Electronic Journal of Statistics
Vol. 9 (2015) 869–897
ISSN: 1935-7524
DOI: 10.1214/15-EJS1022

A consistency property of the AIC for multivariate linear models when the dimension and the sample size are large*

Hirokazu Yanagihara†, **Hirofumi Wakaki**‡ **and Yasunori Fujikoshi**§

Department of Mathematics, Graduate School of Science, Hiroshima University, 1-3-1 Kagamiyama, Higashi-Hiroshima, Hiroshima 739-8626, Japan
e-mail: yanagi@math.sci.hiroshima-u.ac.jp; wakaki@math.sci.hiroshima-u.ac.jp; fujikoshi_y@yahoo.co.jp

Abstract: It is common knowledge that Akaike's information criterion (AIC) is not a consistent model selection criterion, and Bayesian information criterion (BIC) is. These have been confirmed from an asymptotic selection probability evaluated from a large-sample 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, there are cases that the AIC for selecting variables in multivariate linear models is consistent, but the BIC is not. The AIC and BIC are included in a family of information criteria defined by adding a penalty term expressing the complexity of the model to a negative twofold maximum log-likelihood. By clarifying the condition of the penalty term to ensure the consistency, we derive conditions for consistency of the AIC, BIC and other information criteria under the high-dimensional asymptotic framework.

MSC 2010 subject classifications: Primary 62J05; secondary 62E20.
Keywords and phrases: AIC, bias-corrected AIC, BIC, consistent AIC, high-dimensional asymptotic framework, multivariate linear model, selection probability, variable selection.

Received September 2013.

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 k nonstochastic explanatory variables, where n is the sample size, and it is assumed that $n - p - k - 1 > 0$. In order to ensure the

*The authors wish to thank the referees, the associate editor and the editor for their helpful suggestions.
†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, #25540012, 2013–2015.
‡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), #24500343, 2012–2014.
§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), #25330038, 2013–2015.

possibility of estimating the model, we also assume that $\mathrm{rank}(\boldsymbol{X}) = k\ (< n)$. 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 $\boldsymbol{X}$. Of course, it holds that $\boldsymbol{X}_\omega = \boldsymbol{X}$ and $k_\omega = k$. Also, we let k_A denote the number of elements of a set A, i.e., $k_A = \#(A)$. Then the following multivariate linear regression model with k_j explanatory variables is considered as the candidate model:

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

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. Here, $\boldsymbol{A} \otimes \boldsymbol{B}$ denotes the Kronecker product of an $m \times n$ matrix $\boldsymbol{A}$ and a $p \times q$ matrix $\boldsymbol{B}$, which is an $mp \times nq$ matrix defined by

$$\boldsymbol{A} \otimes \boldsymbol{B} = \begin{pmatrix} a_{11}\boldsymbol{B} & \cdots & a_{1n}\boldsymbol{B} \\ \vdots & \ddots & \vdots \\ a_{m1}\boldsymbol{B} & \cdots & a_{mn}\boldsymbol{B} \end{pmatrix},$$

where a_{ij} is the (i, j)th element of $\boldsymbol{A}$ (see, e.g., [18, chap. 16]). In particular, the model with $\boldsymbol{X}_\omega$ (namely $\boldsymbol{X}$) is called 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 of (1.1) is one of basic models of multivariate analysis. This model is introduced in many multivariate statistical textbooks (see, e.g., [31, chap. 9], [33, chap. 4]), and even now is widely used in chemometrics, engineering, econometrics, psychometrics, and many other fields, for the prediction of multiple responses from a set of explanatory variables (see, e.g., [10, 24, 25, 40]). Since it is important to specify factors affecting response variables in regression analysis, searching for the optimal subset j is essential.

Akaike's information criterion (AIC), proposed by [1, 2], is widely used for selecting the best model. The AIC was proposed as an asymptotic unbiased estimator of the risk function assessed by the expected Kullback-Leibler (KL) loss [20] under the assumption that the candidate model includes the true model. One purpose of a model selection method based on 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 (see, e.g., [27, 29, 39]). A Bayesian information criterion (BIC) proposed by [26] and a consistent AIC (CAIC) proposed by [6] are also widely used for model selection purposes. 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, i.e., the probability of selecting the true model goes to 1 asymptotically, although

for the AIC is not. When using the AIC for 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 the following asymptotic framework that represents an ordinary asymptotic procedure [11, 12, 22, 28]:

- A large-sample (LS) asymptotic framework: the sample size is approaching ∞ under a fixed number of parameters. In this paper, $\lim_{n\to\infty}$ means a limit as $n \to \infty$ under the condition that the number of parameters is fixed.

In the case of multivariate linear models, although there are many bias-corrected AICs for the risk function (see, e.g., [3, 14, 17, 37, 38]), such a bias-corrected AIC is still not consistent for model selection.

In recent years, high-dimensional data analysis has been attracting the attention of many researchers. It is known that the LS asymptotic framework gives a poor approximation when the dimension is large. However, the following asymptotic framework gives a better approximation than the LS asymptotic framework when the dimension and the sample size are large, and sometimes even when the dimension is not so large [13, 15, 16]:

- A high-dimensional (HD) asymptotic framework: the sample size and the dimension of the response variables simultaneously approach ∞ under the condition that $c_{n,p} = p/n \to c_0 \in [0, 1)$. For simplicity, we will write "$(n, p) \to \infty$ simultaneously under the condition that $c_{n,p} \to c_0$" as "$c_{n,p} \to c_0$", and $\lim_{c_{n,p}\to c_0}$ means a limit under the HD asymptotic framework. It should be emphasized that we assume that p always goes to ∞ in the HD asymptotic framework. Hence, the notation $c_{n,p} \to 0$ does not mean the LS asymptotic framework.

When the HD asymptotic framework is used for evaluating the asymptotic probability of selecting the true model, there is a possibility that the AIC can become consistent. In fact, in this paper, we will prove that a variable selection method based on the AIC becomes consistent in multivariate linear models under a HD asymptotic framework. The AIC is included in a family of information criteria defined by adding a penalty term expressing the complexity of the model to a negative twofold maximum likelihood. By clarifying the condition of the penalty term to satisfy the consistency property, we will also prove that a variable selection method based on the bias-corrected AIC ($\mathrm{AIC_c}$), as proposed by [3], becomes consistent under more non-restrictive situation than that based on the AIC, and those based on the BIC and the CAIC are not necessarily consistent when $c_0 \in (0, 1)$. Additionally, we derive a sufficient condition to satisfy the consistency of the family of information criteria under an asymptotic framework such that the number of candidate models may approach ∞.

In this paper, $o(x)$, $O(x)$, $o_p(x)$, and $O_p(x)$ used in a vector or matrix having finite dimension or size 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 Landau notations indicate the orders as $n \to \infty$ under a fixed number of parameters

when the LS asymptotic framework is considered. Meanwhile, those Landau notations are also used for the orders as $c_{n,p} \to c_0$ when the HD asymptotic framework is considered. As stated already, we deal with not a strong consistency but a weak consistency. Hence, throughout the paper, the word "consistency" means weak consistency.

The remainder of the paper is organized as follows: In Section 2, we present the necessary notation for evaluating an asymptotic selection probability. In Section 3, the asymptotic probability of selecting the true model is calculated under the HD asymptotic framework. In Section 4, we compare with variable selection methods based on the AIC, $\mathrm{AIC_c}$, BIC and CAIC 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 asymptotic selection probability. First, we describe several classes of the set j. Let $\mathcal{J}$ be a set of candidate models denoted by $\mathcal{J} = \{j_1, \ldots, j_K\}$, where K is the number of candidate models. 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., $\mathcal{J}_- = \mathcal{J}_+^c \cap \mathcal{J}$. We use the same terminology, "overspecified model" and "underspecified model", as was used by [14].

Estimations for the unknown parameters $\boldsymbol{\Theta}_j$ and $\boldsymbol{\Sigma}_j$ in the model (1.1) are carried out by the maximum likelihood method, 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., $\boldsymbol{P}_j = \boldsymbol{X}_j(\boldsymbol{X}_j'\boldsymbol{X}_j)^{-1}\boldsymbol{X}_j'$. A family of information criteria in the model (1.1) is

$$\mathrm{IC}_m(j) = \mathcal{L}(j) + np(\log 2\pi + 1) + m(j), \tag{2.1}$$

where $\mathcal{L}(j) = n \log\det(\hat{\boldsymbol{\Sigma}}_j)$ and $m(j)$ is a positive constant expressing a penalty for the complexity of the model (1.1). An information criterion included in this family is specified by an individual penalty term $m(j)$. This family contains AIC, $\mathrm{AIC_c}$, BIC and CAIC as a special case.

$$m(j) = \begin{cases} 2\{pk_j + p(p+1)/2\} & (\mathrm{AIC}) \\ 2n\{pk_j + p(p+1)/2\}/(n - k_j - p - 1) & (\mathrm{AIC_c}) \\ \{pk_j + p(p+1)/2\}\log n & (\mathrm{BIC}) \\ \{pk_j + p(p+1)/2\}(1 + \log n) & (\mathrm{CAIC}) \end{cases}. \tag{2.2}$$

When $p = 1$, the $\mathrm{AIC_c}$ coincides with the bias-corrected AIC proposed by [32]. [9] 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 [30, pp. 18–20]. This property indicates that, for all the overspecified models, the $\mathrm{AIC_c}$ is better than the AIC at estimating the risk function. The best subsets of ω is chosen by minimizing $\mathrm{IC}_m(j)$, i.e., it is presented as

$$\hat{j}_m = \arg\min_{j \in \mathcal{J}} \mathrm{IC}_m(j).$$

Next, we describe a noncentrality matrix that plays a critical role for proving consistency. In fact, asymptotic behaviors of elements or eigen values of a noncentrality matrix are one of important factors that determines whether an information criterion is consistent or not. The noncentrality matrix is 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

$$j_+ = j \cup j_*, \ (j \in \mathcal{J}). \tag{2.3}$$

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

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

This indicates that the rank of $\boldsymbol{X}_*'(\boldsymbol{I}_n - \boldsymbol{P}_j)\boldsymbol{X}_*$ is independent of p if k_* is an independent of p. Let the rank of the noncentrality matrix be denoted by γ_j. It follows from the inequality $\mathrm{rank}(\boldsymbol{\Theta}_*\boldsymbol{\Sigma}_*^{-1}\boldsymbol{\Theta}_*') \le \min\{p, k_*\}$ and a knowledge of an elementary linear algebra that

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

It notes that $\gamma_j = k_{j_+} - k_j$ if $\boldsymbol{\Theta}_*$ is a full row rank matrix. Since the noncentrality matrix is a positive semidefinite matrix, and its rank is γ_j, it is decomposed as

$$\boldsymbol{\Sigma}_*^{-1/2}\boldsymbol{\Theta}_*'\boldsymbol{X}_*'(\boldsymbol{I}_n - \boldsymbol{P}_j)\boldsymbol{X}_*\boldsymbol{\Theta}_*\boldsymbol{\Sigma}_*^{-1/2} = \boldsymbol{\Gamma}_j\boldsymbol{\Gamma}_j', \tag{2.4}$$

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 $p \ge k_*$. If we assume that the orders of elements of $\boldsymbol{X}'\boldsymbol{X}$ are $O(n)$ and elements of $\boldsymbol{\Theta}_*$ and $\boldsymbol{\Sigma}_*$ are independent of n, which are common assumptions in papers dealing with an asymptotic theory on the regression model [14, 17],

the orders of elements of $\boldsymbol{\Gamma}_j\boldsymbol{\Gamma}_j'$ are $O(n)$. Let $(\boldsymbol{A})_{ab}$ denote the (a,b)th element of a matrix $\boldsymbol{A}$. Notice that

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

Hence, if we assume that all the orders of the elements of $\boldsymbol{\Gamma}_j\boldsymbol{\Gamma}_j'$ are $O(n)$, all the orders of the elements of $\boldsymbol{\Gamma}_j'\boldsymbol{\Gamma}_j$ are uniformly equal, and γ_j is constant, then all the orders of the elements of $\boldsymbol{\Gamma}_j'\boldsymbol{\Gamma}_j$ are $O(np)$. 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 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.

Let $\lambda_{j,1} \geq \cdots \geq \lambda_{j,\gamma_j} > 0$ be eigen values of $\boldsymbol{\Gamma}_j'\boldsymbol{\Gamma}_j$. In order to evaluate the probability of selecting the model j by the IC_m, we introduce the following assumptions:

Assumption 1. The true model is included in the set of candidate models, i.e., $j_* \in \mathcal{J}$.

