Some Properties of Estimation Criteria for Dimensionality in Principal Component Analysis

Yasunori Fujikoshi* and Tetsuro Sakurai**

*Department of Mathematics, Graduate School of Science, Hiroshima University, 1-3-1 Kagamiyama, Higashi Hiroshima, Hiroshima 739-8626, Japan

**Center of General Education, Tokyo University of Science, Suwa, 5000-1 Toyohira, Chino, Nagano 391-0292, Japan

Abstract

Principal component analysis is a method for reduction of dimensionality of data in the form of N observations of p variables. In this paper we consider to estimate the number of the largest characteristic roots of the covariance matrix of p variables, which is called dimensionality, in a covariance structure such that the remainder characteristic roots are the same. Our purpose is to examine properties of the estimation criteria AIC and BIC based on model selection criteria by Akaike (1973) and Schwart (1978). Under large-sample asymptotic framework, we evaluate the bias term as an estimator of AIC-type risk. Further, we note that AIC is not consistent, but BIC is consistent. For high-dimensional case, it is conjectured that there are cases that AIC is consistent, but BIC is not consistent, based on simulation study.

AMS 2000 subject classification: primary 62H12; secondary 62H25

Key Words and Phrases: AIC, BIC, Consistency property, Dimensionality, High-dimensional framework, Large-sample framework, Number of relevant components, Principal component analysis.

1. Introduction

Many methods have been proposed to determine the number of relevant components or dimensionality in principal component analysis (see, e.g. Jolliffe

(2002), Ferré (1995)). It may be noted that the methods should depend on the aims of statistical analysis and the models considered. As an approach for determining the number of relevant components in principal component analysis, we consider a covariance structure model such that

$$M_j : \lambda_1 > \cdots > \lambda_j > \lambda_{j+1} = \cdots = \lambda_p = \lambda, \tag{1.1}$$

where $\lambda_1 \geq \cdots \geq \lambda_p$ are the characteristic roots of the covariance matrix Σ of p-dimensional random vector $\boldsymbol{x}$. If M_j is true, we can say that the number of relevant principal components or the dimensionality in principal component analysis is j.

Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ be a random sample of size $N = n+1$ from a p-dimensional normal population $\mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Based on the sample, we consider to estimate the dimensionality by selecting an appropriate model from the set $\{M_0, M_1, \ldots, M_{p-1}\}$. Especially, we consider two estimation criteria AIC and BIC based on the model selection criteria by Akaike (1973) and Schwarz (1978). Our purpose is to study asymptotic properties of AIC and BIC.

The two criteria are given in Section 2 by evaluating the bias term in the estimation of AIC-type risk. In Section 3, it is pointed that, under large-sample asymptotic framework, AIC is not consistent, but BIC is consistent. In Section 4, high-dimensional properties of the criteria are studied by simulation experiment. It is conjectured that there are cases that AIC is consistent, but BIC is not consistent. In Section 5, we discuss our conclusions. In Appendix we proof an asymptotic result for the bias term in estimation of AIC-type risk.

2. AIC and BIC

In general, AIC for a model M is defined (Akaike (1973)) as

$$\mathrm{AIC} = -2 \log \hat{L} + 2d, \tag{2.1}$$

where $\hat{L}$ is the maximum likelihood under M, and d is the number of independent parameters under M. First we shall obtain AIC for M_j by using (2.1).

Let the likelihood of $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)'$ denote by $L(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and the maximum likelihood estimators of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ under M_j by $\hat{\boldsymbol{\mu}}_j$ and $\hat{\boldsymbol{\Sigma}}_j$, respectively. Then, $\hat{\boldsymbol{\mu}}_j = \bar{\boldsymbol{x}} = (1/N)\sum_{i=1}^{N} \boldsymbol{x}_i$, and we have

$$-2\log L(\hat{\boldsymbol{\mu}}_j, \boldsymbol{\Sigma}) = N\log|\boldsymbol{\Sigma}| + n\mathrm{tr}\boldsymbol{\Sigma}^{-1}\mathbf{S} + pN\log 2\pi,$$

where $\mathbf{S}$ is the sample covariance matrix given by

$$\mathbf{S} = \frac{1}{n}\sum_{i=1}^{N}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})', {}^{q}\, uadn = N - 1.$$

