The Annals of Statistics
2018, Vol. 46, No. 3, 1050–1076
https://doi.org/10.1214/17-AOS1577

CONSISTENCY OF AIC AND BIC IN ESTIMATING THE NUMBER OF SIGNIFICANT COMPONENTS IN HIGH-DIMENSIONAL PRINCIPAL COMPONENT ANALYSIS

BY ZHIDONG BAI[1], KWOK PUI CHOI[2] AND YASUNORI FUJIKOSHI[3]

Northeast Normal University, National University of Singapore and Hiroshima University

In this paper, we study the problem of estimating the number of significant components in principal component analysis (PCA), which corresponds to 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 [In *2nd International Symposium on Information Theory* (1973) 267–281, Akadémia Kiado] and Schwarz [*Estimating the dimension of a model* **6** (1978) 461–464] under a high-dimensional asymptotic framework. Using random matrix theory techniques, we derive sufficient conditions for the criterion to be strongly consistent for the case when the dominant population eigenvalues are bounded, and 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, and results show that the sufficient conditions in our theorems are essential.

1. Introduction. Principal component analysis (PCA) is a widely used technique for reducing the dimensionality of data which are 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 $\mathbf{\Sigma}$ of a p-dimensional random vector $\mathbf{y}$. As an approach to determine the dimensionality, we consider a spike covariance structure model proposed by Johnstone (2001) 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 $\lambda_1 = \cdots = \lambda_p = \lambda$.

Received October 2015; revised January 2017.

[1]Supported by NSFC 11571067 and 11471140.

[2]Supported by the Singapore Ministry of Education Academic Research Fund R-155-000-141-112.

[3]Supported by the Ministry of Education, Science, Sports, and Culture, a Grant-in-Aid for Scientific Research (C), #25330038, 2013–2015.

MSC2010 subject classifications. 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.

If M_k is true, we say that the true dimensionality or the true number of significant components is k. The model M_k was used by Bai, Miao and Rao (1990) in the problem of estimating the direction of signals. Spike models find wide applications in A.I. such as face, handwriting and speech recognition, in wireless communication, statistical learning, etc. For further results and applications, see Paul (2007) and Johnstone and Lu (2009) and the references therein. The number, k, in this 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 $\mathbf{y}_1, \ldots, \mathbf{y}_n$ be a random sample of size n from a p-dimensional population with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, and let $\mathbf{S}_n$ be the sample covariance matrix given by

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

where $\bar{\mathbf{y}} = (1/n)\sum_{i=1}^{n}\mathbf{y}_i$. 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, a traditional way is to test $M_0, M_1, \ldots$, sequentially until an M_j is accepted according to certain selection criterion. Here, 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

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

we consider [see, e.g., Fujikoshi, Ulyanov and Shimizu (2010), Fujikoshi and Sakurai (2016a)]

$$\mathrm{AIC}_j = n\log(\ell_{1p}\cdots\ell_{jp}) + n(p-j)\log\bar{\ell}_{jp} + 2d_j + C_{p,n}, \tag{1.3}$$

$$\mathrm{BIC}_j = n\log(\ell_{1p}\cdots\ell_{jp}) + n(p-j)\log\bar{\ell}_{jp} + (\log n)d_j + C_{p,n}, \tag{1.4}$$

where $\ell_{1p} > \cdots > \ell_{pp}$ are the sample eigenvalues of $\mathbf{S}_n$, and for $1 \le j \le p-1$, $\bar{\ell}_{jp}$ 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 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. Instead of AIC_j and BIC_j, it is equivalent to consider

$$\mathrm{A}_j = \frac{1}{n}(\mathrm{AIC}_j - \mathrm{AIC}_{p-1}), \qquad \mathrm{B}_j = \frac{1}{n}(\mathrm{BIC}_j - \mathrm{BIC}_{p-1}). \tag{1.7}$$

Motivated by numerous modern data structure in which $p > n$, we extend our study to cover this situation in Section 3.2. We modify the definition of (1.5) to (3.11), and propose to use the modified criteria $\tilde{\mathrm{A}}_j$ and $\tilde{\mathrm{B}}_j$.

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 example, Shibata (1976), Nishii (1984), Nishii, Bai and Krishnaiah (1988) and Gunderson and Muirhead (1997). Similar selection consistency results in PCA are shown by Zhao, Krishnaiah and Bai (1986) and Fujikoshi and Sakurai (2016a). However, under some high-dimensional models, different consistency behaviour of these model selection criteria are noted. For example, in a regression model, AIC is asymptotically efficient when the true model is infinite. These were discussed by Shibata (1976), Shao (1997), Yang (2005), etc. Kim, Kwon and Choi (2012) studied model selection consistency when the number of regressors exceeds the sample size. Further, in the high-dimensional multivariate model with p-variate such that both p and n tend to infinity, consistency properties of model selection criteria have been shown: 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. Fujikoshi and Sakurai (2016b) discussed consistency properties of model selection criteria for estimating the reduced rank in the multivariate linear model.

In general, methods based on model selection criteria encounter computational problems as the number of candidate models increases. For such cases, we need to use some conventional methods such as stepwise methods. Methods based on model selection criteria in the context of this paper do not suffer a computational problem, since the number of candidate models is p. If we can know the dimensionality k, it is expected that the statistical interpretation will becomes easy. Further, it makes the estimation problem of principal components more efficient. Recently, there are many works on modifying principal components based on the penalized method for high dimension, which are called sparse principal components. See, for example, Jolliffe, Trendafilov and Uddin (2003), Zou, Hastie and Tibshirani (2006) and Johnstone and Lu (2009).

Our purpose is to study the consistency of the estimation criteria AIC and BIC under a high-dimensional asymptotic framework where $p, n \to \infty$ such that $p/n \to c > 0$. 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 want to highlight some results in this paper. For $0 < c < 1$, Theorem 3.1 states that if the largest population eigenvalue remains bounded, AIC is strongly consistent under the "gap condition" (C3), but BIC is not. Furthermore, if the dominant k population eigenvalues tend to infinity, AIC is always strongly consistent regardless of whether the gap condition holds. If the dominant k population eigenvalues tend to infinity with a rate faster than $\log n$, Theorem 3.2 shows that BIC is 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 the 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 RMT and state the two new lemmas. Main results on strong consistency of AIC and BIC are stated and proved in Section 3, first for $c < 1$ (Section 3.1) and then for $c > 1$ (Section 3.2). In Section 4, we present the results of our simulation studies. They show that the gap condition (C3) for $0 < c < 1$ or (C5) for $c > 1$; and the finite fourth moment condition are essential for the selection consistency of AIC. We end our paper with some concluding remarks and conjectures in Section 5. Proofs of Lemmas 2.2 and 2.3 are given in the Appendix.