Assumption 2. $\lim_{n\to\infty} n^{-1}\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{R}_0$ exists and is positive definite, and $\lim_{n\to\infty} n^{-1}\boldsymbol{\Gamma}_j\boldsymbol{\Gamma}_j' = \boldsymbol{\Psi}_{j,0}$ exists and is not the zero matrix for all $j \in \mathcal{J}_-$.

Assumption 3. For all $j \in \mathcal{J}_-$, γ_j is constant, and $\limsup_{c_{n,p}\to c_0}(np)^{-1}\lambda_{j,1} < \infty$ and $\liminf_{c_{n,p}\to c_0}(np)^{-1}\lambda_{j,\gamma_j} > 0$.

For $\boldsymbol{R}_0$ in assumption 2, we write a limiting value of $n^{-1}\boldsymbol{X}_j'\boldsymbol{X}_\ell$ as $\boldsymbol{R}_{j,\ell,0}$ for $j,\ell \in \mathcal{J}$. It is clear that $\boldsymbol{R}_{j,\ell,0}$ is a submatrix of $\boldsymbol{R}_0$, and $\boldsymbol{R}_{j,\ell,0}$ also exists if $\boldsymbol{R}_0$ exists. Moreover, it notes that $\boldsymbol{\Psi}_{j,0}$ still depends on p because $\boldsymbol{\Psi}_{j,0}$ is the convergent value under the LS asymptotic framework.

3. Main results

In this section, we evaluate an asymptotic probability of selecting a model by the IC_m in (2.1). First, we describe the asymptotic selection probabilities of selecting the true model j_* under the ordinary asymptotic framework, i.e., the LS asymptotic framework. Using the ideas of [11, 12, 22, 28], we obtain the following Theorem 3.1 (the proof is given in Appendix A.1):

Theorem 3.1. *Suppose that assumptions 1 and 2 hold. A variable selection method based on the IC_m is consistent when $n \to \infty$ if the following conditions are satisfied simultaneously:*

C1-1. *For all $j \in \mathcal{J}_-$,*

$$\lim_{n\to\infty}\frac{m(j)-m(j_*)}{n} = 0.$$

C1-2. *For all $j \in \mathcal{J}_+\backslash\{j_*\}$,*

$$\lim_{n\to\infty}\{m(j)-m(j_*)\} = \infty.$$

If one of the above two conditions is not satisfied, a variable selection method based on the IC_m is not consistent when $n \to \infty$. Additionally, when $m(j) = O(1)$ as $n \to \infty$ and $\lim_{n\to\infty}\{m(j) - m(\ell)\} = pm_0(k_j - k_\ell)$ for all $j, \ell \in \mathcal{J}_+$, the asymptotic probability of selecting the model j by the IC_m is

$$\lim_{n\to\infty} P(\hat{j}_m = j) = \begin{cases} 0 & (j \in \mathcal{J}_-) \\ P(\cap_{\ell\in\mathcal{J}_+\setminus\{j\}}(\boldsymbol{z}_\ell'\boldsymbol{z}_\ell - \boldsymbol{z}_j'\boldsymbol{z}_j) < m_0 p(k_\ell - k_j)) & (j \in \mathcal{J}_+) \end{cases}, \tag{3.1}$$

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

These results include the results of [22, 35] etc. as a special case for $p = 1$. Theorem 3.1 points out a well-known fact that, when $n \to \infty$, the AIC and the AIC_c are not consistent and the BIC and the CAIC are consistent in model selection. However, when behaviors of the information criteria are evaluated under the HD framework, we obtain new properties, as in Theorem 3.2 (the proof is given in Appendix A.2).

Theorem 3.2. *Suppose that assumptions 1 and 3 are satisfied. Then, a variable selection method based on the IC_m is consistent when $c_{n,p} \to c_0$ if the following conditions are satisfied simultaneously:*

C2-1. *For all $j \in \mathcal{J}_-$,*

$$\lim_{c_{n,p}\to c_0} \frac{m(j) - m(j_*)}{n \log p} > -\gamma_j.$$

C2-2. *For all $j \in \mathcal{J}_+\setminus\{j_*\}$,*

$$\lim_{c_{n,p}\to c_0} \frac{m(j) - m(j_*)}{p} > -\frac{1}{c_0}(k_j - k_*)\log(1 - c_0).$$

If the sign "$>$" becomes "$<$" in one of the above two conditions, a variable selection method based on the IC_m is not consistent when $c_{n,p} \to c_0$.

It notes that $\lim_{c\to 0} c^{-1}\log(1-c) = -1$ and $c^{-1}\log(1-c)$ is a monotonically decreasing function in $0 \le c < 1$. From Theorem 3.2, consistency properties of specific criteria are clarified as the following corollary (the proof is given in Appendix A.4):

Corollary 3.1. *Suppose that assumptions 1 and 3 are satisfied.*

(*i*) *When $c_{n,p} \to c_0$, a variable selection method based on the AIC is consistent if $c_0 \in [0, c_a)$, and is not consistent if $c_0 \in (c_a, 1)$, where c_a (≈ 0.797) is a constant satisfying*

$$\log(1 - c_a) + 2c_a = 0. \tag{3.2}$$

(*ii*) *When $c_{n,p} \to c_0$, a variable selection method based on the AIC_c is consistent*

(*iii*) *When* $c_{n,p} \to c_0$*, variable selection methods based on the BIC and the CAIC are consistent if* $c_0 \in [0, c_{\rm b})$*, and are not consistent if* $c_0 \in (c_{\rm b}, 1)$*, where* $c_{\rm b} = \min\{1, \min_{j \in \mathcal{S}_-} \gamma_j/(k_* - k_j)\}$ *and* $\mathcal{S}_- = \{j \in \mathcal{J}_- | k_* - k_j > 0\}$.

Corollary 3.1 shows that, there is no restriction of c_0 in the condition for consistency of the $\mathrm{AIC_c}$ although it is restricted in the AIC. This indicates that it is possible that the bias correction to the risk function has a positive effect on selection of the true model. Moreover, Corollary 3.1 indicates that the BIC and the CAIC are not always consistent in variable selection when $c_{n,p} \to c_0$. If $\mathbf{\Theta}_*$ is the full row 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 $\mathbf{\Theta}_*$ is the full row rank matrix. In contrast, if $c_0 = 0$ then $\gamma_j > c_0(k_* - k_j)$ is satisfied. Therefore, we can see that variable selection methods based on the BIC and the CAIC are consistent as $c_{n,p} \to c_0$ if $\mathbf{\Theta}_*$ is the full row rank matrix, or $c_{n,p}$ converges to 0. However, if $\mathbf{\Theta}_*$ is not the full row rank matrix and $c_0 \in (0, 1)$, we cannot determine as if variable selection methods based on the BIC and the CAIC are consistent as $c_{n,p} \to c_0$.

In order to clarify the condition to ensure inconsistency, assumption 3 is assumed in Theorem 3.2, i.e., we assume that the orders of eigen values of $\mathbf{\Gamma}_j' \mathbf{\Gamma}_j$ are uniformly the same and γ_j is independent of n and p. If the aim is only to derive a sufficient condition for consistency, such a strong assumption like assumption 3 is unnecessary. In fact, for evaluating consistency, there is no need to assume the same orders for all the eigen values of $\mathbf{\Gamma}_j' \mathbf{\Gamma}_j$. Whether an information criterion is consistent strongly depends on the orders of divergence speeds of several eigen values. Hence, we clarify condition of the orders of divergence speeds of eigen values of $\mathbf{\Gamma}_j' \mathbf{\Gamma}_j$ to ensure a consistency. In addition, most recently, many researchers pay close attention to "big data analysis", and thus study on a theory of a variable selection when the number of candidate models approaches ∞ (see, e.g., [19]). Hence, we derive a sufficient condition for the consistency by using the following asymptotic framework:

- A high-dimensional and large-model (HD-LM) asymptotic framework: the HD asymptotic framework under the condition that the following equations are satisfied:

$$\max_{j \in \mathcal{J}} \frac{k_j}{n} \to 0, \quad {}^{\exists} l > 0 \ s.t. \ K = o(p^l), \tag{3.3}$$

  where K is the number of candidate models. In the HD-LM asymptotic framework, p always goes to ∞, and it makes no difference whether K is constant or K goes to ∞. This indicates that the HD asymptotic framework is a special case of the HD-LM asymptotic framework. For simplicity, we will write "$(n, p) \to \infty$ simultaneously under the HD-LM asymptotic framework" as "$c_{n,p} \to c_0$ under LM", and $\lim_{c_{n,p} \to c_0, \mathrm{LM}}$ and $\liminf_{c_{n,p} \to c_0, \mathrm{LM}}$ mean a limit and a limit inferior under the HD-LM asymptotic framework, respectively.

Theorem 3.3. *Suppose that assumption 1 holds. A variable selection method based on the* IC_m *is consistent under the HD-LM asymptotic framework if the following conditions are satisfied:*

C3-1. *For sufficiently large n, there exist positive constants δ_1, δ_2 such that for all $j \in \mathcal{J}_-$ there exists an integer $q \in [1, \gamma_j]$ such that $\lambda_{j,q}/q^2 > n^{\delta_1}$ and*

$$\log \frac{\beta_{j,q}}{n - p - k_{j_+} + (q+1)/2} - \frac{1}{qn}\{m(j_+) - m(j)\} > \delta_2,$$

where j_+ is given by (2.3), $\gamma_j = \mathrm{rank}(\mathbf{\Gamma}_j)$, and $\beta_{j,q}$ is the geometric mean of the largest q eigen values of $\mathbf{\Gamma}_j\mathbf{\Gamma}_j'$, i.e.,

$$\beta_{j,q} = \left(\prod_{i=1}^{q} \lambda_{j,i}\right)^{1/q}. \tag{3.4}$$

C3-2. *For sufficiently large n, there exists a positive constant δ such that for all $j \in \mathcal{J}_+ \backslash \{j_*\}$,*

$$\frac{m(j) - m(j_*)}{p(k_j - k_*)} + \frac{1}{c_{n,p}} \log(1 - c_{n,p}) > \delta.$$

The proof is given in Appendix A.5. Roughly speaking, the existence of δ_2 in condition C3-1 is related to the order of noncentrality matrix. Assumption 3 is equivalent to the condition that the orders of all the eigen values of $\mathbf{\Gamma}_j'\mathbf{\Gamma}_j$ are $O(np)$. However, the condition $\lambda_{j,q}/q^2 > n^{\delta_1}$ indicates that the orders of the eigen values of $\mathbf{\Gamma}_j'\mathbf{\Gamma}_j$ do not need to be the same orders uniformly. Moreover, in Theorem 3.3, k_*, γ_j and K do not have to be bounded. Hence, Theorem 3.3 can be applied to more non-restrictive situations than Theorem 3.2.

Although we have derived sufficient conditions for consistency in Theorem 3.3, it is hard to check from the conditions whether an information criterion considered is consistent. Hence, in order to establish an easy-to understand formula, we rewrite the conditions by using a limit inferior. Besides, by using 1 as q, we simplify condition C3-1 although the sufficient conditions become restrictive.

Corollary 3.2. *Suppose that assumption 1 holds. A variable selection method based on the IC_m is consistent under the HD-LM asymptotic framework if the following conditions are satisfied:*

C3-1′. $\inf_{j \in \mathcal{J}_-} \liminf_{c_{n,p} \to c_0, \mathrm{LM}} \frac{\log \lambda_{j,1}}{\log n} > 0$, *and*

$$\inf_{j \in \mathcal{J}_-} \liminf_{c_{n,p} \to c_0, \mathrm{LM}} \left\{ \log \frac{\lambda_{j,1}}{n} - \frac{m(j_+) - m(j)}{n} \right\} > \log(1 - c_0).$$

C3-2. $\inf_{j \in \mathcal{J}_+ \backslash \{j_*\}} \liminf_{c_{n,p} \to c_0, \mathrm{LM}} \frac{m(j) - m(j_*)}{p(k_j - k_*)} > -\frac{1}{c_0} \log(1 - c_0)$.

The proof is given in Appendix A.9. It notes that not "min" but "inf" is used for conditions C3-1′ and -2 because the number of candidate models may go to ∞. Although we cannot check whether condition C3-1′ is satisfied from an actual data, we can derive the order of the divergence speed of the maximum eigen value of $\mathbf{\Gamma}_j'\mathbf{\Gamma}_j$ to ensure a consistency. If the size of the order is small, we can consider that a possibility that an information criterion is consistent is high.

Hence, the size of the order helps to assess a quality of an information criterion in the sense of a possibility to have consistency.

