c_n) - \eta\left(c, \frac{\alpha}{\sigma^2}\right) - (p-m)\log(\beta_2), -2\log(1-c) + \frac{2c}{\beta_2}\left(\frac{1}{\beta_2} - 2\right)\right).$$

4.4 Complementary proofs

Proof of (4) in main paper

The general theory of the m.l.e. for the PPCA model (1) in the classical setting has been developed in Anderson and Amemiya (1988) with in particular the following result.

Proposition 2. *Let $\Theta = (\theta_{ij})_{1 \le i,j \le p} = \mathbf{\Psi} - \mathbf{\Lambda}(\mathbf{\Lambda}'\mathbf{\Psi}^{-1}\mathbf{\Lambda})^{-1}\mathbf{\Lambda}'$. If $(\theta_{ij}^2)_{1 \le i,j \le p}$ is nonsingular, if $\mathbf{\Lambda}$ and $\mathbf{\Psi}$ are identified by the condition that $\mathbf{\Lambda}'\mathbf{\Psi}\mathbf{\Lambda}$ is diagonal and the diagonal elements are different and ordered, if $\mathbf{S}_n \to \mathbf{\Lambda}\mathbf{\Lambda}' + \mathbf{\Psi}$ in probability and if $\sqrt{n}(\mathbf{S}_n - \mathbf{\Sigma})$ has a limiting distribution, then $\sqrt{n}(\widehat{\mathbf{\Lambda}} - \mathbf{\Lambda})$ and $\sqrt{n}(\widehat{\mathbf{\Psi}} - \mathbf{\Psi})$ have a limiting distribution. The covariance of $\sqrt{n}(\widehat{\mathbf{\Psi}}_{ii} - \mathbf{\Psi}_{ii})$ and $\sqrt{n}(\widehat{\mathbf{\Psi}}_{jj} - \mathbf{\Psi}_{jj})$ in the limiting distribution is $2\mathbf{\Psi}_{ii}^2\mathbf{\Psi}_{jj}^2\xi^{ij}$ $(1 \le i,j \le p)$, where $(\xi^{ij}) = (\theta_{ij}^2)^{-1}$.*

To prove the CLT (4) in main paper, by Proposition 2, we know that the inverse of the Fisher information matrix is $\mathcal{I}^{-1}(\psi_{11}, \ldots, \psi_{pp}) = (2\psi_{ii}^2\psi_{jj}^2\xi^{ij})_{ij}$. We have to change the parametrization: in our case, we have $\psi_{11} = \cdots = \psi_{pp}$. Let $g : \mathbb{R} \to \mathbb{R}^p$, $a \mapsto (a, \ldots, a)$. The information matrix in this new parametrization becomes

$$\mathcal{I}(\sigma^2) = J'\mathcal{I}(g(\sigma^2))J,$$

where J is the Jacobian matrix of g. As

$$\mathcal{I}(g(\sigma^2)) = \frac{1}{2\sigma^8}(\theta_{ij}^2)_{ij},$$

we have

$$\mathcal{I}(\sigma^2) = \frac{1}{2\sigma^8} \sum_{i,j=1}^{p} \theta_{ij}^2,$$

and

$$\begin{aligned} \Theta = (\theta_{ij})_{ij} &= \boldsymbol{\Psi} - \boldsymbol{\Lambda}(\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda})^{-1}\boldsymbol{\Lambda}' \\ &= \sigma^2(\mathbf{I}_p - \boldsymbol{\Lambda}(\boldsymbol{\Lambda}'\boldsymbol{\Lambda})^{-1}\boldsymbol{\Lambda}'). \end{aligned}$$

By hypothesis, we have $\boldsymbol{\Lambda}'\boldsymbol{\Lambda} = \mathrm{diag}(d_1^2, \ldots, d_m^2)$. Consider the Singular Value Decomposition of $\boldsymbol{\Lambda}$, $\boldsymbol{\Lambda} = \mathbf{UDV}$, where $\mathbf{U}$ is a $p \times p$ matrix such that $\mathbf{UU}' = \mathbf{I}_p$, $\mathbf{V}$ is a $m \times m$ matrix such that $\mathbf{V}'\mathbf{V} = \mathbf{I}_m$, and $\mathbf{D}$ is a $p \times m$ diagonal matrix with $d_1, \ldots, d_m$ as diagonal elements. As $\boldsymbol{\Lambda}'\boldsymbol{\Lambda}$ is diagonal, $\mathbf{V} = \mathbf{I}_m$, so $\boldsymbol{\Lambda} = \mathbf{UD}$. By elementary calculus, one can find that