Let $\ell_1 > \ldots > \ell_p$ be the characteristic roots of $\mathbf{S}$ and $\boldsymbol{h}_i, i = 1, \ldots, p$ the corresponding orthonormal characteristic vectors. These are expressed in matrix notation as $\mathbf{L} = \mathrm{diag}(\ell_1, \ldots, \ell_p)$ and $\mathbf{H} = (\boldsymbol{h}_1, \ldots, \boldsymbol{h}_p)$. Then, the maximum likelihood estimator of $\boldsymbol{\Sigma}$ under M_j is given (Anderson (1963)) as

$$\hat{\boldsymbol{\Sigma}}_j = \frac{n}{N}\mathbf{S}_j, \quad \mathbf{S}_j = \mathbf{H}\begin{pmatrix} \mathbf{L}_1 & \mathbf{O} \\ \mathbf{O} & \bar{\ell}_{(p-j)}\mathbf{I}_{p-j} \end{pmatrix}\mathbf{H}', \tag{2.2}$$

where

$$\mathbf{L}_1 = \mathrm{diag}(\ell_1, \ldots, \ell_j), \quad \bar{\ell}_{(p-j)} = \frac{1}{p-j}\sum_{i=j+1}^{p}\ell_i.$$

Therefore, we have

$$\begin{aligned} \mathrm{AIC}_j = {} & N\log \ell_1 \cdots \ell_j + N(p-j)\log \bar{\ell}_{(p-j)} + 2d_j \\ & + N\log\left(\frac{n}{N}\right)^p + Np\left(\log 2\pi + 1\right). \end{aligned} \tag{2.3}$$

Here, d_j is the number of independent parameters under M_j, which is evaluated as follows. Let $\boldsymbol{\gamma}_i$ be the characteristic vector of $\boldsymbol{\Sigma}$ corresponding to the i-th root λ_i such that $\boldsymbol{\gamma}_i$'s are orthonormal. Let

$$\begin{aligned} &\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_p), \quad \boldsymbol{\Lambda}_1 = \mathrm{diag}(\lambda_1, \ldots, \lambda_j), \\ &\boldsymbol{\Gamma} = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_p) = (\boldsymbol{\Gamma}_1\ \boldsymbol{\Gamma}_2), \quad \boldsymbol{\Gamma}_1 = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_j). \end{aligned}$$

Then, we can express $\boldsymbol{\Sigma}$ under M_j as

$$\begin{aligned} \boldsymbol{\Sigma} &= \boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{\Gamma}' \\ &= (\boldsymbol{\Gamma}_1\ \boldsymbol{\Gamma}_2)\begin{pmatrix} \boldsymbol{\Lambda}_1 & \mathbf{O} \\ \mathbf{O} & \lambda\mathbf{I}_{p-j} \end{pmatrix}(\boldsymbol{\Gamma}_1\ \boldsymbol{\Gamma}_2)' \\ &= \boldsymbol{\Gamma}_1(\boldsymbol{\Lambda}_1 - \lambda\mathbf{I}_j)\boldsymbol{\Gamma}_1' + \lambda\mathbf{I}_p. \end{aligned} \tag{2.4}$$

For the derivation of the last expression, we use $\mathbf{\Gamma}_2\mathbf{\Gamma}_2' = \mathbf{I}_p - \mathbf{\Gamma}_1\mathbf{\Gamma}_1'$. Therefore, the dimensionality of $\{\mathbf{\Sigma}, \boldsymbol{\mu}\}$ under M_j is equal to the dimensionality of $\{\mathbf{\Gamma}_1, \mathbf{\Lambda}, \lambda, \boldsymbol{\mu}\}$ which is given by

$$d_j = pj - \frac{1}{2}j(j+1) + j + 1 + p. \tag{2.5}$$

In general, BIC for a model selection M is defined (see Schwarz (1978)) as

$$\text{BIC} = -2\log\hat{L} + (\log n)d. \tag{2.6}$$

Therefore, we can write BIC for M_j as

$$\begin{aligned}\text{BIC}_j = {} & N\log\ell_1\cdots\ell_j + N(p-j)\log\bar{\ell}_{(p-j)} + (\log n)d_j \\ & + N\log\left(\frac{n}{N}\right)^p + Np\left(\log 2\pi + 1\right).\end{aligned} \tag{2.7}$$