Even if k_* is not bounded, Theorem 3.3 and Corollary 3.2 hold. However, in order to clarify the sufficient conditions to ensure consistency, we consider the simple case that k_* is bounded. Then, conditions to satisfy consistency properties of specific criteria are simplified as the following corollary (the proof is given in Appendix A.10):

Corollary 3.3. *Suppose that assumption 1 is satisfied, and k_* is bounded.*

(*i*) *A variable selection method based on the AIC is consistent when $c_{n,p} \to c_0$ under LM if $c_0 \in [0, c_{\mathrm{a}})$ and*

$$\inf_{j \in \mathcal{J}_-} \liminf_{c_{n,p} \to c_0, \mathrm{LM}} \log \frac{\lambda_{j,1}}{n} > \log(1 - c_0) + 2k_* c_0, \tag{3.5}$$

where c_{a} is the constant given by (3.2).

(*ii*) *A variable selection method based on the AIC_c is consistent when $c_{n,p} \to c_0$ under LM if*

$$\inf_{j \in \mathcal{J}_-} \liminf_{c_{n,p} \to c_0, \mathrm{LM}} \log \frac{\lambda_{j,1}}{n} > \log(1 - c_0) + k_* c_0 \left\{ \frac{1}{1 - c_0} + \frac{1}{(1 - c_0)^2} \right\}. \tag{3.6}$$

(*iii*) *Variable selection methods based on the BIC and the CAIC are consistent when $c_{n,p} \to c_0$ under LM if*

$$\inf_{j \in \mathcal{J}_-} \liminf_{c_{n,p} \to c_0, \mathrm{LM}} \frac{\log(\lambda_{j,1}/n)}{\log n} > k_* c_0. \tag{3.7}$$

An example of the noncentrality matrix is shown in Appendix A.11. From Corollary 3.3, we can see that the $\mathrm{AIC_c}$ is consistent if $\lim_{c_{n,p} \to c0, \mathrm{LM}} \log(\lambda_{j,1}/n) = \infty$, and the BIC and the CAIC are consistent if $\lim_{c_{n,p} \to c0, \mathrm{LM}} \log(\lambda_{j,1}/n)/\log n = \infty$. Moreover, the AIC is consistent if $c_0 < c_{\mathrm{a}}$ and $\lim_{c_{n,p} \to c0, \mathrm{LM}} \log(\lambda_{j,1}/n) = \infty$. Hence, the AIC and the $\mathrm{AIC_c}$ has a superiority over the BIC and the CAIC in the sense of a possibility to have a consistency. Moreover, although the AIC is consistent under the restriction $c_0 < c_{\mathrm{a}}$, there is no such a restriction in $\mathrm{AIC_c}$. Consequently, we can judge that the $\mathrm{AIC_c}$ has a superiority over the AIC, BIC and CAIC in the sense of a possibility to have a consistency.

4. Numerical study

In this section, we compare with the probabilities of selecting the true model by AIC, $\mathrm{AIC_c}$, BIC and CAIC in (2.2), which were evaluated by Monte Carlo simulations based on 10,000 replications under several different values of n and p. A set of candidate models was $\mathcal{J} = \{j_1, \ldots, j_k\}$, where $j_\alpha = \{1, \ldots, \alpha\}$ $(\alpha = 1, \ldots, k)$. A 1000×156 matrix $\boldsymbol{M}\boldsymbol{\Phi}(156)^{1/2}$ was generated, where an each element of $\boldsymbol{M}$ was independent and identically chosen from $U(-1, 1)$, and $\boldsymbol{\Phi}(q)$ is a $q \times q$ symmetric matrix whose the (a, b)th element was defined by $(0.8)^{|a-b|}$. Using this matrix, we constructed an $n \times k$ matrix of explanatory variables $\boldsymbol{X}$ as

TABLE 1
Selection probabilities of the true model (%)

	Case 1						Case 2 ($c_0 = 0.02$)					
n	p	k	AIC	AIC_c	BIC	CAIC	p	k	AIC	AIC_c	BIC	CAIC
100	2	10	75.8	84.5	98.7	99.5	2	10	75.8	84.5	98.7	99.5
200	2	10	79.0	83.2	99.5	99.9	4	10	86.0	90.3	99.9	100.0
500	2	10	79.4	81.2	99.8	99.9	10	10	96.3	97.4	100.0	100.0
1000	2	10	79.1	79.9	99.9	100.0	20	10	99.3	99.6	100.0	100.0
∞	2	10	80.2	80.2	100.0	100.0	∞	10	100.0	100.0	100.0	100.0
	Case 3						Case 4 ($c_0 = 0.1$)					
n	p	k	AIC	AIC_c	BIC	CAIC	p	k	AIC	AIC_c	BIC	CAIC
100	10	10	92.9	99.3	100.0	100.0	10	10	92.9	99.3	100.0	100.0
200	10	10	95.1	98.3	100.0	100.0	20	10	98.6	99.9	100.0	100.0
500	10	10	96.3	97.5	100.0	100.0	50	10	100.0	100.0	100.0	100.0
1000	10	10	96.5	97.1	100.0	100.0	100	10	100.0	100.0	100.0	100.0
∞	10	10	96.8	96.8	100.0	100.0	∞	10	100.0	100.0	100.0	100.0
	Case 5						Case 6 ($c_0 = 0.3$)					
n	p	k	AIC	AIC_c	BIC	CAIC	p	k	AIC	AIC_c	BIC	CAIC
100	30	10	97.0	40.1	0.0	0.0	30	10	97.0	40.1	0.0	0.0
200	30	10	99.4	100.0	100.0	55.8	60	10	99.8	100.0	0.0	0.0
500	30	10	99.7	100.0	100.0	100.0	150	10	100.0	100.0	0.0	0.0
1000	30	10	99.8	99.9	100.0	100.0	300	10	100.0	100.0	0.0	0.0
∞	30	10	99.9	99.9	100.0	100.0	∞	10	100.0	100.0	0.0	0.0
	Case 7 ($c_0 = 0.0$)						Case 8 ($c_0 = 0.0$)					
100	30	10	97.0	40.1	0.0	0.0	30	10	97.0	40.1	0.0	0.0
200	32	10	99.4	100.0	100.0	15.8	40	10	99.8	100.0	59.5	0.0
500	35	10	99.9	100.0	100.0	100.0	50	10	100.0	100.0	100.0	100.0
1000	40	10	100.0	100.0	100.0	100.0	60	10	100.0	100.0	100.0	100.0
∞	∞	10	100.0	100.0	100.0	100.0	∞	10	100.0	100.0	100.0	100.0
	Case 9 ($c_0 = 0.1$)						Case 10 ($c_0 = 0.1$)					
100	10	10	92.9	99.3	100.0	100.0	10	10	92.9	99.3	100.0	100.0
200	20	14	98.8	99.9	100.0	100.0	20	31	98.7	100.0	100.0	100.0
500	50	22	100.0	100.0	100.0	100.0	50	84	100.0	100.0	100.0	100.0
1000	100	31	100.0	100.0	100.0	100.0	100	156	100.0	100.0	100.0	100.0
∞	∞	∞	100.0	100.0	—	—	∞	∞	100.0	100.0	—	—
	Case 11 ($c_0 = 0.3$)						Case 12 ($c_0 = 0.3$)					
100	30	10	97.0	40.1	0.0	0.0	30	10	97.0	40.1	0.0	0.0
200	60	14	99.8	100.0	0.0	0.0	60	31	99.7	100.0	0.0	0.0
500	150	22	100.0	100.0	0.0	0.0	150	84	100.0	100.0	0.0	0.0
1000	300	31	100.0	100.0	0.0	0.0	300	156	100.0	100.0	0.0	0.0
∞	∞	∞	100.0	100.0	—	—	∞	∞	100.0	100.0	—	—

a submatrix of $\boldsymbol{M}\boldsymbol{\Phi}(156)^{1/2}$ from rows 1 to n and columns 1 to k. The true model was determined by $\boldsymbol{\Theta}_* = (1, 1, 3, -4, 5)'\mathbf{1}_p'$, $j_* = \{1, 2, 3, 4, 5\}$, and $\boldsymbol{\Sigma}_* = \boldsymbol{\Phi}(p)$, where $\mathbf{1}_p$ was the p-dimensional vector of ones. Thus, j_α with $\alpha = 1, \ldots, 4$ was the underspecified model, and j_α with $\alpha \geq 5$ was the overspecified model.

In our numerical study, $\gamma_j = 1$ and $\max_{j \in \mathcal{S}_-}(k_* - k_j) = 4$ hold. 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 BIC and the CAIC were not consistent in variable selection when $c_0 > 1/4$ under the fixed k.

Table 1 shows the probability of selecting the true model by the AIC, AIC_c, BIC, and CAIC. For $n = \infty$ or $p = \infty$, we list the theoretical values obtained

from Theorems 3.1, 3.2 and 3.3. A symbol "—" means that theoretical values are unclear because the sufficient condition for consistency does not hold. In the table, Cases 1, 3, and 5 are the results when $n \to \infty$ under fixed p and $k = 10$, and Cases 2, 4, 6, 7, and 8 are the results when $(n, p) \to \infty$ under a fixed $k = 10$ and with $c_0 = 0.02, 0.1, 0.3, 0.0$, and 0.0. Moreover, Cases 9, 10, 11 and 12 are the results when $(n, p) \to \infty$ and with $c_0 = 0.1$ and 0.3, and $k = 10 + [n^{1/2} - 10^{1/2}]$ and $k = 10 + [n^{3/4} - 10^{3/4}]$, where [ ] is the Gauss' symbol. 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 considered, 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 considered, the higher the selection probabilities became, with the exception of Case 6. This was because variable selection methods based on 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. Moreover, we can see that above tendencies were satisfied even if the number of explanatory variables becomes large.

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 there is the case 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 the HD asymptotic framework. Although the AIC becomes consistent under the restriction $c_0 < c_\mathrm{a}$, the $\mathrm{AIC_c}$ becomes consistent without the restriction of c_0. This indicates that it is possible that correcting the bias to the risk function may have a positive effect on the selection of the true model. It is a well-known fact that variable selection methods based on the BIC and the CAIC are consistent if we approximate the probability of selecting the true model using the LS 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 the HD asymptotic framework.

It is known that the LS asymptotic theory gives a poor approximation when the dimension is large. The HD asymptotic theory gives a better approximation than the LS asymptotic theory when the sample size and the dimension are large, and sometimes even when the dimension is not so large. Hence, the consistency property of the AIC that we demonstrated will be useful for high-dimensional data analysis. Usually, the HD 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 HD asymptotic theory to high-dimensional data.

From the simulation study, we found that, the larger the dimension and sample size considered, the higher the selection probabilities became. 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 $\hat{\boldsymbol{\Sigma}}_j$ becomes singular when $p > n$. Unfortunately, $n > p$ is not always satisfied in the actual data. If our results can be extended to the case of $n \leq p$, we clarify the conditions to satisfy consistency property in many infinite-dimensional statistics, e.g., the time series analysis (see [4, 5]), spatiotemporal geostatistical analysis (see [7, 8]) and functional data analysis (see [23]). The singularity of $\hat{\boldsymbol{\Sigma}}_j$ can be avoided by using a ridge-type estimator of the covariance matrix, as demonstrated by [36]. 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 3.1

Recall that the LS asymptotic framework is used for proving Theorem 3.1. We can see that $\hat{\boldsymbol{\Sigma}}_j \xrightarrow{p} \boldsymbol{\Sigma}_*$ as $n \to \infty$ holds when $j \in \mathcal{J}_+$ and $\hat{\boldsymbol{\Sigma}}_j \xrightarrow{p} \boldsymbol{\Sigma}_*^{1/2}\boldsymbol{\Psi}_{j,0}\boldsymbol{\Sigma}_*^{1/2} + \boldsymbol{\Sigma}_*$ as $n \to \infty$ holds when $j \in \mathcal{J}_-$, where $\boldsymbol{\Psi}_{j,0} = \lim_{n\to\infty} n^{-1}\boldsymbol{\Gamma}_j\boldsymbol{\Gamma}_j'$ and $\boldsymbol{\Gamma}_j$ is given by (2.4). Notice that $\boldsymbol{\Psi}_{j,0}$ is a positive semidefinite matrix. When $\lim_{n\to\infty}\{m(j) - m(j_*)\}/n = 0$ for all $j \in \mathcal{J}_-$, we have

$$\begin{aligned}\frac{1}{n}\{\mathrm{IC}_m(j) - \mathrm{IC}_m(j_*)\} &\xrightarrow{p} \log\det(\boldsymbol{\Sigma}_*^{1/2}\boldsymbol{\Psi}_{j,0}\boldsymbol{\Sigma}_*^{1/2} + \boldsymbol{\Sigma}_*) - \log\det(\boldsymbol{\Sigma}_*)\\ &= \log\det(\boldsymbol{I}_p + \boldsymbol{\Psi}_{j,0}) > 0.\end{aligned}$$

