High-Dimensional Consistency of AIC and BIC for Estimating the Number of Significant Components in Principal Component Analysis

Zhidong Bai[1], Yasunori Fujikoshi[2] and Kwok Pui Choi[3]

[1]*School of Mathematics and Statistics and KLAS, Northeast Normal University, Changchun, Jilin 130024, China*

[2]*Department of Mathematics, Graduate School of Science, Hiroshima University, Higashi Hiroshima, Hiroshima 739-8626, Japan*

[3]*Department of Statistics and Applied Probability, National University of Singapore, 6 Science Drive 2, 117546, Singapore*

Abstract

In this paper we study the problem of estimating the number of significant components in principal component analysis (PCA) which denotes the number of dominant eigenvalues of the covariance matrix of p variables. Our purpose is to examine the consistency of the estimation criteria AIC and BIC based on the model selection criteria by Akaike (1973) and Schwarz (1978) under a high-dimensional asymptotic framework. Using random matrix theory, we derive sufficient conditions for the criteria to be strongly consistent for the case when the dominant population eigenvalues are finite, and the case when the dominant eigenvalues tend to infinity. Moreover, the asymptotic results are obtained without normality assumption on the population distribution. Simulation studies are also conducted.

AMS 2000 subject classification: primary 62H12; secondary 62H30
Key Words and Phrases: AIC, BIC, Consistency, Dimensionality, High-dimensional framework, Number of significant components,

Principal component analysis, Random matrix theory, Signal processing, Spiked model.

1 Introduction

Principal component analysis (PCA) is a widely used technique for reducing the dimensionality of data in the form of n observations of p variables. An important issue in the application of PCA is to determine the number of significant components (see, e.g. Jolliffe (2002), Ferré (1995)) which is also called the dimensionality in PCA. Let $\lambda_1 \geq \cdots \geq \lambda_p$ be the population eigenvalues of the covariance matrix $\boldsymbol{\Sigma}$ of a p-dimensional random vector $\boldsymbol{y}$. As an approach for determining the dimensionality, we consider a covariance structure model in which the number of dominant eigenvalues is k, that is,

$$M_k : \quad \lambda_k > \lambda_{k+1} = \cdots = \lambda_p = \lambda. \tag{1.1}$$

Here M_0 refers to $\lambda_1 = \cdots = \lambda_p = \lambda$.

If M_k is true, we say that the true dimensionality or the true number of significant components is k. The model M_k has also been considered by Bai, Miao and Rao (1995) as a signal processing model and by Johnstone (2001) as a spiked model. The number, k, in these work is respectively referred to as the number of signals and the number of spikes.

In general, the number of significant components, k, is unknown, and we need to estimate it. Specifically, let $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ be a random sample of size n from a p-dimensional population with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, and let S_n be the sample covariance matrix given by

$$\mathsf{S}_n = \frac{1}{n-1} \sum_{i=1}^{n} (\boldsymbol{y}_i - \bar{\boldsymbol{y}})(\boldsymbol{y}_i - \bar{\boldsymbol{y}})^\top, \tag{1.2}$$

where $\bar{\boldsymbol{y}} = (1/n) \sum_{i=1}^{n} \boldsymbol{y}_i$. Based on the sample, we estimate the dimensionality by selecting an appropriate model from the set $\{M_0, M_1, \ldots, M_{p-1}\}$. In particular, we consider two estimation criteria AIC and BIC based on the

decision rules of Akaike (1973) and that of Schwarz (1978) respectively. We shall discuss $p < n$ first. With

$$C_{p,n} = n \log \left((n-1)/n \right)^p + np\{1 + \log(2\pi)\},$$

these are given by (see, e.g., Fujikoshi, Ulyanov and Shimizu (2010), Fujikoshi and Sakurai (2015))

$$\mathrm{AIC}_j = n \log(\ell_1 \cdots \ell_j) + n(p-j) \log \bar{\ell}_{jp} + 2d_j + C_{p,n}, \tag{1.3}$$

$$\mathrm{BIC}_j = n \log(\ell_1 \cdots \ell_j) + n(p-j) \log \bar{\ell}_{jp} + (\log n) d_j + C_{p,n}, \tag{1.4}$$

where $\ell_{1p} > \ldots > \ell_{pp}$ are the sample eigenvalues of $\mathbf{S}_n$, and for $1 \leq j \leq p-1$, $\bar{\ell}_{jp}$ is the arithmetic mean of $\ell_{j+1,p}, \ldots, \ell_{pp}$, that is,

$$\bar{\ell}_{jp} := \frac{1}{p-j} \sum_{t=j+1}^{p} \ell_{tp}. \tag{1.5}$$

Furthermore, d_j denotes the number of independent parameters under for $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}$ under M_j which is given by

$$\begin{aligned} d_j &= pj - \frac{1}{2} j(j+1) + j + 1 + p \\ &= (j+1)(p+1-j/2). \end{aligned} \tag{1.6}$$

Then the AIC and BIC select respectively the number of significant components according to

$$\hat{k}_{\mathrm{A}} = \arg\min_j \mathrm{AIC}_j \quad \text{and} \quad \hat{k}_{\mathrm{B}} = \arg\min_j \mathrm{BIC}_j.$$

When we are interested in only the first q models M_j, $j = 0, 1, \ldots, q-1$, then the criteria are defined by considering the minimum only with respect to $j = 0, 1, \ldots, q-1$. We call q the number of candidate models. We also use

$$\mathrm{A}_j = \frac{1}{n}(\mathrm{AIC}_j - \mathrm{AIC}_{p-1}), \quad \mathrm{B}_j = \frac{1}{n}(\mathrm{BIC}_j - \mathrm{BIC}_{p-1})$$

instead of AIC_j and BIC_j.

Motivated by numerous modern data structure in which $p > n$, we extend our study to cover this situation in Section 4. We modify the definition of (1.5) to (4.2), and propose to use the modified criteria $\tilde{\mathrm{A}}_j$ and $\tilde{\mathrm{B}}_j$ as defined in Section 4.

In general, under a large-sample asymptotic framework, in which p is fixed and n goes to infinity, it has been pointed out in various models that AIC is not consistent, but BIC is. See, for examples, Shibata (1976), Nishii (1984), Nishii, Bai and Krishnaiah (1988), and Gunderson and Muirhead (1997). Similar selection consistency results in PCA is shown by Fujikoshi and Sakurai (2015). However, under the high-dimensional framework, in which both p and n tend to infinity, Fujikoshi, Sakurai and Yanagihara (2014) and Yanagihara, Wakaki and Fujikoshi (2015) showed that in multivariate regression model there are cases in which AIC is consistent, but BIC is not.

Our purpose is to study the consistency of the estimation criteria AIC and BIC under a high-dimensional asymptotic framework $p/n \to c > 0$. It is assumed that the true number of significant components, k, is fixed; and that the number of candidate models q satisfies $q > k$. We provide complete proofs for $0 < c < 1$ case, and sketch how these proofs can be extended in a similar way for $c > 1$. The case $c = 1$ is more intricate (see the remark at the end of Section 4) and will not be explored in this study.

For $0 < c < 1$, Theorem 3.1 states that if the dominant k population eigenvalues are bounded, AIC is strongly consistent under the "gap condition" (3.4), but BIC is not. Furthermore, if the dominant k population eigenvalues tend to infinity, AIC is always strongly consistent. If the dominant k population eigenvalues tend to infinity with a rate faster than $\log n$, BIC is shown to be strongly consistent as well. These results are extended to $c > 1$.

Our main results are obtained by techniques from random matrix theory (RMT). An attractive feature of our results is that we require very mild distributional assumption on the population: finite fourth moment. In particular, the results hold without assuming normality. Two new results, Lemmas 2.2

and 2.3, on the limiting behaviors of the sample eigenvalues are of independent interests. The first describes the limiting behaviors of the sample eigenvalues when the dominant population eigenvalues tend to infinity. The second is concerned with monotonicity of the ratio of quantiles of Marčenko-Pastur (MP) distribution.

This paper is organized as follows. In Section 2, we recall some basic results on random matrix theory and state the two new lemmas. Main results on strong consistency of AIC and BIC are stated and proved in Section 3. In Section 4 the results are extended to the case $c > 1$. In Section 5, we present our simulation studies. They show that the gap condition and the finite fourth moment condition are essential for the selection consistency of AIC, in addition to demonstrate perfect agreement with our Theorems. We end our paper with concluding remarks in Section 6, and a conjecture. Proofs of Lemmas 2.2 and 2.3 are given in the Appendix.

2 Preliminaries

In this section we recall some basic results in random matrix theory. For more details, see Bai and Silverstein (2010) and Yao, Zheng and Bai (2015). Moreover, we obtain new results on the limiting behaviors of the sample eigenvalues of $\mathbf{S}_n$ when the population spiked eigenvalues tend to infinity, and a monotonicity property of a ratio of quantiles of the MP distribution.

2.1 Marčenko-Pastur (MP) law

Let $\{x_{ij}, i = 1, \ldots, p; j = 1, \ldots, n\}$ be a double array of iid random variables with mean 0 and variance 1. Write $\boldsymbol{x}_k = (x_{1k}, \ldots, x_{pk})^\top$ and $\mathbf{X} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)$. Suppose that $p/n \to c \in (0, 1)$ and define the sample covariance matrix as

$$\mathbf{S}_n = \frac{1}{n-1} \sum_{j=1}^{n} (\boldsymbol{x}_j - \bar{\boldsymbol{x}})(\boldsymbol{x}_j - \bar{\boldsymbol{x}})^\top,$$

where $\bar{\boldsymbol{x}} = \frac{1}{n} \sum_{k=1}^{n} \boldsymbol{x}_k$.

Denote the eigenvalues of $\mathbf{S}_n$ as $\ell_{1p} > \ell_{2p} > \cdots > \ell_{pp} > 0$. Define the empirical spectral distribution (ESD) of $\mathbf{S}_n$ by

$$F_n(x) = \frac{1}{p}\sum_{i=1}^{p} I_{(-\infty,x]}(\ell_{ip}),$$

where $I_A(\cdot)$ is the indicator function. With probability 1, $F_n(x) \xrightarrow{w} F_c(x)$, where

$$F_c'(x) = f_c(x) = \begin{cases} \frac{1}{2\pi x c}\sqrt{(b-x)(x-a)}, & \text{if } x \in (a,b), \\ 0, & \text{otherwise,} \end{cases}$$

where $a = (1-\sqrt{c})^2$ and $b = (1+\sqrt{c})^2$.