$$\boldsymbol{\Lambda}(\boldsymbol{\Lambda}'\boldsymbol{\Lambda})^{-1}\boldsymbol{\Lambda}' = \mathrm{diag}(\underbrace{1, \ldots, 1}_{m}, \underbrace{0, \ldots, 0}_{p-m}),$$

so

$$\Theta = \sigma^2 \mathrm{diag}(\underbrace{0, \ldots, 0}_{m}, \underbrace{1, \ldots, 1}_{p-m}).$$

Finally,

$$\mathcal{I}(\sigma^2) = \frac{1}{2\sigma^8}(p-m)\sigma^4 = \frac{p-m}{2\sigma^4}\ ,$$

and the asymptotic variance of $\widehat{\sigma}^2$ is

$$s^2 = \mathcal{I}^{-1}(\sigma^2) = \frac{2\sigma^4}{p-m}.$$

Proof of (17) in main paper

By Proposition 1, for $g(x) = x$, by using the variable change $x = \sigma^2(1 + c - 2\sqrt{c}\cos\theta)$, $0 \leq \theta \leq \pi$, we have

$$\begin{aligned} m(g) &= \frac{g(a(c)) + g(b(c))}{4} - \frac{1}{2\pi}\int_{a(c)}^{b(c)} \frac{x}{\sqrt{4c\sigma^4 - (x - \sigma^2 - c\sigma^2)^2}}\,\mathrm{d}x,\ j = 1, \ldots, k \\ &= \frac{\sigma^2(1+c)}{2} - \frac{\sigma^2}{2\pi}\int_0^{\pi}(1 + c - 2\sqrt{c}\cos\theta)\,\mathrm{d}\theta \\ &= 0. \end{aligned}$$

Proof of (18) in main paper

Let $\underline{s}(z)$ be the Stieltjes transform of $(1-c)1_{[0,\infty)} + cF_{c,\delta_1}$. One can show that

$$\underline{m}(z) = \frac{1}{\sigma^2}\underline{s}\left(\frac{z}{\sigma^2}\right).$$

Then, in Proposition 1, we have

$$v(f_j, f_l) = -\frac{1}{2\pi^2}\oint\oint \frac{f_j(\sigma^2 z_1) f_l(\sigma^2 z_2)}{(\underline{s}(z_1)-\underline{s}(z_2))^2}\,\mathrm{d}\underline{s}(z_1)\,\mathrm{d}\underline{s}(z_2),\ j,l = 1,\ldots,k. \tag{10}$$

For $g(x) = x$, we have

$$\begin{aligned} v(g) &= -\frac{1}{2\pi^2}\oint\oint \frac{g(\sigma^2 z_1) g(\sigma^2 z_2)}{(\underline{s}(z_1)-\underline{s}(z_2))^2}\,\mathrm{d}\underline{s}(z_1)\,\mathrm{d}\underline{s}(z_2) \\ &= -\frac{\sigma^4}{2\pi^2}\oint\oint \frac{z_1 z_2}{(\underline{s}(z_1)-\underline{s}(z_2))^2}\,\mathrm{d}\underline{s}(z_1)\,\mathrm{d}\underline{s}(z_2) \\ &= 2c\sigma^4, \end{aligned}$$

where $-\frac{1}{2\pi^2}\oint\oint \frac{z_1 z_2}{(\underline{s}(z_1)-\underline{s}(z_2))^2}\,\mathrm{d}\underline{s}(z_1)\,\mathrm{d}\underline{s}(z_2) = 2c$ is calculated in Bai et al. (2009) (it corresponds to $v(z_1, z_2)$, Section 5, proof of (3.4)).