2. Notation and preliminary lemmas. In this section, we introduce our notation and some of the conditions for our main results to hold. Let $\mathbf{y}_1, \ldots, \mathbf{y}_n$ denote a sample of size n from a p-dimensional population of mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Let $\lambda_1 \geq \cdots \lambda_k > \lambda_{k+1} = \cdots = \lambda_p = 1$ be the eigenvalues of $\boldsymbol{\Sigma}$. Since the context is clear, for the sake of simplifying notation, we suppress the dependence of p on n, and $\boldsymbol{\Sigma}, \lambda_1, \lambda_2 \ldots$, on p. As we are concerned with the sample eigenvalues of the covariance matrix of the $\mathbf{y}_j$'s, (1.2), we can assume $\boldsymbol{\mu} = \mathbf{0}$ without loss of generality. Furthermore, we assume $\mathbf{y}_j = \boldsymbol{\Sigma}^{1/2}\mathbf{x}_j$ for $j = 1, \ldots, n$ where $\mathbf{x}_k = (x_{1k}, \ldots, x_{pk})^\top$, and $\{x_{ij}, i = 1, \ldots, p; j = 1, \ldots, n\}$ is a double array of i.i.d. random variables with mean 0, variance 1 with finite fourth moment. Let $\bar{\mathbf{x}} = \frac{1}{n}\sum_{k=1}^n \mathbf{x}_k$, the covariance matrix of the $\mathbf{x}_j$'s is similarly denoted as

$$\mathbf{S}_{n\mathbf{x}} = \frac{1}{n-1}\sum_{j=1}^{n}(\mathbf{x}_j - \bar{\mathbf{x}})(\mathbf{x}_j - \bar{\mathbf{x}})^\top. \tag{2.1}$$

Assume that

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

(C2) $$\lambda_1 \geq \cdots \geq \lambda_k > \lambda_{k+1} = \cdots = \lambda_p = \lambda = 1.$$

We wish to point out that (C2) implies that the empirical spectral distribution (ESD) of $\mathbf{\Sigma}$ converges weakly to the limiting spectral distribution (LSD) $H(\cdot) = I(1 \leq \cdot)$ as $p \to \infty$.

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

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

where $I_A(\cdot)$ is the indicator function of A. With probability 1, $F_n(x) \xrightarrow{w} F_c(x)$. Here, for $0 < c \leq 1$, F_c is given as

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

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,

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

where a and b are the same as in the case $0 < c \leq 1$. We remark that $\int_a^b f_c(t)\,dt = 1$ or $1/c$ according to $c \leq 1$ or $c > 1$, respectively.

By a result of Silverstein (1995), the ESD of $\mathbf{S}_{n\mathbf{x}}$ also converges weakly to F_c as $n \to \infty$ under condition (C2) where the number of spikes, k, is fixed or $k = o(n)$. 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-quantile of the MP law, that is, $F_c(\mu_\alpha) = \alpha$.

The ith largest eigenvalue, λ_i, of $\mathbf{\Sigma}$ is said to be a distant spiked eigenvalue if $\psi'(\lambda_i) > 0$ where $\psi(x) = x + cx/(x-1)$. Equivalently, λ_i is a distant spiked eigenvalue if $\lambda_i > 1 + \sqrt{c}$.

Our definition above and Lemma 2.1 below are a special case of a more general definition and result in Bai and Yao (2012).

LEMMA 2.1. *Let ℓ_{ip} denote the ith largest eigenvalue of $\mathbf{S}_n$ in (1.2). Suppose that $E(x_{11}^4) < \infty$, (C1) and (C2) hold, and that λ_1 is bounded*:

(i) *If λ_i is a distant spiked eigenvalue, then $\ell_{ip} \xrightarrow{a.s.} \psi(\lambda_i) = \lambda_i + \frac{c\lambda_i}{\lambda_i - 1}$.*

(ii) *If λ_i is not a distant spiked eigenvalue and $i/p \to \alpha$, then $\ell_{ip} \xrightarrow{a.s.} \mu_{1-\alpha}$ and the convergence is uniform in $0 \le \alpha \le 1$.*

A new limiting result for the distant spiked eigenvalues is needed when the spiked population eigenvalues tend to infinity. Intuitively, if $\lambda_j \to \infty$, $\psi(\lambda_j) \sim \lambda_j$. Under the assumption of finite fourth moment, this is indeed the case and is summarized in the following lemma. We remark that both Lemmas 2.1 and 2.2 hold for general LSD function H. For the purpose of this paper, it suffices to state the special case when LSD, $H(\cdot) = I(1 \le \cdot)$. For more details, see Bai and Yao (2012) and the references therein.

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*:

(i) *For any $j \le k$, $\lim_{n\to\infty} \ell_{jp}/\lambda_j = 1$ a.s.*
(ii) *If λ_i is not a distant spiked eigenvalue and $i/p \to \alpha$ as $n \to \infty$, then $\lim_{n\to\infty} \ell_{ip} = \mu_{1-\alpha}$ a.s. and the convergence is uniform in $0 \le \alpha \le 1$.*

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

Lemma 2.3 below is about the monotonicity property of the ratios of the quantiles of the MP law. The proof of this lemma is given in the Appendix.

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

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

where

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

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

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 (i) k is fixed, (ii) $p/n \to c \in (0, 1)$, (iii) $\lambda_i = O(n)$, $i = 1, \ldots, k$ and (iv) $\mathbf{y}$ is normal.

3. Main results. In this section, we derive the consistency of two estimation criteria $\hat{k}_{\mathrm{A}}$ and $\hat{k}_{\mathrm{B}}$ based on AIC and BIC. We shall study the case $c < 1$ first in Section 3.1 and the case $c > 1$ in Section 3.2. We conclude this Section 3 with some discussions and further remarks in Section 3.3.

3.1. *The case $c < 1$.* Throughout this subsection, we assume $0 < c < 1$. 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 λ_i/λ, $i = 1, \ldots, k$.

Recall A_j and B_j in (1.7), we have

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

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

The decision rule of AIC (resp., BIC) selects the model $\hat{k}_{\mathrm{A}}$ (resp., $\hat{k}_{\mathrm{B}}$) by

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

When we are interested in models M_j, $j = 0, 1, \ldots, q-1$, then the criteria are restricted to minimize over $j = 0, 1, \ldots, q-1$.

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

The consistency properties of AIC and BIC criteria for the high-dimensional case are derived based on the log-likelihood for the models. However, unlike the finite dimensional case, they do not rely on the quadratic expansion of the log-likelihood. In fact, the quadratic expansion does not hold for high dimension because the residuals do not tend to zero. For high-dimensional settings, we exploit the fact that the log-likelihood can be written as a function of eigenvalues of the sample covariance matrix so that we may tap into the techniques of random matrix theory to derive the limiting properties of AIC and BIC.