If $c > 1$, F_c has a point mass $1 - 1/c$ at the origin, that is,

$$F_c(x) = \begin{cases} 0, & \text{if } x < 0, \\ 1 - 1/c, & \text{if } 0 \le x < a, \\ 1 - 1/c + \int_a^x f_c(t)dt, & \text{if } x \ge a. \end{cases}$$

We remark that $\int_a^b f_c(t)dt = 1$ or $1/c$ in accordance with $c < 1$ or $c \ge 1$ respectively.

From the MP law, we have the easy consequence that if $i/p \to \alpha \in (0,1)$, then $\ell_{ip} \xrightarrow{a.s.} \mu_{1-\alpha}$, where μ_α is the α-quantitle of the MP law, that is $F_c(\mu_\alpha) = \alpha$.

2.2 Limits of eigenvalues under spiked model

Let $\{x_{ij}, i = 1, \ldots, p; j = 1, \ldots, n\}$ be a double array of iid random variables with mean 0 and variance 1. Write $\boldsymbol{x}_k = (x_{1k}, \ldots, x_{pk})^\top$ and $\mathbf{Y} = (\boldsymbol{y}_1, \cdots, \boldsymbol{y}_n) = \boldsymbol{\Sigma}^{1/2}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$. Define the sample covariance matrix of $\mathbf{Y}$ as

$$\mathbf{S}_n = \frac{1}{n-1}\sum_{j=1}^{n}(\boldsymbol{y}_j - \bar{\boldsymbol{y}})(\boldsymbol{y}_j - \bar{\boldsymbol{y}})^\top, \tag{2.1}$$

where $\bar{\boldsymbol{y}} = \frac{1}{n}\sum_{k=1}^{n}\boldsymbol{y}_k$. Write the eigenvalues of $\mathbf{S}_n$ as $\ell_{1p} > \ell_{2p} > \cdots > \ell_{pp} > 0$. Define the ESD of $\mathbf{S}_n$ by

$$F_n(y) = \frac{1}{p}\sum_{i=1}^{p} I_{(-\infty,y]}(\ell_{ip}).$$

Assume that

(C1) $p/n \to c > 0$.

(C2) The ESD of $\mathbf{\Sigma}$, $H_p = F^{\mathbf{\Sigma}}$ converges weakly to H as $p \to \infty$.

Under (C1) and (C2), with probability 1, $F_n(y) \xrightarrow{w} F^{c,H}(y)$, where the Stieltjes transform $\underline{s}$ of $\underline{F}^{c,H}(y) = (1-c)\delta_0 + cF^{c,H}(y)$ is the unique solution to the equation

$$z = -\frac{1}{\underline{s}} + c\int \frac{t\ dH(t)}{1+t\underline{s}},$$

on the upper complex plane for every z with $\mathrm{Im}(z) > 0$. Define

$$\varphi(x) = x\left(1 + c\int \frac{t\ dH(t)}{x-t}\right). \tag{2.2}$$

By Silverstein and Choi (1995), if a nonzero point x does not belong to the support of H, then $\varphi(x)$ is an inner point of the complement of the support of $F^{c,H}$ if and only if

$$\varphi'(x) > 0.$$

Suppose the eigenvalues of $\mathbf{\Sigma}$ are $\lambda_1, \ldots, \lambda_k, 1, \ldots, 1$. In the following, for simplicity, we consider the case

(C3) $\lambda_1 \geq \cdots \geq \lambda_k > \lambda_{k+1} = \cdots = \lambda_p = \lambda = 1$.

Denote by μ_F the support of a distribution F. The i-th largest eigenvalue, λ_i, of $\mathbf{\Sigma}$ is called a distant spiked eigenvalue if $\varphi'(\lambda_i) > 0$. Bai and Yao (2012) proved the following lemma.

Lemma 2.1. *Let ℓ_{ip} denote the i-th largest eigenvalue of $\mathbf{S}_n$ in* (2.1)*. Suppose that* $\mathrm{E}(x_{11}^4) < \infty$, (C1), (C2), (C3) *hold, and λ_1 is bounded.*

(1) *If λ_i is a distant spiked eigenvalue, then $\ell_{ip} \xrightarrow{a.s.} \varphi(\lambda_i)$.*

(2) *If λ_i is not a distant spiked eigenvalue and $i/p \to \alpha$, then $\ell_{ip} \xrightarrow{a.s.} \mu_{1-\alpha}^{c,H}$ and the convergence is uniform in $0 \leq \alpha \leq 1$. Here $\mu_\alpha^{c,H}$ denotes the α-th quantile of the limiting spectral distribution* (LSD)*, $F^{c,H}$.*

Baik and Silverstein (2006) shows the special case where $H(\theta) = I(1 \leq \theta)$. In this case, $\varphi(x) = x\{1 + c/(x-1)\}$, which will be denoted by ψ for the rest of this paper. In this case, an eigenvalue λ of $\mathbf{\Sigma}$ satisfying $\lambda > 1 + \sqrt{c}$ is a distant spiked eigenvalue.

In Bai and Yao (2012), it is assumed that the spectral norm (that is, the largest singular value of $\mathbf{\Sigma}^{1/2}$) is bounded. Therefore, when the spiked population eigenvalues tend to infinity, we need to establish a new limiting result for the distant spiked eigenvalues. Intuitively, if $\lambda_j \to \infty$, $\varphi(\lambda_j) \sim \lambda_j$. Under the assumption of finite 4-th moment, this is indeed the case and is summarized in the following lemma.

Lemma 2.2. *In the same setup of* Lemma 2.1*, instead of assuming* λ_1 *bounded, we assume that* $\lambda_k \to \infty$ *as* $p \to \infty$*. We have the following results.*

(1) *For any* $j \leq k$, $\lim_{n\to\infty} \ell_{jp}/\lambda_j = 1$ *a.s.*

(2) *If* λ_i *is not a distant spiked eigenvalue and* $i/p \to \alpha$ *as* $n \to \infty$*, then* $\lim_{n\to\infty} \ell_{ip} = \mu^{c,H}_{1-\alpha}$ *a.s. and the convergence is uniform in* $0 \leq \alpha \leq 1$*. Here* $\mu^{c,H}_{\alpha}$ *denotes the* α*-th quantile of the limiting spectral distribution* (LSD)*,* $F^{c,H}$*.*

The proof of Lemma 2.2 is given in the Appendix. Note that Lemmas 2.1 and 2.2 are true for both cases $0 < c < 1$ and $c \geq 1$. The only difference is $\mu_{1-t} = 0$ when $t > 1/c$ if $c > 1$.

The asymptotic framework that the largest k population eigenvalues tending to infinity was introduced in Schott (2006), and in Fujikoshi et al. (2007). In fact, they derived the asymptotic distributions of test statistics for testing the hypothesis $\lambda_{k+1} = \cdots = \lambda_p$ under the assumptions that (1) k is fixed, (2) $p/n \to c \in (0, 1)$, (3) $\lambda_i = \mathrm{O}(n)$, $i = 1, \ldots, k$, and (4) $\boldsymbol{y}$ is normal.

2.3 Monotonicity property of a ratio of quantiles of MP law

Let $\mathbf{S}_n$ be the sample covariance in (2.1) with the population covariance matrix $\mathbf{\Sigma} = \mathbf{I}_p$, and let $\ell_{1p} > \ell_{2p} > \cdots > \ell_{pp} > 0$ be the eigenvalues of $\mathbf{S}_n$. Consider the ratios

$$R_i = \frac{\ell_{ip}}{\frac{1}{p-i}\sum_{t=i+1}^{p} \ell_{tp}} = \frac{\ell_{ip}}{\bar{\ell}_{ip}}, \quad i = 1, 2, \ldots, p-1.$$

Monotonicity of the ratio of quantiles of MP law in Lemma 2.3 below leads us to conjecture that in most cases

$$R_1 > R_2 > \cdots > R_{p-1},$$

hold almost surely.

Lemma 2.3. *Let μ_α be the α-th quantile of the MP distribution, that is, $F_c(\mu_\alpha) = \alpha$. We define*

$$x(t) = \frac{\mu_{1-t}}{\bar{\mu}_{1-t}}, \quad 0 \le t \le \min\{1, 1/c\}$$

where

$$\bar{\mu}_{1-t} = \begin{cases} \frac{1}{1-t}\int_0^{1-t} \mu_s ds = \frac{1}{1-t}\int_a^{\mu_{1-t}} x f_c(x) dx, & \text{if } 0 < c < 1; \\ \\ \frac{c}{1-ct}\int_{1-1/c}^{1-t} \mu_s ds = \frac{c}{1-ct}\int_a^{\mu_{1-t}} x f_c(x) dx, & \text{if } c \ge 1. \end{cases}$$

Then (i) *when $c < 1$, $x(t)$ strictly decreases from b to 1 as t increases from 0 to 1; and* (ii) *when $c \ge 1$, $x(t)$ strictly decreases from b/c to 1 as t increases from 0 to $1/c$.*

Lemma 2.3 is used in the proof of consistency of AIC and BIC. The proof of this lemma is given in the Appendix.

3 Main results

In this section we derive consistency of two estimation criteria $\hat{k}_A$ and $\hat{k}_B$ based on AIC and BIC. Throughout this section, we assume $0 < c < 1$. The case for $c \geq 1$ will be dealt with in the next section.

Suppose that the true number of significant components (or true dimensionality or the true number of spikes) is k. AIC and BIC being scale invariant so when we consider the distributions of AIC and BIC, we may assume, without loss of generality, that the population eigenvalues are

$$\lambda_{k+1} = \cdots = \lambda_p = 1. \tag{3.1}$$

Here λ_i should be read as $\lambda_i/\lambda, i = 1, \ldots, k$.

We may consider the A_j and B_j below instead of AIC_j and BIC_j:

$$\begin{aligned}
\mathrm{A}_j &= \frac{1}{n}\left(\mathrm{AIC}_j - \mathrm{AIC}_{p-1}\right) \\
&= (p-j)\log \bar{\ell}_{jp} - \sum_{i=j+1}^{p} \log \ell_{ip} - (p-j-1)(p-j+2)/n, \\
\mathrm{B}_j &= \frac{1}{n}\left(\mathrm{BIC}_j - \mathrm{BIC}_{p-1}\right) \\
&= (p-j)\log \bar{\ell}_{jp} - \sum_{i=j+1}^{p} \log \ell_{ip} - \frac{(p-j-1)(p-j+2)}{2n}\log n.
\end{aligned}$$

Here $\mathrm{A}_{p-1} = 0$ and $\mathrm{B}_{p-1} = 0$. Then, the decision rule of AIC (respectively, BIC) selects the model $\hat{k}_A$ (respectively, $\hat{k}_B$) by

$$\hat{k}_A = \arg\min_j \mathrm{A}_j \quad \text{and} \quad \hat{k}_B = \arg\min_j \mathrm{B}_j.$$

When we are interested in models M_j, $j = 0, 1, \ldots, q-1$, then, the criteria are defined by considering the minimum with respect to $j = 0, 1, \ldots, q-1$.