Since we select the model with the minimum value of AIC or BIC, instead of AIC_j and BIC_j we may use the following A_j and B_j:

$$\begin{aligned}\text{A}_j &= \text{AIC}_j - \text{AIC}_{p-1}, \quad j = 0, 1, \ldots, p-1 \\ &= -N\left\{\sum_{i=j+1}^{p}\log\ell_i - (p-j)\log\bar{\ell}_{p-j}\right\} - 2q_j, \\ \text{B}_j &= \text{BIC}_j - \text{BIC}_{p-1}, \quad j = 0, 1, \ldots, p-1 \\ &= -N\left\{\sum_{i=j+1}^{p}\log\ell_j - (p-j)\log\bar{\ell}_{p-j}\right\} - (\log n)q_j,\end{aligned}$$

where $q_j = \frac{1}{2}(p-j-1)(p-j+2)$, $\text{A}_{p-1} = 0$ and $\text{B}_{p-1} = 0$. Note that the first part in A_j and B_j

$$T_j = -N\left\{\sum_{i=j+1}^{p}\log\ell_i - (p-j)\log\bar{\ell}_{p-j}\right\}$$

is a likelihood ratio test statistic for testing M_j. Further, the null distribution of T_j is asymptotically distributed (Anderson (1963)) as a chi-square distribution with q_j degrees of freedom under large-sample framework. Based on A_j and B_j, we estimate the dimensionality by

$$\hat{j}_A = \arg\min_j \text{A}_j, \quad \text{and} \quad \hat{j}_B = \arg\min_j \text{B}_j,$$

respectively.

Now we derive the term "$2d_j$" in connection with Akaike-type risk. Let $f(\mathbf{X}; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$ be the the density function of $\mathbf{X}$ under M_j. Then the AIC-type risk of M_j is given as

$$R_A = \mathrm{E}^*_{\mathbf{Y}} \mathrm{E}^*_{\mathbf{Z}} [-2 \log f(\mathbf{Z}; \hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j)], \tag{2.8}$$

which is based on Kullback-Leibler information. Here $\mathbf{Z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_N)'$ has the same distribution as $\mathbf{X}$ and is independent of $\mathbf{X}$, and E^* denotes the expectation to the true model M_*. The true model is assumed that $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ are independently and identically distributed as $\mathrm{N}(\boldsymbol{\mu}_*, \boldsymbol{\Sigma}_*)$. Note that $\mathbf{Z}$ may be regarded a future random matrix for $\mathbf{X}$. Now we estimate R_A by $-2 \log f(\mathbf{X}; \hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j)$. Let us denote the bias term in the estimation by "$-b_A(M_j)$". Then

$$b_A(M_j) = \mathrm{E}^*_{\mathbf{X}} \mathrm{E}^*_{\mathbf{Z}} \left\{ -2 \log f(\mathbf{Z}; \hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j) + 2 \log f(\mathbf{X}; \hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j) \right\}. \tag{2.9}$$

We can write

$$\begin{aligned} b_A(M_j) &= \mathrm{E}^*_{\mathbf{X}} \mathrm{E}^*_{\mathbf{Z}} \left\{ \sum_{i=1}^{N} \mathrm{tr} \hat{\boldsymbol{\Sigma}}_j^{-1} (\boldsymbol{z}_i - \hat{\boldsymbol{\mu}}_j)(\boldsymbol{z}_i - \hat{\boldsymbol{\mu}}_j) - n \mathrm{tr} \hat{\boldsymbol{\Sigma}}_j^{-1} \mathbf{S} \right\} \\ &= N(N+1) n^{-1} \mathrm{E}^*_{\mathbf{X}} \left\{ \mathrm{tr} \mathbf{S}_j^{-1} \boldsymbol{\Sigma}_* \right\} - Np. \end{aligned} \tag{2.10}$$

For the special cases M_{p-1} and M_0, it is possible to obtain their exact expressions. In fact, note that $\mathbf{S}_{p-1} = \mathbf{S}$. We have

$$\begin{aligned} b_A(M_{p-1}) &= N(N+1) \mathrm{E}^*_{\mathbf{X}} (\mathrm{tr} \mathbf{S}^{-1} \boldsymbol{\Sigma}_*) - Np \\ &= \frac{pN(N+1)}{n-p-1} - Np. \end{aligned}$$