3.1.1. *AIC.* Suppose that λ_1 is finite. Assuming that λ_i's ($1 \le i \le k$) are distant spiked eigenvalues, by Lemma 2.1, we have for $i = 1, \ldots, k$, $\ell_{ip} \xrightarrow{a.s.} \psi_i$, where

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

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

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

It is easy to see that $h(x) > 0$, for $x > m(c)$. We consider the following condition:

$$\psi_k > m(c), \tag{C3}$$

which is equivalent to

$$\gamma(c) \equiv \psi_k - 1 - \log \psi_k - 2c > 0. \tag{3.4}$$

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

THEOREM 3.1. *Suppose the conditions* (C1) *with* $0 < c < 1$, *and* (C2) *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* $\hat{k}_{\mathrm{A}}$ *based on* AIC:

(i) *Suppose that* λ_1 *is bounded. If the gap condition* (C3) *does not hold, then* $\hat{k}_{\mathrm{A}}$ *is not consistent. If the gap condition* (C3) *holds, then* $\hat{k}_{\mathrm{A}}$ *is strongly consistent.*
(ii) *Suppose that* $\lambda_k \to \infty$. *Then* $\hat{k}_{\mathrm{A}}$ *is strongly consistent.*

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

$$\bar{\ell}_{ip} = \frac{1}{(p-i)} \sum_{t=i+1}^{p} \ell_{tp} \xrightarrow{a.s.} \int_a^b t f_c(t)\, dt = 1. \tag{3.5}$$

This implies

$$\begin{aligned} \mathrm{A}_j - \mathrm{A}_k &= \sum_{i=j+1}^{k} (\mathrm{A}_{i-1} - \mathrm{A}_i) \\ &= \sum_{i=j+1}^{k} \left[(p-i+1) \log \left\{ 1 - \frac{1}{p-i+1} (1 - \ell_{ip}/\bar{\ell}_{ip}) \right\} \right. \\ &\qquad \left. + \log \bar{\ell}_{ip} - \log \ell_{ip} - 2(p-i+1)/n \right] \\ &\sim \sum_{i=j+1}^{k} (\psi_i - 1 - \log \psi_i - 2c). \end{aligned} \tag{3.6}$$

If the gap condition (C3) does not hold, or equivalently, $\psi_k - 1 - \log \psi_k - 2c < 0$, then for sufficiently large n, $\mathrm{A}_{k-1} - \mathrm{A}_k < 0$ by (3.6), and hence $\hat{k}_{\mathrm{A}}$ is not consistent.

If the gap condition (C3) holds, that is, $\psi_k > m(c)$ (which implies $\lambda_k > 1 + \sqrt{c}$), then $\psi_i > \psi_k$. Since h is increasing, therefore, the summands in (3.6) are decreasing in i. For $0 \le j < k$, and for sufficiently large n, apply (3.6) to conclude $\mathrm{A}_j - \mathrm{A}_k \ge (k-j)(\psi_k - 1 - \log\psi_k - 2c) > 0$. In other words,

$$\hat{k}_A \ge k \qquad \text{a.s.} \tag{3.7}$$

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

$$\begin{aligned}
\mathrm{A}_j - \mathrm{A}_k &= \sum_{i=k+1}^{j} (\mathrm{A}_i - \mathrm{A}_{i-1}) \\
&= \sum_{i=k+1}^{j} \Bigg[-(p-i+1)\log\Bigg\{1 - \frac{1}{p-i+1}(1 - \ell_{ip}/\bar{\ell}_{ip})\Bigg\} \\
&\qquad - \log\bar{\ell}_{ip} + \log\ell_{ip} + 2(p-i+1)/n \Bigg] \\
&\sim \sum_{i=k+1}^{j} \{(1 - \ell_{ip}/\bar{\ell}_{ip}) + \log(\ell_{ip}/\bar{\ell}_{ip}) + 2c(1 - i/p)\}.
\end{aligned} \tag{3.8}$$

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

$$\begin{aligned}
\mathrm{A}_j - \mathrm{A}_k &\sim (j-k)(1 - b + \log b + 2c) \\
&= (j-k)\{c - 2\sqrt{c} + 2\log(1+\sqrt{c})\} > 0.
\end{aligned}$$

Combining this with (3.7), we complete the proof of (i).

To prove (ii), we first note that, for $k < j = o(p)$, the proof proceeds in the same manner as in the proof of (i) and will not be repeated here. For $j < k$, as in the proof of (i),

$$\mathrm{A}_j - \mathrm{A}_k \sim \sum_{i=j+1}^{k} [\ell_{ip}/\bar{\ell}_{ip} - 1 - \log(\ell_{ip}/\bar{\ell}_{ip}) - 2c].$$

When $\lambda_k \to \infty$, as $\frac{1}{p}\sum_{t=k+1}^{p} \ell_{tp} \sim \int_a^b x f_c(x)\,dx = 1$ and $\ell_{ip} \sim \lambda_i$ for $i \le k$ by Lemma 2.2. Thus, as $n \to \infty$

$$\frac{\ell_{i,p}}{\bar{\ell}_{ip}} \sim \frac{\lambda_i}{\frac{\lambda_{i+1}+\cdots+\lambda_k}{p} + 1} \ge \frac{\lambda_i}{(k-i)\lambda_i/p + 1} \to \infty.$$

So, n large enough, $\mathrm{A}_j - \mathrm{A}_k > 0$. This completes the proof of (ii). $\square$

3.1.2. *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* $0 < c < 1$, *and* (C2) *hold. We have the following consistency results of the estimation criterion* $\hat{k}_{\mathrm{B}}$ *based on* BIC:

(i) *Suppose that* $\lambda_k / \log n \to 0$. *Then* $\hat{k}_{\mathrm{B}}$ *is not consistent.*
(ii) *Suppose that* $\lambda_k / \log n \to \infty$. *Then* $\hat{k}_{\mathrm{B}}$ *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 OF THEOREM 3.2. 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

$$\mathrm{B}_j - \mathrm{B}_k \sim \sum_{i=j+1}^{k} (\psi_i - 1 - \log \psi_i - c \log n). \tag{3.9}$$

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

If $\lambda_k / \log n \to \infty$, then for sufficiently large n, by (3.9),

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

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

Consider $k < j$, analogous to the derivation of (3.8), we have

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

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 $j > 2p/3$,

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

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 (ii), and hence the theorem. □

3.2. *The case $c > 1$.* We consider the case where $p, n \to \infty$ such that $p > n$ and $p/n \to c > 1$. Observe that 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. Since $n < p$, it is not possible to infer the smallest population eigenvalues $\lambda_n, \lambda_{n+1}, \ldots, \lambda_p > 0$, and so in this subsection we assume (C1) with $c > 1$ and (C4) hold where

$$\lambda_{n-1} = \lambda_n = \cdots = \lambda_p = \lambda. \tag{C4}$$

Assumption (C4) is rather natural at least in a high-dimensional PCA setting. Under (C4), we have, for $j = 0, 1, \ldots, n-2$,

$$\widetilde{M}_j : \lambda_j > \lambda_{j+1} = \cdots = \lambda_{n-1} \quad \Leftrightarrow \quad M_j : \lambda_j > \lambda_{j+1} = \cdots = \lambda_p. \tag{3.10}$$

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

$$\bar{\ell}_{jp} := \frac{1}{n-1-j} \sum_{t=j+1}^{n-1} \ell_{tp}, \qquad j = 1, 2, \ldots, n-1. \tag{3.11}$$

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 $\hat{k}_{\tilde{\mathrm{A}}}$ (or $\hat{k}_{\tilde{\mathrm{B}}}$), respectively, by

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

We abbreviate quasi-AIC and quasi-BIC to qAIC and qBIC, respectively.

