A Consistency Property of the AIC for Multivariate Linear Models When the Dimension and the Sample Size are Large

(Last Modified: April 1, 2012)

Hirokazu YANAGIHARA[1], Hirofumi WAKAKI[2] AND Yasunori FUJIKOSHI[3]

Department of Mathematics, Graduate School of Science, Hiroshima University

1-3-1 Kagamiyama, Higashi-Hiroshima, Hiroshima 739-8626, Japan

Abstract

It is common knowledge that the Akaike's information criterion (AIC) is not a consistent model selection criterion. This inconsistency property has been confirmed from an asymptotic selection probability evaluated from a large-sample asymptotic framework. However, when a high-dimensional asymptotic framework, such that the dimension of the response variables and the sample size are approaching ∞, is used for evaluating the selection probability, we can prove a consistency property of the AIC for selecting variables in multivariate linear models. This means that the probability of selecting the true model by the AIC goes to 1 as the sample size and the dimension simultaneously approach ∞. The consistency property is also checked numerically by conducting a Monte Carlo simulation.

AMS 2010 *subject classifications*: Primary 62J05; Secondary 62E20.

Key words: AIC, Bias-corrected AIC, BIC, Consistent AIC, High-dimensional asymptotic framework, High-dimensional data, Multivariate linear model, Selection probability, Variable selection.

1. Introduction

Let $\boldsymbol{Y}$ be an $n \times p$ observation matrix of p response variables, and let $\boldsymbol{X}$ be an $n \times k$ observation matrix of full rank k, where k is the number of nonstochastic explanatory

[1]Corresponding author, *E-mail*: yanagi@math.sci.hiroshima-u.ac.jp. The first author's research is partially supported by the Ministry of Education, Science, Sports, and Culture, a Grant-in-Aid for Challenging Exploratory Research, #22650058, 2010-2012.

[2]The second author's research is partially supported by the Ministry of Education, Science, Sports, and Culture, a Grant-in-Aid for Scientific Research (C), #21500278, 2009-2011.

[3]The third author's research is partially supported by the Ministry of Education, Science, Sports, and Culture, a Grant-in-Aid for Scientific Research (C), #22500259, 2010-2012.

variables, n is the sample size, and it is assumed that $n - p - k - 1 > 0$. In order to ensure the possibility of estimating the model, we also assume that $\mathrm{rank}(\boldsymbol{X}) = k$ for all $n \geq k$. Suppose that j denotes a subset of $\omega = \{1, \ldots, k\}$ containing k_j elements, and $\boldsymbol{X}_j$ denotes the $n \times k_j$ matrix consisting of the columns of $\boldsymbol{X}$ indexed by the elements of j. For example, if $j = \{1, 2, 4\}$, then $\boldsymbol{X}_j$ consists of the first, second, and fourth columns of $\boldsymbol{X}$. Also, in general, we will let k_A denote the number of elements of a set A, i.e., $k_A = \#(A)$. Of course, it holds that $\boldsymbol{X}_\omega = \boldsymbol{X}$ and $k_\omega = k$. We then consider the following candidate model with k_j explanatory variables:

$$\boldsymbol{Y} \sim N_{n \times p}(\boldsymbol{X}_j \boldsymbol{\Theta}_j, \boldsymbol{\Sigma}_j \otimes \boldsymbol{I}_n), \tag{1.1}$$

where $\boldsymbol{\Theta}_j$ is a $k_j \times p$ unknown matrix of regression coefficients, and $\boldsymbol{\Sigma}_j$ is a $p \times p$ unknown covariance matrix. We call the model with $\boldsymbol{X}_\omega$ (namely $\boldsymbol{X}$) the full model. We will assume that the data are generated from the following true model:

$$\boldsymbol{Y} \sim N_{n \times p}(\boldsymbol{X}_{j_*} \boldsymbol{\Theta}_*, \boldsymbol{\Sigma}_* \otimes \boldsymbol{I}_n), \tag{1.2}$$

where j_* is a set of integers indicating the subset of explanatory variables in the true model. Henceforth, for simplicity, we represent $\boldsymbol{X}_{j_*}$ and k_{j_*} as $\boldsymbol{X}_*$ and k_*, respectively.

The multivariate linear regression model in (1.1) is one of basic models of multivariate analysis. This model is introduced in many multivariate statistical textbooks (see, e.g., Srivastava 2002, chap. 9; Timm 2002, chap. 4), and even now is widely used in chemometrics, engineering, econometrics, psychometrics, and many other fields, for the predication of multiple responses to a set of explanatory variables (see, e.g., Yoshimoto, Yanagihara, and Ninomiya 2005; Dien et al. 2006; Saxén and Sundell 2006; Sárbu et al. 2008). Since it is important to specify the factors affecting response variables in regression analysis, searching for the optimal subset j is essential.

The Akaike's information criterion (AIC), proposed by Akaike (1973, 1974), is widely used for selecting the best model. In the case of regression analysis, the best model for a subset of explanatory variables is chosen. The AIC was proposed as an asymptotic unbiased estimator of the risk function assessed by the expected Kullback-Leibler (KL) loss (Kullback and Leibler 1951) under the assumption that the candidate model includes the true model. One purpose of model selection using the AIC is to choose a model that makes the risk function small. For that purpose, using the AIC for model selection will be asymptotically efficient when the true model is infinite (Shibata 1980; Shao 1997;

Yang 2005). The Bayesian information criterion (BIC) proposed by Schwarz (1978) and a consistent AIC proposed by Bozdogan (1987) are also widely used for model selection. It is a well-known fact that, when the true model is included in a set of the candidate models, these two criteria are consistent in model selection, although the AIC is not. When using the AIC in model selection, this inconsistency property sometimes becomes a target for criticism, although the purpose of the AIC is not to choose the true model. The inconsistency property of the AIC is confirmed from the asymptotic probability of selecting the model, which is evaluated from a large-sample asymptotic framework that represents an ordinary asymptotic procedure (Shibata 1976; Nishii 1984; Fujikoshi 1983, 1985). In the case of multivariate linear models, although there are many bias-corrected AICs for the risk function (see, e.g., Bedrick and Tsai 1994; Fujikoshi and Satoh 1997; Fujikoshi, Yanagihara, and Wakaki 2005; Yanagihara 2006; and Yanagihara, Kamo, and Tonda 2011), the bias-corrected AIC is still not consistent for model selection.

However, there is a possibility that the AIC can acquire a consistency property when another asymptotic framework is used for evaluating the asymptotic probability of selecting the true model. In fact, in this paper we will prove that a selection method using the AIC is consistent for selecting variables in multivariate linear models under a high-dimensional asymptotic framework. More precisely, we show that the probability of selecting the true model by the AIC goes to 1 as the sample size and the dimension of the response variables simultaneously approach ∞ under the condition that $c_{n,p} = p/n \to c_0 \in [0, 1)$. Furthermore, we will also prove that a selection using the bias-corrected AIC, as proposed by Bedrick and Tsai (1994), satisfies the consistency in a wider range than that using the AIC. We find that variable selections using the BIC and the consistent AIC do not become consistent when $c_{n,p} \to c_0 \in (0, 1)$. In this paper, $\lim_{c_{n,p} \to c_0}$ means a limit as $(n, p) \to \infty$ simultaneously under the condition that $c_{n,p} \to c_0$. We assume that p is not constant in the high-dimensional asymptotic framework.

In this paper, $o(x)$, $O(x)$, $o_p(x)$, and $O_p(x)$ used in a vector or matrix mean that the orders of all the elements in that vector or matrix are $o(x)$, $O(x)$, $o_p(x)$, and $O_p(x)$, respectively. Furthermore, the notations o, O, o_p, and O_p indicate the orders as $n \to \infty$ when the large-sample asymptotic framework is considered. Meanwhile, those are the orders as $c_{n,p} \to c_0$ when the high-dimensional asymptotic framework is used.

The present paper is organized as follows: In Section 2, we present the necessary notation for evaluating a selection probability. In Section 3, the asymptotic probability

of selecting the true model is calculated under a high-dimensional asymptotic framework. In Section 4, we verify the adequacy of our claim by conducting numerical experiments. In Section 5, we discuss our conclusions. Technical details are provided in Appendix.

2. Preliminaries

In this section, we present and discuss the notation that we used for evaluating the selection probability. First, we describe several classes of j that express subsets of $\boldsymbol{X}$ in the candidate model. Let $\mathcal{J}$ be a set of candidate models denoted by $\mathcal{J} = \{j_1, \ldots, j_m\}$. We then separate $\mathcal{J}$ into two sets, one of which is a set of overspecified models, candidate models that include the true model, i.e., $\mathcal{J}_+ = \{j \in \mathcal{J} | j_* \subseteq j\}$, and the other is a set of underspecified models that are not the overspecified models, i.e., $\mathcal{J}_- = \mathcal{J}_+^c \cap \mathcal{J}$. Thus, the true model j_* can be regarded as the smallest overspecified model. We use the same terminologies, "overspecified model" and "underspecified model," as were used by Fujikoshi and Satoh (1997).

Estimations for the unknown parameters $\boldsymbol{\Theta}_j$ and $\boldsymbol{\Sigma}_j$ in the model (1.1) are carried out by the maximum likelihood estimation, i.e., $\boldsymbol{\Theta}_j$ and $\boldsymbol{\Sigma}_j$ are estimated by

$$\hat{\boldsymbol{\Theta}}_j = (\boldsymbol{X}_j'\boldsymbol{X}_j)^{-1}\boldsymbol{X}_j'\boldsymbol{Y}, \quad \hat{\boldsymbol{\Sigma}}_j = \frac{1}{n}\boldsymbol{Y}'(\boldsymbol{I}_n - \boldsymbol{P}_j)\boldsymbol{Y},$$

where $\boldsymbol{P}_j$ is the projection matrix to the subspace spanned by the columns of $\boldsymbol{X}_j$, i.e., $\boldsymbol{P}_j = \boldsymbol{X}_j(\boldsymbol{X}_j'\boldsymbol{X}_j)^{-1}\boldsymbol{X}_j'$. Then, the AIC and the bias-corrected AIC ($\mathrm{AIC_c}$, Bedrick and Tsai 1994) in the model (1.1) are defined by

$$\mathrm{AIC}(j) = n\log|\hat{\boldsymbol{\Sigma}}_j| + np(\log 2\pi + 1) + 2\left\{k_j p + \frac{1}{2}p(p+1)\right\}, \tag{2.1}$$

$$\mathrm{AIC_c}(j) = n\log|\hat{\boldsymbol{\Sigma}}_j| + np(\log 2\pi + 1) + \frac{2n}{n - k_j - p - 1}\left\{k_j p + \frac{1}{2}p(p+1)\right\}. \tag{2.2}$$