For M_0, $\mathbf{S}_0 = p^{-1} \mathrm{tr} \mathbf{S}$. Assume that $M_* \subset M_0$. Then, noting that we may write $\boldsymbol{\Sigma}_* = \sigma_*^2 \mathbf{I}_p$,

$$\begin{aligned} b_A(M_0) &= N(N+1) p \mathrm{E}^*_{\mathbf{X}} (\mathrm{tr} \mathbf{S}^{-1} \boldsymbol{\Sigma}_*) - Np \\ &= \frac{p^2 N(N+1)}{np-2} - Np. \end{aligned}$$

These imply that

$$b_A(M_0) = 2(p+1) + \mathrm{O}(n^{-1}),$$
$$b_A(M_{p-1}) = 2\left\{\frac{1}{2}p(p+1) + p\right\} + \mathrm{O}(n^{-1}).$$

In general, under the assumption $M_* \subset M_j$, it is shown that

$$b_A(M_j) = 2d_j + \mathrm{O}(n^{-1}). \tag{2.11}$$

For the derivation, see Appendix. Therefore, we can propose

$$-2\log f(\mathbf{X}; \hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j) + 2d_j,$$

which is AIC_j in (2.3), as a refinement of a naive estimator $-2\log f(\mathbf{X}; \hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j)$ of R_A.

3. Large-Sample Properties

In this section we examine consistency of the criteria $\hat{j}_A$ and $\hat{j}_B$ under large-sample asymptotic framework. Our assumptions are summarized as follows.

A1; p is fixed, $\quad n \to \infty$.

A2; The minimum model including the true model M_* is M_j.

A3; The characteristic roots of Σ under the true model satisfies

$$\lambda_1 > \cdots > \lambda_j > \lambda_{j+1} = \cdots = \lambda_p = \lambda.$$

Note that $n\mathbf{S}$ is distributed as a Wishart distribution $\mathrm{W}_p(n, \boldsymbol{\Sigma})$. We use the following lemma (Anderson (1963)).

Lemma 1. *Let $\ell_1 > \cdots > \ell_p > 0$ be the ordered characteristic roots of the sample covariance matrix $\mathbf{S}$, based on a normal sample of size $N = n - 1$. Suppose that the characteristic roots λ_i's of $\boldsymbol{\Sigma}$ satisfy* (1.1)*. Consider the transformed characteristic roots defined by*

$$y_i = \sqrt{n}(\ell_i - \lambda_i), \quad i = 1, 2, \ldots, p.$$

Then, under a large-sample A1 *it holds that*

(1) *The limiting distributions of* $y_1, \ldots, y_j$ *and* $\{y_{k+1}, \ldots, y_p\}$ *are independent.*

(2) *For* $i = 1, \ldots, j$, *the limiting distribution of* y_i *is* $\mathrm{N}(0, 2\lambda_i^2)$.

(3) *The limiting joint density of* $y_{j+1}, \ldots, y_p$ *is given by*

$$\frac{K(p-j)}{\lambda^{(p-j)(p-j+1)/2}} e^{-\sum_{i=1}^{p-j} y_{j+i}^2/(4\lambda^2)} \prod_{i<k} (y_{j+i} - y_{j+k}),$$

where $y_{j+1} > \cdots > y_p$ *and* 0 *otherwise, and*

$$1/K(p-j) = 2^{(p-j)(p-j+3)/4} \prod_{i=1}^{p-j} \Gamma[\frac{1}{2}(p-j+1-i)].$$

Theorem 1. *Suppose that the assumptions* A2 *and* A3 *are satisfied. Then, under large-sample asymptotic framework* A1 *it holds that*

(1) $\hat{j}_A$ *is not consistent, and*

$$\lim_{n\to\infty} P(\hat{j}_A = j) = P(Q_{j,j+1} > 0, \ldots, Q_{j,p} > 0), \tag{3.1}$$

where

$$\begin{aligned} 2Q_{jk} = &- \left(y_{j+1}^2 + \cdots + y_k^2\right) - \frac{1}{p-k} \left(y_{k+1} + \cdots + y_p\right)^2 \\ &+ \frac{1}{p-j} \left(y_{j+1} + \cdots + y_p\right)^2 + 2(k-j)(2p+1-k-j), \end{aligned}$$

and the probability of the right side of (3.1) *is evaluated with respect to* $(y_{j+1}, \ldots, y_p)$ *whose density is given by Lemma* 1 (3).