Finally, as $c > 1$, the gap condition (3.4) is modified to

$$\tilde{\gamma}(c) := \psi_k/c - 1 - \log(\psi_k/c) - 2/c > 0. \tag{C5}$$

We shall first provide some intuition about our proposed criteria before proceeding to the theorems (which say that $\hat{k}_{\tilde{\mathrm{A}}}$ and $\hat{k}_{\tilde{\mathrm{B}}}$ possess similar consistency properties as that of $\hat{k}_{\mathrm{A}}$ and $\hat{k}_{\mathrm{B}}$) and their proofs. The AIC and BIC criteria for case $c = p/n > 1$ cannot be interpreted by the likelihood ratio framework since

the determinants involved in the likelihood ratio statistic are all zeros. Therefore, it is impossible to derive the criteria following the arguments used in papers of Akaike and Schwarz. The criteria proposed above are motivated by the asymptotics of spiked eigenvalues of sample covariance and comparing to the case for $c < 1$.

THEOREM 3.3. *Suppose the conditions* (C1) *with* $c > 1$ *and* (C4) *hold, and that the number of candidate models* $q = o(p)$. *We have the following results on the consistency of the estimation criterion* $\hat{k}_{\tilde{\mathrm{A}}}$ *based on* qAIC:

(i) *Suppose that* λ_1 *is bounded. If the modified gap condition* (C5) *fails,* $\hat{k}_{\tilde{\mathrm{A}}}$ *is not consistent. If the modified gap condition* (C5) *holds,* $\hat{k}_{\tilde{\mathrm{A}}}$ *is strongly consistent.*
(ii) *Suppose that* $\lambda_k \to \infty$. *Then* $\hat{k}_{\tilde{\mathrm{A}}}$ *is strongly consistent.*

THEOREM 3.4. *Suppose the conditions* (C1) *with* $c > 1$ *and* (C4) *hold. We have the following results on the consistency of the estimation criterion* $\hat{k}_{\tilde{\mathrm{B}}}$ *based on* qBIC:

(i) *Suppose that* $\lambda_k / \log n \to 0$. *Then* $\hat{k}_{\tilde{\mathrm{B}}}$ *is not consistent.*
(ii) *Suppose that* $\lambda_k / \log n \to \infty$. *Then* $\hat{k}_{\tilde{\mathrm{B}}}$ *is strongly consistent.*

We shall sketch the proofs of Theorems 3.3 and 3.4 below. For $j < k$, we have

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

When λ_1 is bounded, 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 (C5) is satisfied,

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

When $\lambda_k \to \infty$, 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

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

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 $\tilde{\mathrm{A}}_j - \tilde{\mathrm{A}}_k > 0$. Combining these results, we have proved that $\min_{j \neq k}(\tilde{\mathrm{A}}_j - \tilde{\mathrm{A}}_k) > 0$. Similarly, one can prove that $\min_{j \neq k}(\tilde{\mathrm{B}}_j - \tilde{\mathrm{B}}_k) > 0$, when $\lambda_k / \log n \to \infty$.

3.3. *Discussions and further remarks*. Intuitively, if the largest eigenvalues are not well separated from the support of the MP law, there is no way to identify the true model from all candidate models. Since the penalty term in AIC is fixed, to prevent AIC from under-estimating the true model, one requires the spiked eigenvalues are large enough as compared with the penalty term leading to the gap condition Specifically, for $0 < c < 1$, if the gap condition (C3) does not hold, then according to the Tracy–Widom law there will be a positive probability that the AIC criterion will under-estimate the model. In fact, if $\psi_k \leq m(c)$, then $\psi_k - 1 - \log(\psi_k) - 2c \leq 0$. So $\ell_{kp} < \psi_k$ implies $\ell_{kp} - 1 - \log(\ell_{kp}) - 2c < 0$, that is, the model is under-estimated. By Bai and Yao (2012), we have

$$\begin{aligned} P(\text{AIC is under-estimated}) &\geq P(\ell_{kp} - \psi_k < 0) \\ &= P\big(\sqrt{n}(\ell_{kp} - \psi_k) < 0\big) \quad \to \quad \Phi(0) = \frac{1}{2}. \end{aligned}$$

We provide some intuition why different sufficient conditions are needed for AIC and BIC to achieve consistency. The penalty for BIC tends to ∞ with a rate of $\log n$, thus BIC will surely under-estimate the model unless the spiked eigenvalues tend to infinity at a rate faster than $\log n$. However, the penalty for AIC has a fixed magnitude, so a sufficient condition for the consistency of AIC does not even require the spiked eigenvalues tend to infinity.

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 3.3 holds for $c = 1$.

4. Simulation studies. In our experiments, we define p-variate $\mathbf{y}$ as

$$\mathbf{y} = \mathbf{\Sigma}^{1/2}(x_1, \ldots, x_p)^\top, \tag{4.1}$$

where $\mathbf{\Sigma} = \operatorname{diag}(\widetilde{\lambda}_1, \ldots, \widetilde{\lambda}_k, \widetilde{\lambda}, \ldots, \widetilde{\lambda})$ and $x_1, \ldots, x_p$ are i.i.d. with mean 0 and variance 1. Clearly, the covariance of $\mathbf{y}$ is given by $\mathbf{\Sigma}$.

We conducted a number of simulation studies to examine the effects on the consistency of $\hat{k}_{\mathrm{A}}$ and $\hat{k}_{\mathrm{B}}$ 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.

4.1. *Simulation studies for* $0 < c < 1$. We set $p/n = 1/3$, that is, $c = 1/3$. So $m(c) = 2.636$. For the distribution of x_i's, we consider the following five cases: (i) standard normal distribution (D1); (ii) standardized t distribution with 4 degrees of freedom (D2), that is, $x_i \sim t_4/\sqrt{\mathrm{Var}(t_4)}$; (iii) standardized t distribution with 5 degrees of freedom (D3); (iv) standardized t distribution with 10 degrees of freedom (D4); and (v) standardized chi-square distribution with 3 degrees of freedom (D5).