When $p = 1$, the $\mathrm{AIC_c}$ in (2.2) coincides with the bias-corrected AIC proposed by Sugiura (1978). Davies, Neath, and Cavanaugh (2006) showed that Sugiura's bias-corrected AIC is a uniformly minimum-variance unbiased estimator (UMVUE) of the risk function consisting of the expected KL loss when the candidate model includes the true model. By extending the result to the multivariate case, this property can be proved even when $p > 1$. The detailed proof is omitted because it can be obtained from the Lehman-Scheffé theorem and the fact that $\hat{\boldsymbol{\Theta}}_j$ and $\hat{\boldsymbol{\Sigma}}_j$ are complete sufficient statistics. Complete efficiencies of $\hat{\boldsymbol{\Theta}}_j$ and $\hat{\boldsymbol{\Sigma}}_j$ can be derived by slightly modifying the results of Siotani, Hayakawa,

and Fujikoshi (1985, pp. 18-20). This property indicates that, for all the overspecified models, the $\mathrm{AIC_c}$ is better than the AIC at esimating the risk function. On the other hand, the BIC proposed by Schwarz (1978) and the consistent AIC (CAIC, Bozdogan 1987), are also well-known information criteria for model selection. The BIC and the CAIC in the model (1.1) are defined by

$$\mathrm{BIC}(j) = n \log |\hat{\boldsymbol{\Sigma}}_j| + np(\log 2\pi + 1) + \left\{ k_j p + \frac{1}{2} p(p+1) \right\} \log n, \tag{2.3}$$

$$\mathrm{CAIC}(j) = n \log |\hat{\boldsymbol{\Sigma}}_j| + np(\log 2\pi + 1) + \left\{ k_j p + \frac{1}{2} p(p+1) \right\} (1 + \log n). \tag{2.4}$$

Four information criteria are defined, each one by adding a penalty term based on the complexity of the model to -2 times the maximum log-likelihood of the model. Thus, each criterion is specified by an individual penalty term. The best subsets of ω, chosen by minimizing the AIC, the $\mathrm{AIC_c}$, the BIC, and the CAIC, are written as

$$\hat{j}_{\mathrm{a}} = \arg\min_{j \in \mathcal{J}} \mathrm{AIC}(j), \quad \hat{j}_{\mathrm{c}} = \arg\min_{j \in \mathcal{J}} \mathrm{AIC_c}(j),$$
$$\hat{j}_{\mathrm{b}} = \arg\min_{j \in \mathcal{J}} \mathrm{BIC}(j), \quad \hat{j}_{\mathrm{o}} = \arg\min_{j \in \mathcal{J}} \mathrm{CAIC}(j).$$

Next, we deal with a noncentrality matrix defined by

$$\boldsymbol{\Sigma}_*^{-1/2} \boldsymbol{\Theta}_*' \boldsymbol{X}_*' (\boldsymbol{I}_n - \boldsymbol{P}_j) \boldsymbol{X}_* \boldsymbol{\Theta}_* \boldsymbol{\Sigma}_*^{-1/2}.$$

In order to decompose the noncentrality matrix, the minimum overspecified model including j is prepared as

$$j_+ = j \cup j_*, \ (j \in \mathcal{J}). \tag{2.5}$$

If j_* is arranged as $j_* = \{\{j_* \cap j\}, \{j_* \cap j^c\}\}$, $(\boldsymbol{I}_n - \boldsymbol{P}_j)\boldsymbol{X}_* = (\boldsymbol{O}_{n,k_{j_* \cap j}}, \boldsymbol{X}_{j_* \cap j^c})$ is satisfied, where $\boldsymbol{O}_{k,p}$ is a $k \times p$ matrix of zeros. It is easy to see that $\boldsymbol{X}_{j_* \cap j^c}$ is a full column rank matrix because it is assumed that $\boldsymbol{X}$ is a full column rank matrix. Hence, the rank of $\boldsymbol{X}_*'(\boldsymbol{I}_n - \boldsymbol{P}_j)\boldsymbol{X}_*$ is calculated as

$$\mathrm{rank}(\boldsymbol{X}_*'(\boldsymbol{I}_n - \boldsymbol{P}_j)\boldsymbol{X}_*) = k_{j_* \cap j^c} = k_{j_+} - k_j < k_*, \ (^\forall j \in \mathcal{J}_-).$$

Let the rank of the noncentrality matrix be denoted by γ_j, and let us assume that it is independent of n and p. From the inequality $\mathrm{rank}(\boldsymbol{\Theta}_* \boldsymbol{\Sigma}_*^{-1} \boldsymbol{\Theta}_*') \leq \min\{p, k_*\}$ and a knowledge of an elementary linear algebra, we can see that

$$\gamma_j \leq \min\{\mathrm{rank}(\boldsymbol{X}_*'(\boldsymbol{I}_n - \boldsymbol{P}_j)\boldsymbol{X}_*), \mathrm{rank}(\boldsymbol{\Theta}_* \boldsymbol{\Sigma}_*^{-1} \boldsymbol{\Theta}_*')\} \leq \min\{p, k_{j_+} - k_j\}.$$

It should be kept in mind that $\gamma_j = k_{j_+} - k_j$ if $\boldsymbol{\Theta}_*\boldsymbol{\Sigma}_*^{-1}\boldsymbol{\Theta}_*'$ is a full-rank matrix. Since the noncentrality matrix is a positive semidefinite matrix, and its rank is γ_j, it is decomposed as

$$\boldsymbol{\Sigma}_*^{-1/2}\boldsymbol{\Theta}_*'\boldsymbol{X}_*'(\boldsymbol{I}_n - \boldsymbol{P}_j)\boldsymbol{X}_*\boldsymbol{\Theta}_*\boldsymbol{\Sigma}_*^{-1/2} = \boldsymbol{\Gamma}_j\boldsymbol{\Gamma}_j', \tag{2.6}$$

where $\boldsymbol{\Gamma}_j$ is a $p \times \gamma_j$ matrix. $\boldsymbol{\Gamma}_j$ is a full column rank matrix in the case of large p, at least $p \geq k_*$. We will assume $\boldsymbol{X}'\boldsymbol{X} = O(n)$ and that the order of elements of $\boldsymbol{\Gamma}_j\boldsymbol{\Gamma}_j'$ is $O(n)$, which is a common assumption in papers dealing with an asymptotic theory on the regression model (Fujikoshi and Satoh 1997; Fujikoshi, Yanagihara and Wakaki 2005). Notice that

$$\sum_{a=1}^{p}(\boldsymbol{\Gamma}_j\boldsymbol{\Gamma}_j')_{aa} = \mathrm{tr}(\boldsymbol{\Gamma}_j\boldsymbol{\Gamma}_j') = \mathrm{tr}(\boldsymbol{\Gamma}_j'\boldsymbol{\Gamma}_j) = \sum_{a=1}^{\gamma_j}(\boldsymbol{\Gamma}_j'\boldsymbol{\Gamma}_j)_{aa},$$

where $(\boldsymbol{A})_{ab}$ denotes the (a,b)th element of a matrix $\boldsymbol{A}$. Hence, if we assume that all the orders of the elements of $\boldsymbol{\Gamma}_j\boldsymbol{\Gamma}_j'$ are $O(n)$ and all the orders of the elements of $\boldsymbol{\Gamma}_j'\boldsymbol{\Gamma}_j$ are uniformally equal, $(\boldsymbol{\Gamma}_j'\boldsymbol{\Gamma}_j)_{aa} = O(np)$ holds because γ_j does not depend on n or p. From this fact and the inequality $\{(\boldsymbol{\Gamma}_j'\boldsymbol{\Gamma}_j)_{ab}\}^2 \leq (\boldsymbol{\Gamma}_j'\boldsymbol{\Gamma}_j)_{aa}(\boldsymbol{\Gamma}_j'\boldsymbol{\Gamma}_j)_{bb}$, $(\boldsymbol{\Gamma}_j'\boldsymbol{\Gamma}_j)_{ab} = O(np)$ is also obtained. Consequently, it is natural to assume that $\boldsymbol{\Gamma}_j'\boldsymbol{\Gamma}_j = O(np)$ when $\boldsymbol{X}'\boldsymbol{X} = O(n)$ is assumed.

Finally, in order to evaluate the probability of selecting the model j by the AIC, the $\mathrm{AIC_c}$, the BIC, and the CAIC, we prepare the following assumptions:

Assumption A1 : The true model is included in the set of candidate models, i.e., $j_* \in \mathcal{J}$.

Assumption A2 : None of the elements of $\boldsymbol{\Theta}_*$ and $\boldsymbol{\Sigma}_*$ depend on the sample size n, and $\boldsymbol{\Sigma}_*$ is positive definite for all p.

Assumption A3 : $\lim_{n\to\infty} n^{-1}\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{M}$ exists and is positive definite.

Assumption A4 : $\lim_{c_{n,p}\to c_0}(np)^{-1}\boldsymbol{\Gamma}_j\boldsymbol{\Gamma}_j' = \boldsymbol{\Delta}_{j,0}$ exists and is positive definite.

For $\boldsymbol{M}$ in A3, we write a limiting value of $n^{-1}\boldsymbol{X}_j'\boldsymbol{X}_\ell$ as $\boldsymbol{M}_{j,\ell}$ for $j,\ell \in \mathcal{J}$. It is clear that $\boldsymbol{M}_{j,\ell}$ is a submatrix of $\boldsymbol{M}$, and $\boldsymbol{M}_{j,\ell}$ also exists if $\boldsymbol{M}$ exists.