(2) $\hat{j}_B$ *is consistent.*

Proof. Without loss of generality we may assume that $\lambda = 1$, and for $i < j$, λ_i should be regarded as λ_i/λ. For $k > j$,

$$\begin{aligned} \mathrm{A}_k - \mathrm{A}_j = &N \left(\ell_{j+1} + \cdots + \ell_k\right) \\ &+ N(p-k) \log \frac{1}{p-k} \left(\ell_{k+1} + \cdots + \ell_p\right) \\ &- N(p-j) \log \frac{1}{p-j} \left(\ell_{j+1} + \cdots + \ell_p\right) \\ &+ (k-j)(2p+1-k-j). \end{aligned}$$

Further, from Lemma 1, for $k > j$

$$\begin{aligned}\log \ell_k &= \log\left(1 + \frac{1}{\sqrt{n}} y_k\right) \\ &= \frac{1}{\sqrt{n}} y_k - \frac{1}{2}\left(\frac{1}{\sqrt{n}} y_k\right)^2 + \cdots .\end{aligned}$$

Therefore, we can see that

$$\mathrm{A}_k - \mathrm{A}_j = Q_{jk} + \mathrm{O}(n^{-1/2}), \tag{3.2}$$

On the other hand, for $i < j$ we have

$$\lim_{n\to\infty} \frac{1}{N}(\mathrm{A}_i - \mathrm{A}_j) = -\log RM(\lambda_{i+1}, \ldots, \lambda_p) > 0, \tag{3.3}$$

where

$$RM(\lambda_{i+1}, \ldots, \lambda_p) = \frac{\prod_{a=1}^{p-i} \lambda_{i+a}}{\left(\frac{1}{p-i}\sum_{a=1}^{p-i} \lambda_{i+a}\right)^{p-i}}.$$

The results (3.2) and (3.3) implies the first result.

Based on the above results on A_j, it easy to see that

(B1) for $k > j$;

$$\lim_{n\to\infty} (\log n)^{-1} (\mathrm{B}_k - \mathrm{B}_j) = (k-j)(2p+1-k-j) > 0.$$

(B2) for $i < j$;

$$\lim_{n\to\infty} \frac{1}{N} (\mathrm{B}_i - \mathrm{B}_j) = -\log RM(\lambda_{i+1}, \ldots, \lambda_p) > 0.$$

These imply the second result.

□

In order to confirm the results in Theorem 1 and see the speed of convergence, we tried a numerical experiment. In our numerical experiment we define p-variate $\boldsymbol{x}$ as

$$\boldsymbol{x} = \boldsymbol{\Lambda}^{1/2} \begin{pmatrix} z_1 \\ \vdots \\ z_p \end{pmatrix}, \tag{3.4}$$

where $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_j, \lambda, \ldots, \lambda)$ and $z_1, \ldots, z_p$ are independently and identically distributed as $\mathrm{N}(0, 1)$. The population characteristic roots were set as

$$\lambda_1 = 20 > \lambda_2 = 10 > \lambda_3 = \lambda_4 = \lambda_5 = \lambda = 3,$$

and hence the true dimensionality is $j = 2$. Let M_j denote by j simply. Let the minimum model including the true model denote by $\mathcal{F}_*$. Further, let the sets of underspecified and overspecified models by $\mathcal{F}_-$ and $\mathcal{F}_+$, respectively. In our setting,

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

Then, we obtained the probabilities selecting $\mathcal{F}_-, \mathcal{F}_*$ and $\mathcal{F}_+$ by Monte Carlo simulations with 10^4 repetition. The results are given in Table 1.

Table 1.

		A_j			B_j		
n	p	$\mathcal{F}_-$	$\mathcal{F}_*$	$\mathcal{F}_+$	$\mathcal{F}_-$	$\mathcal{F}_*$	$\mathcal{F}_+$
30	5	5.2	77.0	17.8	22.2	74.7	3.2
60	5	0.1	83.1	16.8	2.2	96.8	1.0
90	5	0.0	84.4	15.6	0.1	99.3	0.6
120	5	0.0	83.8	16.2	0.0	99.6	0.4
150	5	0.0	84.7	15.4	0.0	99.8	0.2
180	5	0.0	84.4	15.6	0.0	99.7	0.3
210	5	0.0	84.7	15.3	0.0	99.9	0.2
240	5	0.0	84.6	15.5	0.0	99.9	0.1
270	5	0.0	84.6	15.4	0.0	99.9	0.1
300	5	0.0	85.1	14.9	0.0	99.9	0.1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1000	5	0.0	85.5	14.5	0.0	100.0	0.0
5000	5	0.0	84.3	15.7	0.0	100.0	0.0

From Table 1 we can see the following tendencies.