In general, a criterion $\hat{k}$ for estimating the true number of significant components k is said to be consistent (or strongly consistent) if $\lim_{n\to\infty} P(\hat{k} = k) = 1$ (respectively, $P(\lim_{n\to\infty} \hat{k} = k) = 1$).

In this section we assume conditions (C1) with $0 < c < 1$, (C2), (C3) and

(C4) $\quad \lambda_k > 1 + \sqrt{c}$.

3.1 AIC

Suppose that λ_1 is finite. Then, from Lemma 2.1 and (C4) it follows that for $i = 1, \ldots, k$, $\ell_{ip} \xrightarrow{a.s.} \psi_i$, where

$$\psi_i \equiv \psi(\lambda_i) = \lambda_i + \frac{c\lambda_i}{\lambda_i - 1},\ i = 1, 2, \ldots, k. \tag{3.2}$$

Consider a function $h(x) = x - 1 - \log x - 2c,\ x \geq 1$. Let $x = m(c)$ be the only solution to the equation

$$m = 1 + \log m + 2c, \quad m > 1. \tag{3.3}$$

Then, it is easily seen that $h(x) > 0$, for $x > m(c)$. We consider a condition

$$\psi_k > m(c), \tag{3.4}$$

which is equivalent to

$$\gamma(c) \equiv \psi_k - 1 - \log \psi_k - 2c > 0. \tag{3.5}$$

Condition $\psi_k > m(c)$ or $\gamma(c) > 0$ is called the gap condition.

Theorem 3.1. *Suppose the conditions* (C1) *with* $0 < c < 1$, (C2)–(C4) *hold, and that the number of candidate models, q, satisfies $q = o(p)$. We have the following results on the consistency of the estimation criterion $\hat{k}_A$ based on* AIC.

(1) *Suppose that λ_1 is finite. If the gap condition* (3.4) *(i.e., $\psi_k < m(c)$) does not hold, then $\hat{k}_A$ is not consistent. If the gap condition (3.4) holds, then $\hat{k}_A$ is strongly consistent.*

(2) *Suppose that λ_k tends to infinity, and $\lambda_1 = O(p)$. Then, $\hat{k}_A$ is strongly consistent.*

Proof. Suppose that λ_1 is finite. We first consider the case where $j < k$. Noting that for $i \in [j, k)$, $\ell_{ip} \xrightarrow{a.s.} \psi_i := \psi(\lambda_i) = \lambda_i + c\lambda_i/(\lambda_i - 1)$ and

$$\bar{\ell}_{ip} = \frac{1}{(p-i)} \sum_{t=i+1}^{p} \ell_{tp} \xrightarrow{a.s.} \int_a^b t f_c(t)\ dt = 1. \tag{3.6}$$

This implies

$$
\begin{aligned}
A_j - A_k &= \sum_{i=j+1}^{k} (A_{i-1} - A_i) \\
&= \sum_{i=j+1}^{k} \Bigg[(p-i+1) \log \left\{ 1 - \frac{1}{p-i+1} (1 - \ell_{ip}/\bar{\ell}_{ip}) \right\} \\
&+ \log \bar{\ell}_{ip} - \log \ell_{ip} - 2(p-i+1)/n \Big] \\
&\sim \sum_{i=j+1}^{k} (\psi_i - 1 - \log \psi_i - 2c) \,. \qquad (3.7)
\end{aligned}
$$

If the gap condition (3.4) does not hold, or equivalently, $\psi_k - 1 - \log \psi_k - 2c < 0$, then for sufficiently large n, $A_{k-1} - A_k < 0$ by (3.7) and hence $\hat{k}_A$ is not consistent.

We next consider $\psi_k > m(c)$. For $0 \le j < k$, and for sufficiently large n, apply (3.7) to conclude

$$
A_j - A_k \ge (k-j) \, (\psi_k - 1 - \log \psi_k - 2c) > 0.
$$

In other words,

$$
\hat{k}_A \ge k \quad a.s. \qquad (3.8)
$$

Next we consider the case where $k < j = o(p)$. We have

$$
\begin{aligned}
A_j - A_k &= \sum_{i=k+1}^{j} (A_i - A_{i-1}) \\
&= \sum_{i=k+1}^{j} \Bigg[- (p-i+1) \log \left\{ 1 - \frac{1}{p-i+1} (1 - \ell_{ip}/\bar{\ell}_{ip}) \right\} \\
&\qquad - \log \bar{\ell}_{ip} + \log \ell_{ip} + 2(p-i+1)/n \Bigg] \\
&\sim \sum_{i=k+1}^{j} \left\{ \left(1 - \ell_{ip}/\bar{\ell}_{ip}\right) + \log \left(\ell_{ip}/\bar{\ell}_{ip}\right) + 2c \, (1 - i/p) \right\} . \qquad (3.9)
\end{aligned}
$$

For $k < i \le j$, $\ell_{jp} \le \ell_{ip} \le \ell_{k+1,p}$. From part (2) in Lemma 2.1, $\ell_{k+1,p}$ and

$\ell_{jp} \xrightarrow{a.s.} \mu_1 = b$ as $n \to \infty$, so $\ell_{ip} \xrightarrow{a.s.} b$. It implies almost surely that

$$\begin{aligned} A_j - A_k &\sim (j-k)(1-b+\log b+2c) \\ &= (j-k)\left\{c - 2\sqrt{c} + 2\log(1+\sqrt{c})\right\} > 0. \end{aligned}$$

Combining the above with (3.8), we complete the proof of (1).

To prove (2), the case for $k < j = o(p)$ is the same as that of (1) as $\bar{\ell}_{ip} \to 1$ remains to hold.

For $j < k$, as in the proof of (1),

$$A_j - A_k \sim \sum_{i=j+1}^{k} \left[\ell_{ip}/\bar{\ell}_{ip} - 1 - \log(\ell_{ip}/\bar{\ell}_{ip}) - 2c\right].$$

When $\lambda_k \to \infty$ and $\lambda_2 = o(n)$, we still have $\bar{\ell}_{ip} \to 1$ (because $\ell_{tp}/p \to 0$ for $t = j+1, ..., k$), and thus $\ell_{ip}/\bar{\ell}_{ip} \sim \lambda_i \to \infty$. Hence $A_j - A_k \sim \sum_{i=j+1}^{k} [\lambda_i - 1 - \log\lambda_i - 2c] > 0$.

When $\lambda_k \to \infty$ and $\lambda_i/p \to \eta_{i-1}$, we have $\bar{\ell}_{ip} \to 1 + \eta_{k-1} + \cdots + \eta_i$, and thus $\ell_{ip}/\bar{\ell}_{ip} \sim \lambda_i/(1 + \eta_{k-1} + ... + \eta_i) \to \infty$. Hence

$$A_j - A_k \sim \sum_{i=j+1}^{k} \left[\frac{\lambda_i}{1+\eta_{k-1}+\cdots+\eta_i} - 1 - \log\left(\frac{\lambda_i}{1+\eta_{k-1}+...+\eta_i}\right) - 2c\right] > 0.$$

□

3.2 BIC

In general, BIC is consistent under a large-sample asymptotic framework. However, under a high-dimensional asymptotic framework, BIC is not necessarily consistent. By the method of proof similar to that of Theorem 3.1 for AIC, we obtain the following theorem.

Theorem 3.2. *Suppose the conditions* (C1) *with* $0 < c < 1$, (C2)–(C4) *hold. We have the following results on the consistency of the estimation criterion* $\hat{k}_B$ *based on* BIC.

(1) *Suppose that* $\lambda_k/\log n \to 0$. *Then,* $\hat{k}_B$ *is not consistent.*

(2) *Suppose that* $\lambda_k/\log n \to \infty$. *Then,* $\hat{k}_B$ *is strongly consistent.*

Remark. Since the penalty in BIC tends to infinity as $n \to \infty$, no further condition on the number of candidate models: $q = o(p)$ is required in Theorem 3.2.

Proof. We first consider the case where $j < k$. Note that for $i \in [j, k)$, $\ell_{ip} \xrightarrow{a.s.} \psi_i$. Similar to the AIC argument, we have

$$\mathrm{B}_j - \mathrm{B}_k \sim \sum_{i=j+1}^{k} (\psi_i - 1 - \log \psi_i - c \log n). \tag{3.10}$$

If $\lambda_k/\log n \to 0$, or equivalently, $\psi_k/\log n \to 0$, then $\mathrm{B}_{k-1} - \mathrm{B}_k \sim \psi_k - 1 - \log \psi_k - c \log n < 0$. This proves (1).

If $\lambda_k/\log n \to \infty$, then for sufficiently large n, by (3.10),

$$\mathrm{B}_j - \mathrm{B}_k \geq (k - j)\,(\psi_k - 1 - \log \psi_k - c \log n) > 0 \ a.s.$$

for any $1 \leq j < k$. That is, $\hat{k}_{\mathrm{B}} \geq k$ *a.s.*

Consider $k < j$, analogous to the derivation of (3.9), we have

$$\mathrm{B}_j - \mathrm{B}_k \sim \sum_{i=k+1}^{j} [1 - x(i/p) + \log x(i/p) + c(1 - i/p) \log n]$$

where $x(t)$ is defined in Lemma 2.3.

We first consider the case where $k < j \leq 2p/3$. Lemma 2.3 implies the monotonicity of $1 - x(t) + \log x(t)$. Therefore, when n is large enough, $\ell_{jp}/\bar{\ell}_{jp} \sim b$ and

$$\mathrm{B}_j - \mathrm{B}_k > (j - k)\,[1 - b + \log b + (c/3) \log n] > 0.$$

When $j > 2p/3$,

$$\begin{aligned} \mathrm{B}_j - \mathrm{B}_k &\geq ([2p/3] - k)\,[1 - b + \log b + (c/3) \log n] \\ &\quad + (j - [2p/3])\,(1 - b + \log b) \\ &> ([2p/3] - k)\,[(c/3) \log n - 2(b - 1 - \log b)] > 0 \end{aligned}$$

where we used the fact that $j - [2p/3] < [2p/3] - k$. So $\min\{\mathrm{B}_j,\, j \neq k\} > \mathrm{B}_k$ a.s. This completes the proof of (2). □

We end this section with the following conjecture, and a sketch of evidence for supporting this conjecture.

Conjecture 1: Theorem 3.1 continues to hold when the candidate models are $\{M_0, M_1, \ldots,\, M_{p-1}\}$. In other words, the condition that the number of candidate models is $o(p)$ is superfluous.