This result implies that $\lim_{n\to\infty} P(\mathrm{IC}_m(j_*) > \mathrm{IC}_m(j)) = 0$ for any $j \in \mathcal{J}_-$. Thus, we obtain

$$\begin{aligned}&\lim_{n\to\infty} P(\hat{j}_m = j)\\ &= \begin{cases} 0 & (j \in \mathcal{J}_-)\\ \lim\limits_{n\to\infty} P(\cap_{\ell\in\mathcal{J}_+\setminus\{j\}}\{\mathrm{IC}_m(\ell) > \mathrm{IC}_m(j)\}) & (j \in \mathcal{J}_+) \end{cases}.\end{aligned} \tag{A.1}$$

From here to the end of proof, we assume $j, \ell \in \mathcal{J}_+$. 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

$$\boldsymbol{\mathcal{E}} = (\boldsymbol{Y} - \boldsymbol{X}_*\boldsymbol{\Theta}_*)\boldsymbol{\Sigma}_*^{-1/2}. \tag{A.2}$$

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 $\mathrm{IC}_m(j)$ can be expanded as

$$\mathcal{L}(j) = n\log\det(\boldsymbol{\Sigma}_*) + \sqrt{n}\mathrm{tr}(\boldsymbol{V}) - \{\mathrm{tr}(\boldsymbol{V}^2)/2 + \mathrm{tr}(\boldsymbol{Z}_j'\boldsymbol{Z}_j)\} + O_p(n^{-1/2}).$$

Let $\boldsymbol{z}_j$ be a k_jp-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., [18, 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

$$\mathrm{IC}_m(\ell) - \mathrm{IC}_m(j) = -(\boldsymbol{z}_\ell'\boldsymbol{z}_\ell - \boldsymbol{z}_j'\boldsymbol{z}_j) + m(\ell) - m(j) + O_p(n^{-1/2}). \tag{A.3}$$

Hence, when $\lim_{n\to\infty}\{m(j) - m(j_*)\} = \infty$ holds for all $j \in \mathcal{J}_+\backslash\{j_*\}$ we derive

$$\frac{1}{m(j) - m(j_*)}\{\mathrm{IC}_m(j) - \mathrm{IC}_m(j_*)\} \xrightarrow{p} 1 > 0. \tag{A.4}$$

On the other hand, when $\lim_{n\to\infty} m(j) = m_0\{pk_j + p(p+1)/2\} < \infty$ holds, $\lim_{n\to\infty}\{m(\ell) - m(j)\} = m_0p(k_\ell - k_j)$ is satisfied. Consequently, by combining this result, and (A.3) and (A.4) with (A.1), Theorem 3.1 is proved.

A.2. The proof of Theorem 3.2

At first, we describe the lemma which is used for proving Theorems 3.2 and 3.3 (the proof of lemma is given after this subsection).

Lemma A.1. *Let $T = -l(pq)^{-1}\log\Lambda$ where Λ is distributed according to the Wilks' lambda distribution $\Lambda_q(p, l+q)$, and let $\kappa_T^{(s)}$ be the sth order cumulant of T. Suppose that $p/l \to \alpha$ (constant) and $q/l \to 0$. If $\alpha > 0$ then*

$$\kappa_T^{(1)} \to \frac{1}{\alpha}\log(1+\alpha),$$
$$(ql)^{s-1}\kappa_T^{(s)} \to \frac{2^{s-1}(s-2)!}{c}\left\{1 - \left(\frac{1}{1+c}\right)^{s-1}\right\} \quad (s \geq 2).$$

If $\alpha = 0$ then $\kappa_T^{(1)} \to 1$ and $(pq)^{s-1}\kappa_T^{(s)} \to 2^{s-1}(s-1)!$. Hence whether $\alpha = 0$ or $\alpha > 0$ for any positive integer m and any positive value δ,

$$P(|T - \kappa_T^{(1)}| > \delta) \leq \frac{1}{\delta^{2m}}E[(T - \kappa_T^{(1)})^{2m}] = O((pq)^{-m}).$$

Recall that the HD asymptotic framework is used for proving Theorem 3.2. First, we consider the case of $j \in \mathcal{J}_-$. Let

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

Notice that $\mathrm{rank}(\boldsymbol{\mathcal{A}}_j) = \gamma_j$ because $\boldsymbol{\mathcal{A}}_j'\boldsymbol{\mathcal{A}}_j = \boldsymbol{\Gamma}_j\boldsymbol{\Gamma}_j'$ and $\mathrm{rank}(\boldsymbol{\Gamma}_j) = \gamma_j$, where $\boldsymbol{\Gamma}_j$ is given by (2.4). By using a singular value decomposition, $\boldsymbol{\mathcal{A}}_j$ can be rewritten as

$$\boldsymbol{\mathcal{A}}_j = \boldsymbol{H}_j\boldsymbol{L}_j^{1/2}\boldsymbol{G}_j',$$

where $\boldsymbol{H}_j$ and $\boldsymbol{G}_j$ are $n \times \gamma_j$ and $p \times \gamma_j$ matrices satisfying $\boldsymbol{H}_j'\boldsymbol{H}_j = \boldsymbol{I}_{\gamma_j}$ and $\boldsymbol{G}_j'\boldsymbol{G}_j = \boldsymbol{I}_{\gamma_j}$, respectively, and $\boldsymbol{L}_j$ is a $\gamma_j \times \gamma_j$ diagonal matrix whose diagonal elements are squared singular values of $\boldsymbol{\mathcal{A}}_j$. By using $\boldsymbol{\mathcal{A}}_j$ and $\boldsymbol{\mathcal{E}}$ given by (A.2), we have

$$n\boldsymbol{\Sigma}_*^{-1/2}\hat{\boldsymbol{\Sigma}}_j\boldsymbol{\Sigma}_*^{-1/2} = \boldsymbol{W}_1 + \boldsymbol{W}_2 + \boldsymbol{W}_3, \quad n\boldsymbol{\Sigma}_*^{-1/2}\hat{\boldsymbol{\Sigma}}_{j_+}\boldsymbol{\Sigma}_*^{-1/2} = \boldsymbol{W}_1, \tag{A.5}$$

where

$$\begin{aligned}
&\boldsymbol{W}_1 = \boldsymbol{\mathcal{E}}'(\boldsymbol{I}_n - \boldsymbol{P}_{j_+})\boldsymbol{\mathcal{E}}, \quad \boldsymbol{W}_2 = \boldsymbol{\mathcal{E}}'(\boldsymbol{P}_{j_+} - \boldsymbol{H}_j\boldsymbol{H}_j' - \boldsymbol{P}_j)\boldsymbol{\mathcal{E}},\\
&\boldsymbol{W}_3 = (\boldsymbol{\mathcal{A}}_j + \boldsymbol{\mathcal{E}})'\boldsymbol{H}_j\boldsymbol{H}_j'(\boldsymbol{\mathcal{A}}_j + \boldsymbol{\mathcal{E}}).
\end{aligned}$$

It follows from the equations $\boldsymbol{P}_{j_*}\boldsymbol{X}_* = \boldsymbol{X}_*$ and $\boldsymbol{P}_j\boldsymbol{P}_{j_+} = \boldsymbol{P}_j$ that

$$\begin{aligned}
\boldsymbol{\mathcal{A}}_j'\boldsymbol{P}_{j_+} &= \boldsymbol{\Sigma}_*^{-1/2}\boldsymbol{\Theta}_*'\boldsymbol{X}_*'(\boldsymbol{P}_{j_+} - \boldsymbol{P}_j) = \boldsymbol{\Sigma}_*^{-1/2}\boldsymbol{\Theta}_*'\boldsymbol{X}_*'\boldsymbol{P}_{j_+}(\boldsymbol{I}_n - \boldsymbol{P}_j)\\
&= \boldsymbol{\Sigma}_*^{-1/2}\boldsymbol{\Theta}_*'\boldsymbol{X}_*'(\boldsymbol{I}_n - \boldsymbol{P}_j) = \boldsymbol{\mathcal{A}}_j'.
\end{aligned}$$

Using this result and $\boldsymbol{\mathcal{A}}_j'\boldsymbol{P}_j = \boldsymbol{O}_{p,n}$ yields $\boldsymbol{H}_j'\boldsymbol{P}_j = \boldsymbol{O}_{\gamma_j,n}$ and $\boldsymbol{H}_j'\boldsymbol{P}_{j_+} = \boldsymbol{H}_j'$. These imply that

$$\begin{aligned}
&\boldsymbol{H}_j\boldsymbol{H}_j'(\boldsymbol{I}_n - \boldsymbol{P}_{j_+}) = \boldsymbol{O}_{n,n}, \ \boldsymbol{H}_j\boldsymbol{H}_j'(\boldsymbol{P}_{j_+} - \boldsymbol{H}_j\boldsymbol{H}_j' - \boldsymbol{P}_j) = \boldsymbol{O}_{n,n},\\
&(\boldsymbol{I}_n - \boldsymbol{P}_{j_+})(\boldsymbol{P}_{j_+} - \boldsymbol{H}_j\boldsymbol{H}_j' - \boldsymbol{P}_j) = \boldsymbol{O}_{n,n},\\
&(\boldsymbol{P}_{j_+} - \boldsymbol{H}_j\boldsymbol{H}_j' - \boldsymbol{P}_j)^2 = \boldsymbol{P}_{j_+} - \boldsymbol{H}_j\boldsymbol{H}_j' - \boldsymbol{P}_j.
\end{aligned}$$

From the above results and the multivariate version of the Cochran theorem (see, e.g., [30, chap. 2.8]), we can see that $\boldsymbol{W}_1$, $\boldsymbol{W}_2$, and $\boldsymbol{W}_3$ are $p \times p$ mutually independent random matrices distributed according to the Wishart or the noncentral Wishart distributions;

$$\boldsymbol{W}_1 \sim W_p(n - k_{j_+}, \boldsymbol{I}_p), \quad \boldsymbol{W}_2 \sim W_p(d_j, \boldsymbol{I}_p), \quad \boldsymbol{W}_3 \sim W_p(\gamma_j, \boldsymbol{I}_p; \boldsymbol{\Gamma}_j\boldsymbol{\Gamma}_j'), \tag{A.6}$$

where $d_j = k_{j_+} - k_j - \gamma_j$. It follows from (A.5) and (A.6), and the property of the Wishart distributions (see [16, p. 57 th. 3.2.4]) that

$$\begin{aligned}
\frac{1}{n}\{\mathcal{L}(j) - \mathcal{L}(j_+)\} &= \log\frac{\det(\boldsymbol{W}_1 + \boldsymbol{W}_2 + \boldsymbol{W}_3)}{\det(\boldsymbol{W}_1 + \boldsymbol{W}_2)} + \log\frac{\det(\boldsymbol{W}_1 + \boldsymbol{W}_2)}{\det(\boldsymbol{W}_1)}\\
&= -\log\frac{\det(\boldsymbol{U}_1)}{\det(\boldsymbol{U}_1 + \boldsymbol{U}_2)} - \log\frac{\det(\boldsymbol{U}_3)}{\det(\boldsymbol{U}_3 + \boldsymbol{U}_4)},
\end{aligned} \tag{A.7}$$

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

$$\begin{aligned} &\boldsymbol{U}_1 \sim W_{\gamma_j}(n-k_j-p, \boldsymbol{I}_{\gamma_j}), \qquad \boldsymbol{U}_2 \sim W_{\gamma_j}(p, \boldsymbol{I}_{\gamma_j}; \boldsymbol{\Gamma}_j'\boldsymbol{\Gamma}_j), \\ &\boldsymbol{U}_3 \sim W_{d_j}(n-k_j-\gamma_j-p, \boldsymbol{I}_{d_j}), \ \boldsymbol{U}_4 \sim W_{d_j}(p, \boldsymbol{I}_{d_j}). \end{aligned} \tag{A.8}$$

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

$$\frac{1}{n-k_j-p}\boldsymbol{U}_1 \xrightarrow{p} \boldsymbol{I}_{\gamma_j}, \quad \frac{1}{n-k_j-\gamma_j-p}\boldsymbol{U}_3 \xrightarrow{p} \boldsymbol{I}_{d_j}, \quad \frac{1}{p}\boldsymbol{U}_4 \xrightarrow{p} \boldsymbol{I}_{d_j}. \tag{A.9}$$

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 $\boldsymbol{Z} \sim N_{p\times\gamma_j}(\boldsymbol{O}_{p,\gamma_j}, \boldsymbol{I}_{p\gamma_j})$. Let $\boldsymbol{\Gamma}_j'\boldsymbol{\Gamma}_j)^{1/2} = \boldsymbol{\Delta}_j$. Then, we have

$$\boldsymbol{\Delta}_j^{-1}\boldsymbol{U}_2\boldsymbol{\Delta}_j^{-1} = \boldsymbol{\Delta}_j^{-1}\left(\boldsymbol{Z}'\boldsymbol{Z} + \boldsymbol{\Gamma}_j'\boldsymbol{Z} + \boldsymbol{Z}'\boldsymbol{\Gamma}_j + \boldsymbol{\Delta}_j^2\right)\boldsymbol{\Delta}_j^{-1}. \tag{A.10}$$