- As the sample size increases, the probabilities of selecting the true model by $\hat{j}_A$ are increasing, but they do not converge to 1 and select the overspecified models with the probabilities $0.145 \sim 0.155$.

- As the sample size increases, the probabilities of selecting the true model by $\hat{j}_B$ tend to 1.

4. High-Dimensional Properties-Simulation Study

In this section we examine consistency of $\hat{j}_A$ and $\hat{j}_B$ in high-dimensional situations such that $p/n \to c \in (0, 1)$. The p-dimensional variate $\boldsymbol{x}$ was constructed in the same way as in (3.4). For the characteristic roots of $\boldsymbol{\Sigma}$, we considered the following three cases;

Case 1: $\lambda_1 = 30 > \lambda_2 = 20 > \lambda_3 = 13 > \lambda_4 = 8 > \lambda_5 = \cdots = \lambda_p = \lambda = 3$

Case 2: $\lambda_1 = 30 > \lambda_2 = 22 > \lambda_3 = 16 > \lambda_4 = 10 > \lambda_5 = \cdots = \lambda_p = \lambda = 3$

Case 3: $\lambda_1 = 30 \times \sqrt{\dfrac{p}{10}},\ \lambda_2 = 20 \times \sqrt{\dfrac{p}{10}},\ \lambda_3 = 13 \times \sqrt{\dfrac{p}{10}},\ \lambda_4 = 8 \times \sqrt{\dfrac{p}{10}},$
$\lambda_5 = \cdots = \lambda_p = \lambda = 3$

For Case 3, it holds that $\lim_{p\to\infty} \lambda_i = \infty, \quad i = 1, 2, 3$. In our setting,

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

Then, we obtained the probabilities selecting $\mathcal{F}_-$, $\mathcal{F}_*$ and $\mathcal{F}_+$ by Monte Carlo simulations with 10^4 repetitions.

Table 2. Selection probabilities of $\hat{j}_A$

		Case 1			Case 2			Case 3		
n	p	$\mathcal{F}_-$	$\mathcal{F}_*$	$\mathcal{F}_+$	$\mathcal{F}_-$	$\mathcal{F}_*$	$\mathcal{F}_+$	$\mathcal{F}_-$	$\mathcal{F}_*$	$\mathcal{F}_+$
30	10	23.0	53.3	0.6	9.2	65.3	25.5	0.0	71.1	28.9
60	20	10.5	77.3	0.0	1.5	84.4	14.1	0.0	84.4	15.6
90	30	5.7	86.9	0.0	0.2	91.0	8.8	0.0	91.1	8.9
120	40	2.9	92.2	0.0	0.0	94.7	5.3	0.0	94.2	5.8
150	50	1.4	95.4	0.0	0.0	96.4	3.6	0.0	96.0	4.0
180	60	1.0	97.1	0.0	0.0	97.6	2.4	0.0	97.2	2.8
210	70	0.4	98.3	0.0	0.0	98.6	1.4	0.0	98.4	1.6
240	80	0.3	98.6	0.0	0.0	98.9	1.1	0.0	98.9	1.1
270	90	0.2	99.3	0.0	0.0	99.3	0.7	0.0	99.2	0.8
300	100	0.1	99.5	0.0	0.0	99.6	0.4	0.0	99.6	0.4