Evidence 1. From the proof in Theorem 3.1, indeed we have shown for $k < j < p$,

$$\begin{aligned} \mathrm{A}_j - \mathrm{A}_k &\sim \sum_{i=k+1}^{j} \left[1 - \frac{\ell_{ip}}{\bar{\ell}_{ip}} + \log\left(\frac{\ell_{ip}}{\bar{\ell}_{ip}}\right) + 2c\left(1 - \frac{i}{p}\right)\right] \\ &\quad - \sum_{i=k+1}^{p} \frac{1}{p-i+1}\left(1 - \frac{\ell_{ip}}{\bar{\ell}_{ip}}\right)^2 \\ &= \sum_{i=k+1}^{j} g_i. \end{aligned}$$

By the MP law and the boundedness of ℓ_{1p} under finite fourth moment condition, it can be shown that

$$\sum_{i=k+1}^{j} g_i = (1 + o_{a.s.}(1)) \sum_{i=k+1}^{j} \hat{g}_i$$

where

$$\hat{g}_i = 1 - x(i/p) = \log x(i/p) + 2c(1 - i/p)$$

and

$$x(t) = \frac{\mu_{1-t}}{\frac{1}{1-i/p}\int_a^{\mu_{1-t}} t f_c(s) ds}.$$

It remains to consider the case where $j > k$ and $j/p \to \alpha \in (0, 1)$. For this case, it can be shown that

$$\mathrm{A}_j - \mathrm{A}_k \sim \int_0^{\alpha} \hat{g}(t) dt =: I(c, \alpha).$$

Proof of $I(c,1) > 0$. Note that

$$I(c,1) = 1 - \int_0^1 x(t)dt + \int_0^1 \log x(t)\ dt + c. \tag{3.11}$$

Let $u = \mu_{1-t}$, then

$$\int_0^1 x(t)dt = \int_a^b \frac{u f_c(u) F_c(u)}{\int_a^s f_c(s)\ ds}\ du = -\int_a^b f_c(u) \log\left(\int_a^u s f_c(s) ds\right) du; \tag{3.12}$$

and

$$\begin{aligned} \int_0^1 \log x(t)dt &= \int_a^b f_c(u) \log\left(\frac{u F_c(u)}{\int_a^u s f_c(s) ds}\right) du \\ &= \int_a^b f_c(u) \log u du - 1 + \int_0^1 x(t)dt. \end{aligned} \tag{3.13}$$

By (3.11)-(3.13), we have

$$\begin{aligned} I(c,1) &= c + \int_a^b f(u) \log u\ du \\ &= c + \frac{1}{\pi}\int_{-\pi}^{\pi} \frac{\sin^2\theta}{1 + c - 2\sqrt{c}\cos\theta} \log(1 + c - 2\sqrt{c}\cos\theta)\ d\theta \\ &= \frac{(1-c)}{c}[-\log(1-c) - c] > 0. \end{aligned}$$

We used contour integration in the last step.

Evidence 2. We can show that $\hat{g}(t)$ is positive in the neighbourhood of 0. Numerical calculation for various values of c shows that $\hat{g}(t)$ has at most one zero. We have not been able to prove this. If this were true, then we could prove that

$$I(c,\alpha) > 0 \tag{3.14}$$

and Conjecture 1 would be proved.

Proof of (3.14). If $\hat{g}(t)$ has no zero, then (3.14) holds trivially. If $\hat{g}(t)$ has one zero in $(0,1)$, we denote this zero by t_0. If $0 < \alpha \leq t_0$, then (3.14) holds trivially. If $\alpha > t_0$, then $I(c,\alpha) > I(c,1) > 0$.

4 The case $c > 1$

Increasing number of scientific studies lead to datasets in which $p > n$. Motivated by this scenario, we consider the case when $p, n \to \infty$ such that $p/n \to c \in (1, \infty)$. Then the smallest $p - (n-1)$ eigenvalues of $\mathbf{S}_n$ are zero, that is,

$$\ell_{n-1,p} > \ell_{np} = \cdots = \ell_{pp} = 0.$$

It is still of interest to estimate the true number of significant components in (1.1) under this setting. As $n < p$, it is not possible to infer the smallest population eigenvalues $\lambda_n, \lambda_{n+1}, \ldots, \lambda_p > 0$, and so in this section we assume (C1) with $c > 1$, (C3), (C4) and (C5) hold where

(C5) $\lambda_{n-1} = \lambda_n = \cdots = \lambda_p = \lambda$.

The assumption (C5) is rather natural at least in a high-dimensional PCA setting. Under (C5), we have, for $j = 0, 1, \ldots, n-2$,

$$\widetilde{M}_j;\ \lambda_j > \lambda_{j+1} = \cdots = \lambda_{n-1} \quad \Leftrightarrow \quad M_j :\ \lambda_j > \lambda_{j+1} = \cdots = \lambda_p. \tag{4.1}$$

First, we modify the definition of $\bar{\ell}_{jp}$ in (1.5) to

$$\bar{\ell}_{jp} := \frac{1}{n-1-j} \sum_{t=j+1}^{n-1} \ell_{tp}, \tag{4.2}$$

for $i = 1, 2, \ldots, n-1$.

Second, for selecting a model from the set of models $M_0, M_1, \ldots, M_{n-2}$, we consider the following modified criteria $\tilde{\mathrm{A}}_j$ and $\tilde{\mathrm{B}}_j$ obtained from replacing the p and n in A_j and B_j by $n-1$ and p respectively:

$$\tilde{\mathrm{A}}_j = (n-1-j) \log \bar{\ell}_{jp} - \sum_{i=j+1}^{n-1} \log \ell_{ip} - \frac{(n-j-2)(n-j+1)}{p},$$

$$\tilde{\mathrm{B}}_j = (n-1-j) \log \bar{\ell}_{jp} - \sum_{i=j+1}^{n-1} \log \ell_{ip} - \frac{(n-j-2)(n-j+1)}{2p} \log p.$$

Here $\tilde{\mathrm{A}}_{n-2} = 0$, $\tilde{\mathrm{B}}_{n-2} = 0$. Similar to the case where $c < 1$, we propose the quasi-AIC (or quasi-BIC) rule to select the model $\hat{k}_{\tilde{A}}$ (or $\hat{k}_{\tilde{B}}$) respectively by

$$\hat{k}_{\tilde{A}} = \arg\min(\tilde{\mathrm{A}}_j, j \le n-2) \quad \text{and} \quad \hat{k}_{\tilde{B}} = \arg\min(\tilde{\mathrm{B}}_j, j \le n-2).$$

Third, as $c > 1$, the gap condition (3.5) is modified to

$$\tilde{\gamma}(c) := \psi_k/c - 1 - \log(\psi_k/c) - 2c^{-1} > 0. \tag{4.3}$$

The following theorems show that $\hat{k}_{\tilde{A}}$ and $\hat{k}_{\tilde{B}}$ possess similar consistency properties as $\hat{k}_A$ and $\hat{k}_B$ do.

Theorem 4.1. *Suppose the conditions* (C1) *with* $c > 1$, (C2)–(C5) *hold, and that the number of candidate models* $q = o(p)$. *We have the following results on the consistency of the estimation criterion* $\hat{k}_{\tilde{A}}$ *based on* AIC.

(1) *Suppose that* λ_1 *is finite. If the modified gap condition* (4.3) *fails,* $\hat{k}_A$ *is not consistent. If the modified gap condition* (4.3) *holds,* $\hat{k}_{\tilde{A}}$ *is strongly consistent.*

(2) *Suppose that* λ_k *tends to infinity and* $\lambda_1 = O(p)$. *Then,* $\hat{k}_{\tilde{A}}$ *is strongly consistent.*

Theorem 4.2. *Suppose the conditions* (C1) *with* $c > 1$, (C2)–(C5) *hold. We have the following results on the consistency of the estimation criterion* $\hat{k}_{\tilde{B}}$ *based on* BIC.

(1) *Suppose that* $\lambda_k/\log n \to 0$. *Then,* $\hat{k}_{\tilde{B}}$ *is not consistent.*

(2) *Suppose that* $\lambda_k/\log n$ *tends to infinity. Then,* $\hat{k}_{\tilde{B}}$ *is strongly consistent.*

We shall sketch the proofs of Theorems 4.1 and 4.2 below. For $j < k$, we have

$$\tilde{A}_j - \tilde{A}_k = \sum_{i=j+1}^{k} \left[(n-i) \log \left\{ 1 - \frac{1}{n-i} \left(1 - \frac{\ell_{ip}}{\bar{\ell}_{ip}} \right) \right\} - \log \frac{\ell_{ip}}{\bar{\ell}_{ip}} - \frac{2(n-i)}{p} \right].$$

When $\lambda_1 = o(n)$, we have $\bar{\ell}_{i,n-1} \sim c\int_a^b t f_c(t)dt = c$, and hence if the modified gap condition (4.3) is satisfied,

$$\tilde{A}_j - \tilde{A}_k \sim \sum_{i=j+1}^{k} \left(\psi_i/c - 1 - \log(\psi_i/c) - 2c^{-1}\right) \geq (k-j)\tilde{\gamma}(c) > 0.$$

When $\lambda_1 = O(n)$, the same inequality can be obtained without the modified gap condition $\tilde{\gamma}(c) > 0$.

We next consider the case where $i \in [k+1, n-2]$. Similarly, we have

$$\begin{aligned}
&\tilde{A}_j - \tilde{A}_k \\
&= \sum_{i=k+1}^{j} \left[-(n-i)\log\left\{1 - \frac{1}{n-i}\left(1 - \frac{\ell_{ip}}{\bar{\ell}_{ip}}\right)\right\} + \log\frac{\ell_{ip}}{\tilde{\ell}_{ip}} + \frac{2(n-i)}{p}\right] \\
&\sim \sum_{i=k+1}^{j} \left\{\tilde{g}(i/n) + o(1)\right\},
\end{aligned}$$

when $j = o(p)$, and we used the approximation $\ell_{ip}/\bar{\ell}_{ip} \sim b/c$. Here

$$\tilde{g}(t) = \log\tilde{x}(t) - \tilde{x}(t) + 1 + 2c^{-1}(1-t),$$

and

$$\tilde{x}(t) = \frac{\mu_{1-t}}{\frac{c}{1-ct}\int_{1-1/c}^{1-t}\mu_s ds} \geq 1,$$

and $o(1)$ is uniformly in $i \in [k+1, n-2]$. Similar to the case where $c < 1$, one can prove $\tilde{A}_j - \tilde{A}_k > 0$. Combining these results, we have proved that $\min_{j\neq k}(\tilde{A}_j - \tilde{A}_k) > 0$. Similarly one can prove that $\min_{j\neq k}(\tilde{B}_j - \tilde{B}_k) > 0$, when $\lambda_k/\log n \to \infty$.