3. Main Results

In this section, we evaluate the asymptotic probability of selecting a model by the AIC, the $\mathrm{AIC_c}$, the BIC, and the CAIC. First, we describe the asymptotic selection

probabilities under the ordinary asymptotic framework, i.e., the large-sample asymptotic framework. Using the ideas of Shibata (1976), Nishii (1984), and Fujikoshi (1983; 1985), we obtain the following Theorem 1 (the proof is given in Appendix A.1):

THEOREM 1: *Suppose that the assumptions* A1, A2, *and* A3 *hold. Then, as* $n \to \infty$, *the asymptotic probability of selecting the model* j *by the AIC or the* AIC_c *is*

$$\begin{aligned} &\lim_{n\to\infty} P(\hat{j}_{\rm a} = j) = \lim_{n\to\infty} P(\hat{j}_{\rm c} = j) \\ &= \begin{cases} 0 & (j \in \mathcal{J}_-) \\ P(\cap_{\ell\in\mathcal{J}_+\setminus\{j\}}(\boldsymbol{z}_\ell'\boldsymbol{z}_\ell - \boldsymbol{z}_j'\boldsymbol{z}_j) < 2p(k_\ell - k_j)) & (j \in \mathcal{J}_+) \end{cases}, \end{aligned} \tag{3.1}$$

where $\boldsymbol{z}_j \sim N_{k_jp}(\boldsymbol{0}_{k_jp}, \boldsymbol{I}_{k_jp})$, $Cov[\boldsymbol{z}_j, \boldsymbol{z}_\ell] = \boldsymbol{I}_p \otimes \boldsymbol{M}_{j,j}^{-1/2}\boldsymbol{M}_{j,\ell}\boldsymbol{M}_{\ell,\ell}^{-1/2}$, *and* $\boldsymbol{0}_p$ *is the* p*-dimensional vector of zeros.*

These results include the results of Nishii (1984) as a special case. When the candidate models are nested, the probability of selecting the true model j_* by the AIC or the $\rm AIC_c$ becomes simple, as is shown in Corollary 1, as follows (a short proof is given in Appendix A.2).

COROLLARY 1: *Suppose that the assumptions* A1, A.2, *and* A3 *hold. When the candidate models are nested, i.e.,* $\mathcal{J} = \{\{1\}, \{1,2\}, \ldots, \{1,\ldots,k\}\}$, *as* $n \to \infty$, *the asymptotic probability of selecting the true model* $j_* = \{1,\ldots,k_*\}$ *by the AIC or the* AIC_c *is*

$$\lim_{n\to\infty} P(\hat{j}_{\rm a} = j_*) = \lim_{n\to\infty} P(\hat{j}_{\rm c} = j_*) = \sum_{[k-k_*]} \prod_{i=1}^{k-k_*} \frac{F_{h_ip}(2h_ip)^{h_i}}{ih_i}, \tag{3.2}$$

where $F_p(x)$ *is the distribution function of the chi-square distribution with* p *degrees of freedom, the summation* $\sum_{[\alpha]}$ *extends over all the* α*-tuples* $(h_1,\ldots,h_\alpha)$ *of non-negative integers with the property* $\sum_{i=1}^{\alpha} ih_i = \alpha$. *Specific forms of* (3.2) *are, e.g.,*

$$\begin{aligned} &F_p(2p),\ (k-k_*=1), \quad \frac{1}{2}\{F_p(2p)^2 + F_{2p}(4p)\},\ (k-k_*=2), \\ &\frac{1}{6}\{F_p(2p)^3 + 3F_p(2p)F_{2p}(4p) + 2F_{3p}(6p)\},\ (k-k_*=3), \\ &\frac{1}{24}\{F_p(2p)^4 + 6F_p(2p)^2F_{2p}(4p) + 3F_{2p}(4p)^2 + 8F_p(2p)F_{3p}(6p) + 6F_{4p}(8p)\},\ (k-k_*=4). \end{aligned}$$

Table 1 shows the values of the probability expression (3.2) for several choices of p and $k - k_*$. From (3.2), we can see that as the number of candidate models increases,

TABLE 1. Values of the equation (3.2) (%)

$p\backslash k-k_*$	1	2	3	4	5	6	7	8
1	84.27	78.74	76.02	74.46	73.49	72.85	72.41	72.10
2	86.47	82.80	81.32	80.60	80.22	80.01	79.89	79.81
3	88.84	86.36	85.51	85.17	85.01	84.93	84.89	84.88
4	90.84	89.14	88.65	88.47	88.40	88.38	88.36	88.36
5	92.48	91.30	91.00	90.91	90.88	90.87	90.87	90.87
6	93.80	92.98	92.80	92.76	92.74	92.74	92.74	92.74
7	94.88	94.30	94.20	94.17	94.17	94.16	94.16	94.16
8	95.76	95.35	95.29	95.28	95.27	95.27	95.27	95.27
9	96.48	96.19	96.15	96.15	96.15	96.15	96.14	96.14
10	97.07	96.87	96.84	96.84	96.84	96.84	96.84	96.84
11	97.56	97.42	97.40	97.40	97.40	97.40	97.40	97.40
12	97.97	97.86	97.85	97.85	97.85	97.85	97.85	97.85
13	98.30	98.22	98.22	98.22	98.22	98.22	98.22	98.22
14	98.58	98.52	98.52	98.52	98.52	98.52	98.52	98.52
15	98.81	98.77	98.77	98.77	98.77	98.77	98.77	98.77
16	99.00	98.97	98.97	98.97	98.97	98.97	98.97	98.97
17	99.16	99.14	99.14	99.14	99.14	99.14	99.14	99.14
18	99.29	99.28	99.28	99.28	99.28	99.28	99.28	99.28
19	99.41	99.40	99.40	99.40	99.40	99.40	99.40	99.40
20	99.50	99.49	99.49	99.49	99.49	99.49	99.49	99.49

and as $n \to \infty$, the asymptotic probability of selecting the true model by the AIC or the $\mathrm{AIC_c}$ decreases. Moreover, since $F_\beta(2\beta)$ is a monotonically increasing function with respect to $\beta \geq 1$, the asymptotic selection probability always increases with increasing p. These theoretical results can be confirmed with the data in Table 1. Theorem 1 points out that, when $n \to \infty$, the AIC and the $\mathrm{AIC_c}$ are not consistent in the selection of variables. However, when the behaviors of the AIC and the $\mathrm{AIC_c}$ are evaluated under a high-dimensional framework, we obtain new information, as in Theorem 2 (the proof is given in Appendix A.3).

THEOREM 2: *Suppose that the assumptions* A1*,* A2*, and* A4 *are satisfied.*

(1) *If* $c_{n,p} \to c_0 \in [0, c_{\mathrm{a}})$ *holds, where* c_{a} (≈ 0.797) *is a constant satisfying* $\log(1 - c_{\mathrm{a}}) + 2c_{\mathrm{a}} = 0$*, then the asymptotic probability of selecting the true model* j_* *by the AIC is*

$$\lim_{c_{n,p} \to c_0} P(\hat{j}_{\mathrm{a}} = j_*) = 1.$$

(2) *If* $c_{n,p} \to c_0 \in [0, 1)$ *holds, then the asymptotic probability of selecting the true model*

j_* by the AIC$_c$ is

$$\lim_{c_{n,p}\to c_0} P(\hat{j}_{\rm c} = j_*) = 1.$$

Theorem 2 shows that, when $c_{n,p} \to c_0$, the AIC and the AIC$_{\rm c}$ are consistent in model selection if $c_0 \in [0, c_a)$ for the AIC, and if $c_0 \in [0, 1)$ for the AIC$_{\rm c}$. Therefore, the range of values for (n, p) that satisfy consistency is wider for the AIC$_{\rm c}$ than it is for the AIC. This indicates that it is possible that the bias correction to the risk function has a positive effect on model selection.

In Theorem 2, it seems that the existence of a limiting value of $(np)^{-1}\boldsymbol{\Gamma}_j'\boldsymbol{\Gamma}_j$ as $c_{n,p} \to c_0$ is a strong assumption. Next, we consider weakening assumption A4 in Theorem 3. For a matrix $\boldsymbol{A}$, let $\lambda_{\min}(\boldsymbol{A})$ and $\lambda_{\max}(\boldsymbol{A})$ indicate the minimum and maximum eigenvalues, respectively. Then, we will replace assumption A4 with the following assumption:

ASSUMPTION A4$'$: $\lim_{c_{n,p}\to c_0} \eta_{n,p} = \infty$, and $\lim_{c_{n,p}\to c_0} \lambda_{\max}(\boldsymbol{\Gamma}_j'\boldsymbol{\Gamma}_j)/(n^2\eta_{n,p}^2) = 0$, where $\eta_{n,p} = \lambda_{\min}(\boldsymbol{\Gamma}_j'\boldsymbol{\Gamma}_j)/n$.

Even if we use assumption A4$'$ instead of A4, the consistencies of the AIC and the AIC$_{\rm c}$ hold, as is shown in the following theorem (the proof is given in Appendix A.4):

THEOREM 3: *Even when assumption* A4 *is replaced with assumption* A4$'$*, Theorem 2 still holds, i.e., the AIC is consistent when $c_{n,p} \to c_0 \in [0, c_{\rm a})$, and the AIC$_c$ is consistent when $c_{n,p} \to c_0 \in [0, 1)$.*

Notice that the upper bound of $\lambda_{\max}(\boldsymbol{\Gamma}_j'\boldsymbol{\Gamma}_j)$ is given by

$$\lambda_{\max}(\boldsymbol{\Gamma}_j'\boldsymbol{\Gamma}_j) \le \lambda_{\max}(\boldsymbol{\Theta}_*'\boldsymbol{\Sigma}_*^{-1}\boldsymbol{\Theta}_*')\lambda_{\max}(\boldsymbol{X}_*'(\boldsymbol{I}_n - \boldsymbol{P}_j)\boldsymbol{X}_*).$$

Additionally, if we assume that $\boldsymbol{\Theta}_*\boldsymbol{\Sigma}_*^{-1}\boldsymbol{\Theta}_*'$ is a full-rank matrix, the lower bound of $\lambda_{\min}(\boldsymbol{\Gamma}_j'\boldsymbol{\Gamma}_j)$ is given by

$$\lambda_{\min}(\boldsymbol{\Gamma}_j'\boldsymbol{\Gamma}_j) \ge \lambda_{\min}(\boldsymbol{\Theta}_*'\boldsymbol{\Sigma}_*^{-1}\boldsymbol{\Theta}_*')\lambda_{\min}^{\dagger}(\boldsymbol{X}_*'(\boldsymbol{I}_n - \boldsymbol{P}_j)\boldsymbol{X}_*),$$

where $\lambda_{\min}^{\dagger}(\boldsymbol{A})$ denotes the minimum nonzero eigenvalue of $\boldsymbol{A}$. Therefore, if $\boldsymbol{\Theta}_*\boldsymbol{\Sigma}_*^{-1}\boldsymbol{\Theta}_*'$ is a full-rank matrix, the assumption A4$'$ holds when the following equations are satisfied:

$$\lim_{p\to\infty} \lambda_{\min}(\boldsymbol{\Theta}_*\boldsymbol{\Sigma}_*^{-1}\boldsymbol{\Theta}_*') = \infty, \quad \lim_{c_{n,p}\to c_0} \frac{\lambda_{\max}(\boldsymbol{\Theta}_*\boldsymbol{\Sigma}_*^{-1}\boldsymbol{\Theta}_*')}{n\lambda_{\min}(\boldsymbol{\Theta}_*\boldsymbol{\Sigma}_*^{-1}\boldsymbol{\Theta}_*')^2} = 0,$$
$$\liminf_{n\to\infty} \frac{1}{n}\lambda_{\min}^{\dagger}(\boldsymbol{X}_*'(\boldsymbol{I}_n - \boldsymbol{P}_j)\boldsymbol{X}_*) > 0, \quad \limsup_{n\to\infty} \frac{1}{n}\lambda_{\max}(\boldsymbol{X}_*'(\boldsymbol{I}_n - \boldsymbol{P}_j)\boldsymbol{X}_*) < \infty.$$

Using the above equations, it is easy to check if assumption A4 holds.