Table 3. Selection probabilities of $\hat{j}_B$

		Case 1			Case 2			Case 3		
n	p	$\mathcal{F}_-$	$\mathcal{F}_*$	$\mathcal{F}_+$	$\mathcal{F}_-$	$\mathcal{F}_*$	$\mathcal{F}_+$	$\mathcal{F}_-$	$\mathcal{F}_*$	$\mathcal{F}_+$
30	10	72.2	27.2	0.6	44.3	54.5	1.2	0.2	97.7	2.2
60	20	87.7	12.3	0.0	49.7	50.4	0.0	0.0	100.0	0.0
90	30	95.9	4.1	0.0	60.3	39.8	0.0	0.0	100.0	0.0
120	40	99.0	1.0	0.0	70.5	29.5	0.0	0.0	100.0	0.0
150	50	99.7	0.3	0.0	79.6	20.4	0.0	0.0	100.0	0.0
180	60	100.0	0.0	0.0	86.4	13.6	0.0	0.0	100.0	0.0
210	70	100.0	0.0	0.0	91.5	8.5	0.0	0.0	100.0	0.0
240	80	100.0	0.0	0.0	94.5	5.5	0.0	0.0	100.0	0.0
270	90	100.0	0.0	0.0	97.0	3.0	0.0	0.0	100.0	0.0
300	100	100.0	0.0	0.0	98.4	1.7	0.0	0.0	100.0	0.0

From Tables 2 and 3 we can see the following tendencies.

- $\hat{j}_A$ is consistent for all the three cases.
- $\hat{j}_B$ is consistent for Case 3, but it is not consistent for Cases 1 and 2 and select underspecified models.

5. Concluding Remarks

In this paper we consider two estimation criteria $\hat{j}_A$ and $\hat{j}_B$ based on AIC_j and BIC_j, which are equivalent to A_j and B_j, for estimating the number of different characteristic roots in covariance structure model in (1.1). Under large-sample asymptotic framework, it was shown that the AIC_j is an asymptotic unbiased estimator of the AIC-type risk R_A defined by (2.8). Further, it was shown that $\hat{j}_j$ is not consistent, but $\hat{j}_B$ is consistent, as in autoregressive model, linear regression model and discriminant analysis (see Shibat (1976), Nishii (1984), Fujikoshi (1983), Fujikoshi et al. (2010), etc.). This property was confirmed by theoretical arguments as well as numerical experiments. Next we study asymptotic behaviors of $\hat{j}_A$ and $\hat{j}_B$ in high-dimensional asymptotic framework such that $p/n \to c \in (0, 1)$ by numerical experiments. It was pointed that there are cases that A is consistent, but

B is not consistent, and both are consistent. It is hoped that these high-dimensional properties are theoretically proved under nonnormality as well as normality.

Appendix

A Derivation of an Asymptotic Result (2.11) for the Bias Term (2.9)

In this section we give an outline of deriving an asymptotic result (2.11) for the bias term (2.9). More precisely, our purpose is to give an outline of deriving

$$\mathrm{E}(\mathrm{tr}\mathbf{S}_j^{-1}\boldsymbol{\Sigma}_*) = p + \frac{2}{n}(d_j - 2) + \mathrm{O}(n^{-2}), \tag{A1}$$

under $M_* \subset M_j$, where d_j and $\mathbf{S}_j$ are given by (2.5) and (2.2), respectively. For a notational simplicity, we express $\boldsymbol{\Sigma}_*$ as $\boldsymbol{\Sigma}$ in (2.4). In following, we assume that $\lambda_1 > \cdots > \lambda_j > \lambda_{j+1} = \cdots = \lambda_p = \lambda$. Using

$$\begin{aligned}
\mathrm{tr}\mathbf{S}_j^{-1}\boldsymbol{\Sigma} &= \mathrm{tr}\mathbf{H}\mathbf{L}^{-1}\mathbf{H}'\boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{\Gamma}' \\
&= \mathrm{tr}\mathbf{H}\begin{pmatrix} \mathbf{L}_1 & O \\ O & \bar{\ell}_{(p-j)}\mathbf{I}_{p-j} \end{pmatrix}\mathbf{H}'\boldsymbol{\Gamma}\begin{pmatrix} \boldsymbol{\Lambda}_1 & O \\ O & \lambda\mathbf{I}_{p-j} \end{pmatrix}\boldsymbol{\Gamma}',
\end{aligned}$$