Remark. When $c = 1$, the behavior of the smallest eigenvalue is not well understood and we may not have the property that $x(t)$ decreases to 1 as t increases to 1. However, if the number of candidate models is $o(p)$, Theorem 3.1 or Theorem 4.1 holds for $c = 1$.

We also end this section with another conjecture below.

Conjecture 2: Theorem 4.1 continues to hold when the candidate models are $\{M_0, M_1, \ldots, M_{n-2}\}$.

5 Simulation studies

When the population eigenvalues remain bounded, we impose the gap condition and the finite fourth moment condition to establish the consistency of the AIC and BIC; whereas when $\lambda_k \to \infty$ at a rate faster than $\log n$, we only need finite fourth moment. We conducted a number of simulation studies to examine the effects on the consistency of $\hat{k}_A$ and $\hat{k}_B$ when the gap condition or the finite fourth moment condition does not hold. Moreover, when these conditions are met, we are interested to gain some insight at the rate of convergence.

5.1 Simulation studies for $0 < c < 1$

In our experiments, we define p-variate $\boldsymbol{y}$ as

$$\boldsymbol{y} = \boldsymbol{\Lambda}^{1/2}(x_1, \ldots, x_p)^\top, \tag{5.1}$$

where $\boldsymbol{\Lambda} = \mathrm{diag}(\widetilde{\lambda}_1, \ldots, \widetilde{\lambda}_k, \widetilde{\lambda}, \ldots, \widetilde{\lambda})$ and $x_1, \ldots, x_p$ are iid with mean 0 and variance 1. We also set $p/n = 1/3$, that is, $c = 1/3$. So $m(c) = 2.636$. It follows that the covariance of $\boldsymbol{y}$ is given by $\boldsymbol{\Lambda}$. As for the distribution of x_i, we consider the following five cases. Note that the fourth moment of x_i exists only for cases D1, D3–D5; whereas for case D2, which is a standardized t_4 distribution with finite moments up to order 3.

D1 Standard normal distribution: $x_i \sim N(0, 1)$
D2 Standardized t distribution with 4 d.f.: $x_i \sim t_4/\sqrt{\mathrm{Var}(t_4)}$
D3 Standardized t distribution with 5 d.f.: $x_i \sim t_5/\sqrt{\mathrm{Var}(t_5)}$
D4 Standardized t distribution with 10 d.f.: $x_i \sim t_{10}/\sqrt{\mathrm{Var}(t_{10})}$
D5 Standardized chi-square distribution with 3 d.f.: $x_i \sim (\chi^2_3 - 3)/\sqrt{\mathrm{Var}(\chi^2_3)}$

For the eigenvalues of $\boldsymbol{\Sigma}$, we considered the following three cases. In the first case, L1, the gap condition fails: $\lambda_4 = 5/3$ and $\psi_4 = 2.5$ which is less than $m = 2.636$. In L2, the gap condition holds. In L3, the spiked eigenvalues $(1 \le i \le 4)$ tend to infinity at a rate $n^{1/2}$ which is faster than $\log n$. Here $\alpha_p = \sqrt{p/10}$.

	$\widetilde{\lambda}_1$	$\widetilde{\lambda}_2$	$\widetilde{\lambda}_3$	$\widetilde{\lambda}_4$	$\widetilde{\lambda}_5$	$\cdots$	$\widetilde{\lambda}_p$
L1	30	20	13	5	3	$\cdots$	3
L2	30	22	16	10	3	$\cdots$	3
L3	$30\alpha_p$	$20\alpha_p$	$13\alpha_p$	$8\alpha_p$	3	$\cdots$	3

In our framework, λ_i is taken to be $\widetilde{\lambda}_i/\widetilde{\lambda}$. The true number of significant components in all cases, L1–L3, is the same: $k = 4$. We simply use j to denote M_j. Let the minimum model including the true model be denoted by $\mathcal{F}_*$. Furthermore, let the sets of under-specified and over-specified models be denoted by $\mathcal{F}_-$ and $\mathcal{F}_+$ respectively. In our simulation studies,

$$\mathcal{F}_- = \{0, 1, 2, 3\}, \quad \mathcal{F}_* = \{4\}, \quad \mathcal{F}_+ = \{5, 6, \ldots, p\}.$$

The selection percentages of selecting $\mathcal{F}_-$, $\mathcal{F}_*$ and $\mathcal{F}_+$ by Monte Carlo simulations with 10^4 repetitions were computed. For space consideration, we only report the selection percentages for standardized t_4 and t_5 under cases L1–L3 in Tables 2–4 below. We first summarize our findings in Table 1 below. Here, "Y" denotes $\hat{k}_A$ (or $\hat{k}_B$) is consistent; and "N" for inconsistent.

Table 1: Summary of consistency of AIC and BIC. Here GC stands for gap condition.

	L1 (GC fails)		L2 (GC holds)		L3 (GC holds)	
	AIC	BIC	AIC	BIC	AIC	BIC
D1	N	N	Y	N	Y	Y
D3	N	N	Y	N	Y	Y
D4	N	N	Y	N	Y	Y
D5	N	N	Y	N	Y	Y
D2	N	N	N	N	N	Y

We highlight some observations from Tables 1-4 as follows.

(i) In the standardized t_4 case, D2, $\hat{k}_A$ is not consistent across our choices of eigenvalues L1–L3. The finite fourth moment condition is essential for Theorem 3.1 to hold. Moreover, when AIC does not specify the true number of significant components correctly, it tends to over-specify it.

(ii) In the standardized t_4 case, D2, $\hat{k}_B$ is not consistent for L1 and L2 cases with tendency to under-specify the true number of significant components. Interestingly, when eigenvalues tend to infinity fast enough as in L3, our simulation results suggest that $\hat{k}_B$ is consistent even the finite fourth moment condition fails.

(iii) Our simulation studies suggest that the rate of convergence to the true number of significant components depend on the difference of the spiked population eigenvalues and λ_{k+1}. Convergence is faster for greater difference. Moreover, when $\hat{k}_A$ and $\hat{k}_B$ are both consistent, selection percentages of $\hat{k}_B$ converges faster to 100 than $\hat{k}_A$ does.

Table 2. Selection percentages of $\hat{k}_A$ and $\hat{k}_B$ for population eigenvalues in L1

		Standardized t_4				Standardized t_5			
		$\hat{k}_A$		$\hat{k}_B$		$\hat{k}_A$		$\hat{k}_B$	
n	p	$\mathcal{F}_-$	$\mathcal{F}_*$	$\mathcal{F}_-$	$\mathcal{F}_*$	$\mathcal{F}_-$	$\mathcal{F}_*$	$\mathcal{F}_-$	$\mathcal{F}_*$
30	10	37.1	27.9	83.3	13.3	43.5	28.9	88.5	10.1
60	20	28.1	34.4	91.3	8.0	42.4	35.1	96.5	3.4
90	30	26.4	35.9	94.0	5.6	43.4	37.7	98.4	1.6
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
270	90	26.9	39.0	96.0	3.9	59.1	34.0	99.5	0.5
300	100	28.8	39.4	96.6	3.3	62.3	31.9	99.4	0.6
900	300	29.3	40.0	96.6	3.3	63.8	31.5	99.5	0.5

Table 3. Selection percentages of $\hat{k}_A$ and $\hat{k}_B$ for population eigenvalues in L2

		Standardized t_4				Standardized t_5			
		$\hat{k}_A$		$\hat{k}_B$		$\hat{k}_A$		$\hat{k}_B$	
n	p	$\mathcal{F}_-$	$\mathcal{F}_*$	$\mathcal{F}_-$	$\mathcal{F}_*$	$\mathcal{F}_-$	$\mathcal{F}_*$	$\mathcal{F}_-$	$\mathcal{F}_*$
30	10	11.7	35.8	51.9	38.3	11.2	43.1	51.7	42.0
60	20	2.9	34.5	56.4	38.9	3.2	50.8	56.0	42.7
90	30	1.2	34.4	62.1	35.2	1.2	55.7	63.2	36.1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
270	90	0.0	36.5	81.8	17.2	0.0	77.9	87.9	12.0
300	100	0.0	38.3	83.3	15.9	0.0	79.9	90.0	9.9
900	300	0.0	50.6	94.5	5.3	0.0	92.7	99.3	0.7

Table 4. Selection percentages of $\hat{k}_A$ and $\hat{k}_B$ for population eigenvalues in L3

		Standardized t_4				Standardized t_5			
		$\hat{k}_A$		$\hat{k}_B$		$\hat{k}_A$		$\hat{k}_B$	
n	p	$\mathcal{F}_-$	$\mathcal{F}_*$	$\mathcal{F}_-$	$\mathcal{F}_*$	$\mathcal{F}_-$	$\mathcal{F}_*$	$\mathcal{F}_-$	$\mathcal{F}_*$
30	10	21.7	32.0	69.8	24.3	23.1	37.9	70.9	25.4
60	20	1.3	35.2	38.7	55.3	1.2	51.1	36.0	61.4
90	30	0.1	32.1	16.8	76.1	0.0	55.1	12.8	85.3
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
270	90	0.0	35.2	0.0	94.1	0.0	76.7	0.0	99.3
300	100	0.0	37.4	0.0	94.0	0.0	79.7	0.0	99.3
900	300	0.0	49.6	0.0	95.7	0.0	92.3	0.0	99.6

5.2 Simulation studies for $c > 1$

For the case where $n, p \to \infty$ such that $p/n \to c > 1$, we consider the consistency properties of $\hat{k}_{\tilde{A}}$ and $\hat{k}_{\tilde{B}}$ under the population eigenvalues

$$\text{L4: } \tilde{\lambda}_1 = 30 > \tilde{\lambda}_2 = 20 > \tilde{\lambda}_3 = 13 > \tilde{\lambda}_4 = 8 > \tilde{\lambda}_5 = \cdots = \tilde{\lambda}_p = 1$$

in addition to L1, L2 and L3. The variables $x_1, \ldots, x_p$ in (5.1) were chosen to be iid from the standard normal distribution. We set $p/n = 3$ and hence $c = 3$. Note that L1 satisfies the gap condition (4.3), but not the condition

(C4); L2 satisfies the condition (C4), but not the gap condition. So, we do not expect that $\hat{k}_{\tilde{A}}$ and $\hat{k}_{\tilde{B}}$ are consistent in L1 and L2. On the other hand, from Theorems 4.1 and 4.2 it is expected that $\hat{k}_{\tilde{A}}$ is consistent in L3 and L4; and $\hat{k}_{\tilde{B}}$ is consistent in L3. Simulation results, as shown in Table 5, confirm the results of Theorems 4.1 and 4.2.