Before concluding this section, we describe the consistencies of the BIC and the CAIC. Let $\mathcal{S}_- = \{j \in \mathcal{J}_- | k_* - k_j > 0\}$. Then, the probabilities of selecting the true model j_* by the BIC or the CAIC are given in the following theorems (the proofs are given in Appendices A.5 and A.6, respectively):

THEOREM 4: *Suppose that assumptions* A1*,* A2*, and* A3 *hold. Then, as* $n \to \infty$*, the asymptotic probability of selecting the true model* j_* *by the BIC or the CAIC is*

$$\lim_{n\to\infty} P(\hat{j}_{\mathrm{b}} = j_*) = \lim_{n\to\infty} P(\hat{j}_{\mathrm{o}} = j_*) = 1.$$

THEOREM 5: *Suppose that assumptions* A1*,* A2*,* A4*, and* A5 *hold, and* $\gamma_j > c_0(k_* - k_j)$ *is satisfied for all* $j \in \mathcal{S}_-$*. If* $c_{n,p} \to c_0 \in [0, c_{\mathrm{b}})$ *holds, where* $c_{\mathrm{b}} = \min\{1, \min_{j\in\mathcal{S}_-} \gamma_j/(k_* - k_j)\}$*, then the asymptotic probability of selecting the true model* j_* *by the BIC or the CAIC is*

$$\lim_{c_{n,p}\to c_0} P(\hat{j}_{\mathrm{b}} = j_*) = \lim_{c_{n,p}\to c_0} P(\hat{j}_{\mathrm{o}} = j_*) = 1.$$

Theorem 4 confirms the well-known fact, that the BIC and the CAIC are consistent in variable selection when $n \to \infty$, is also satisfied in the multivariate linear regression model. However, Theorem 5 indicates that the BIC and the CAIC are not always consistent in variable selection when $c_{n,p} \to c_0$. If $\mathbf{\Theta}_*\mathbf{\Sigma}_*^{-1}\mathbf{\Theta}_*'$ is a full-rank matrix, γ_j becomes $k_{j_+} - k_j$. Since $c_0 < 1$ and $k_{j_+} - k_j > k_* - k_j$ for all $j \in \mathcal{S}_-$, $\gamma_j > c_0(k_* - k_j)$ is satisfied if $\mathbf{\Theta}_*\mathbf{\Sigma}_*^{-1}\mathbf{\Theta}_*'$ is a full-rank matrix. In contrast, if $c_0 = 0$ then $\gamma_j > c_0(k_* - k_j)$ is satisfied. Therefore, we can see that variable selections using the BIC and the CAIC are consistent as $c_{n,p} \to c_0$ if $\mathbf{\Theta}_*\mathbf{\Sigma}_*^{-1}\mathbf{\Theta}_*'$ is a full-rank matrix, or $c_{n,p}$ converges to 0. However, if $\mathbf{\Theta}_*\mathbf{\Sigma}_*^{-1}\mathbf{\Theta}_*'$ is not a full-rank matrix and $c_0 \in (0, 1)$, we cannot determine if variable selection using the BIC and the CAIC are consistent as $c_{n,p} \to c_0$.

4. Numerical Study

In this section, we numerically examine the validity of our claim. The probability of selecting the true model by the AIC in (2.1), the $\mathrm{AIC}_{\mathrm{c}}$ in (2.2), the BIC in (2.3), and the CAIC in (2.4), was evaluated by Monte Carlo simulations with 10,000 iterations. The ten candidate models $j_\alpha = \{1, \ldots, \alpha\}$ $(\alpha = 1, \ldots, 10)$, with several different values of n and

TABLE 2. Selection probabilities of the true model (%)

	Case 1					Case 2 ($c_0 = 0.01$)				
n	p	AIC	AIC_c	BIC	CAIC	p	AIC	AIC_c	BIC	CAIC
100	2	75.7	81.5	76.7	66.0	2	75.7	81.5	76.7	66.0
200	2	79.6	83.2	98.6	98.0	4	87.1	91.0	95.1	88.3
500	2	79.7	81.3	99.8	99.9	10	96.1	97.4	100.0	100.0
1000	2	81.0	81.7	99.9	100.0	20	99.4	99.6	100.0	100.0
∞	2	80.2	80.2	100.0	100.0	∞	100.0	100.0	100.0	100.0
	Case 3					Case 4 ($c_0 = 0.1$)				
n	p	AIC	AIC_c	BIC	CAIC	p	AIC	AIC_c	BIC	CAIC
100	10	86.5	73.5	5.2	0.2	10	86.5	73.5	5.2	0.2
200	10	95.0	98.2	67.8	37.3	20	98.5	99.8	18.1	0.8
500	10	96.2	97.5	100.0	100.0	50	100.0	100.0	99.1	69.6
1000	10	96.7	97.0	100.0	100.0	100	100.0	100.0	100.0	100.0
∞	10	96.8	96.8	100.0	100.0	∞	100.0	100.0	100.0	100.0
	Case 5					Case 6 ($c_0 = 0.3$)				
n	p	AIC	AIC_c	BIC	CAIC	p	AIC	AIC_c	BIC	CAIC
100	30	89.9	0.0	0.0	0.0	30	89.9	0.0	0.0	0.0
200	30	99.4	99.6	1.1	0.0	60	99.8	21.3	0.0	0.0
500	30	99.8	100.0	99.9	97.4	150	100.0	100.0	0.0	0.0
1000	30	99.9	99.9	100.0	100.0	300	100.0	100.0	0.0	0.0
∞	30	99.9	99.9	100.0	100.0	∞	100.0	100.0	0.0	0.0
	Case 7 ($c_0 = 0.0$)					Case 8 ($c_0 = 0.0$)				
n	p	AIC	AIC_c	BIC	CAIC	p	AIC	AIC_c	BIC	CAIC
100	30	89.9	0.0	0.0	0.0	30	89.9	0.0	0.0	0.0
200	32	99.6	99.6	0.3	0.0	40	99.7	97.4	0.0	0.0
500	35	99.9	100.0	99.8	93.9	50	100.0	100.0	99.1	69.7
1000	40	99.9	100.0	100.0	100.0	60	100.0	100.0	100.0	100.0
∞	∞	100.0	100.0	100.0	100.0	∞	100.0	100.0	100.0	100.0

p, were prepared for Monte Carlo simulations. We generated $z_1, \ldots, z_n \sim i.i.d.\ U(-1, 1)$. Using $z_1, \ldots, z_n$, we constructed a $n \times 10$ matrix of explanatory variables $\boldsymbol{X}$ where the (a, b)th element was defined by z_a^{b-1} $(a = 1, \ldots, n; b = 1, \ldots, 10)$. The true model was determined by $\boldsymbol{\Theta}_* = (1, -2, 3, -4, 5)'\mathbf{1}_p'$, $j_* = \{1, 2, 3, 4, 5\}$, and $\boldsymbol{\Sigma}_*$, where the (i, j)th element was defined by $(0.8)^{|a-b|}$ $(a = 1, \ldots, p; b = 1, \ldots, p)$. Here $\mathbf{1}_p$ was the p-dimensional vector of ones. Thus, j_1, j_2, j_3, and j_4 were underspecified models, and j_5, j_6, j_7, j_8, j_9, and j_{10} were overspecified models.

In our numerical study, $\gamma_j = 1$ and $\max(k_* - k_j) = 4$ hold for all $j \in \mathcal{S}_-$. This implies that when $c_0 > 1/4$, the inequality $\gamma_j > c_0(k_* - k_j)$ was not always satisfied for

all $j \in \mathcal{S}_-$. Thus, the probability of selecting j_* by the BIC and the CAIC converged to 0 as $c_{n,p} \to c_0 \in (1/4, 1)$. This means that the BIC and the CAIC were not consistent in variable selection when $c_0 > 1/4$.

Table 2 shows the probability of selecting the true model by the AIC, the $\mathrm{AIC_c}$, the BIC, and the CAIC. For $n = \infty$ or $p = \infty$, we list the theoretical values obtained from Corollary 1 and Theorems 2, 4, and 5. In the table, Cases 1, 3, and 5 are the results when $n \to \infty$ under a fixed p, and Cases 2, 4, 6, 7, and 8 are the results when $(n, p) \to \infty$ and with $c_0 = 0.01, 0.1, 0.3, 0.0$, and 0.0. From the table, we can see that in the cases of the AIC and the $\mathrm{AIC_c}$, the greater the dimension and sample size were, the greater the probabilities became. Compared with the results obtained from the AIC and the $\mathrm{AIC_c}$, probabilities by the $\mathrm{AIC_c}$ tended to be higher than those by the AIC when n was not small. In the cases of the BIC and the CAIC, the greater the dimension and sample size were, the higher the selection probabilities became, with the exception of Case 6. This was because variable selection using the BIC and the CAIC were not consistent in Case 6. Additionally, when n was small and p was large, the selection probabilities of the BIC and the CAIC were both very low. However, if the BIC and the CAIC were consistent in variable selection, these probabilities became high as n and p increased.

We simulated several other models and obtained similar results. Since the theoretical difference between using the AIC and the $\mathrm{AIC_c}$ occurs when $c_{n,p} > 0.8$, we should list the numerical results for such a case. However, when $c_{n,p}$ is close to 1, the convergence of selection probabilities was extremely slow . Thus, we do not show simulation results for dimensions close to the sample size.

5. Conclusion and Discussion

In this paper, we demonstrated that the AIC for the multivariate linear regression model is consistent in variable selection when we approximate the probability of selecting the true model using a high-dimensional asymptotic framework. The AIC and the bias-corrected AICs are sometimes pilloried for inconsistency, although the value of the AIC is not in choosing the true model. The results presented in this paper will help to dispel the undeserved negative reputation of the AIC. Moreover, a range of the parameters necessary for the $\mathrm{AIC_c}$ to satisfy consistency is wider than that for the AIC. This indicates that it is possible that correcting the bias to the risk function may have a positive effect on the

model selection. It is a well-known fact that variable selections using the BIC and the CAIC are consistent if we approximate the probability of selecting the true model using a large-sample asymptotic framework. However, we found that there is a possibility that the BIC and the CAIC become inconsistent if we approximate the probability of selecting the true model using a high-dimensional asymptotic framework.