we have

$$\begin{aligned}
\mathrm{tr}\mathbf{S}_j^{-1}\boldsymbol{\Sigma} = {} & \mathrm{tr}\mathbf{L}_1^{-1}\mathbf{G}'(\boldsymbol{\Lambda}_1 - \lambda\mathbf{I}_j)\mathbf{G} - \bar{\ell}_{(p-j)}^{-1}\mathrm{tr}\mathbf{G}'(\boldsymbol{\Lambda}_1 - \lambda\mathbf{I}_j)\mathbf{G} \\
& + \lambda\mathrm{tr}\mathbf{L}_1^{-1} + \bar{\ell}_{(p-j)}^{-1}\mathrm{tr}(\boldsymbol{\Lambda}_1 - \lambda\mathbf{I}_j) + \bar{\ell}_{(p-j)}^{-1}(p-j)\lambda,
\end{aligned} \tag{A2}$$

where $\mathbf{G} = \boldsymbol{\Gamma}_1'\mathbf{H}_1$. Let $\tilde{\mathbf{S}} = \boldsymbol{\Gamma}'\mathbf{S}\boldsymbol{\Gamma}$, and

$$\tilde{\mathbf{S}} = \boldsymbol{\Lambda} + \frac{1}{\sqrt{n}}\mathbf{V}, \quad \mathbf{V} = (v_{\alpha\beta}).$$

Then, letting $\mathbf{F} = \boldsymbol{\Gamma}'\mathbf{H}$, we have

$$\tilde{\mathbf{S}}\mathbf{F} = \mathbf{F}\mathbf{L}.$$

We use perturbation expansions of the characteristic roots and vectors of $\tilde{\mathbf{S}}$. For the results, see, e.g., Siotani et al. (1985). Since the characteristic roots of $\tilde{\mathbf{S}}$ are the same as the ones of $\mathbf{S}$, we have, for $a = 1, \ldots, j$,

$$\ell_a = \lambda_a + \frac{1}{\sqrt{n}} v_{aa} + \frac{1}{n} \sum_{b \neq a}^{p} \lambda_{ab} v_{ab}^2 + \cdots , \tag{A3}$$

where $\lambda_{ab} = 1/(\lambda_a - \lambda_b)$. Note that $\mathbf{G}$ is a submatrix of $\mathbf{F}$, and the columns of $\mathbf{F}$ are the characteristic vectors of $\tilde{\mathbf{S}}$. Therefore, we can expand $\mathbf{G}$ as

$$\mathbf{G} = \mathbf{I}_j + \frac{1}{\sqrt{n}} \mathbf{G}^{(1)} + \frac{1}{n} \mathbf{G}^{(2)} + \cdots , \tag{A4}$$

where $\mathbf{G}^{(1)} = (g_{ab}^{1)}),\ \mathbf{G}^{(2)} = (g_{ab}^{(2)})$ and

$$\begin{aligned}
g_{aa}^{(1)} &= 0, \\
g_{ja}^{(1)} &= -\lambda_{ja} v_{ja},\ j \neq a, \\
g_{aa}^{(2)} &= -\frac{1}{2} \sum_{b \neq a} \lambda_{ab}^2 v_{ab}^2, \\
g_{ja}^{(2)} &= -\lambda_{ja} \left[\lambda_{ja} v_{ja} v_{aa} + \sum_{b \neq a} \lambda_{ab} v_{jb} v_{ba} \right],\ j \neq a.
\end{aligned}$$

Substituting (A3) and (A4) into (A2) we can get (2.11), after much computation. The computations can be carried out by using $\mathrm{E}(v_{ab}) = 0$ and

$$\mathrm{E}(v_{ab} v_{cd}) = \begin{cases} \lambda_a^2, & a = b = c = d, \\ \lambda_a \lambda_b, & \{a, b\} = \{c, d\}, a \neq b, \\ 0, & \text{others.} \end{cases}$$

Acknowledgements

The first author's research is partially supported by the Ministry of Education, Science, Sports, and Culture, a Grant-in-Aid for Scientific Research (C), #25330038, 2013-2015.

References

[1] Akaike, H. (1973). Informaiton 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] Anderson, T.W. (1963). Asymptotic theory for principal components. *Ann. Math. Statist.*, **34**, 122–148.

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

[4] Fujikoshi, Y. (1983). A criterion for variable selection in multiple discriminant analysis. *Hiroshima Math. J.*, **13**, 203–214.

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

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

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

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

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

[10] Siotani, M., Hayakawa, T. and Fujikoshi, Y.(1985). *Modern Multivariate Statistical Analysis: A Graduate Course and Aandbook*. American Sciences Press, Columbus, Ohio.