Table 5. Selection percentages of $\hat{k}_{\tilde{A}}$ and $\hat{k}_{\tilde{B}}$ for eigenvalues in L1–L4

		L1				L2			
		$\hat{k}_{\tilde{A}}$		$\hat{k}_{\tilde{B}}$		$\hat{k}_{\tilde{A}}$		$\hat{k}_{\tilde{B}}$	
n	p	$\mathcal{F}_-$	$\mathcal{F}_*$	$\mathcal{F}_-$	$\mathcal{F}_*$	$\mathcal{F}_-$	$\mathcal{F}_*$	$\mathcal{F}_-$	$\mathcal{F}_*$
10	30	92.2	4.7	97.5	1.9	88.6	7.0	95.6	3.1
20	60	93.8	5.3	100.0	0.1	86.7	11.5	99.8	0.2
30	90	94.7	5.1	100.0	0.0	85.9	13.2	100.0	0.0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
90	270	98.5	1.5	100.0	0.0	86.0	14.0	100.0	0.0
100	300	98.9	1.1	100.0	0.0	86.3	13.7	100.0	0.0
300	900	100.0	0.0	100.0	0.0	94.1	5.9	100.0	0.0

		L3				L4			
		$\hat{k}_{\tilde{A}}$		$\hat{k}_{\tilde{B}}$		$\hat{k}_{\tilde{A}}$		$\hat{k}_{\tilde{B}}$	
n	p	$\mathcal{F}_-$	$\mathcal{F}_*$	$\mathcal{F}_-$	$\mathcal{F}_*$	$\mathcal{F}_-$	$\mathcal{F}_*$	$\mathcal{F}_-$	$\mathcal{F}_*$
10	30	74.8	16.8	85.6	11.1	47.1	37.0	59.0	33.2
20	60	28.6	61.7	71.8	28.1	15.0	73.8	49.6	50.3
30	90	5.4	87.4	46.9	53.1	5.2	87.8	47.6	52.4
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
90	270	0.0	99.3	0.4	99.6	0.0	99.3	67.9	32.1
100	300	0.0	99.4	0.1	99.9	0.0	99.5	72.2	27.8
300	900	0.00	100.0	0.0	100	0.0	100	99.7	0.3

6 Concluding remarks

In this paper we consider the consistency problem in estimating the number of dominant eigenvalues in (1.1) which is called the number of significant components or the dimensionality in PCA. High-dimensional properties are

studied for two estimation criteria $\hat{k}_A$ and $\hat{k}_B$ based on AIC_j and BIC_j, which are equivalent to A_j and B_j. When the true number of significant components is $o(p)$, we give sufficient conditions in Theorems 3.1 and 3.2 for the criteria $\hat{k}_A$ and $\hat{k}_B$ to be strongly consistent under a high-dimensional asymptotic framework such that $p/n \to c \in (0, 1)$. We emphasize that the consistency properties of the AIC and BIC criteria differ substantially from those in a large-sample asymptotic framework. In a large-sample asymptotic framework, in general, $\hat{k}_A$ is not consistent, but $\hat{k}_B$ is consistent. When $n < p$, we propose quasi-AIC and quasi-BIC decision rules $\hat{k}_{\tilde{A}}$ and $\hat{k}_{\tilde{B}}$. Further, their consistency properties are summarized in Theorems 4.1 and 4.2.

These theorems were proved by random matrix theory techniques. We were also led to discover some interesting limiting results in sample eigenvalues when the population eigenvalues tend to infinity (see Lemma 2.2); and monotonicity property on ratio of quantiles of the MP law (see Lemma 2.3).

Appendix

A Additional lemmas and proof of lemmas

A.1 Two Additional Lemmas

We need two additional lemmas to prove Lemma 2.2. The first Lemma A.4 below which is a modification of Lemma 2 from Bai and Yin (1993).

Lemma A.4. *Let x be a random variable with $E|x|^{(1+\beta)/\alpha} < \infty$ for some $\alpha > 1/2$, $\beta \geq 0$. Let $\{x_{ij}\}$ be a double array of random variables such that $P(|x_{ij}| > t) \leq KP(|x| > t)$ for all i, j, $t > 0$, and a fixed constant K. For each j fixed, we assume further that $x_{1j}, \ldots, x_{nj}$ are independent. For $1/2 < \alpha \leq 1$, we require further that x_{ij}'s have the same mean. Then for any constant $0 < M < \infty$, we have*

$$\lim_{n\to\infty} \sup_{j \leq Mn^{\beta}} \left| n^{-\alpha} \sum_{i=1}^{n} (x_{ij} - \nu) \right| = 0 \ a.s. \tag{A.1}$$

Here

$$\nu = \begin{cases} E(x_{11}), & \text{if } 1/2 < \alpha \leq 1, \\ \text{any constant}, & \text{if } \alpha > 1. \end{cases}$$

Proof. The proof of the Lemma is the same as the proof for the sufficient part of Lemma 2 of Bai and Yin (1993) by noticing that the independence between rows of random variables was in fact not used in the latter. Details are omitted. □

Write $\mathbf{S}_{n\boldsymbol{x}} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^\top$ and $\boldsymbol{\Sigma} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^\top$, where $\mathbf{U} = (\mathbf{U}_1, \mathbf{U}_2) = (\mathbf{u}_1, \cdots, \mathbf{u}_p)$ is a p-dimensional orthogonal matrix with $\mathbf{U}_1$ of dimension $p \times k$ and $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_k, 1, \ldots, 1)$ is a diagonal matrix of eigenvalues of $\boldsymbol{\Sigma}$. Then,

$$\begin{aligned} \mathbf{S}_n &= \mathbf{U}\boldsymbol{\Lambda}^{1/2}\mathbf{U}^\top\mathbf{S}_{n\boldsymbol{x}}\mathbf{U}\boldsymbol{\Lambda}^{1/2}\mathbf{U}^\top \\ &= \mathbf{U}\begin{pmatrix} \boldsymbol{\Lambda}_1^{1/2}\mathbf{U}_1^\top\mathbf{S}_{n\boldsymbol{x}}\mathbf{U}_1\boldsymbol{\Lambda}_1^{1/2} & \Lambda_1^{1/2}\mathbf{U}_1^\top\mathbf{S}_{n\boldsymbol{x}}\mathbf{U}_2 \\ \mathbf{U}_2^\top\mathbf{S}_{n\boldsymbol{x}}\mathbf{U}_1\boldsymbol{\Lambda}_1^{1/2} & \mathbf{U}_2^\top\mathbf{S}_{n\boldsymbol{x}}\mathbf{U}_2 \end{pmatrix}\mathbf{U}^\top, \end{aligned} \tag{A.2}$$

where $\boldsymbol{\Lambda}_1 = \text{diag}(\lambda_1, \ldots, \lambda_k)$. Since $\mathbf{U}_2\mathbf{U}_2^\top$ has $p-k$ eigenvalues 1 and k eigenvalues 0, we know that the ESD of $\mathbf{U}_2^\top\mathbf{S}_{nx}\mathbf{U}_2$ tends to MP law by Silverstein (1995) and its largest eigenvalue tends to b and smallest eigenvalue tends to a by Bai and Silverstein (1998).

Lemma A.5. *Under the assumption of Lemma 2.2, we have*

$$\max_{j \leq k} \left|\mathbf{u}_j^\top\mathbf{S}_{n\boldsymbol{x}}\mathbf{u}_j - 1\right| \to 0 \ a.s.$$

Proof. It suffices to show that $\mathbf{u}_1^\top\mathbf{S}_{n\boldsymbol{x}}\mathbf{u}_1 \to 1$ *a.s.* Without loss of generality, we may assume the means of the random entries are 0. Let $\mathbf{u}_1 =$

$(u_1, \cdots, u_p)^\top$, then we have

$$\begin{aligned}
|\mathbf{u}_1' \mathbf{S}_{n\boldsymbol{x}} \mathbf{u}_1 - 1| &= \left| \frac{1}{n-1} \sum_{j=1}^{p} u_i^2 \sum_{j=1}^{n} (x_{ij}^2 - 1) \right. \\
&\quad \left. + \frac{1}{n-1} \sum_{j_1 \neq j_2} u_{j_1} u_{j_2} \sum_{i=1}^{n} x_{ij_1} x_{ij_2} - \frac{n}{n-1} (\mathbf{u}_1^\top \bar{\mathbf{x}})^2 \right| \\
&\leq \sup_{j \leq p} \left| \frac{1}{n-1} \sum_{i=1}^{n} (x_{ij}^2 - 1) \right| \\
&\quad + \left| \frac{1}{n-1} \sum_{i=1}^{n} \sum_{j_1 \neq j_2} u_{j_1} u_{j_2} x_{ij_1} x_{ij_2} \right| + \frac{n}{n-1} \sup_{j \leq p} \left| \frac{1}{n} \sum_{i=1}^{n} x_{ij} \right|^2 .
\end{aligned}$$

By Lemma A.4, the first and the third term tends to 0 with probability 1. That the second term tends to 0 with probability 1 can be proved by noticing that

$$\begin{aligned}
& E \left| \frac{1}{n-1} \sum_{i=1}^{n} \sum_{j_1 \neq j_2} u_{j_1} u_{j_2} x_{ij_1} x_{ij_2} \right|^4 \\
= \; & \frac{1}{(n-1)^4} \left[\sum_{i=1}^{n} E \left(\sum_{j_1 \neq j_2} u_{j_1} u_{j_2} x_{ij_1} x_{ij_2} \right)^4 \right. \\
& \left. + 3 \sum_{i_1 \neq i_2} E \left(\sum_{j_1 \neq j_2} u_{j_1} u_{j_2} x_{i_1 j_1} x_{i_1 j_2} \right)^2 \left(\sum_{j_1 \neq j_2} u_{j_1} u_{j_2} x_{i_2 j_1} x_{i_2 j_2} \right)^2 \right] \\
\leq \; & \frac{n}{(n-1)^4} \left[24 \sum_{\substack{j_1, j_2, j_3, j_4 \\ distinct}} u_{j_1}^2 u_{j_2}^2 u_{j_3}^2 u_{j_4}^2 + 24 \sum_{\substack{j_1, j_2, j_3 \\ distinct}} u_{j_1}^3 u_{j_2}^3 u_{j_3}^2 E x_{11}^3 E x_{11}^3 \right. \\
& \left. + 8 \sum_{j_1 \neq j_2} u_{j_1}^4 u_{j_2}^4 E x_{11}^4 E x_{11}^4 + 12n(n-1) \right] \leq \frac{K}{n^2},
\end{aligned}$$

for some constant K. The proof is complete. □

A.2 Proof of (1) of Lemma 2.2

First we prove that $\liminf \ell_{ip}/\lambda_i \geq 1$ a.s. for $i \leq k$. We note that

$$\ell_{ip}/\lambda_i = \lambda_i^{-1} \inf_{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{i-1}} \sup_{\boldsymbol{u} \perp \boldsymbol{v}_1, \ldots, \boldsymbol{v}_{i-1}, \|\mathbf{u}\|=1} \boldsymbol{u}^\top \mathbf{S}_n \boldsymbol{u}$$

For any given $\boldsymbol{v}_1, \cdots, \boldsymbol{v}_{i-1}$ there exists a vector $\mathbf{u}$ in the linear space spanned by $\mathbf{u}_1, \cdots, \mathbf{u}_i$ which is orthogonal to $\boldsymbol{v}_1, \cdots, \boldsymbol{v}_{i-1}$ denoted by $\mathbf{u} = \sum_{j=1}^{i} a_j \mathbf{u}_j$ with $\sum_{j=1}^{i} a_j^2 = 1$.