Recall that that $\limsup_{c_{n,p}\to c_0} np\lambda_{j,\gamma}^{-1} < \infty$ is derived from assumption 3. It follows from the above result and $E[\boldsymbol{Z}'\boldsymbol{Z}] = p\boldsymbol{I}_{\gamma_j}$ that

$$E\left[\mathrm{tr}\left(\boldsymbol{\Delta}_j^{-1}\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{\Delta}_j^{-1}\right)\right] = p\mathrm{tr}(\boldsymbol{\Delta}_j^{-2}) \le \frac{p\gamma_j}{\lambda_{j,\gamma_j}} \to 0.$$

This equation implies that

$$\boldsymbol{\Delta}_j^{-1}\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{\Delta}_j^{-1} \xrightarrow{p} \boldsymbol{O}_{\gamma_j,\gamma_j}. \tag{A.11}$$

Moreover, it is easy to see that $E[\boldsymbol{\Gamma}_j'\boldsymbol{Z}] = \boldsymbol{O}_{\gamma_j,\gamma_j}$ and $E[\boldsymbol{Z}'\boldsymbol{\Gamma}_j] = \boldsymbol{O}_{\gamma_j,\gamma_j}$, and

$$E\left[\mathrm{tr}\left\{(\boldsymbol{\Delta}_j^{-1}\boldsymbol{\Gamma}_j'\boldsymbol{Z}\boldsymbol{\Delta}_j^{-1})'(\boldsymbol{\Delta}_j^{-1}\boldsymbol{\Gamma}_j'\boldsymbol{Z}\boldsymbol{\Delta}_j^{-1})\right\}\right] = \mathrm{tr}(\boldsymbol{I}_{\gamma_j})\mathrm{tr}(\boldsymbol{\Delta}_j^{-2}) \le \frac{\gamma_j}{\lambda_{j,\gamma_j}} \to 0.$$

These equations imply that

$$\boldsymbol{\Delta}_j^{-1}\boldsymbol{\Gamma}_j'\boldsymbol{Z}\boldsymbol{\Delta}_j^{-1} \xrightarrow{p} \boldsymbol{O}_{\gamma_j,\gamma_j}, \quad \boldsymbol{\Delta}_j^{-1}\boldsymbol{Z}'\boldsymbol{\Gamma}_j\boldsymbol{\Delta}_j^{-1} \xrightarrow{p} \boldsymbol{O}_{\gamma_j,\gamma_j}. \tag{A.12}$$

From (A.10), (A.11) and (A.12), we derive the convergence in probability of $\boldsymbol{\Delta}_j^{-1}\boldsymbol{U}_2\boldsymbol{\Delta}_j^{-1}$ as

$$\boldsymbol{\Delta}_j^{-1}\boldsymbol{U}_2\boldsymbol{\Delta}_j^{-1} \xrightarrow{p} \boldsymbol{I}_{\gamma_j}. \tag{A.13}$$

Notice that

$$E\left[\mathrm{tr}\left(\boldsymbol{\Delta}_j^{-1}\boldsymbol{U}_1\boldsymbol{\Delta}_j^{-1}\right)\right] = (n-k_j-p)\mathrm{tr}(\boldsymbol{\Delta}_j^{-2}) \le \frac{(n-k_j-p)\gamma_j}{\lambda_{j,\gamma_j}} \to 0.$$

This equation implies that

$$\boldsymbol{\Delta}_j^{-1}\boldsymbol{U}_1\boldsymbol{\Delta}_j^{-1} \xrightarrow{p} \boldsymbol{O}_{\gamma_j,\gamma_j}. \tag{A.14}$$

Combining the equations (A.9), (A.13) and (A.14) yields

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

Using the results of the convergence in probability, the first and second terms in (A.7) are expanded as

$$\begin{aligned} -\log\frac{\det(\boldsymbol{U}_1)}{\det(\boldsymbol{U}_1+\boldsymbol{U}_2)} &= \log\left(\frac{p}{1-c_{n,p}-k_j/n}\right)^{\gamma_j} + \log\det\{(np)^{-1}\boldsymbol{\Delta}_j^2\} \\ &\quad - \log\frac{\det\{\boldsymbol{U}_1/(n-k_j-p)\}}{\det\{\boldsymbol{\Delta}_j^{-1}(\boldsymbol{U}_1+\boldsymbol{U}_2)\boldsymbol{\Delta}_j^{-1}\}} \\ &= \gamma_j\log p - \gamma_j\log(1-c_0) \\ &\quad + \log\det\{(np)^{-1}\boldsymbol{\Delta}_j^2\} + o_p(1), \end{aligned} \tag{A.15}$$

and

$$\begin{aligned} -\log\frac{\det(\boldsymbol{U}_3)}{\det(\boldsymbol{U}_3+\boldsymbol{U}_4)} &= -\log\frac{\det\{\boldsymbol{U}_3/(n-k_j-\gamma_j-p)\}}{\det\{(\boldsymbol{U}_3+\boldsymbol{U}_4)/(n-k_j-\gamma_j-p)\}} \\ &= -d_j\log(1-c_0)+o_p(1). \end{aligned} \tag{A.16}$$

Notice that

$$\gamma_j\log\left(\frac{\lambda_{j,\gamma_j}}{np}\right) \le \log\det\left(\frac{1}{np}\boldsymbol{\Delta}_j^2\right) \le \gamma_j\log\left(\frac{\lambda_{j,1}}{np}\right).$$

It follows from the above result and assumption 3 that

$$\liminf_{c_{n,p}\to c_0}\log\det\left(\frac{1}{np}\boldsymbol{\Delta}_j^2\right) > 0, \quad \limsup_{c_{n,p}\to c_0}\log\det\left(\frac{1}{np}\boldsymbol{\Delta}_j^2\right) < \infty.$$

Therefore, we have

$$\frac{\log\det\left\{(np)^{-1}\boldsymbol{\Delta}_j^2\right\}}{\log p} \to 0.$$

Using the above equation after substituting the equations (A.15) and (A.16) into (A.7) yields

$$\frac{1}{n\log p}\{\mathcal{L}(j)-\mathcal{L}(j_+)\} \xrightarrow{p} \gamma_j > 0. \tag{A.17}$$

Using the same idea as in the derivation of (A.16), it can be shown that

$$\frac{1}{n\log p}\{\mathcal{L}(j_+)-\mathcal{L}(j_*)\} \xrightarrow{p} 0. \tag{A.18}$$

From the results (A.17) and (A.18), when condition C2-1 holds, the difference between the information criteria of the model j and the true model j_* is con-

vergent as

$$\begin{aligned}&\frac{1}{n\log p}\{\mathrm{IC}_m(j)-\mathrm{IC}_m(j_*)\}\\&=\frac{1}{n\log p}\{\mathcal{L}(j)-\mathcal{L}(j_+)+\mathcal{L}(j_+)-\mathcal{L}(j_*)+m(j)-m(j_*)\}\\&\xrightarrow{p}\gamma_j+\lim_{c_{n,p}\to c_0}\frac{m(j)-m(j_*)}{n\log p}>0.\end{aligned}\tag{A.19}$$

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

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

It follows from the property of the Wishart distributions (see [16, p. 57 th. 3.2.4]) that

$$\mathcal{L}(j)-\mathcal{L}(j_*)=n\log\Lambda,\tag{A.20}$$

where Λ is distributed according to the Wilks' lambda distribution $\Lambda_{r_j}(p,n-k_*-p)$ and $r_j=k_j-k_*$. By using Lemma A.1, we have

$$\begin{aligned}\frac{1}{p}\{\mathcal{L}(j)-\mathcal{L}(j_*)\}&=\left(-\frac{nr_j}{n-k_*-p-r_j}\right)\left(-\frac{n-k_*-p-r_j}{pr_j}\log\Lambda\right)\\&\xrightarrow{p}-\frac{r_j}{1-c_0}\left(\frac{c_0}{1-c_0}\right)^{-1}\log\left(1+\frac{c_0}{1-c_0}\right)=\frac{r_j}{c_0}\log(1-c_0).\end{aligned}$$

Therefore, when condition C2-2 holds, the difference between the information criteria of the model j and the true model j_* is convergent as

$$\begin{aligned}\frac{1}{p}\{\mathrm{IC}_m(j)-\mathrm{IC}_m(j_*)\}&=\frac{1}{p}\{\mathcal{L}(j)-\mathcal{L}(j_*)+m(j)-m(j_*)\}\\&\xrightarrow{p}\frac{r_j}{c_0}\log(1-c_0)+\lim_{c_{n,p}\to c_0}\frac{m(j)-m(j_*)}{p}>0.\end{aligned}\tag{A.21}$$

Consequently, from (A.19) and (A.21), Theorem 3.2 is proved.

A.3. The proof of Lemma A.1

The limiting values are easily obtained by using the following bounds for the cumulants of $-\log\Lambda$. Let $\kappa^{(s)}$ be the sth cumulant of $-\log\Lambda$ then

$$b^{(s)}(l+1,p,q)<\kappa^{(s)}<b^{(s)}(l-1/2,p,q),\quad(s=1,2,\dots),$$

$$b^{(1)}(l,p,q)=l\log\left(\frac{l}{l+q}\right)-(l+p)\log\left(\frac{l+p}{l+p+q}\right)-q\log\left(\frac{l+q}{l+p+q}\right),$$

$$b^{(2)}(l,p,q)=2\log\left\{1+\frac{pq}{(l+p+q)l}\right\},$$

$$b^{(s)}(l,p,q)=\frac{2^{s-1}(s-3)!}{l^{s-2}}\left\{1-\left(\frac{l}{l+q}\right)^{s-2}-\left(\frac{l}{l+p}\right)^{s-2}+\left(\frac{l}{l+p+q}\right)^{s-2}\right\}.$$

These bounds for $s \geq 2$ are given by [34]. The bound for $\kappa^{(1)}$ is also obtained by using the same method. The order of the $2m$th order central moment is given by

$$E[(T-\kappa_T^{(1)})^{2m}] = \sum_{l=1}^{m} \frac{(2m)!}{l!} \sum_{\substack{s_1+\cdots+s_l=2m \\ s_1\geq 2,\ldots,s_l\geq 2}} \frac{\kappa_T^{(s_1)}\cdots\kappa_T^{(s_l)}}{s_1!\cdots s_l!} = \sum_{l=1}^{m} O((pq)^{-(2m-l)}).$$

A.4. The proof of Corollary 3.1

Recall that the HD asymptotic framework is used for proving Corollary 3.1. First, we consider the cases of the AIC and the $\mathrm{AIC_c}$. Notice that $m(j)-m(j_*)$ in the $\mathrm{AIC_c}$ can be expanded as

$$\begin{aligned} m(j)-m(j_*) &= \frac{p(k_j-k_*)(2-c_{n,p}-1/n)}{\{1-c_{n,p}-(k_j+1)/n\}\{1-c_{n,p}-(k_*+1)/n\}} \\ &= \frac{(k_j-k_*)(2-c_{n,p})p}{(1-c_{n,p})^2} + O(pn^{-1}). \end{aligned} \tag{A.22}$$

Hence, differences between the penalty terms of the AICs and the $\mathrm{AIC_c}$s are convergent as

$$\lim_{c_{n,p}\to c_0} \frac{1}{n\log p}\{m(j)-m(j_*)\} = 0.$$

This indicates that condition C2-1 holds in AIC and $\mathrm{AIC_c}$. Furthermore, it follows from the equality (A.22) that

$$\lim_{c_{n,p}\to c_0} \frac{1}{p}\{m(j)-m(j_*)\} = \begin{cases} 2(k_j-k_*) & (\mathrm{AIC}) \\ (k_j-k_*)\{(1-c_0)^{-1}+(1-c_0)^{-2}\} & (\mathrm{AIC_c}) \end{cases}.$$

Notice that $c^{-1}\log(1-c)+2$ and $c^{-1}\log(1-c)+(1-c)^{-1}+(1-c)^{-2}$ are monotonically decreasing and increasing functions in $0 \leq c < 1$, respectively. Hence, when $j \in \mathcal{J}\backslash\{j_*\}$, the penalty terms in the $\mathrm{AIC_c}$ always satisfy condition C2-2 and these in AIC satisfy condition C2-2 if $c_0 \in [0, c_{\mathrm{a}})$, where c_{a} is a constant satisfying $\log(1-c_{\mathrm{a}})+2c_{\mathrm{a}} = 0$.