It is known that the large-sample asymptotic theory gives a poor approximation when the dimension is large. The high-dimensional asymptotic theory gives a better approximation than the large-sample asymptotic theory when the sample size is large, and sometimes even when the dimension is not so large (Fujikoshi and Seo 1998; Fujikoshi and Sakurai 2009; Fujikoshi, Shimizu, and Ulyanov 2010). Hence, the consistency property of the AIC that we demonstrated will be useful for high-dimensional data analysis, which recently has been attracting the attention of many researchers. Usually, the high-dimensional asymptotic theory is used to improve the approximations of the distributions of statistics. However, the results in this paper suggest a possibility that new insight can be provided by applying the high-dimensional asymptotic theory to high-dimensional data.

From the simulation study, we found that, the larger the dimension and sample size, the higher the selection probabilities. This numerical result naturally implies that using multiple response variables at the same time as the model selection can increase the probability of selecting the true model. In other words, we should not select variables using only each response variable. That is a strong reason to apply the model selection procedure based on the multivariate linear regression model to high-dimensional data.

In this paper, we considered the case of $n > p$ because $\hat{\mathbf{\Sigma}}_j$ becomes singular when $p > n$. However, using a ridge-type estimator of the covariance matrix, the singularity can be avoided, as demonstrated by Yamamura, Yanagihara, and Srivastava (2010). We can expect that an AIC consisting of such a ridge-type estimator will be consistent in model selection.

Appendix

A.1. The Proof of Theorem 1

Since $\mathrm{AIC_c} = \mathrm{AIC} + O(n^{-1})$ when p is fixed, it is enough to show only the case of the AIC for proving Theorem 1. The selection probability of a model j selected by the AIC

is

$$P(\hat{j}_{\rm a} = j) = P(\cap_{\ell \in \mathcal{J} \setminus \{j\}} \{{\rm AIC}(\ell) > {\rm AIC}(j)\})$$
$$= P(\{\cap_{\ell \in \mathcal{J}_- \setminus \{j\}} \{{\rm AIC}(\ell) > {\rm AIC}(j)\}\} \cap \{\cap_{\ell \in \mathcal{J}_+ \setminus \{j\}} \{{\rm AIC}(\ell) > {\rm AIC}(j)\}\}). \quad \text{(A.1)}$$

Notice that $n^{-1}\boldsymbol{\Gamma}_j\boldsymbol{\Gamma}_j'$ is convergent when assumption A3 holds, where $\boldsymbol{\Gamma}_j$ is given by (2.6). Let $\boldsymbol{\Psi}_{\ell,0} = \lim_{n\to\infty} n^{-1}\boldsymbol{\Gamma}_\ell\boldsymbol{\Gamma}_\ell'$. When $\ell_1 \in \mathcal{J}_+$ and $\ell_2 \in \mathcal{J}_-$, $\hat{\boldsymbol{\Sigma}}_{\ell_1} \xrightarrow{p} \boldsymbol{\Sigma}_*$ and $\hat{\boldsymbol{\Sigma}}_{\ell_2} \xrightarrow{p} \boldsymbol{\Sigma}_*^{1/2}\boldsymbol{\Psi}_{\ell_2,0}\boldsymbol{\Sigma}_*^{1/2} + \boldsymbol{\Sigma}_*$ as $n \to \infty$. Since $\boldsymbol{\Psi}_{\ell,0}$ is a positive semidefinite matrix, we have

$$\frac{1}{n}\{{\rm AIC}(\ell_2) - {\rm AIC}(\ell_1)\} \xrightarrow{p} \log|\boldsymbol{\Sigma}_*^{1/2}\boldsymbol{\Psi}_{\ell_2,0}\boldsymbol{\Sigma}_*^{1/2} + \boldsymbol{\Sigma}_*| - \log|\boldsymbol{\Sigma}_*| = \log|\boldsymbol{I}_p + \boldsymbol{\Psi}_{\ell_2,0}| > 0.$$

This result implies that

$$\lim_{n\to\infty} P({\rm AIC}(\ell_2) > {\rm AIC}(\ell_1)) = 1, \quad \lim_{n\to\infty} P({\rm AIC}(\ell_1) > {\rm AIC}(\ell_2)) = 0.$$

Using the above two equalities and a basic probability theorem, we have

$$\lim_{n\to\infty} P(\cap_{\ell \in \mathcal{J}_- \setminus \{j\}} \{{\rm AIC}(\ell) > {\rm AIC}(j)\}) = 1, \ (j \in \mathcal{J}_+),$$
$$\lim_{n\to\infty} P(\cap_{\ell \in \mathcal{J}_+ \setminus \{j\}} \{{\rm AIC}(\ell) > {\rm AIC}(j)\}) = 0, \ (j \in \mathcal{J}_-).$$

Thus, from the above equalities and (A.1), we obtain the following result.

$$\lim_{n\to\infty} P(\hat{j}_{\rm a} = j) = \begin{cases} 0 & (j \in \mathcal{J}_-) \\ \lim_{n\to\infty} P(\cap_{\ell \in \mathcal{J}_+ \setminus \{j\}} \{{\rm AIC}(\ell) > {\rm AIC}(j)\}) & (j \in \mathcal{J}_+) \end{cases}. \quad \text{(A.2)}$$

From here to the end of proof, we assume $j \in \mathcal{J}_+$. Let $\boldsymbol{V}$ and $\boldsymbol{Z}_j$ be the $p \times p$ and the $k_j \times p$ matrices defined by

$$\boldsymbol{V} = \frac{1}{\sqrt{n}}(\boldsymbol{\mathcal{E}}'\boldsymbol{\mathcal{E}} - n\boldsymbol{I}_p), \quad \boldsymbol{Z}_j = (\boldsymbol{X}_j'\boldsymbol{X}_j)^{-1/2}\boldsymbol{X}_j'\boldsymbol{\mathcal{E}},$$

where $\boldsymbol{\mathcal{E}} = (\boldsymbol{Y} - \boldsymbol{X}_{j_*}\boldsymbol{\Theta}_*)\boldsymbol{\Sigma}_*^{-1/2}$. It is well known that $\boldsymbol{V}$ has an asymptotic normality as $n \to \infty$, and $\boldsymbol{Z}_j \sim N_{k_j \times p}(\boldsymbol{O}_{k_j,p}, \boldsymbol{I}_{k_j p})$. Furthermore, using

$$\boldsymbol{\Sigma}_*^{-1/2}\hat{\boldsymbol{\Sigma}}_j\boldsymbol{\Sigma}_*^{-1/2} = \frac{1}{n}\boldsymbol{\mathcal{E}}'(\boldsymbol{I}_n - \boldsymbol{P}_j)\boldsymbol{\mathcal{E}} = \frac{1}{n}(\boldsymbol{\mathcal{E}}'\boldsymbol{\mathcal{E}} - \boldsymbol{Z}_j'\boldsymbol{Z}_j),$$

we have

$$\boldsymbol{\Sigma}_*^{-1/2}\hat{\boldsymbol{\Sigma}}_j\boldsymbol{\Sigma}_*^{-1/2} = \boldsymbol{I}_p + \frac{1}{\sqrt{n}}\boldsymbol{V} - \frac{1}{n}\boldsymbol{Z}_j'\boldsymbol{Z}_j.$$

From the above expression, the first term of the AIC(j) can be expanded as

$$n\log|\hat{\boldsymbol{\Sigma}}_j| = n\log|\boldsymbol{\Sigma}_*| + \sqrt{n}{\rm tr}(\boldsymbol{V}) - \{{\rm tr}(\boldsymbol{V}^2) + {\rm tr}(\boldsymbol{Z}_j\boldsymbol{Z}_j)\} + O_p(n^{-1/2}).$$

Let $\boldsymbol{z}_j$ be a $k_j p$-dimensional random vector defined by $\boldsymbol{z}_j = \mathrm{vec}(\boldsymbol{Z}_j)$, where $\mathrm{vec}(\boldsymbol{A})$ is an operator that transforms a matrix to a vector by stacking the first to the last columns of $\boldsymbol{A}$, i.e., $\mathrm{vec}(\boldsymbol{A}) = (\boldsymbol{a}_1', \ldots, \boldsymbol{a}_m')'$ when $\boldsymbol{A} = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_m)$ (see, e.g., Harville 1997, chap. 16.2). Then, it follows from the expansion and the equality $\mathrm{tr}(\boldsymbol{Z}_j'\boldsymbol{Z}_j) = \boldsymbol{z}_j'\boldsymbol{z}_j$ that

$$\mathrm{AIC}(\ell) - \mathrm{AIC}(j) = -(\boldsymbol{z}_\ell'\boldsymbol{z}_\ell - \boldsymbol{z}_j'\boldsymbol{z}_j) + 2p(k_\ell - k_j) + O_p(n^{-1/2}). \tag{A.3}$$

Consequently, by combining (A.3) with (A.2), Theorem 1 is proved.

A.2. The Proof of Corollary 1

Let $\boldsymbol{z}_j$ be the same random vector as in Theorem 1. Notice that when $\ell_1 \subset \ell_2 \subset \ell_3$ and $\ell_1, \ell_2, \ell_3 \in \mathcal{J}_+$, $\boldsymbol{z}_{\ell_3}'\boldsymbol{z}_{\ell_3} - \boldsymbol{z}_{\ell_1}'\boldsymbol{z}_{\ell_1}$ is distributed according to the chi-square distribution with $p(k_{\ell_3} - k_{\ell_1})$ degrees of freedom, and $\boldsymbol{z}_{\ell_2}'\boldsymbol{z}_{\ell_2} - \boldsymbol{z}_{\ell_1}'\boldsymbol{z}_{\ell_1}$ and $\boldsymbol{z}_{\ell_3}'\boldsymbol{z}_{\ell_3} - \boldsymbol{z}_{\ell_2}'\boldsymbol{z}_{\ell_2}$ are independently distributed according to the chi-square distributions with $p(k_{\ell_2} - k_{\ell_1})$ and $p(k_{\ell_3} - k_{\ell_2})$ degrees of freedoms, respectively. Using these properties, when the candidate models are nested, the distribution in (3.1) is rewritten as

$$P\left(\max_{\alpha=1,\ldots,k-k_*} \textstyle\sum_{i=1}^{\alpha}(w_i - 2p) < 0\right), \tag{A.4}$$

where $w_1, \ldots, w_{k-k_*}$ are independently and identically distributed according to the chi-square distribution with p degrees of freedom. Using lemma 1 of Shibata (1976), the probability (A.4) is explicitly evaluated as (3.2).