Proof of (5)

By Proposition 1, for $\sigma^2 = 1$ and $g(x) = \log(x)$, by using the variable change $x = 1 + c - 2\sqrt{c}\cos\theta$, $0 \le \theta \le \pi$, we have

$$\begin{aligned} m(g) &= \frac{g(a(c)) + g(b(c))}{4} - \frac{1}{2\pi}\int_{a(c)}^{b(c)} \frac{x}{\sqrt{4c - (x-1-c)^2}}\,\mathrm{d}x,\ j = 1,\ldots,k \\ &= \frac{\log(1-c)}{2} - \frac{1}{2\pi}\int_0^{\pi} \log(1 + c - 2\sqrt{c}\cos\theta)\,\mathrm{d}\theta \\ &= \frac{\log(1-c)}{2} - \frac{1}{4\pi}\int_0^{2\pi} \log|1 - \sqrt{c}e^{i\theta}|^2\,\mathrm{d}\theta \\ &= \frac{\log(1-c)}{2}, \end{aligned}$$

where $\int_0^{2\pi} \log|1 - \sqrt{c}e^{i\theta}|^2\,\mathrm{d}\theta = 0$ is calculated in Bai and Silverstein (2010).

Proof of (6)

By Proposition 1 and (10), for $\sigma^2 = 1$ and $g(x) = x$, we have

$$\begin{aligned} v(g) &= -\frac{1}{2\pi^2}\oint\oint \frac{g(z_1)g(z_2)}{(\underline{s}(z_1)-\underline{s}(z_2))^2}\,\mathrm{d}\underline{s}(z_1)\,\mathrm{d}\underline{s}(z_2) \\ &= -\frac{1}{2\pi^2}\oint\oint \frac{\log(z_1)\log(z_2)}{(\underline{s}(z_1)-\underline{s}(z_2))^2}\,\mathrm{d}\underline{s}(z_1)\mathrm{d}\underline{s}(z_2) \\ &= -2\log(1-c_n), \end{aligned}$$

where the last integral is calculated in Bai and Silverstein (2010).

Proof of (8)

F_{c_n,δ_1} is the Marčenko-Pastur distribution of index c_n. By using the variable change $x = 1 + c_n - 2\sqrt{c_n}\cos\theta$, $0 \le \theta \le \pi$, we have

$$\begin{aligned} \int \log(x)dF_{c_n,\delta_1}(x) &= \int_{a(c_n)}^{b(c_n)} \frac{\log x}{2\pi x c_n}\sqrt{(b(c_n)-x)(x-a(c_n))}\,\mathrm{d}x \\ &= \frac{1}{2\pi c_n}\int_0^{\pi} \frac{\log(1+c_n-2\sqrt{c_n}\cos\theta)}{1+c_n-2\sqrt{c_n}\cos\theta}4c_n\sin^2\theta\,\mathrm{d}\theta \\ &= \frac{1}{2\pi}\int_0^{2\pi} \frac{2\sin^2\theta}{1+c_n-2\sqrt{c_n}\cos\theta}\log|1-\sqrt{c_n}e^{i\theta}|^2\,\mathrm{d}\theta \\ &= \frac{c_n-1}{c_n}\log(1-c_n)-1, \end{aligned}$$

where the last integral is calculated in Bai and Silverstein (2010).

Proof of (7)

In the normal case with $\sigma^2 = 1$, Zheng (2012) gives the following equivalent expression of (3):

$$v(f_j, f_l) = -\lim_{r\to 1^+}\frac{\kappa}{4\pi^2}\oint\oint_{|\xi_1|=|\xi_2|=1} f_j(|1+h\xi_1|^2)f_l(|1+h\xi_2|^2)\frac{1}{(\xi_1-r\xi_2)^2}\,\mathrm{d}\xi_1\,\mathrm{d}\xi_2,$$

where $\kappa = 2$ in the real case and $h = \sqrt{c}$ in our case. We take $f_j(x) = \log(x)$ and $f_l(x) = x$, so we need to calculate

$$v(\log(x), x) = -\lim_{r\to 1^+}\frac{1}{2\pi^2}\oint\oint_{|\xi_1|=|\xi_2|=1} |1+\sqrt{c}\xi_2|^2\frac{\log(|1+\sqrt{c}\xi_1|^2)}{(\xi_1-r\xi_2)^2}\,\mathrm{d}\xi_1\,\mathrm{d}\xi_2.$$