Next, we consider the cases of the BIC and the CAIC. When $j \in \mathcal{J}_+\backslash\{j_*\}$, the differences between the penalty terms of the BICs and the CAICs are

$$\lim_{c_{n,p}\to c_0} \frac{1}{p\log n}\{m(j)-m(j_*)\} = k_j-k_* > 0.$$

Thus, condition C2-2 holds. Moreover, it is easy to obtain

$$\frac{1}{n\log p}\{m(j)-m(j*)\} = \begin{cases} c_{n,p}(k_j-k_*)\left(-\frac{\log c_{n,p}}{\log p}+1\right) & (\mathrm{BIC}) \\ c_{n,p}(k_j-k_*)\left(\frac{1-\log c_{n,p}}{\log p}+1\right) & (\mathrm{CAIC}). \end{cases}$$

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

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

When $j \in \mathcal{S}_-^c \cap \mathcal{J}_-$, condition C2-1 is satisfied because $c_0(k_j - k_*) \geq 0$ holds. When $j \in \mathcal{S}_-$, condition C2-1 is satisfied if $c_0 < \gamma_j/(k_* - k_j)$ holds for all $j \in \mathcal{S}_-$.

A.5. The proof of Theorem 3.3

At first, we describe two lemmas which are used for proving Theorem 3.3 (the proofs of lemmas are given after this subsection).

Lemma A.2. *Let χ^2_n be a random variable distributed according to the chi-square distribution with n degrees of freedom. If $z > n$,*

$$P(\chi^2_n > z) < \exp\left\{-\frac{z}{2}\left(1 - \frac{n}{z} - \frac{n}{z}\log\frac{n}{z}\right)\right\} < \exp\left[-\frac{z}{2}\left\{1 - \left(\frac{n}{z}\right)^2\right\}\right].$$

Lemma A.3. *Let $T = q^{-1}\log\det(\boldsymbol{V})$, where $\boldsymbol{V} \sim W_q(n, \boldsymbol{I}_q)$, and let $\kappa_T^{(s)}$ be the sth order cumulant of T. Then*

$$\log(n-q) < \kappa_T^{(1)} < \log\left(n - \frac{q-1}{2}\right). \tag{A.23}$$

Moreover if $q/n \to 0$ as $n \to \infty$

$$(qn)^{s-1}\kappa_T^{(s)} \to 2^{s-1}(s-2)! \quad (s \geq 2), \tag{A.24}$$

and hence for any positive integer l,

$$P(|T - \kappa_T^{(1)}| > \delta) \leq \frac{1}{\delta^{2l}} E[(T - \kappa_T^{(1)})^{2l}] = O((qn)^{-l}).$$

Recall that the HD-LM asymptotic framework is used for proving Theorem 3.3. First, we consider the case of $j \in \mathcal{J}_-$. Let $d_j = k_{j_+} - k_j$. As in the proof of Theorem 3.2, represent

$$\frac{1}{n}\{\mathcal{L}(j) - \mathcal{L}(j_+)\} = \log\frac{\det(\boldsymbol{W}_1 + \boldsymbol{W}_2)}{\det(\boldsymbol{W}_1)} = \log\frac{\det(\boldsymbol{U}_1 + \boldsymbol{U}_2)}{\det(\boldsymbol{U}_1)}, \tag{A.25}$$

where $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ are independent, also $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$ are independent, and

$$\begin{aligned} &\boldsymbol{W}_1 \sim W_p(n - k_{j_+}, \boldsymbol{I}_p), \qquad \boldsymbol{W}_2 \sim W_p(d_j, I_p; \boldsymbol{\Gamma}_j\boldsymbol{\Gamma}_j'), \\ &\boldsymbol{U}_1 \sim W_{d_j}(n - p - k_j, \boldsymbol{I}_{d_j}), \ \boldsymbol{U}_2 \sim W_{d_j}(p, \boldsymbol{I}_{d_j}; \boldsymbol{\Omega}_j), \end{aligned}$$

with $\boldsymbol{\Omega}_j = \mathrm{diag}(\lambda_{j,1}, \ldots, \lambda_{j,d_j})$. It should be kept in mind that we recycle some notations to denote different random matrices from those in the proof of Theorem 3.2. Let q be the integer for j in condition C3-1. Express

$$\boldsymbol{U}_2 = (\boldsymbol{Z} + \boldsymbol{\Gamma}_{j,q})(\boldsymbol{Z} + \boldsymbol{\Gamma}_{j,q})' + \boldsymbol{U}_3,$$

where $\boldsymbol{Z}$ and $\boldsymbol{U}_3$ are mutually independent randam matrices defined by

$$\boldsymbol{Z} \sim N_{d_j \times q}(\boldsymbol{O}_{d_j,q}, \boldsymbol{I}_{d_j} \otimes \boldsymbol{I}_q), \quad \boldsymbol{U}_3 \sim W_{d_j}(p - q, \boldsymbol{I}_{d_j}; \boldsymbol{\Omega}_j - \boldsymbol{\Gamma}_{j,q}\boldsymbol{\Gamma}_{j,q}'),$$

Here, $\boldsymbol{\Gamma}_{j,q} = (\boldsymbol{\Omega}_{j,q}^{1/2}, \boldsymbol{O}_{q,d_j-q})'$ and $\boldsymbol{\Omega}_{j,q} = \mathrm{diag}(\lambda_{j,1}, \ldots, \lambda_{j,q})$. Then

$$\begin{aligned} \frac{\det(\boldsymbol{U}_1 + \boldsymbol{U}_2)}{\det(\boldsymbol{U}_1)} &> \frac{\det\{\boldsymbol{U}_1 + (\boldsymbol{Z} + \boldsymbol{\Gamma}_{j,q})(\boldsymbol{Z} + \boldsymbol{\Gamma}_{j,q})'\}}{\det(\boldsymbol{U}_1)} \\ &= \frac{\det\{\boldsymbol{V}_1 + (\boldsymbol{Z} + \boldsymbol{\Gamma}_{j,q})'(\boldsymbol{Z} + \boldsymbol{\Gamma}_{j,q})\}}{\det(\boldsymbol{V}_1)}, \end{aligned} \tag{A.26}$$

where

$$\begin{aligned} \boldsymbol{V}_1 =& \{(\boldsymbol{Z} + \boldsymbol{\Gamma}_{j,q})'(\boldsymbol{Z} + \boldsymbol{\Gamma}_{j,q})\}^{1/2} \{(\boldsymbol{Z} + \boldsymbol{\Gamma}_{j,q})' \boldsymbol{U}_1^{-1} (\boldsymbol{Z} + \boldsymbol{\Gamma}_{j,q})\}^{-1} \\ & \times \{(\boldsymbol{Z} + \boldsymbol{\Gamma}_{j,q})'(\boldsymbol{Z} + \boldsymbol{\Gamma}_{j,q})\}^{1/2}. \end{aligned}$$

We can show that $\boldsymbol{V}_1$ and $\boldsymbol{Z}$ are independent, and $\boldsymbol{V}_1 \sim W_q(n - p - k_{j_+} + q, \boldsymbol{I}_q)$ (see [16, p. 57 th. 3.2.4]). Let $\delta_3 = \delta_2/2$ and $h = 1 - \exp(-\delta_2/2)$. Then $0 < h < 1$ and

$$\log \frac{\beta_{q,j}(1-h)}{n - p - k_{j_+} + (q+1)/2} - \frac{1}{qn}\{m(j_+) - m(j)\} > \delta_3, \tag{A.27}$$

If an event

$$A_{j,h} \ : \ \mathrm{tr}\{(\boldsymbol{\Gamma}_{j,q}'\boldsymbol{\Gamma}_{j,q})^{-1/2} \boldsymbol{\Gamma}_{j,q}' \boldsymbol{Z}\boldsymbol{Z}' \boldsymbol{\Gamma}_{j,q} (\boldsymbol{\Gamma}_{j,q}'\boldsymbol{\Gamma}_{j,q})^{-1/2}\} < h^2 \lambda_{j,q}/4,$$

occurs, then for any unit vector $\boldsymbol{b}$

$$\begin{aligned} & |\boldsymbol{b}'(\boldsymbol{\Gamma}_{j,q}'\boldsymbol{\Gamma}_{j,q})^{-1/2}(\boldsymbol{Z}'\boldsymbol{\Gamma}_{j,q} + \boldsymbol{\Gamma}_{j,q}'\boldsymbol{Z})(\boldsymbol{\Gamma}_{j,q}'\boldsymbol{\Gamma}_{j,q})^{-1/2}\boldsymbol{b}| \\ & \quad \le 2\{\boldsymbol{b}'(\boldsymbol{\Gamma}_{j,q}'\boldsymbol{\Gamma}_{j,q})^{-1}\boldsymbol{b}\boldsymbol{b}'(\boldsymbol{\Gamma}_{j,q}'\boldsymbol{\Gamma}_{j,q})^{-1/2}\boldsymbol{\Gamma}_{j,q}'\boldsymbol{Z}\boldsymbol{Z}'\boldsymbol{\Gamma}_{j,q}(\boldsymbol{\Gamma}_{j,q}'\boldsymbol{\Gamma}_{j,q})^{-1/2}\boldsymbol{b}\}^{1/2} \\ & \quad \le 2\left[\lambda_{j,q}^{-1}\mathrm{tr}\left\{(\boldsymbol{\Gamma}_{j,q}'\boldsymbol{\Gamma}_{j,q})^{-1/2}\boldsymbol{\Gamma}_{j,q}'\boldsymbol{Z}\boldsymbol{Z}'\boldsymbol{\Gamma}_{j,q}(\boldsymbol{\Gamma}_{j,q}'\boldsymbol{\Gamma}_{j,q})^{-1/2}\right\}\right]^{1/2} < h. \end{aligned}$$

Hence

$$\begin{aligned} & \det\{\boldsymbol{V}_1 + (\boldsymbol{Z} + \boldsymbol{\Gamma}_{j,q})'(\boldsymbol{Z} + \boldsymbol{\Gamma}_{j,q})\} \\ & > \det(\boldsymbol{Z}'\boldsymbol{\Gamma}_{j,q} + \boldsymbol{\Gamma}_{j,q}'\boldsymbol{Z} + \boldsymbol{\Gamma}_{j,q}'\boldsymbol{\Gamma}_{j,q}) \\ & > \det(\boldsymbol{\Gamma}_{j,q}'\boldsymbol{\Gamma}_{j,q})\det(\boldsymbol{I}_q - h\boldsymbol{I}_q) = \beta_{j,q}^q(1-h)^q, \end{aligned} \tag{A.28}$$

where $\beta_{j,q}$ is given by (3.4). Using (A.25), (A.26), (A.28) and (A.27), we obtain

$$\begin{aligned} & P(\hat{j}_m = j) \\ & \le P(\mathrm{IC}_m(j) - \mathrm{IC}_m(j_+) < 0) \\ & < P(A_{j,h}^c) \\ & \quad + P(q\log\beta_{j,q} + q\log(1-h) + \{m(j) - m(j_+)\}/n < \log\det(\boldsymbol{V}_1)) \\ & < P(A_{j,h}^c) \\ & \quad + P(\log\{n - p - k_{j_+} + (q+1)/2\} + \delta_3 < q^{-1}\log\det(\boldsymbol{V}_1)). \end{aligned} \tag{A.29}$$

Since $\mathrm{tr}\{(\boldsymbol{\Gamma}_{j,q}'\boldsymbol{\Gamma}_{j,q})^{-1/2}\boldsymbol{\Gamma}_{j,q}'\boldsymbol{Z}\boldsymbol{Z}'\boldsymbol{\Gamma}_{j,q}(\boldsymbol{\Gamma}_{j,q}'\boldsymbol{\Gamma}_{j,q})^{-1/2}\}$ is distributed according to the chi-square distribution with q^2 degrees of freedom, using Lemma A.2 and

condition C3-1 we obtain

$$P(A_{j,h}^c) < \exp\left[-\frac{h^2\lambda_{j,q}}{8}\left\{1-\left(\frac{4q^2}{h^2\lambda_{j,q}}\right)^2\right\}\right] < \exp\left[-\frac{h^2q^2n^{\delta_1}}{8}\left\{1-\left(\frac{4}{h^2n^{\delta_1}}\right)^2\right\}\right],$$

for sufficiently large n. Using Lemma A.3, we obtain

$$\begin{aligned} &P(\log\{n-p-k_{j_+}+(q+1)/2\}+\delta_3 < q^{-1}\log\det(\boldsymbol{V}_1)) \\ &< P(q^{-1}|\log\det(\boldsymbol{V}_1)-E[\log\det(\boldsymbol{V}_1)]| > \delta_3) \\ &= O(\{q(n-p-k_{j_+}+q)\}^{-l}). \end{aligned} \tag{A.30}$$

Next, we consider the case of $j \in \mathcal{J}_+\backslash\{j_*\}$. Let Λ be a random variable distributed according to $\Lambda_{r_j}(p, n-k_*-p)$, which is given in (A.20). From the equation in (3.3) and condition C3-2, by using Lemma A.1 we obtain that

$$\begin{aligned} &P(\hat{j}_m = j) \\ &\le P(-(r_jc_{n,p})^{-1}\log\Lambda > (r_jp)^{-1}\{m(j)-m(j_*)\}) \\ &\le P(-(r_jc_{n,p})^{-1}\log\Lambda + c_{n,p}^{-1}\log(1-c_{n,p}) > \delta) \\ &\le P((r_jc_{n,p})^{-1}|\log\Lambda - E[\log\Lambda]| > \delta/2) = O((r_jp)^{-l}), \end{aligned} \tag{A.31}$$

for sufficiently large n. In the last inequality, we used the fact that $E[(r_jc_{n,p})^{-1}\log\Lambda] + c_{n,p}^{-1}\log(1-c_{n,p}) \to 0$.