For the eigenvalues of $\mathbf{\Sigma}$, we considered the following three spectrums (eigenvalues arranged in decreasing order, denoted by $\widetilde{\lambda}_i$'s): (i) $\{30, 20, 13, 5, 3, \ldots, 3\}$ (L1); (ii) $\{30, 22, 16, 10, 3, \ldots, 3\}$ (L2); and (iii) $\{30\alpha_p, 20\alpha_p, 13\alpha_p, 5\alpha_p, 3, \ldots, 3\}$ where $\alpha_p = \sqrt{p/10}$ (L3).

We highlight some salient features of our choices of the distributions and the eigenvalues. The finite fourth moment of x_i's are satisfied in all the distributions above except (D2), in which case, the standardized t distribution with 4 degrees of freedom has only finite moments up to order 3. In L1, the gap condition (C3) fails: $\lambda_4 = 5/3$ and $\psi_4 = 2.5 < m = 2.636$. 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 $\log n$.

In our framework, λ_i is taken to be $\widetilde{\lambda}_i/\widetilde{\lambda}_p$. The true number of significant components in all three cases, L1–L3, is the same: $k = 4$. 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; that is, in our simulation studies,

$$\mathcal{F}_- = \{M_0, M_1, M_2, M_3\}, \qquad \mathcal{F}_* = \{M_4\}, \qquad \mathcal{F}_+ = \{M_5, M_6, \ldots, M_p\}.$$

The selection percentages of $\mathcal{F}_-$, $\mathcal{F}_*$ and $\mathcal{F}_+$ by Monte Carlo simulations with 10^4 repetitions were computed. Since the sum of the three selection percentages is 100, and so for the sake of clarity of the plots we only display the selection percentages of $\mathcal{F}_-$ and $\mathcal{F}_*$. We plot the selection percentages based on AIC criterion (brown dot for $\mathcal{F}_*$, and brown circle for $\mathcal{F}_-$) and BIC criterion (blue solid triangle for $\mathcal{F}_*$, and blue triangle for $\mathcal{F}_-$) on the same graph for easy visual comparison.

We highlight two observations from Figure 1. (i) In D2 case (i.e., the standardized t_4 distribution), $\hat{k}_{\mathrm{A}}$ is not consistent across our choices of eigenvalues, L1–L3 illustrating that 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 D2 case, $\hat{k}_{\mathrm{B}}$ 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

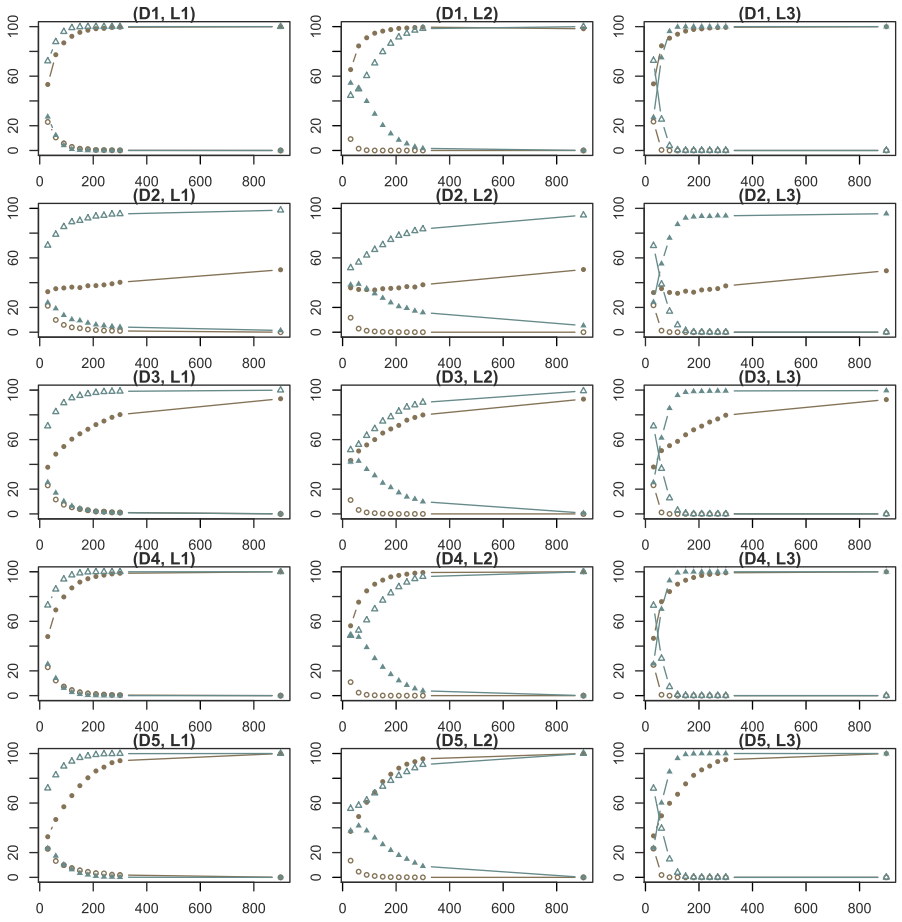

FIG. 1. *Selection percentages under AIC and BIC for $c = 1/3 < 1$ case. Here, the horizontal axes represent the sample sizes, and vertical axes selection percentages. Brown solid circles (resp., brown circles) denote the selection percentages of $\mathcal{F}_*$ (resp., $\mathcal{F}_-$) under AIC decision rule. Similarly, blue solid triangles (resp., blue triangles) for selection percentages of $\mathcal{F}_*$ (resp., $\mathcal{F}_-$) under the BIC decision rule.*

L3, our simulation results suggest that $\hat{k}_{\mathrm{B}}$ is consistent although the finite fourth moment condition fails.

4.2. *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 $\hat{k}_{\tilde{\mathrm{A}}}$ and $\hat{k}_{\tilde{\mathrm{B}}}$ under the population eigenvalues L4: $\{30, 20, 13, 8, 1, \ldots, 1\}$ in addition to L1, L2 and L3 as described in Section 4.1. The variables $x_1, \ldots, x_p$ in (4.1) are chosen to be i.i.d. from the standard normal distribution. In Section 4.1, we set $p/n = c = 1/3$, and so as a natural first step to conduct the simulation studies in this case, we inter-