A.3. The Proof of Theorem 2

First, we consider the case of $j \in \mathcal{J}_-$. Let $\boldsymbol{W}_1, \boldsymbol{W}_2$, and $\boldsymbol{W}_3$ be $p \times p$ mutually independent random matrices distributed according to $W_p(n - k_{j_+}, \boldsymbol{I}_p)$, $W_p(d_j, \boldsymbol{I}_p)$, and $W_p(\gamma_j, \boldsymbol{I}_p; \boldsymbol{\Gamma}_j\boldsymbol{\Gamma}_j')$, respectively, where $d_j = k_{j_+} - k_j - \gamma_j$ and $\boldsymbol{\Gamma}_j$ is given by (2.6). Using these three matrices, we have

$$n\boldsymbol{\Sigma}_*^{-1/2}\hat{\boldsymbol{\Sigma}}_j\boldsymbol{\Sigma}_*^{-1/2} = \boldsymbol{W}_1 + \boldsymbol{W}_2 + \boldsymbol{W}_3, \quad n\boldsymbol{\Sigma}_*^{-1/2}\hat{\boldsymbol{\Sigma}}_{j_+}\boldsymbol{\Sigma}_*^{-1/2} = \boldsymbol{W}_1.$$

It follows from the property of Wishart distributions (see Fujikoshi, Shimizu, and Ulyanov 2010, p. 57 th. 3.2.4) that

$$\begin{aligned}\log|\hat{\boldsymbol{\Sigma}}_j| - \log|\hat{\boldsymbol{\Sigma}}_{j_+}| &= \log\frac{|\boldsymbol{W}_1 + \boldsymbol{W}_2 + \boldsymbol{W}_3|}{|\boldsymbol{W}_1 + \boldsymbol{W}_2|} + \log\frac{|\boldsymbol{W}_1 + \boldsymbol{W}_2|}{|\boldsymbol{W}_1|} \\ &= -\log\frac{|\boldsymbol{U}_1|}{|\boldsymbol{U}_1 + \boldsymbol{U}_2|} - \log\frac{|\boldsymbol{U}_3|}{|\boldsymbol{U}_3 + \boldsymbol{U}_4|},\end{aligned} \tag{A.5}$$

where $\boldsymbol{U}_1$, $\boldsymbol{U}_2$, $\boldsymbol{U}_3$, and $\boldsymbol{U}_4$ are random matrices distributed according to Wishart or noncentral Wishart distributions;

$$\begin{array}{ll} \boldsymbol{U}_1 \sim W_{\gamma_j}(n-k_j-p, \boldsymbol{I}_{\gamma_j}), & \boldsymbol{U}_2 \sim W_{\gamma_j}(p, \boldsymbol{I}_{\gamma_j}; \boldsymbol{\Gamma}_j'\boldsymbol{\Gamma}_j), \\ \boldsymbol{U}_3 \sim W_{d_j}(n-k_j-\gamma_j-p, \boldsymbol{I}_{d_j}), & \boldsymbol{U}_4 \sim W_{d_j}(p, \boldsymbol{I}_{d_j}). \end{array} \tag{A.6}$$

Here, $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$ are mutually independent, and $\boldsymbol{U}_3$ and $\boldsymbol{U}_4$ are also mutually independent. When $c_{n,p} \to c_0 \in [0,1)$, we have

$$\frac{1}{n-k_j-p}\boldsymbol{U}_1 \xrightarrow{p} \boldsymbol{I}_{\gamma_j}, \quad \frac{1}{n-k_j-\gamma_j-p}\boldsymbol{U}_3 \xrightarrow{p} \boldsymbol{I}_{d_j}, \quad \frac{1}{p}\boldsymbol{U}_4 \xrightarrow{p} \boldsymbol{I}_{d_j}. \tag{A.7}$$

From the definition of the noncentral Wishart distribution, a different expression of $\boldsymbol{U}_2$ is given as $\boldsymbol{U}_2 = (\boldsymbol{Z}+\boldsymbol{\Gamma}_j)'(\boldsymbol{Z}+\boldsymbol{\Gamma}_j)$, where $\boldsymbol{Z} \sim N_{p\times\gamma_j}(\boldsymbol{O}_{p,\gamma_j}, \boldsymbol{I}_{p\gamma_j})$. Thus, we derive

$$\mathrm{vec}(\boldsymbol{U}_2) = \mathrm{vec}(\boldsymbol{Z}'\boldsymbol{Z}) + \mathrm{vec}(\boldsymbol{\Gamma}_j'\boldsymbol{Z}) + \mathrm{vec}(\boldsymbol{Z}'\boldsymbol{\Gamma}_j) + \mathrm{vec}(\boldsymbol{\Gamma}_j'\boldsymbol{\Gamma}_j). \tag{A.8}$$

Notice that

$$Cov[\mathrm{vec}(\boldsymbol{\Gamma}_j'\boldsymbol{Z})] = \boldsymbol{I}_{\gamma_j} \otimes \boldsymbol{\Gamma}_j'\boldsymbol{\Gamma}_j, \quad Cov[\mathrm{vec}(\boldsymbol{Z}'\boldsymbol{\Gamma}_j)] = \boldsymbol{\Gamma}_j'\boldsymbol{\Gamma}_j \otimes \boldsymbol{I}_{\gamma_j}. \tag{A.9}$$

Hence, $\mathrm{vec}(\boldsymbol{\Gamma}_j'\boldsymbol{Z}) = O_p(n^{1/2}p^{1/2})$ and $\mathrm{vec}(\boldsymbol{Z}'\boldsymbol{\Gamma}_j) = O_p(n^{1/2}p^{1/2})$ are obtained, because $\boldsymbol{\Gamma}_j'\boldsymbol{\Gamma}_j = O(np)$ is satisfied. Needless to say, $\mathrm{vec}(\boldsymbol{Z}'\boldsymbol{Z}) = O_p(1)$ holds. These results imply the convergence in probability of $\boldsymbol{U}_2$ as

$$\frac{1}{np}\boldsymbol{U}_2 \xrightarrow{p} \boldsymbol{\Delta}_{j,0}, \tag{A.10}$$

where $\boldsymbol{\Delta}_{j,0} = \lim_{c_{n,p}\to c_0}(np)^{-1}\boldsymbol{\Gamma}_j'\boldsymbol{\Gamma}_j$ in the assumption A4. Combining the equations (A.7) and (A.10) yields

$$\frac{1}{np}(\boldsymbol{U}_1+\boldsymbol{U}_2) \xrightarrow{p} \boldsymbol{\Delta}_{j,0}, \quad \frac{1}{n-k_j-\gamma_j-p}(\boldsymbol{U}_3+\boldsymbol{U}_4) \xrightarrow{p} \frac{1}{1-c_0}\boldsymbol{I}_{d_j}.$$

Using the results of the convergence of the probability, the first and second terms in (A.5) are expanded as

$$\begin{aligned} -\log\frac{|\boldsymbol{U}_1|}{|\boldsymbol{U}_1+\boldsymbol{U}_2|} &= \log\left(\frac{p}{1-c_{n,p}-k_j/n}\right)^{\gamma_j} - \log\frac{|\boldsymbol{U}_1/(n-k_j-p)|}{|(\boldsymbol{U}_1+\boldsymbol{U}_2)/(np)|} \\ &= \gamma_j\log p - \gamma_j\log(1-c_0) + \log|\boldsymbol{\Delta}_{j,0}| + o_p(1), \end{aligned} \tag{A.11}$$

and

$$-\log\frac{|\boldsymbol{U}_3|}{|\boldsymbol{U}_3+\boldsymbol{U}_4|} = -\log\frac{|\boldsymbol{U}_3/(n-k_j-\gamma_j-p)|}{|(\boldsymbol{U}_3+\boldsymbol{U}_4)/(n-k_j-\gamma_j-p)|} = -d_j\log(1-c_0)+o_p(1). \tag{A.12}$$

Since $\log|\boldsymbol{\Delta}_{j,0}|$ is a constant, $\lim_{c_{n,p}\to c_0}(\log p)^{-1}\log|\boldsymbol{\Delta}_{j,0}| = 0$ holds. Substituting the equations (A.11) and (A.12) into (A.5) yields

$$\frac{1}{\log p}(\log|\hat{\boldsymbol{\Sigma}}_j| - \log|\hat{\boldsymbol{\Sigma}}_{j_+}|) \xrightarrow{p} \gamma_j > 0. \tag{A.13}$$

Using the same idea as in the derivation of (A.12), it can be shown that

$$\frac{1}{\log p}(\log|\hat{\boldsymbol{\Sigma}}_{j_+}| - \log|\hat{\boldsymbol{\Sigma}}_{j_*}|) \xrightarrow{p} 0. \tag{A.14}$$

Let m_j be the penalty term in the $\mathrm{AIC_c}$ for the model j. Then, we have

$$\begin{aligned} m_j &= \frac{2n}{n-k_j-p-1}\left\{pk_j + \frac{1}{2}p(p+1)\right\} \\ &= \frac{npc_{n,p}}{1-c_{n,p}} + \frac{p}{(1-c_{n,p})^2}\{(2-c_{n,p})k_j + 1\} + O(pn^{-1}). \end{aligned}$$

From the expansion of m_j, $m_j - m_{j_*}$ is expanded as

$$m_j - m_{j_*} = \frac{r_j(2-c_{n,p})p}{(1-c_{n,p})^2} + O(pn^{-1}), \tag{A.15}$$

where $r_j = k_j - k_*$. Hence, differences between the penalty terms of the $\mathrm{AIC_c}$s and the AICs are convergent as

$$\lim_{c_{n,p}\to c_0}\frac{1}{n\log p}\{2(k_j-k_*)p\} = 0, \quad \lim_{c_{n,p}\to c_0}\frac{1}{n\log p}(m_j - m_{j_*}) = 0.$$

Using these results with the results (A.13) and (A.14), the difference between the information criteria of the model j and the true model j_* is convergent as

$$\frac{1}{n\log p}\{\mathrm{AIC}(j) - \mathrm{AIC}(j_*)\} \xrightarrow{p} \gamma_j > 0, \ \frac{1}{n\log p}\{\mathrm{AIC_c}(j) - \mathrm{AIC_c}(j_*)\} \xrightarrow{p} \gamma_j > 0. \tag{A.16}$$