From (A.29), (A.30), (A.31) and the equation in (3.3),

$$\max_{j\in\mathcal{J}\backslash\{j_*\}} P(\hat{j}_m = j) = O(p^{-l}).$$

Hence the equation in (3.3) leads that

$$P(\hat{j}_m = j_*) = 1 - \sum_{j\in\mathcal{J}\backslash\{j_*\}} P(\hat{j}_m = j) \to 1.$$

A.6. The proof of Lemma A.2

Notice that

$$\begin{aligned} P(\chi_n^2 > z) &= \frac{1}{2\Gamma(n/2)}\int_z^\infty e^{-x/2}(x/2)^{n/2-1}dx \\ &= \frac{1}{2\Gamma(n/2)}\int_z^\infty e^{-rx/2}e^{-(1-r)x/2}(x/2)^{n/2-1}dx \\ &< \frac{e^{-rz/2}}{2\Gamma(a/2)}\int_0^\infty e^{-(1-r)x/2}(x/2)^{n/2-1}dx = \frac{e^{-rz/2}}{(1-r)^{n/2}}, \end{aligned}$$

where $\Gamma(x)$ is the gamma function. Taking the minimum with respect to r we get the first inequality. The second inequality can be obtained by the fact that $1+\log x < x$.

A.7. The proof of Lemma A.3

At first, we describe the lemma which is used for proving Lemma A.3 (the proof of lemma is given after this subsection).

Lemma A.4. *Let $\psi(a)$ be the digamma function defined by $\psi(a) = d\log\Gamma(a)/da$. Then, $\psi(z) > \log(z - 1/2)$ holds if $z > 1/2$ and $\psi(z) < \log z$ if $z > 0$.*

Using the fact that $\det(\boldsymbol{V}) \sim \prod_{i=1}^{q} \chi^2_{n-p+i}$, where $\chi^2_{n-p+1}, \ldots, \chi^2_{n-p+q}$ are independent random variables and χ^2_{n-p+i} is distributed according to the chi-square distribution with $n-p+i$ degrees of freedom (see [21, p. 100 th. 3.2.15]), the moment generating function of $\log\det(\boldsymbol{V})$ is given by

$$g(t) = \log E[\det(\boldsymbol{V})^t] = t(q\log 2) + \sum_{i=1}^{q}\sum_{s=1}^{\infty} \frac{t^s}{s!}\psi^{(s-1)}\left(\frac{n-q+i}{2}\right),$$

where $\psi^{(s)}(a)$ is the sth order derivative of $\psi(a)$. Hence the first order cumulant of $\log\det(\boldsymbol{V})$ is given by

$$\kappa^{(1)} = q\log 2 + \sum_{i=1}^{q} \psi\left(\frac{n-q+i}{2}\right).$$

Using Lemma A.4 and the fact that $\log x$ is increasing and concave function of x,

$$\begin{aligned} q\log(n-q) &< \sum_{i=1}^{q} \log(n-q+i-1) < \kappa^{(1)} \\ &< \sum_{i=1}^{q} \log(n-q+i) < q\log\left(n-q+\frac{q+1}{2}\right). \end{aligned}$$

The sth order cumulant of $\log\det(\boldsymbol{V})$ can be expressed as

$$\kappa^{(s)} = \sum_{i=1}^{q}\sum_{k=0}^{\infty} \frac{(-1)^s(s-1)!}{(\frac{n-q+i}{2}+k)^s}, \quad (s \geq 2).$$

Since $f(x,y) = 2^s(n-q+x+2y)^{-s}$ is a decreasing and convex function of x and y,

$$\begin{aligned} &\int_0^{\infty}\left\{\int_1^{q+1} \frac{2^s(s-1)!}{(n-q+x+2y)^s}dx\right\}dy < (-1)^s\kappa^{(s)} \\ &< \int_{-1/2}^{\infty}\left\{\int_{-1/2}^{q+1/2} \frac{2^s(s-1)!}{(n-q+x+2y)^s}dx\right\}dy. \end{aligned}$$

Calculating the integrals and taking the limits, we obtain (A.24). The $2l$th order central moment is the sum of the products of cumulants, $\kappa_T^{(s_1)}\cdots\kappa_T^{(s_k)}$ such that $s_1+\cdots+s_k = 2l$. Hence the order of the $2l$th order central moment is equal to the order of $(\kappa_T^{(2)})^l$.

A.8. The proof of Lemma A.4

The digamma function can be expressed as

$$\psi(z) = -C + \sum_{k=0}^{\infty} \left(\frac{1}{1+k} - \frac{1}{z+k} \right),$$

where C is the Euler's constant defined by $C = \lim_{n\to\infty}(\sum_{k=1}^{n} k^{-1} - \log n)$. Hence

$$\lim_{n\to\infty} \{\log n - \psi(n)\} = \lim_{n\to\infty} \{\log(n-1/2) - \psi(n)\} = 0. \tag{A.32}$$

Since $f(x) = (z+x)^{-2}$ is convex and decreasing function of x,

$$\begin{aligned} \frac{1}{z} &= \int_0^{\infty} \frac{1}{(z+x)^2} dx < \sum_{k=0}^{\infty} \frac{1}{(z+k)^2} = \frac{d}{dz}\psi(z) \\ &< \int_{-1/2}^{\infty} \frac{1}{(z+x)^2} dx = \frac{1}{(z-1/2)}. \end{aligned}$$

Hence $\log z - \psi(z)$ is decreasing function, and $\log(z-1/2) - \psi(z)$ is increasing function of z. Combining these properties with (A.32) we get the desired results.

A.9. The proof of Corollary 3.2

Recall that the HD-LM asymptotic framework is used for proving Corollary 3.2. Notice that $k_j/n \to 0$ holds when the equation in (3.3) holds. It is easy to see that condition C3-2 is rewritten as

$$\inf_{j\in\mathcal{J}_+\setminus\{j_*\}} \liminf_{c_{n,p}\to c_0,\mathrm{LM}} \left\{ \frac{m(j)-m(j_*)}{p(k_j-k_*)} + \frac{1}{c_{n,p}} \log(1-c_{n,p}) \right\} > 0.$$

Recall that $\beta_{j,1} = \lambda_{j,1}$. Hence, when $q = 1$, condition C3-1 is rewritten as

$$\begin{aligned} &\inf_{j\in\mathcal{J}_-} \liminf_{c_{n,p}\to c_0,\mathrm{LM}} \frac{\log \lambda_{j,1}}{\log n} > 0, \\ &\inf_{j\in\mathcal{J}_-} \liminf_{c_{n,p}\to c_0,\mathrm{LM}} \left\{ \log \frac{\lambda_{j,1}}{n-p-k_j+1} - \frac{m(j_+)-m(j)}{n} \right\} > 0. \end{aligned}$$

Notice that

$$\log \frac{\lambda_{j,1}}{n-p-k_j+1} = \log \frac{\lambda_{j,1}}{n} + \log \frac{n}{n-p-k_j+1},$$

and

$$\begin{aligned} \lim_{c_{n,p}\to c_0,\mathrm{LM}} \frac{1}{c_{n,p}} \log(1-c_{n,p}) &= \frac{1}{c_0}\log(1-c_0), \\ \lim_{c_{n,p}\to c_0,\mathrm{LM}} \log \frac{n}{n-p-k_j+1} &= -\log(1-c_0). \end{aligned}$$

Hence, Corollary 3.2 is proved.

A.10. The proof of Corollary 3.3

Recall that the HD-LM asymptotic framework is used for proving Corollary 3.3. Notice that $k_j/n \to 0$ holds for all $j \in \mathcal{J}$ when the equation in (3.3) holds. From the above result and (A.22), we derive a limit of $\{m(j) - m(j_*)\}/\{p(k_j - k_*)\}$ in each criterion as

$$\lim_{c_{n,p}\to c_0,\mathrm{LM}} \frac{m(j) - m(j)}{p(k_j - k_*)} = \begin{cases} 2 & (\mathrm{AIC}) \\ (1-c_0)^{-1} + (1-c_0)^{-2} & (\mathrm{AIC_c}) \\ \infty & (\mathrm{BIC, CAIC}) \end{cases}.$$

Therefore, condition C3-2 in the $\mathrm{AIC_c}$, BIC and CAIC holds, and that in the AIC holds if $c_0 < c_\mathrm{a}$. Moreover, $k_{j_+} - k_j$ is also bounded when k_* is bounded. It follows from $k_{j_+} - k_j \le k_*$ that

$$-\frac{m(j_+) - m(j)}{n} \ge \begin{cases} -2c_{n,p}k_* & (\mathrm{AIC}) \\ -\frac{c_{n,p}(2-c_{n,p}-1/n)}{\{1-c_{n,p}-(k_j+1)/n\}^2} & (\mathrm{AIC_c}) \\ -c_{n,p}k_* \log n & (\mathrm{BIC}) \\ -c_{n,p}k_*(1+\log n) & (\mathrm{CAIC}) \end{cases}.$$

Hence, we have

$$\liminf_{c_{n,p}\to c_0,\mathrm{LM}} -\frac{m(j_+) - m(j)}{n} = \begin{cases} -2k_*c_0 & (\mathrm{AIC}) \\ -k_*c_0\left\{(1-c_0)^{-1} + (1-c_0)^{-2}\right\} & (\mathrm{AIC_c}) \end{cases},$$

$$\liminf_{c_{n,p}\to c_0,\mathrm{LM}} -\frac{m(j_+) - m(j)}{n\log n} = -k_*c_0 \ (\mathrm{BIC, CAIC}).$$

Hence, condition C3-2 in the AIC and the $\mathrm{AIC_c}$ holds if (3.5) and (3.6) are satisfied, respectively, and that in the BIC and the CAIC holds if (3.7) is satisfied. Consequently, Corollary 3.2 is proved.

A.11. Example of the noncentrality matrix

Theoretically, it is natural to describe conditions in terms of the eigen values of the noncentrality matrices, since the unknown parameters $\boldsymbol{\Theta}_*$ and $\boldsymbol{\Sigma}_*$ affect the distribution of each criterion only through the eigen values. However, the effect of each $\boldsymbol{\Theta}_*$ and $\boldsymbol{\Sigma}_*$ may be of interest. So we illustrate the conditions by a two-way MANOVA model with a certain structure for the regression matrix and the covariance matrix of the error term.