By Lemma A.5, we have

$$\mathbf{u}^\top \mathbf{S}_{nx} \mathbf{u} / \lambda_i = \lambda_i^{-1} \sum_{j=1}^{i} \lambda_j a_j^2 \mathbf{u}_j^\top \mathbf{S}_{nx} \mathbf{u}_j \geq \sum_{j=1}^{i} a_j^2 \mathbf{u}_j^\top \mathbf{S}_{nx} \mathbf{u}_j \xrightarrow{a.s.} 1.$$

Next, we shall show that $\limsup \ell_{ip}/\lambda_i \leq 1$ a.s. for $i \leq k$. As before, we have

$$\begin{aligned}
\ell_{ip}/\lambda_i &= \lambda_i^{-1} \inf_{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{i-1}} \sup_{\boldsymbol{u} \perp \boldsymbol{v}_1, \ldots, \boldsymbol{v}_{i-1}, \|\mathbf{u}\|=1} \boldsymbol{u}^\top \mathbf{S}_n \boldsymbol{u} \\
&\leq \lambda_i^{-1} \sup_{\boldsymbol{u} \perp \boldsymbol{u}_1, \ldots, \boldsymbol{u}_{i-1}, \|\boldsymbol{u}\|=1} \boldsymbol{u}^\top \mathbf{S}_n \boldsymbol{u} \\
&= \lambda_i^{-1} \sup_{a \leq 1} \left\{ a^2 \mathbf{u}_i^\top \mathbf{S}_{nx} \mathbf{u}_i + (1-a^2) \sup_{\boldsymbol{u} \perp \boldsymbol{u}_1, \ldots, \boldsymbol{u}_k . \|\boldsymbol{u}\|=1} \boldsymbol{u}^\top \mathbf{S}_n \boldsymbol{u} \right\} \\
&\sim \sup_{|a| \leq 1} \left\{ a^2 + (1-a^2) \lambda_i^{-1} \|\mathbf{U}_2^\top \mathbf{S}_n \mathbf{U}_2\| \right\} \\
&\sim \sup_{|\sum_{t=i}^{k} a_t^2| \leq 1} \left\{ \sum_{t=i}^{k} a_t^2 + (1 - \sum_{t=i}^{k} a_t^2) \lambda_i^{-1} b \right\} = 1,
\end{aligned}$$

where we have used the fact that $\|\mathbf{U}_2^\top \mathbf{S}_n \mathbf{U}_2\| \to b$ which was proved in Bai and Silverstein (1998).

Combining the two conclusions, we conclude that $\ell_{jp}/\lambda_j \xrightarrow{a.s.} 1$.

A.3 Proof of (2) of Lemma 2.2

By (A.2), $\ell_{1p}, \cdots, \ell_{pp}$ are also the eigenvalues of

$$\begin{pmatrix} \boldsymbol{\Lambda}_1^{1/2} \mathbf{U}_1^\top \mathbf{S}_{n\boldsymbol{x}} \mathbf{U}_1 \boldsymbol{\Lambda}_1^{1/2} & \Lambda_1^{1/2} \mathbf{U}_1^\top \mathbf{S}_{n\boldsymbol{x}} \mathbf{U}_2 \\ \mathbf{U}_2^\top \mathbf{S}_{n\boldsymbol{x}} \mathbf{U}_1 \boldsymbol{\Lambda}_1^{1/2} & \mathbf{U}_2^\top \mathbf{S}_{n\boldsymbol{x}} \mathbf{U}_2 \end{pmatrix}.$$

Write the eigenvalues of the matrix $\mathbf{U}_2^\top \mathbf{S}_{n\boldsymbol{x}} \mathbf{U}_2$ as $\tilde{\ell}_{1p}, \cdots, \tilde{\ell}_{p-k,p}$. By Silverstein (1995), the empirical spectral distribution of $\mathbf{U}_2^\top \mathbf{S}_{n\boldsymbol{x}} \mathbf{U}_2$ tends to MP law with probability 1. Thus, if $i/p \to \alpha$, then $\tilde{\ell}_{ip} \xrightarrow{a.s.} \mu_{1-\alpha}$.

On the other hand, by interlacing theorem (see Rao and Rao (1998)), for any $i \in (1, p-k)$, we have

$$\ell_{ip} \geq \tilde{\ell}_{ip} \geq \ell_{k+i,p} \geq \tilde{\ell}_{k+i,p}.$$

Thus, for all $i \geq k+1$, $\ell_{ip} \stackrel{a.s.}{\to} \mu_{1-\alpha}$, where $\alpha = \lim i/p$. This completes the proof of Lemma 2.2.

A.4 Proof of Lemma 2.3

For notational simplicity, we write F_c and f_c as F and f respectively for the rest of this paper. Define $G(t) = F^{-1}(t)$, the t-th quantile of the MP, which is denoted by μ_t earlier.

Note that $G'(t) = \frac{1}{f(G(t))}$. We write $y(t) = tG(t)/\int_0^t G(s)ds$, which is equal to $x(1-t)$. Thus we want to prove that y increases from $y(0) = 1$ to $y(1) = b$. Towards this end, we have

$$\begin{aligned} y'(t) &= \frac{[G(t) + tG'(t)]\int_0^t G(s)ds - t\left[G(t)\right]^2}{\left(\int_0^t G(s)ds\right)^2} \\ &= \frac{\left[f\left(G(t)\right)G(t) + t\right]\int_0^t G(s)ds - tf\left(G(t)\right)\left[G(t)\right]^2}{f\left(G(t)\right)\left(\int_0^t G(s)ds\right)^2}. \end{aligned}$$

So to prove $y'(t) > 0$, it is equivalent to proving that

$$\Delta(t) \equiv \int_0^t G(s)ds - \frac{tf\left(G(t)\right)\left[G(t)\right]^2}{\left[f\left(G(t)\right)G(t) + t\right]} > 0. \tag{A.3}$$

It is easy to see that $\lim_{t\to 0+}\Delta(t) = 0$. If we can show that

$$\Delta'(t) > 0 \quad \text{for } t \in (0,1), \tag{A.4}$$

then $\Delta(t) > \Delta(0+) = 0$, and so $y'(t) > 0$.

We have

$$\begin{aligned}\Delta'(t) &= G(t) - \frac{f(G(t))[G(t)]^2 + \frac{tf'(G(t))[G(t)]^2}{f(G(t))} + 2tG(t)}{[f(G(t))G(t)+t]} \\ &\quad + \frac{t[G(t)]^2[2f(G(t)) + f'(G(t))G(t)]}{[f(G(t))G(t)+t]^2} \\ &= G(t) - \frac{[f(G(t))]^2[G(t)]^2 + tf'(G(t))[G(t)]^2 + 2tf(G(t))G(t)}{f(G(t))[f(G(t))G(t)+t]} \\ &\quad + \frac{t[G(t)]^2[2f(G(t)) + f'(G(t))G(t)]}{[f(G(t))G(t)+t]^2}.\end{aligned}$$

If we let $u = G(t)$, then $u \in (a,b)$ and $t = F(u)$. We can rewrite $\Delta'(t)$ as $u\psi(u)$ where

$$\begin{aligned}\psi(u) &= 1 - \frac{u[f(u)]^2 + uf'(u)F(u) + 2f(u)F(u)}{f(u)[uf(u)+F(u)]} + \frac{uF(u)[2f(u)+uf'(u)]}{[uf(u)+F(u)]^2} \\ &= \frac{\psi_1(u)F(u)}{[uf(u)+F(u)]^2}.\end{aligned}$$

Here

$$\psi_1(u) = \frac{1}{u} - \frac{h'(u)F(u)}{h^2(u)}, \tag{A.5}$$

where

$$h(u) = uf(u) = (2\pi c)^{-1}\sqrt{(b-u)(u-a)}.$$

Finally, to show that $\Delta'(t) > 0$, it remains to show that $\psi_1(u) > 0$ for $u \in (a,b)$.

Since

$$h'(u) = \frac{-u + (b+a)/2}{2\pi c\sqrt{(b-u)(u-a)}} = \frac{1+c-u}{2\pi c\sqrt{(b-u)(u-a)}},$$

we know that $h'(u) < 0$ if $u \geq 1+c$ and hence $\psi_1(u) > 0$. Thus, we need only to prove that $\psi_1(u) > 0$ for $u \in (a, 1+c)$. Rewriting

$$\psi_1(u) = \frac{1+c-u}{[(b-u)(u-a)]^{3/2}}\ \psi_2(u),$$

where

$$\psi_2(u) = \frac{[(b-u)(u-a)]^{3/2}}{u(1+c-u)} - \int_a^u \frac{\sqrt{(b-s)(s-a)}}{s}\ ds, \quad u \in (a, 1+c).$$

Observe that $\psi_2(a) = 0$. Writing $\beta(u) = \sqrt{(b-u)(u-a)}/[u^2(1+c-u)^2]$, it is straightforward to verify that

$$\begin{aligned}
\psi_2'(u) &= \beta(u)\left\{3(1+c-u)^2 - (b-u)(u-a)(1+c-2u) - u(1+c-u)^2\right\} \\
&= \beta(u)\left\{(1+c)u^2 - 2(1-c)^2u + (1+c)(1-c)^2\right\} \\
&= (1+c)\beta(u)\left\{\left[u - \frac{(1-c)^2}{1+c}\right]^2 + 4c(1-c)^2/(1+c)^2\right\} > 0.
\end{aligned}$$

So ψ_2 is increasing on $(a, 1+c)$. As $\psi_2(a) = 0$, therefore $\psi_2(u) > 0$ and thus $\psi_1(u) > 0$ on $(a, 1+c)$. This completes the proof of Lemma 2.3 for $0 < c < 1$.