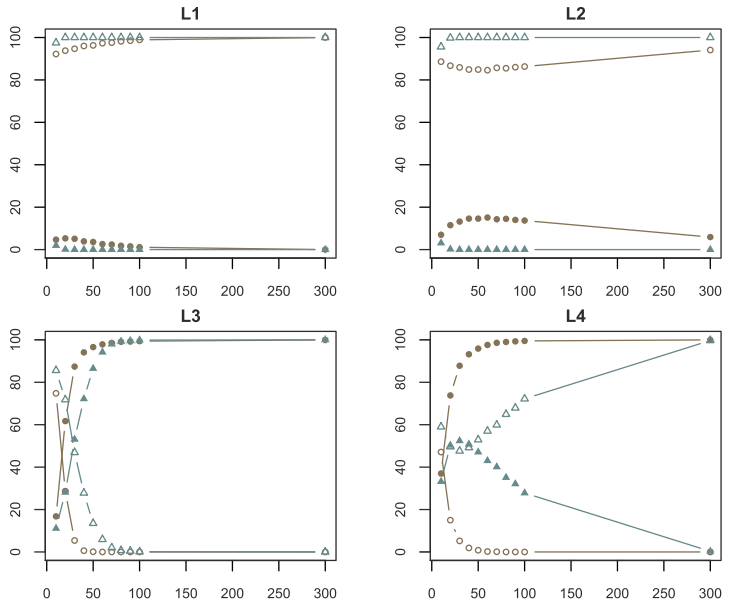

FIG. 2. *Selection percentages under AIC and BIC for $c = 3 > 1$ case. Here, the horizontal axes represent the sample sizes, and vertical axes selection percentages. Brown solid circles (resp., brown circles) denote the selection percentages of $\mathcal{F}_*$ (resp., $\mathcal{F}_-$) under AIC decision rule. Similarly, blue solid triangles (resp., blue triangles) for selection percentages of $\mathcal{F}_*$ (resp., $\mathcal{F}_-$) under BIC decision rule.*

change the roles of p and n. Thus, we consider $p/n = 3$, that is $c = 3$. The true dimension is 4 in all four spectrums.

The simulation results are summarized in the plots as shown in Figure 2. Selection percentages of $\mathcal{F}_*$ and $\mathcal{F}_-$ based on AIC criterion are represented by brown dots and brown circles, respectively. Similarly, we used triangles for the BIC criterion. From the plots, we observe (i) both qAIC and qBIC are inconsistent for spectrums L1 and L2 and both tend to under-estimate the true dimension; (ii) both qAIC and qBIC are consistent for spectrum L3; and (iii) qAIC is consistent, and qBIC is not and tends to under-estimate the true dimension. These observations agree well with the conclusions depicted by Theorems 3.3 and 3.4. Indeed, in L1, L2 and L4, the spiked eigenvalues satisfy the condition in Theorem 3.4(i), we see that simulation results for qBIC in L1, L2 and L4 tally with the conclusion in Theorem 3.4(i). The spiked eigenvalues in L3 satisfies the condition in Theorem 3.4(ii), and the simulation results in L3 and the corresponding conclusion in this theorem agree.

Theorem 3.3 provides explanation for observations (i)–(iii) for qAIC. In L1, λ_4 is not even a distant spiked eigenvalue, and in L2, gap condition (C5) fails. Gap condition (C5) holds in L3 and L4. Likewise, Theorem 3.4 explains these

observations for qBIC: as $n \to \infty$, $\lambda_4/\log n \to 0$ in L1, L2 and L4, whereas $\lambda_4/\log n \to \infty$ in L3.

5. Concluding remarks and conjectures. 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 $\hat{k}_{\mathrm{A}}$ and $\hat{k}_{\mathrm{B}}$ based on AIC_j and BIC_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 $\hat{k}_{\mathrm{A}}$ and $\hat{k}_{\mathrm{B}}$ 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, $\hat{k}_{\mathrm{A}}$ is not consistent, but $\hat{k}_{\mathrm{B}}$ is consistent. When $n < p$, we propose quasi-AIC and quasi-BIC decision rules $\hat{k}_{\tilde{\mathrm{A}}}$ and $\hat{k}_{\tilde{\mathrm{B}}}$. Further, their consistency properties are summarized in Theorems 3.3 and 3.4.

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 (Lemma 2.2); and monotonicity property of the ratios of quantiles of the MP law (Lemma 2.3).

We note that AIC has been proposed as an asymptotic unbiased estimator of the AIC-type risk. Our AIC has been justified under the large-sample asymptotic framework by Fujikoshi and Sakurai (2016b). Results under the high-dimensional framework will be left for future work. The difficulty comes from the fact that the AIC-type risk depends on eigenvectors as well as eigenvalues, and there is no appropriate asymptotic results for eigenvectors under the high-dimensional asymptotic framework. In concluding this paper, we list below some conjectures which arise from this work.

CONJECTURE 1. *Let $\mathbf{S}_n$ be the sample covariance in (1.2) 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*

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

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

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

hold almost surely under some general conditions.

CONJECTURE 2. *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. We shall provide some evidence in Appendix A.5 to support this conjecture.*

CONJECTURE 3. *Theorem 3.3 continues to hold when the candidate models are* $\{M_0, M_1, \ldots, M_{n-2}\}$.

The conjecture below arises from a comment of one of the reviewers.

CONJECTURE 4. *The results concerning AIC remains to hold allowing* $k \to \infty$ *at some rate.*

The proof of the consistency of AIC relies on a well-known result that the spiked eigenvalues tend to known locations, the renowned phase transition theorem. Intuitively, it is not hard to believe that the consistency of AIC should still be true for $k \to \infty$, at least at a certain rate. This will follow if a similar phase transition theorem with infinitely many spiked eigenvalues can be established.

APPENDIX

A.1. Two additional lemmas. We need two additional lemmas to prove Lemma 2.2. Lemma A.1 is a modification of Lemma 2 from Bai and Yin (1993).

LEMMA A.1. *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* $P(|x_{ij}| > t) \leq K P(|x| > t)$ *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*

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

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. □

Recall $\mathbf{S}_{n\mathbf{x}} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top$ in (2.1), and write $\mathbf{\Sigma} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$, where $\mathbf{U} = (\mathbf{U}_1, \mathbf{U}_2) = (\mathbf{u}_1, \ldots, \mathbf{u}_p)$ is a p-dimensional orthogonal matrix with $\mathbf{U}_1$ of dimension $p \times k$ and $\mathbf{\Lambda} = \operatorname{diag}(\lambda_1, \ldots, \lambda_k, 1, \ldots, 1)$ is a diagonal matrix of eigenvalues of $\mathbf{\Sigma}$. Then

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

where $\mathbf{\Lambda}_1 = \operatorname{diag}(\lambda_1, \ldots, \lambda_k)$. Since $\mathbf{U}_2\mathbf{U}_2^\top$ has $p-k$ eigenvalues 1 and k eigenvalues 0, we know that the ESD of $\mathbf{U}_2^\top\mathbf{S}_{n\mathbf{x}}\mathbf{U}_2$ 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.2. *Let $\mathbf{u}_j$, $j = 1, 2, \ldots, k$ denote p-dimensional unit vectors. Under the assumption of Lemma 2.2, we have*

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

PROOF. It suffices to show that $\mathbf{u}_1^\top\mathbf{S}_{n\mathbf{x}}\mathbf{u}_1 \to 1$ a.s. Without loss of generality, we may assume the means of the random entries are 0. Let $\mathbf{u}_1 = (u_1, \ldots, u_p)^\top$, then we have