Suppose there are two factors A and B with a levels and b levels, respectively. We consider the case that a characteristic is observed at time t ($t = 1, \ldots, p$) for m individuals in the cell (l, k) ($l = 1, \ldots, a; k = 1, \ldots, b$). Let $\boldsymbol{y}_{ilk} = (y_{ilk1}, \ldots, y_{ilkp})'$ be p-dimensional observation vector for the ith individual from the cell (l, k) ($i = 1, \ldots, m$). Then two-way MANOVA model is represented as

$$y_{ilkt} = \theta_t + \theta_{lt}^{(\mathrm{A})} + \theta_{kt}^{(\mathrm{B})} + \theta_{lkt}^{(\mathrm{AB})} + \epsilon_{ilkt},$$
$$(i = 1, \dots, m;\ l = 1, \dots, a;\ k = 1, \dots, b;\ t = 1, \dots, p),$$

where $\theta_{lt}^{(\mathrm{A})}$'s and $\theta_{kt}^{(\mathrm{B})}$'s are the main effects of the factor A and B, respectively, $\theta_{lkt}^{(\mathrm{AB})}$'s are the effects of the interaction, and $\boldsymbol{\epsilon} = (\epsilon_{ilk1}, \dots, \epsilon_{ilkp})' \sim N_p(\mathbf{0}_p, \boldsymbol{\Sigma})$. In order to assure the model identifiability, we assume that

$$\theta_{1t}^{(\mathrm{A})} = \theta_{1t}^{(\mathrm{B})} = \theta_{1kt}^{(\mathrm{AB})} = \theta_{l1t}^{(\mathrm{AB})} = 0,\ (i = 1, \dots, p;\ l = 1, \dots, a;\ k = 1, \dots, b).$$

Let

$$\boldsymbol{Y} = (\boldsymbol{y}_{111}, \dots, \boldsymbol{y}_{11m}, \boldsymbol{y}_{121}, \dots, \boldsymbol{y}_{12m}, \dots, \boldsymbol{y}_{ab1}, \dots, \boldsymbol{y}_{abm})',$$
$$\boldsymbol{X} = \left[\mathbf{1}_{abm} \;\middle|\; \underbrace{\begin{bmatrix} \mathbf{0}'_{a-1} \\ \boldsymbol{I}_{a-1} \end{bmatrix} \otimes \mathbf{1}_{bm}}_{\text{main effect of A}} \;\middle|\; \underbrace{\mathbf{1}_a \otimes \begin{bmatrix} \mathbf{0}'_{b-1} \\ \boldsymbol{I}_{b-1} \end{bmatrix} \otimes \mathbf{1}_m}_{\text{main effect of B}} \;\middle|\; \underbrace{\begin{bmatrix} \mathbf{0}'_{a-1} \\ \boldsymbol{I}_{a-1} \end{bmatrix} \otimes \begin{bmatrix} \mathbf{0}'_{b-1} \\ \boldsymbol{I}_{b-1} \end{bmatrix} \otimes \mathbf{1}_m}_{\text{interaction}} \right].$$

Then $\boldsymbol{Y} \sim N_{n \times p}(\boldsymbol{X}\boldsymbol{\Theta}, \boldsymbol{\Sigma} \otimes \boldsymbol{I}_n)$, where $n = abm$ and

$$\boldsymbol{\Theta} = (\boldsymbol{\theta}, \boldsymbol{\theta}_2^{(\mathrm{A})}, \dots, \boldsymbol{\theta}_a^{(\mathrm{A})}, \boldsymbol{\theta}_2^{(\mathrm{B})}, \dots, \boldsymbol{\theta}_b^{(\mathrm{B})}, \boldsymbol{\theta}_{22}^{(\mathrm{AB})}, \boldsymbol{\theta}_{23}^{(\mathrm{AB})}, \dots, \boldsymbol{\theta}_{ab}^{(\mathrm{AB})})',$$
$$\boldsymbol{\theta}_l^{(\mathrm{A})} = (\theta_{l1}^{(\mathrm{A})}, \dots, \theta_{lp}^{(\mathrm{A})})',$$
$$\boldsymbol{\theta}_k^{(\mathrm{B})} = (\theta_{k1}^{(\mathrm{B})}, \dots, \theta_{kp}^{(\mathrm{B})})', \qquad (l = 2, \dots, a;\ k = 2, \dots, b),$$
$$\boldsymbol{\theta}_{lk}^{(\mathrm{AB})} = (\theta_{lkt}^{(\mathrm{AB})}, \dots, \theta_{lkp}^{(\mathrm{AB})})',$$

Suppose that the true model has no interactions and the effects of the factor B are parallel, that is

$$\boldsymbol{\theta}_k^{(\mathrm{B})} = \eta_k \boldsymbol{\theta}_{\mathrm{B}}\ (k = 2, \dots, b)$$

with unknown parameters $\eta_2, \dots, \eta_b$ and a $p \times 1$ vector $\boldsymbol{\theta}_{\mathrm{B}} = (\theta_1^{(\mathrm{B})}, \dots, \theta_p^{(\mathrm{B})})'$. Hence the true model is represented as $j_* = \{1, 2, \dots, a + b - 1\}$ and

$$\boldsymbol{X}_* = \left[\mathbf{1}_{abm} \;\middle|\; \begin{bmatrix} \mathbf{0}'_{a-1} \\ \boldsymbol{I}_{a-1} \end{bmatrix} \otimes \mathbf{1}_{bm} \;\middle|\; \mathbf{1}_a \otimes \begin{bmatrix} \mathbf{0}'_{b-1} \\ \boldsymbol{I}_{b-1} \end{bmatrix} \otimes \mathbf{1}_m \right],$$
$$\boldsymbol{\Theta}_* = (\boldsymbol{\theta}, \boldsymbol{\theta}_2^{(\mathrm{A})}, \dots, \boldsymbol{\theta}_a^{(\mathrm{A})}, \boldsymbol{\theta}_{\mathrm{B}}\boldsymbol{\eta}')',$$

with $\boldsymbol{\eta} = (\eta_2, \dots, \eta_b)'$. In order to distinguish the variations due to the difference among individuals and the observation error we consider the case that $\epsilon_{ilk}^{(t)} = u_{ilk}^{(t)} + v_i$, where v_i, $u_{ilk}^{(t)}$ $(l = 1, \dots, a;\ k = 1, \dots, b;\ t = 1, \dots, p)$ are mutually independent, $u_{ilk}^{(t)} \sim N(0, \sigma^2)$, and $v_i \sim N(0, \tau^2)$. Then

$$\boldsymbol{\Sigma}_* = \sigma^2 \boldsymbol{I}_p + \tau^2 \mathbf{1}_p \mathbf{1}'_p.$$

Under the above setup, the noncentrality matrix of an underspecified model $j = \{1, 2, \dots, a\}$ is given by

$$\boldsymbol{\Gamma}_j \boldsymbol{\Gamma}'_j = am \boldsymbol{\Sigma}_*^{-1/2} \boldsymbol{\theta}_{\mathrm{B}} \boldsymbol{\eta}' (\boldsymbol{I}_{b-1} - b^{-1} \mathbf{1}_{b-1} \mathbf{1}'_{b-1}) \boldsymbol{\eta} \boldsymbol{\theta}'_{\mathrm{B}} \boldsymbol{\Sigma}_*^{-1/2}.$$

Note that the structures of the regression matrix and the covariance matrix are just for the illustration, and the user of the models does not assume these structures. So the number of the unknown parameters of the model j is $k_j p + p(p+1)/2$ with $k_j = a$.

Since $\gamma_j = \mathrm{rank}(\mathbf{\Gamma}_j) = 1$, $q = 1$ in condition C3-1, and some algebraic calculation leads that

$$\beta_{j,q} = \lambda_{j,1} = \mathbf{\Gamma}_j' \mathbf{\Gamma}_j = (np) Q_{\mathrm{B}} S_{\mathrm{B}},$$

where

$$Q_{\mathrm{B}} = \frac{1}{\sigma^2} \frac{1}{p} \sum_{t=1}^{p} (\theta_t^{(\mathrm{B})} - \bar{\theta}_{\mathrm{B}})^2 + \frac{1}{\sigma^2 + p\tau^2} \bar{\theta}_{\mathrm{B}}^2, \quad \bar{\theta}_{\mathrm{B}} = \frac{1}{p} \sum_{t=1}^{p} \theta_t^{(\mathrm{B})},$$

$$S_{\mathrm{B}} = \frac{1}{b} \sum_{k=1}^{b} (\eta_k - \bar{\eta})^2, \quad \eta_1 = 0, \ \bar{\eta} = \frac{1}{b} \sum_{k=1}^{b} \eta_k.$$

If the factor B is a kind of medication for example, the effect is likely to diminish as time goes by. In this case $Q_{\mathrm{B}} \to 0$ as $p \to \infty$. However Theorem 3.3 says that even in this case IC_m has a possibility to have the consistency property. Actually, Corollary 3.3 says that it is sufficient for the AIC and the $\mathrm{AIC}_{\mathrm{c}}$ having consistency if $pQ_{\mathrm{B}} \to \infty$.

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 and F. Csáki) 267–281. Akadémiai Kiadó, Budapest. MR0483125

[2] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automatic Control* **AC-19** 716–723. MR0423716

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

[4] Bosq, D. (2000). *Linear Processes in Function Spaces. Theory and Applications*. Springer-Verlag, New York. MR1783138

[5] Bosq, D. and Blanke, D. (2007). *Inference and Prediction in Large Dimensions*. John Wiley & Sons, Ltd., Paris. MR2364006

[6] Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika* **52** 345–370. MR0914460

[7] Christakos, G. (2000). *Modern Spatiotemporal Geostatistics*. Oxford University Press, New York.

[8] Cressie, N. and Wikle, C. K. (2011). *Statistics for Spatio-Temporal Data*. John Wiley & Sons, Inc., Hoboken. MR2848400

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

[10] Dien, S. J. V., Iwatani, S., Usuda, Y. and 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.
[11] Fujikoshi, Y. (1983). A criterion for variable selection in multiple discriminant analysis. *Hiroshima Math. J.* **13** 203–214. MR0693557
[12] 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. MR0797518
[13] Fujikoshi, Y. and Sakurai, T. (2009). High-dimensional asymptotic expansions for the distributions of canonical correlations. *J. Multivariate Anal.* **100** 231–242. MR2460489
[14] Fujikoshi, Y. and Satoh, K. (1997). Modified AIC and C_p in multivariate linear regression. *Biometrika* **84** 707–716. MR1603952
[15] Fujikoshi, Y. and 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. MR1631003
[16] Fujikoshi, Y., Shimizu, R. and Ulyanov, V. V. (2010). *Multivariate Statistics: High-Dimensional and Large-Sample Approximations*. John Wiley & Sons, Inc., Hoboken, New Jersey. MR2640807
[17] Fujikoshi, Y., Yanagihara, H. and 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. MR2255869
[18] Harville, D. A. (1997). *Matrix Algebra from a Statistician's Perspective*. Springer-Verlag, New York. MR1467237
[19] Kim, Y., Kwon, S. and Choi, H. (2012). Consistent model selection criteria on high dimensions. *J. Mach. Learn. Res.* **13** 1037–1057. MR2930632
[20] Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *Ann. Math. Statist.* **22** 79–86. MR0039968
[21] Muirhead, R. J. (1982). *Aspects of Multivariate Statistical Theory*. John Wiley & Sons, Inc., New York. MR0652932
[22] Nishii, R. (1984). Asymptotic properties of criteria for selection of variables in multiple regression. *Ann. Statist.* **12** 758–765. MR0740928
[23] Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis* (2nd. ed.). Springer, New York. MR2168993
[24] Sârbu, C., Onişor, C., Posa, M., Kevresan, S. and Kuhajda, K. (2008). Modeling and prediction (correction) of partition coefficients of bile acids and their derivatives by multivariate regression methods. *Talanta* **75** 651–657.
[25] Saxén, R. and 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.
[26] Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464. MR0468014

[27] Shao, J. (1997). An asymptotic theory for linear model selection. *Statist. Sinica* **7** 221–264. MR1466682
[28] Shibata, R. (1976). Selection of the order of an autoregressive model by Akaike's information criterion. *Biometrika* **63** 117–126. MR0403130
[29] Shibata, R. (1980). Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. *Ann. Statist.* **8** 147–164. MR0557560
[30] Siotani, M., Hayakawa, T. and Fujikoshi, Y. (1985). *Modern Multivariate Statistical Analysis: A Graduate Course and Handbook*. American Sciences Press, Columbus, Ohio. MR0832440
[31] Srivastava, M. S. (2002). *Methods of Multivariate Statistics*. John Wiley & Sons, New York. MR1915968
[32] Sugiura, N. (1978). Further analysis of the data by Akaike's information criterion and the finite corrections. *Commun. Statist. Theory Methods* **A7** 13–26.
[33] Timm, N. H. (2002). *Applied Multivariate Analysis*. Springer-Verlag, New York. MR1908225
[34] Wakaki, H. (2006). Edgeworth expansion of Wilks' lambda statistic. *J. Multivariate Anal.* **97** 1958–1964. MR2301267
[35] Woodroofe, M. (1982). On model selection and the arc sine laws. *Ann. Statist.* **10** 1182–1194. MR0673653
[36] Yamamura, M., Yanagihara, H. and Srivastava, M. S. (2010). Variable selection in multivariate linear regression models with fewer observations than the dimension. *Japanese J. Appl. Statist.* **39** 1–19.
[37] 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. MR2276149
[38] Yanagihara, H., Kamo, K. and Tonda, T. (2011). Second-order bias-corrected AIC in multivariate normal linear models under nonnormality. *Canad. J. Statist.* **39** 126–146. MR2815342
[39] 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
[40] Yoshimoto, A., Yanagihara, H. and Ninomiya, Y. (2005). Finding factors affecting a forest stand growth through multivariate linear modeling. *J. Jpn. For. Soc.* **87** 504–512 (in Japanese).