Next, we consider the case of $j \in \mathcal{J}_+$. Notice that

$$n\boldsymbol{\Sigma}_*^{-1/2}\hat{\boldsymbol{\Sigma}}_j\boldsymbol{\Sigma}_*^{-1/2} \sim W_p(n-k_j, \boldsymbol{I}_p), \quad n\boldsymbol{\Sigma}_*^{-1/2}\hat{\boldsymbol{\Sigma}}_{j_*}\boldsymbol{\Sigma}_*^{-1/2} \sim W_p(n-k_*, \boldsymbol{I}_p).$$

It follows from the property of Wishart distributions (see Fujikoshi, Shimizu, and Ulyanov 2010, p. 57 th. 3.2.4) that

$$n\log|\hat{\boldsymbol{\Sigma}}_j| - n\log|\hat{\boldsymbol{\Sigma}}_{j_*}| = n\log\frac{|\boldsymbol{B}_1|}{|\boldsymbol{B}_1+\boldsymbol{B}_2|} = -n\log|\boldsymbol{I}_{r_j} + \boldsymbol{B}_2\boldsymbol{B}_1^{-1}|, \tag{A.17}$$

where $r_j = k_j - k_*$, and $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$ are $r_j \times r_j$ independent random matrices with Wishart distributions;

$$\boldsymbol{B}_1 \sim W_{r_j}(n-k_*-p, \boldsymbol{I}_{r_j}), \quad \boldsymbol{B}_2 \sim W_{r_j}(p, \boldsymbol{I}_{r_j}).$$

Notice that as $c_{n,p} \to c_0 \in [0, 1)$,

$$\frac{1}{n - k_* - p}\boldsymbol{B}_1 \xrightarrow{p} \boldsymbol{I}_{r_j}, \quad \frac{1}{p}\boldsymbol{B}_2 \xrightarrow{p} \boldsymbol{I}_{r_j}.$$

Using these results and $p < n$, the right-hand side of (A.17) is expanded as

$$\begin{aligned} -n \log |\boldsymbol{I}_{r_j} + \boldsymbol{B}_2 \boldsymbol{B}_1^{-1}| &= -n \log \left| \boldsymbol{I}_{r_j} + \frac{p}{n - k_* - p} \left(\frac{1}{p}\boldsymbol{B}_2 \right) \left(\frac{1}{n - k_* - p}\boldsymbol{B}_1 \right)^{-1} \right| \\ &= -n \log \left| \boldsymbol{I}_{r_j} + \frac{p}{n - k_* - p} \left\{ \boldsymbol{I}_{r_j} + O_p(p^{-1/2}) \right\} \right|. \end{aligned}$$

Therefore, the equation (A.17) divided by p is evaluated as

$$\begin{aligned} \frac{n}{p}(\log |\hat{\boldsymbol{\Sigma}}_j| - \log |\hat{\boldsymbol{\Sigma}}_{j_*}|) &= -\frac{1}{c_{n,p}} \log |\boldsymbol{I}_{r_j}/(1 - c_{n,p}) + O_p(p^{1/2} n^{-1})| \\ &= \frac{r_j}{c_{n,p}} \log(1 - c_{n,p}) + O_p(p^{-1/2}). \end{aligned}$$

Furthermore, it follows from the equality (A.15) that

$$\frac{1}{p}(m_j - m_{j_*}) = \frac{r_j(2 - c_{n,p})}{(1 - c_{n,p})^2} + O(n^{-1}).$$

From these results, we can see that

$$\begin{aligned} &\frac{1}{p}\left\{ \mathrm{AIC}(j) - \mathrm{AIC}(j_*) \right\} = r_j \left\{ \frac{1}{c_{n,p}} \log(1 - c_{n,p}) + 2 \right\} + O_p(p^{-1/2}), \\ &\frac{1}{p}\left\{ \mathrm{AIC}_{\mathrm{c}}(j) - \mathrm{AIC}_{\mathrm{c}}(j_*) \right\} = r_j \left\{ \frac{1}{c_{n,p}} \log(1 - c_{n,p}) + \frac{1}{1 - c_{n,p}} + \frac{1}{(1 - c_{n,p})^2} \right\} + O_p(p^{-1/2}). \end{aligned}$$

Notice that the $\lim_{c \to 0} c^{-1} \log(1 - c) = -1$, and $c^{-1} \log(1 - c) + 2$ is a monotonically decreasing function in $0 \leq c < 1$. Thus, when $c_0 < c_{\mathrm{a}}$ holds, and c_{a} is a constant satisfying $\log(1 - c_0) + 2c_0 = 0$, $c_0^{-1} \log(1 - c_0) + 2 > 0$ is satisfied. Therefore, when $c_0 < c_{\mathrm{a}}$, we derive

$$\frac{1}{p}\left\{ \mathrm{AIC}(j) - \mathrm{AIC}(j_*) \right\} \xrightarrow{p} r_j \left\{ \frac{1}{c_0} \log(1 - c_0) + 2 \right\} > 0. \tag{A.18}$$

Meanwhile, $c^{-1} \log(1 - c) + (1 - c)^{-1} + (1 - c)^{-2}$ is a monotonically increasing function in $0 \leq c < 1$. Thus, we have

$$\frac{1}{p}\left\{ \mathrm{AIC}_{\mathrm{c}}(j) - \mathrm{AIC}_{\mathrm{c}}(j_*) \right\} \xrightarrow{p} r_j \left\{ \frac{1}{c_0} \log(1 - c_0) + \frac{1}{1 - c_0} + \frac{1}{(1 - c_0)^2} \right\} > 0. \tag{A.19}$$

It follows from the results in (A.16), (A.18). and (A.19) that

$$\begin{aligned} &\lim_{c_{n,p} \to c_0} P(\hat{j}_{\mathrm{a}} = j) = 0, \ (j \in \mathcal{J} \backslash \{j_*\}, \ c_0 \in [0, c_{\mathrm{a}})), \\ &\lim_{c_{n,p} \to c_0} P(\hat{j}_{\mathrm{c}} = j) = 0, \ (j \in \mathcal{J} \backslash \{j_*\}, \ c_0 \in [0, 1)). \end{aligned}$$

Consequently, Theorem 2 is proved.

A.4. The Proof of Theorem 3

The proof in the case of $j \in \mathcal{J}_+$ is the same as that in Theorem 2, so it is sufficient to prove Theorem 3 only in the case of $j \in \mathcal{J}_-$, where we use assumption A4$'$ instead of assumption A4. Thus, we will show that the selection probability of the model $j \in \mathcal{J}_-$ converges to 0. In order to prove this, we use $\boldsymbol{U}_1$, $\boldsymbol{U}_2$, $\boldsymbol{U}_3$, and $\boldsymbol{U}_4$, which are given by (A.6). Notice that $\boldsymbol{\Gamma}_j'\boldsymbol{\Gamma}_j \leq \lambda_{\max}(\boldsymbol{\Gamma}_j'\boldsymbol{\Gamma}_j)\boldsymbol{I}_{\gamma_j}$. Hence, when the assumption A4$'$ holds, $\boldsymbol{\Gamma}_j'\boldsymbol{\Gamma}_j = o(n^2\eta_{n,p}^2)$ holds. Recall that $\boldsymbol{U}_2 = (\boldsymbol{Z}+\boldsymbol{\Gamma}_j)'(\boldsymbol{Z}+\boldsymbol{\Gamma}_j)$, where $\boldsymbol{Z} \sim N_{p\times\gamma_j}(\boldsymbol{O}_{p,\gamma_j}, \boldsymbol{I}_{p\gamma_j})$. From the order of $\boldsymbol{\Gamma}_j'\boldsymbol{\Gamma}_j$ and the results in (A.9), we can see that $\mathrm{vec}(\boldsymbol{\Gamma}_j'\boldsymbol{Z}) = o_p(n\eta_{n,p})$ and $\mathrm{vec}(\boldsymbol{Z}'\boldsymbol{\Gamma}_j) = o_p(n\eta_{n,p})$. These results and equation (A.8) imply the convergence in probability of $\boldsymbol{U}_2$ and $\boldsymbol{U}_1 + \boldsymbol{U}_2$ as

$$\frac{1}{n\eta_{n,p}}(\boldsymbol{U}_2 - \boldsymbol{\Gamma}_j'\boldsymbol{\Gamma}_j) \xrightarrow{p} \boldsymbol{O}_{\gamma_j,\gamma_j}, \quad \frac{1}{n\eta_{n,p}}(\boldsymbol{U}_1 + \boldsymbol{U}_2 - \boldsymbol{\Gamma}_j'\boldsymbol{\Gamma}_j) \xrightarrow{p} \boldsymbol{O}_{\gamma_j,\gamma_j}. \tag{A.20}$$

Let $\boldsymbol{\Delta}_j = (n\eta_{n,p})^{-1}\boldsymbol{\Gamma}_j'\boldsymbol{\Gamma}_j$. It follows from $\lambda_{\min}(\boldsymbol{\Delta}_j) = 1$ that

$$\liminf_{c_{n,p}\to c_0} \log|\boldsymbol{\Delta}_j| \geq 0.$$

The above result and the assumption that $\lim_{c_{n,p}\to c_0}\eta_{n,p} = \infty$ lead us to

$$\liminf_{c_{n,p}\to c_0} \frac{1}{\log\eta_{n,p}} \log|\boldsymbol{\Delta}_j| \geq 0. \tag{A.21}$$

The results in (A.7) and (A.20) give an expansion of $-\log(|\boldsymbol{U}_1|/|\boldsymbol{U}_1+\boldsymbol{U}_2|)$ as

$$\begin{aligned} -\log\frac{|\boldsymbol{U}_1|}{|\boldsymbol{U}_1+\boldsymbol{U}_2|} &= \log\left(\frac{\eta_{n,p}}{1-c_{n,p}-k_j/n}\right)^{\gamma_j} - \log\frac{|\boldsymbol{U}_1/(n-k_j-p)|}{|(\boldsymbol{U}_1+\boldsymbol{U}_2)/(n\eta_{n,p})|} \\ &= \gamma_j\log\eta_{n,p} - \gamma_j\log(1-c_0) + \log|\boldsymbol{\Delta}_j| + o_p(1). \end{aligned} \tag{A.22}$$

Hence, substituting equations (A.12) and (A.22) into (A.5) yields

$$\frac{1}{\log\eta_{n,p}}(\log|\hat{\boldsymbol{\Sigma}}_j| - \log|\hat{\boldsymbol{\Sigma}}_{j_+}| - \log|\boldsymbol{\Delta}_j|) \xrightarrow{p} \gamma_j > 0. \tag{A.23}$$

Using the same idea as in the derivation of (A.12), it can be shown that

$$\frac{1}{\log\eta_{n,p}}(\log|\hat{\boldsymbol{\Sigma}}_{j_+}| - \log|\hat{\boldsymbol{\Sigma}}_{j_*}|) \xrightarrow{p} 0. \tag{A.24}$$

From the result in (A.15), the differences between the penalty terms of the AICs, and the AIC_cs are convergent as

$$\lim_{c_{n,p}\to c_0}\frac{1}{n\log\eta_{n,p}}\{2(k_j-k_*)p\} = 0, \quad \lim_{c_{n,p}\to c_0}\frac{1}{n\log\eta_{n,p}}(m_j - m_{j_*}) = 0.$$

The above results with the results in (A.23) and (A.24) imply the convergence in probability of the differences between the information criteria of the model j and the true model j_* is expressed as

$$\frac{1}{n \log \eta_{n,p}} \{\mathrm{AIC}(j) - \mathrm{AIC}(j_*) - \log |\boldsymbol{\Delta}_j|\} \xrightarrow{p} \gamma_j > 0,$$
$$\frac{1}{n \log \eta_{n,p}} \{\mathrm{AIC_c}(j) - \mathrm{AIC_c}(j_*) - \log |\boldsymbol{\Delta}_j|\} \xrightarrow{p} \gamma_j > 0.$$

Combining the above results with the result in (A.21) yields

$$\lim_{c_{n,p} \to c_0} P(\hat{j}_{\mathrm{a}} = j) = 0, \ \lim_{c_{n,p} \to c_0} P(\hat{j}_{\mathrm{c}} = j) = 0, \ (j \in \mathcal{J}_-).$$

Consequently, Theorem 3 is proved.