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

The first term above converges to 0 with probability 1 by Lemma A.1, and the third term converges to 0 with probability 1 by the simple fact that $E(\mathbf{u}_1^\top\bar{\mathbf{x}})^4 = O(n^{-2})$. The second term converges to 0 with probability 1 because

$$\begin{aligned}&E\left|\frac{1}{n-1}\sum_{i=1}^{n}\sum_{j_1\ne j_2}u_{j_1}u_{j_2}x_{ij_1}x_{ij_2}\right|^4 \\ &\quad = \frac{1}{(n-1)^4}\left[\sum_{i=1}^{n}E\Bigg(\sum_{j_1\ne j_2}u_{j_1}u_{j_2}x_{ij_1}x_{ij_2}\Bigg)^4\right.\end{aligned}$$

$$+3\sum_{i_1\neq i_2}E\Bigg(\sum_{j_1\neq j_2}u_{j_1}u_{j_2}x_{i_1j_1}x_{i_1j_2}\Bigg)^2\Bigg(\sum_{j_1\neq j_2}u_{j_1}u_{j_2}x_{i_2j_1}x_{i_2j_2}\Bigg)^2\Bigg]$$

$$\leq\frac{n}{(n-1)^4}\Bigg[24n\sum_{\substack{j_1,j_2,j_3,j_4\\ \text{distinct}}}u_{j_1}^2u_{j_2}^2u_{j_3}^2u_{j_4}^2+24n\sum_{\substack{j_1,j_2,j_3\\ \text{distinct}}}u_{j_1}^3u_{j_2}^3u_{j_3}^2Ex_{11}^3Ex_{11}^3$$

$$+8n\sum_{j_1\neq j_2}u_{j_1}^4u_{j_2}^4Ex_{11}^4Ex_{11}^4+12n(n-1)\Bigg]\leq\frac{K}{n^2},$$

for some constant K. The proof is complete. □

A.2. Proof of (i) of Lemma 2.2. First, we prove that $\liminf\ell_{ip}/\lambda_i\geq 1$ a.s. for $i\leq k$. We note that

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

For any given $\mathbf{v}_1,\ldots,\mathbf{v}_{i-1}$, there exists a vector $\mathbf{u}$ in the linear space spanned by $\mathbf{u}_1,\ldots,\mathbf{u}_i$ which is orthogonal to $\mathbf{v}_1,\ldots,\mathbf{v}_{i-1}$ 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.2, we have

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

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

$$\begin{aligned}\ell_{ip}/\lambda_i&=\lambda_i^{-1}\inf_{\mathbf{v}_1,\ldots,\mathbf{v}_{i-1}}\sup_{\mathbf{u}\perp\mathbf{v}_1,\ldots,\mathbf{v}_{i-1},\|\mathbf{u}\|=1}\mathbf{u}^\top\mathbf{S}_{n\mathbf{x}}\mathbf{u}\\&\leq\lambda_i^{-1}\sup_{\mathbf{u}\perp\mathbf{u}_1,\ldots,\mathbf{u}_{i-1},\|\mathbf{u}\|=1}\mathbf{u}^\top\mathbf{S}_{n\mathbf{x}}\mathbf{u}\\&=\lambda_i^{-1}\sup_{a\leq1}\Big\{a^2\mathbf{u}_i^\top\mathbf{S}_{n\mathbf{x}}\mathbf{u}_i+(1-a^2)\sup_{\mathbf{u}\perp\mathbf{u}_1,\ldots,\mathbf{u}_k,\|\mathbf{u}\|=1}\mathbf{u}^\top\mathbf{S}_{n\mathbf{x}}\mathbf{u}\Big\}\\&\sim\sup_{|a|\leq1}\{a^2+(1-a^2)\lambda_i^{-1}\|\mathbf{U}_2^\top\mathbf{S}_{n\mathbf{x}}\mathbf{U}_2\|\}\\&\sim\sup_{|\sum_{t=i}^k a_t^2|\leq1}\Bigg\{\sum_{t=i}^k a_t^2+\Bigg(1-\sum_{t=i}^k a_t^2\Bigg)\lambda_i^{-1}b\Bigg\}=1,\end{aligned}$$

where we have used the fact that $\|\mathbf{U}_2^\top\mathbf{S}_{n\mathbf{x}}\mathbf{U}_2\|\to b$ in the second equality. To see this is the case, first note that $\mathbf{U}_2\mathbf{U}_2^\top=\mathbf{I}_{p-k}$ implies $\|\mathbf{U}_2\|\leq 1$, and so $\limsup\|\mathbf{U}_2^\top\mathbf{S}_{n\mathbf{x}}\mathbf{U}_2\|\leq\limsup\|\mathbf{S}_{n\mathbf{x}}\|\to b$ a.s. That $\liminf\|\mathbf{U}_2^\top\mathbf{S}_{n\mathbf{x}}\mathbf{U}_2\|\geq b$ is an easy consequence of the convergence of the empirical distribution of the matrix $\mathbf{U}_2^\top\mathbf{S}_{n\mathbf{x}}\mathbf{U}_2$. So $\|\mathbf{U}_2^\top\mathbf{S}_{n\mathbf{x}}\mathbf{U}_2\|\to b$ a.s. holds regardless of $p<n$ or $p\gg n$.

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

A.3. Proof of (ii) of Lemma 2.2. By (A.2), $\ell_{1p}, \ldots, \ell_{pp}$ are also the eigenvalues of

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

Write the eigenvalues of the matrix $\mathbf{U}_2^\top \mathbf{S}_{n\mathbf{x}}\mathbf{U}_2$ as $\tilde{\ell}_{1p}, \ldots, \tilde{\ell}_{p-k,p}$. By Silverstein (1995), the empirical spectral distribution of $\mathbf{U}_2^\top \mathbf{S}_{n\mathbf{x}}\mathbf{U}_2$ 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 the interlace theorem [see, e.g., Section 10.2 in 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 $i \geq k+1$, $\ell_{ip} \xrightarrow{a.s.} \mu_{1-\alpha}$, 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 tth 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 $y(1) = b$. Toward this end, we have

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

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

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

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

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

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

We have

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

$$= G(t) - \frac{[f(G(t))]^2[G(t)]^2 + tf'(G(t))[G(t)]^2 + 2tf(G(t))G(t)}{f(G(t))[f(G(t))G(t) + t]}$$
$$+ \frac{t[G(t)]^2[2f(G(t)) + f'(G(t))G(t)]}{[f(G(t))G(t) + t]^2}.$$

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

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

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

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

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

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

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 tth 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 $y(1) = b$. Toward this end, for $t \in (\bar{c}, 1)$, we have

$$y'(t) = \frac{[G(t) + (t - \bar{c})G'(t)]\int_{\bar{c}}^{t} G(s)\,ds - (t - \bar{c})[G(t)]^2}{(\int_{\bar{c}}^{t} G(s)\,ds)^2}$$

$$= \frac{[f(G(t))G(t) + (t - \bar{c})]\int_{\bar{c}}^{t} G(s)\,ds - (t - \bar{c})f(G(t))[G(t)]^2}{f(G(t))(\int_{\bar{c}}^{t} G(s)\,ds)^2}.$$

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

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

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

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

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

We have

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

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

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

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

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

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.