We follow the calculations done in Zheng (2012): when $|\xi| = 1$, $|1+\sqrt{c}\xi|^2 = (1+\sqrt{c}\xi)(1+\sqrt{c}\xi^{-1})$, so $\log(|1+\sqrt{c}\xi|^2) = \frac{1}{2}\left(\log(1+\sqrt{c}\xi)^2 + \log(1+\sqrt{c}\xi^{-1})^2\right)$. Consequently,

$$\begin{aligned}
\oint_{|\xi_1|=1} \frac{\log(|1+\sqrt{c}\xi_1|^2)}{(\xi_1 - r\xi_2)^2}\,\mathrm{d}\xi_1 &= \frac{1}{2}\oint_{|\xi_1|=1} \frac{\log(1+\sqrt{c}\xi_1)^2}{(\xi_1 - r\xi_2)^2}\,\mathrm{d}\xi_1 + \frac{1}{2}\oint_{|\xi_1|=1} \frac{\log(1+\sqrt{c}\xi_1^{-1})^2}{(\xi_1 - r\xi_2)^2}\,\mathrm{d}\xi_1 \\
&= \frac{1}{2}\oint_{|\xi_1|=1} \log(1+\sqrt{c}\xi_1)^2\left(\frac{1}{(\xi_1 - r\xi_2)^2} + \frac{1}{(1 - r\xi_1\xi_2)^2}\right)\mathrm{d}\xi_1 \\
&= 0 + i\pi\left(\frac{1}{(r\xi_2)^2}\frac{2\sqrt{c}}{1+\frac{\sqrt{c}}{r\xi_2}}\right) \\
&= 2i\pi\frac{\sqrt{c}}{r\xi_2(r\xi_2+\sqrt{c})}.
\end{aligned}$$

Thus,

$$\begin{aligned}
v(\log(x), x) &= \frac{1}{i\pi}\oint_{|\xi_2|=1} |1+\sqrt{c}\xi_2|^2\frac{\sqrt{c}}{\xi_2(\xi_2+\sqrt{c})}\,\mathrm{d}\xi_2 \\
&= \frac{1}{i\pi}\oint_{|\xi|=1}\left(1+c+c(\xi+\xi^{-1})\right)\frac{\sqrt{c}}{\xi(\xi+\sqrt{c})}\,\mathrm{d}\xi \\
&= \frac{1}{i\pi}\oint_{|\xi|=1}\left(\frac{\sqrt{c}(1+c)}{\xi(\xi+\sqrt{c})} + \frac{c}{\xi+\sqrt{c}} + \frac{c}{\xi^2(\xi+\sqrt{c})}\right)\mathrm{d}\xi \\
&= 2(1+c-(1+c)+c+1-1) \\
&= 2c.
\end{aligned}$$

References

T. W. Anderson and Y. Amemiya. The asymptotic normal distribution of estimators in factor analysis under general conditions. *Ann. Statist.*, 16(2):759–771, 1988.

Z. Bai and J. W. Silverstein. CLT for linear spectral statistics of large-dimensional sample covariance matrices. *Ann. Probab.*, 32(1A):553–605, 2004.

Z. Bai and J. W. Silverstein. *Spectral analysis of large dimensional random matrices.* Springer Series in Statistics. Springer, New York, second edition, 2010.

Z. Bai, D. Jiang, J. Yao, and S. Zheng. Corrections to LRT on large-dimensional covariance matrix by RMT. *Ann. Statist.*, 37(6B):3822–3840, 2009.

Z. Bai, J. Chen, and J. Yao. On estimation of the population spectral distribution from a high-dimensional sample covariance matrix. *Aust. N. Z. J. Stat.*, 52(4):423–437, 2010.

I. M. Johnstone. High dimensional statistical inference and random matrices. In *International Congress of Mathematicians. Vol. I*, pages 307–333. Eur. Math. Soc., Zürich, 2007.

I. M. Johnstone and D. M. Titterington. Statistical challenges of high-dimensional data. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.*, 367(1906):4237–4253, 2009.

M. O. Ulfarsson and V. Solo. Dimension estimation in noisy PCA with SURE and random matrix theory. *IEEE Trans. Signal Process.*, 56(12):5804–5816, 2008.

Q. Wang and J. Yao. On the sphericity test with large-dimensional observations. *Electron. J. Stat.*, 7:2164–2192, 2013.

Q. Wang, J. W. Silverstein, and J. Yao. A note on the clt of the lss for sample covariance matrix from a spiked population model. *J. Multivariate Anal.*, 130:194–207, 2014.

S. Zheng. Central limit theorems for linear spectral statistics of large dimensional f-matrices. *Ann. Inst. Henri Poincaré Probab. Stat.*, 48(2):444–476, 2012.

S. Zheng, Z. Bai, and J. Yao. Substitution principle for CLT of linear spectral statistics of high-dimensional sample covariance matrices with applications to hypothesis testing. *Ann. Statist.*, 43(2):546–591, 2015.