A.5. The Proof of Theorem 4

Notice that the differences between the penalty terms of the BICs and the CAICs are convergent as

$$\lim_{n \to \infty} \frac{1}{n} \{p(k_j - k_*) \log n\} = \lim_{n \to \infty} \frac{1}{n} \{p(k_j - k_*)(1 + \log n)\} = 0,$$

and

$$\lim_{n \to \infty} \frac{1}{\log n} \{p(k_j - k_*) \log n\} = \lim_{n \to \infty} \frac{1}{\log n} \{p(k_j - k_*)(1 + \log n)\} = p(k_j - k_*).$$

Thus, using the same method as in the proof of Theorem 1, we have

$$\begin{aligned} &\frac{1}{n} \{\mathrm{BIC}(j) - \mathrm{BIC}(j_*)\} \xrightarrow{p} \log |\boldsymbol{I}_p + \boldsymbol{\Psi}_{j,0}| > 0, \\ &\frac{1}{n} \{\mathrm{CAIC}(j) - \mathrm{CAIC}(j_*)\} \xrightarrow{p} \log |\boldsymbol{I}_p + \boldsymbol{\Psi}_{j,0}| > 0, \end{aligned} \quad (j \in \mathcal{J}_-),$$

and

$$\begin{aligned} &\frac{1}{\log n} \{\mathrm{BIC}(j) - \mathrm{BIC}(j_*)\} \xrightarrow{p} p(k_j - k_*) > 0, \\ &\frac{1}{\log n} \{\mathrm{CAIC}(j) - \mathrm{CAIC}(j_*)\} \xrightarrow{p} p(k_j - k_*) > 0. \end{aligned} \quad (j \in \mathcal{J}_+ \backslash \{j_*\}).$$

These results imply that

$$\lim_{n \to \infty} P(\hat{j}_{\mathrm{b}} = j_*) = \lim_{n \to \infty} P(\hat{j}_{\mathrm{o}} = j_*) = 1.$$

Consequently, Theorem 4 is proved.

A.6. The Proof of Theorem 5

Notice that the differences between the penalty terms of the BICs and the CAICs are convergent as

$$\lim_{c_{n,p}\to c_0} \frac{1}{p\log n}\{p(k_j-k_*)\log n\} = \lim_{c_{n,p}\to c_0} \frac{1}{p\log n}\{p(k_j-k_*)(1+\log n)\} = r_j.$$

Thus, using the same method as in the proof of Theorem 2, we have

$$\begin{aligned} &\frac{1}{p\log n}\{\mathrm{BIC}(j)-\mathrm{BIC}(j_*)\} \xrightarrow{p} r_j > 0, \\ &\frac{1}{p\log n}\{\mathrm{CAIC}(j)-\mathrm{CAIC}(j_*)\} \xrightarrow{p} r_j > 0, \end{aligned} \qquad (j\in\mathcal{J}_+\backslash\{j_*\}). \tag{A.25}$$

Moreover, it is easy to obtain

$$\begin{aligned} \frac{1}{n\log p}\{p(k_j-k_*)\log n\} &= c_{n,p}r_j\left(-\frac{\log c_{n,p}}{\log p}+1\right), \\ \frac{1}{n\log p}p\{(k_j-k_*)(\log n+1)\} &= c_{n,p}r_j\left(\frac{1-\log c_{n,p}}{\log p}+1\right). \end{aligned}$$

Since $\lim_{c\to 0} c\log c = 0$ holds, we derive

$$\lim_{c_{n,p}\to c_0} \frac{1}{n\log p}\{p(k_j-k_*)\log n\} = \lim_{c_{n,p}\to c_0} \frac{1}{n\log p}\{p(k_j-k_*)(1+\log n)\} = c_0 r_j.$$

Therefore, if $\gamma_j > c_0(k_*-k_j)$ is satisfied for all $j\in\{j\in\mathcal{J}_-|k_*-k_j>0\}$, it follows in the same way as in the proof of Theorem 3 that

$$\begin{aligned} &\frac{1}{n\log p}\{\mathrm{BIC}(j)-\mathrm{BIC}(j_*)\} \xrightarrow{p} \gamma_j + c_0 r_j > 0, \\ &\frac{1}{n\log p}\{\mathrm{CAIC}(j)-\mathrm{CAIC}(j_*)\} \xrightarrow{p} \gamma_j + c_0 r_j > 0, \end{aligned} \qquad (j\in\mathcal{J}_-). \tag{A.26}$$

The equations (A.25) and (A.26) imply that

$$\lim_{c_{n,p}\to c_0} P(\hat{j}_{\mathrm{b}} = j_*) = \lim_{c_{n,p}\to c_0} P(\hat{j}_{\mathrm{o}} = j_*) = 1.$$

Consequently, Theorem 5 is proved.

References

[1] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *2nd. International Symposium on Information Theory* (eds. B. N. Petrov & F. Csáki), 267–281, Akadémiai Kiadó, Budapest.

[2] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automatic Control*, **AC-19**, 716–723.

[3] Bedrick, E. J. & Tsai, C.-L. (1994). Model selection for multivariate regression in small samples. *Biometrics*, **50**, 226–231.

[4] Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): the general theory and its analytical extensions. *Psychometrika*, **52**, 345–370.

[5] Davies, S. J., Neath, A. A. & Cavanaugh, J. E. (2006). Estimation optimality of corrected AIC and modified C_p in linear regression model. *International Statist. Review*, **74**, 161–168.

[6] Dien, S. J. V., Iwatani, S.. Usuda, Y. & Matsui, K. (2006). Theoretical analysis of amino acid-producing *Eschenrichia coli* using a stoixhiometrix model and multivariate linear regression. *J. Biosci. Bioeng.*, **102**, 34–40.

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

[8] Fujikoshi, Y. (1985). Selection of variables in two-group discriminant analysis by error rate and Akaike's information criteria. *J. Multivariate Anal.*, **17**, 27–37.

[9] Fujikoshi, Y. & Sakurai, T. (2009). High-dimensional asymptotic expansions for the distributions of canonical correlations. *J. Multivariate Anal.*, **100**, 231–242.

[10] Fujikoshi, Y. & Satoh, K. (1997). Modified AIC and C_p in multivariate linear regression. *Biometrika*, **84**, 707–716.

[11] Fujikoshi, Y. & Seo, T. (1998). Asymptotic approximations for EPMC's of the linear and the quadratic discriminant functions when the sample sizes and the dimension are large. *Random Oper. Stochastic Equations*, **6**, 269–280.

[12] Fujikoshi, Y., Shimizu, R. & Ulyanov, V. V. (2010). *Multivariate Statistics*: *High-Dimensional and Large-Sample Approximations*. John Wiley & Sons, Inc., Hoboken, New Jersey.

[13] Fujikoshi, Y., Yanagihara, H. & Wakaki, H. (2005). Bias corrections of some criteria for selection multivariate linear regression models in a general case. *Amer. J. Math. Management Sci.*, **25**, 221–258.

[14] Harville, D. A. (1997). *Matrix Algebra from a Statistician's Perspective*. Springer-Verlag, New York.

[15] Kullback, S. & Leibler, R. A. (1951). On information and sufficiency. *Ann. Math. Statist.*, **22**, 79–86.

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

[17] Sârbu, C., Onişor, C., Posa, M., Kevresan, S. & Kuhajda, K. (2008). Modeling and prediction (correction) of partition coefficients of bile acids and their derivatives by multivariate regression methods. *Talanta*, **75**, 651–657.

[18] Saxén, R. & Sundell, J. (2006). ^{137}Cs in freshwater fish in Finland since 1986 – a statistical analysis with multivariate linear regression models. *J. Environ. Radioactiv.*, **87**, 62–76.

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

[20] Shao, J. (1997). An asymptotic theory for linear model selection. *Statist. Sinica*, **7**, 221–264.

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

[22] Shibata, R. (1980). Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. *Ann. Statist.*, **8**, 147–164.

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

[24] Srivastava, M. S. (2002). *Methods of Multivariate Statistics*. John Wiley & Sons, New York.

[25] Sugiura, N. (1978). Further analysis of the data by Akaike's information criterion and the finite corrections. *Commun. Statist. Theory Methods*, **A7**, 13–26.

[26] Timm, N. H. (2002). *Applied Multivariate Analysis*. Springer-Verlag, New York.

[27] Yamamura, M., Yanagihara, H., & Srivastava, M. S. (2010). Variable selection in multivariate linear regression models with fewer observations than the dimension. *Japanese J. Appl. Statist.*, **39**, 1–19.

[28] Yanagihara, H. (2006). Corrected version of *AIC* for selecting multivariate normal linear regression models in a general nonnormal case. *J. Multivariate Anal.*, **97**, 1070–1089.

[29] Yanagihara, H., Kamo, K. & Tonda, T. (2011). Second-order bias-corrected AIC in multivariate normal linear models under nonnormality. *Canad. J. Statist.*, **39**, 126–146.

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

[31] Yoshimoto, A., Yanagihara, H. & Ninomiya, Y. (2005). Finding factors affecting a forest stand growth through multivariate linear modeling. *J. Jpn. For. Soc.*, **87**, 504–512 (in Japanese).