A.5. Evidences in support of Conjecture 2. Evidence 1. From the proof in Theorem 3.1, indeed we have shown for $k < j < p$,

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

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

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

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

$$I(c, 1) = 1 - \int_0^1 x(t)\,dt + \int_0^1 \log x(t)\,dt + c. \tag{A.9}$$

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

$$\int_0^1 x(t)\,dt = \int_a^b \frac{u f_c(u) F_c(u)}{\int_a^s f_c(s)\,ds}\,du = -\int_a^b f_c(u) \log\Big(\int_a^u s f_c(s)\,ds\Big)\,du; \tag{A.10}$$

and

$$\begin{aligned}\int_0^1 \log x(t)\,dt &= \int_a^b f_c(u) \log\Big(\frac{u F_c(u)}{\int_a^u s f_c(s)\,ds}\Big)\,du \\ &= \int_a^b f_c(u) \log u\,du - 1 + \int_0^1 x(t)\,dt.\end{aligned} \tag{A.11}$$

By (A.9)–(A.11), we have

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

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

$$I(c, \alpha) > 0 \tag{A.12}$$

and Conjecture 2 would be proved. □

PROOF OF (A.12). If $\hat{g}(t)$ has no zero, then (A.12) 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 (A.12) holds trivially. If $\alpha > t_0$, then $I(c, \alpha) > I(c, 1) > 0$. □

Acknowledgments. We thank an anonymous Associate Editor and three anonymous reviewers for careful reading of our manuscript and many helpful comments which improved the presentation of this paper. We also thank Dr. T. Sakurai, Dr. N. H. Tran and Ms. Siqin Zhou for their help in the simulation studies and numerical computation.

REFERENCES

AKAIKE, H. (1973). Information theory and an extension of the maximum likelihood principle. In 2*nd International Symposium on Information Theory* (B. N. Petrov and F. Csáki, eds.) 267–281, Budapest: Akadémia Kiado. MR0483125

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* (S. Haykins, ed.) **2** 327–347. Prentice Hall, Englewood, Cliffs, NJ.

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. MR1617051

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

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

FERRÉ, L. (1995). Selection of components in principal component analysis: A comparison of methods. *Comput. Statist. Data Anal.* **19** 669–682. MR1342614

FUJIKOSHI, Y. and SAKURAI, T. (2016a). Some properties of estimation criteria for dimensionality in principal component anlysis. *Amer. J. Math. Management Sci.* **35** 133–142.

FUJIKOSHI, Y. and SAKURAI, T. (2016b). High-dimensional consistency of rank estimation criteria in multivariate linear model. *J. Multivariate Anal.* **149** 199–212. MR3507324

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. MR3130429

FUJIKOSHI, Y., ULYANOV, V. V. and SHIMIZU, R. (2010). *Multivariate Statistics*: *High-Dimensional and Large-Sample Approximations*. Wiley, Hoboken, NJ. MR2640807

FUJIKOSHI, Y., YAMADA, T., WATANABE, D. and SUGIYAMA, T. (2007). Asymptotic distribution of the LR statistic for equality of the smallest eigenvalues in high-dimensional principal component analysis. *J. Multivariate Anal.* **98** 2002–2008. MR2396951

GUNDERSON, B. K. and MUIRHEAD, R. J. (1997). On estimating the dimensionality in canonical correlation analysis. *J. Multivariate Anal.* **62** 121–136. MR1467877

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

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

JOLLIFFE, I. T. (2002). *Principal Component Analysis*, 2nd ed. Springer, New York. MR2036084

JOLLIFFE, I. T., TRENDAFILOV, N. T. and UDDIN, M. (2003). A modified principal component technique based on the LASSO. *J. Comput. Graph. Statist.* **12** 531–547. MR2002634

KIM, Y., KWON, S. and CHOI, H. (2012). Consistent model selection criteria on high dimensions. *J. Mach. Learn. Res.* **13** 1037–1057. MR2930632

NISHII, R. (1984). Asymptotic properties of criteria for selection of variables in multiple regression. *Ann. Statist.* **12** 758–765. MR0740928

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. MR0991240

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

RAO, C. R. and RAO, M. B. (1998). *Matrix Algebra and Its Applications to Statistics and Econometrics*. World Scientific, River Edge, NJ. MR1660868

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. MR2256563

SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464. MR0468014

SHAO, J. (1997). An asymptotic theory for linear model selection. *Statist. Sinica* **7** 221–264. MR1466682

SHIBATA, R. (1976). Selection of the order of an autoregressive model by Akaike's information criterion. *Biometrika* **63** 117–126. MR0403130

SILVERSTEIN, J. W. (1995). Strong convergence of the empirical distribution of eigenvalues of large-dimensional random matrices. *J. Multivariate Anal.* **55** 331–339. MR1370408

YANAGIHARA, H., WAKAKI, H. and FUJIKOSHI, Y. (2015). A consistency property of the AIC for multivariate linear models when the dimension and the sample size are large. *Electron. J. Stat.* **9** 869–897. MR3338666

YANG, Y. (2005). Can the strengths of AIC and BIC be shared? A conflict between model indentification and regression estimation. *Biometrika* **92** 937–950. MR2234196

ZHAO, L. C., KRISHNAIAH, P. R. and BAI, Z. D. (1986). On detection of the number of signals in presence of white noise. *J. Multivariate Anal.* **20** 1–25. MR0862239

ZOU, H., HASTIE, T. and TIBSHIRANI, R. (2006). Sparse principal component analysis. *J. Comput. Graph. Statist.* **15** 265–286. MR2252527

Z. BAI
KLAS MOE
AND SCHOOL OF MATHEMATICS AND STATISTICS
NORTHEAST NORMAL UNIVERSITY
CHANGCHUN 130024
CHINA
E-MAIL: baizd@nenu.edu.cn

K. P. CHOI
DEPARTMENT OF STATISTICS
AND APPLIED PROBABILITY
NATIONAL UNIVERSITY OF SINGAPORE
SINGAPORE 117546
REPUBLIC OF SINGAPORE
E-MAIL: stackp@nus.edu.sg

Y. FUJIKOSHI
DEPARTMENT OF MATHEMATICS
GRADUATE SCHOOL OF SCIENCE
HIROSHIMA UNIVERSITY
HIGASHI-HIROSHIMA
HIROSHIMA 739-8526
JAPAN
E-MAIL: fujikoshi_y@yahoo.co.jp