The proof for $c > 1$ is similar and goes as follows. We still write F_c and f_c as F and f for brevity. We let $\bar{c} = 1 - 1/c$. Define $G(t) = F^{-1}(t)$ for $t \in (1-1/c, 1)$ and $G(t) = a$ when $t \in (0, 1-1/c)$, the t-th quantile of the MP, which is denoted by μ_t earlier.

Note that $G'(t) = \frac{1}{f(G(t))}$, when $t > 1 - 1/c$ and $= 0$ otherwise. We write $y(t) = (t-\bar{c})G(t)/\int_{\bar{c}}^t G(s)ds$ when $t \in (\bar{c}, 1)$, which is equal to $x(1-t)$. Thus we want to prove that y increases from $y(\bar{c}) = 1$ to $y(1) = b$. Towards this end, for $t \in (\bar{c}, 1)$, we have

$$\begin{aligned}
y'(t) &= \frac{[G(t) + (t-\bar{c})G'(t)]\int_{\bar{c}}^t G(s)ds - (t-\bar{c})\left[G(t)\right]^2}{\left(\int_{\bar{c}}^t G(s)ds\right)^2} \\
&= \frac{\left[f\left(G(t)\right)G(t) + (t-\bar{c})\right]\int_{\bar{c}}^t G(s)ds - (t-\bar{c})f\left(G(t)\right)\left[G(t)\right]^2}{f\left(G(t)\right)\left(\int_{\bar{c}}^t G(s)ds\right)^2}.
\end{aligned}$$

So to prove $y'(t) > 0$ when $t \in (\bar{c}, 1)$, it is equivalent to proving that

$$\Delta(t) \equiv \int_{\bar{c}}^t G(s)ds - \frac{(t-\bar{c})f\left(G(t)\right)\left[G(t)\right]^2}{\left[f\left(G(t)\right)G(t) + t - \bar{c}\right]} > 0. \tag{A.6}$$

It is easy to see that $\lim_{t\downarrow\bar{c}} \Delta(t) = 0$. If we can show that

$$\Delta'(t) > 0 \quad \text{for } t \in (\bar{c}, 1), \tag{A.7}$$

then $\Delta(t) > \Delta(\bar{c}+) = 0$, and so $y'(t) > 0$.

We have

$$\begin{aligned}
& \Delta'(t) \\
= \; & G(t) - \frac{f(G(t))[G(t)]^2 + \frac{(t-\bar{c})f'(G(t))[G(t)]^2}{f(G(t))} + 2(t-\bar{c})G(t)}{\left[f\left(G(t)\right)G(t) + t - \bar{c}\right]} \\
& + \frac{(t-\bar{c})\left[G(t)\right]^2\left[2f\left(G(t)\right) + f'(G(t))G(t)\right]}{\left[f\left(G(t)\right)G(t) + t - \bar{c}\right]^2} \\
= \; & G(t) + \frac{(t-\bar{c})\left[G(t)\right]^2\left[2f\left(G(t)\right) + f'(G(t))G(t)\right]}{\left[f\left(G(t)\right)G(t) + t - \bar{c}\right]^2} \\
& - \frac{[f(G(t))]^2[G(t)]^2 + (t-\bar{c})f'(G(t))[G(t)]^2 + 2(t-\bar{c})f(G(t))G(t)}{f(G(t))\left[f\left(G(t)\right)G(t) + t - \bar{c}\right]}.
\end{aligned}$$

If we let $u = G(t)$, then $u \in (a, b)$ and $t = F(u)$. We can rewrite $\Delta'(t)$ as $u\psi(u)$ where

$$\begin{aligned}
\psi(u) &= 1 - \frac{u[f(u)]^2 + uf'(u)F(u) + 2f(u)F(u)}{f(u)[uf(u) + F(u)]} + \frac{uF(u)[2f(u) + uf'(u)]}{[uf(u) + F(u)]^2} \\
&= \frac{\psi_1(u)F(u)}{[uf(u) + F(u)]^2}.
\end{aligned}$$

Here

$$\psi_1(u) = \frac{1}{u} - \frac{h'(u)F(u)}{h^2(u)}, \tag{A.8}$$

where

$$h(u) = uf(u) = (2\pi c)^{-1}\sqrt{(b-u)(u-a)}.$$

Finally, to show that $\Delta'(t) > 0$, it remains to show that $\psi_1(u) > 0$ for $u \in (a, b)$.

Since

$$h'(u) = \frac{-u + (b+a)/2}{2\pi c\sqrt{(b-u)(u-a)}} = \frac{1 + c - u}{2\pi c\sqrt{(b-u)(u-a)}},$$

we know that $h'(u) < 0$ if $u \geq 1 + c$ and hence $\psi_1(u) > 0$. Thus, we need only to prove that $\psi_1(u) > 0$ for $u \in (a, 1 + c)$. Rewriting

$$\psi_1(u) = \frac{1 + c - u}{[(b-u)(u-a)]^{3/2}}\ \psi_2(u),$$

where

$$\psi_2(u) = \frac{[(b-u)(u-a)]^{3/2}}{u(1+c-u)} - \int_a^u \frac{\sqrt{(b-s)(s-a)}}{s}\, ds, \quad u \in (a, 1+c).$$

Observe that $\psi_2(a) = 0$. Writing $\beta(u) = \sqrt{(b-u)(u-a)}/[u^2(1+c-u)^2]$, it is straightforward to verify that

$$\begin{aligned}
\psi_2'(u) &= \beta(u)\left\{3(1+c-u)^2 - (b-u)(u-a)(1+c-2u) - u(1+c-u)^2\right\} \\
&= \beta(u)\left\{(1+c)u^2 - 2(1-c)^2 u + (1+c)(1-c)^2\right\} \\
&= (1+c)\beta(u)\left\{\left[u - \frac{(1-c)^2}{1+c}\right]^2 + 4c(1-c)^2/(1+c)^2\right\} > 0.
\end{aligned}$$

So ψ_2 is increasing on $(a, 1+c)$. As $\psi_2(a) = 0$, therefore $\psi_2(u) > 0$ and thus $\psi_1(u) > 0$ on $(a, 1+c)$. This completes the proof of Lemma 2.3.

Acknowledgements

We wish to thank Dr Tetsuro Sakurai for his help in the simulation study in Section 5; and Dr NH Tran for some numerical computation. The first author is partially supported by NSFC 11171057 and by PCSIRT; the second author by the Ministry of Education, Science, Sports, and Culture, a Grant-in-Aid for Scientific Research (C), #25330038, 2013-2015; and the third author by the Singapore Ministry of Education Academic Research Fund R-155-000-141-112.

References

[1] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory*, (B. N. Petrov and F.Csáki,eds.), 267–81, Budapest: Akadémia Kiado.

[2] Bai, Z.D., Miao, B. Q. and Rao, C. R. (1990). Estimation of direction of arrival of signals: Asymptotic results. In *Advances in Spectrum Analysis and Array Processing*, Vol.1 (ed. S. Haykin), New York, 327-347.

[3] Bai, Z.D. and Silverstein, J.W. (1998). No eigenvalues outside the support of the limiting spectral distribution of large-dimensional sample covariance matrices. *Ann. Probab.*, **26**, 316–345.

[4] Bai, Z.D. and Silverstein, J.W. (2010). *Spectral Analysis of Large Dimensional Random Matrices, 2nd ed.*. Springer, New York.

[5] Bai, Z.D. and Yao, J. F. (2012). On sample eigenvalues in a generalized spiked population model. *J. Multivariate Anal.*, **106**, 167–177.

[6] Bai, Z.D. and Yin, Y.Q. (1993). Limits of the smallest eigenvalue of a large dimensional sample covariance matrix. *Ann. Probab.*, **21**, pp 1275-1294.

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

[8] Ferré, L. (1995). Selection of components in principal component analysis: A comparison of methods. *Comput. Statist. Data Anal.*, **19**, 669-682.

[9] Fujikoshi, Y. and Sakurai, T. (2015). Some properties of estimation criteria for dimensionality in principal component anlysis. *Hiroshima Statistical Research Group*, *Technical Report*, 15-11.

[10] Fujikoshi, Y., Sakurai, T. and Yanagihara, H. (2014). Consistency of high-dimensional AIC-type and C_p-type criteria in multivariate linear regression. *J. Multivariate Anal.*, **123**, 184–200.

[11] Fujikoshi, Y., Ulyanov, V. V. and Shimizu, R. (2010). *Multivariate Statistics: High-Dimensional and Large-Sample Approximations*. Wiley, Hobeken, N.J.

[12] 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. *J. Multivariate Anal.*, **98**, 2002-2008.

[13] Gunderson, B. K. and Muirhead, R. J. (1997). On estimating the dimensionality in canonical correlation analysis. *J. Multivariate Anal.*, **62**, 121–136.

[14] Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal component analysis, *Ann. Statist.*, **29**, 295–327.

[15] Jolliffe, I. T. (2002). *Principal Component Analysis* (2nd ed.). Springer, New York.

[16] Nishii, R. (1984). Asymptotic properties of criteria for selection of variables in multiple regression. *Ann. Statist.*, **12**, 758–765.

[17] Nishii, R., Bai, Z. D. and Krishnaiah, P. R. (1988). Strong consistency of the information criterion for model selection in multivariate analysis, *Hiroshima Math. J.*, **18**, 451-462.

[18] Rao, C. R. and Rao, M. B. (1998). *Matrix Algebra and Its Applications to Statistics and Economics.* World Scientific.

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

[20] Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.*, **6**, 461–464.

[21] Shibata, R. (1976). Selection of the order of an autoregressive model by Akaike's information criterion. *Biomertrika*, **63**, 117–126.

[22] Silverstein, J. W. (1995). Strong convergence of the empirical distribution of eigenvalues of large dimensional random matrices. *J. Multivariate Anal.*, **54**, 331-339.

[23] Silverstein, J. W. and S. I. Choi (1995). Analysis of the limiting spectral distribution of large dimensional random matrices. *J. Multivariate Anal.*, **54**, 295-309.

[24] Yanagihara, H., Wakaki, H. and Fujikoshi, Y. (2015). A consistency property of AIC for multivariate linear model when the dimension and the sample size are large. *Electronic J. Statist.*, **9**, 869-897.

[25] Yao, J., Zheng, S. and Bai, Z. (2015). *Large Sample Covariance Matrices and High-Dimensional Data Analysis*. Cambridge University Press.