J. Japan Statist. Soc.
Vol. 45 No. 1 2015 21–56

CONDITIONS FOR CONSISTENCY OF A LOG-LIKELIHOOD-BASED INFORMATION CRITERION IN NORMAL MULTIVARIATE LINEAR REGRESSION MODELS UNDER THE VIOLATION OF THE NORMALITY ASSUMPTION

Hirokazu Yanagihara*

In this paper, we clarify conditions for consistency of a log-likelihood-based information criterion in multivariate linear regression models with a normality assumption. Although normality is assumed for the distribution of the candidate model, we frame the situation so that the assumption of normality may be violated. The conditions for consistency are derived from two types of asymptotic theory; one is based on a large-sample asymptotic framework in which only the sample size approaches ∞, and the other is based on a high-dimensional asymptotic framework in which the sample size and the dimension of the vector of response variables simultaneously approach infinity. In both cases, our results are independent of any indicator measuring a discrepancy between the true distribution and the normal distribution, e.g., skewness, kurtosis and other higher-order cumulants of the true distribution.

Key words and phrases: AIC, assumption of normality, bias-corrected AIC, BIC, consistent AIC, high-dimensional asymptotic framework, HQC, large-sample asymptotic framework, multivariate linear regression model, nonnormality, selection probability, variable selection.

1. Introduction

The multivariate linear regression model is one of the basic models of multivariate analysis. It is introduced in many multivariate statistics textbooks (see, e.g., Srivastava (2002, chap. 9), Timm (2002, chap. 4)), and is still widely used in chemometrics, engineering, econometrics, psychometrics, and many other fields, for the predication of multiple correlated responses to a set of explanatory variables (see, e.g., Yoshimoto *et al.* (2005), Dien *et al.* (2006), Saxén and Sundell (2006), Sárbu *et al.* (2008)). Let $\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)'$ be an $n \times p$ matrix of p correlated response variables, and let $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)'$ be an $n \times k$ matrix of k nonstochastic centralized explanatory variables ($\boldsymbol{X}'\mathbf{1}_n = \mathbf{0}_k$), where n is the sample size, $\mathbf{1}_n$ is an n-dimensional vector of ones, and $\mathbf{0}_k$ is a k-dimensional vector of zeros. In order to ensure the possibility of estimating the model and the existence of an information criterion, we assume that $\mathrm{rank}(\boldsymbol{X}) = k$ ($< n-1$) and $n - p - k - 2 > 0$. 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 $\boldsymbol{X}$

Received July 30, 2013. Revised April 4, 2015. Accepted April 6, 2015.

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

indexed by the elements of j, where k_A denotes the number of elements in a set A, i.e., $k_A = \#(A)$. For example, if $j = \{1, 2, 4\}$, then $\boldsymbol{X}_j$ consists of the first, second, and fourth columns of $\boldsymbol{X}$, and $k_j = 3$.

Since response variables are correlated, it is natural to be interested in the inter-relationships of the response variables and in simultaneous inference. We then consider the following multivariate linear regression model which incorporates the same subset of explanatory variables as the candidate model:

$$\boldsymbol{Y} \sim N_{n\times p}(\boldsymbol{1}_n\boldsymbol{\mu}' + \boldsymbol{X}_j\boldsymbol{\Theta}_j, \boldsymbol{\Sigma}_j \otimes \boldsymbol{I}_n), \tag{1.1}$$

where $\boldsymbol{\mu}$ is a p-dimensional unknown vector of location parameters, $\boldsymbol{\Theta}_j$ is a $k_j \times p$ unknown matrix of regression coefficients, and $\boldsymbol{\Sigma}_j$ is a $p \times p$ unknown covariance matrix. In this paper, we identify the candidate model by the set j and call the candidate model in (1.1) the model j. In particular, we represent the true subset of explanatory variables by a set j_* and call the model j_* the true model.

Since it is important to specify the factors being relevant to the response variables in a regression analysis, searching for the optimal subset j, i.e., variable selection, is essential. A variable selection through the candidate models in (1.1) is equivalent to extracting the subset of relevant explanatory variables that are active in at least one response variable, which is referred to as the support union problem (see Obozinski *et al.* (2008, 2011)), and a result from the variable selection helps to provide a faster and more cost-effective model, and a better understanding of the underlying process that generated the data.

A log-likelihood-based information criterion, which is defined by adding a penalty term that expresses the complexity of the model to a negative twofold maximum log-likelihood, is widely used for selecting the best subset of explanatory variables. The family of log-likelihood-based information criteria contains many widely known information criteria, e.g., Akaike's information criterion (AIC) proposed by Akaike (1973, 1974), the bias-corrected AIC ($\mathrm{AIC_c}$) proposed by Bedrick and Tsai (1994), the Bayesian information criterion (BIC) proposed by Schwarz (1978), the consistent AIC (CAIC) proposed by Bozdogan (1987), and the Hannan-Quinn information criterion (HQC) proposed by Hannan and Quinn (1979). We focus on selecting variables by minimizing the log-likelihood-based information criterion.

An important aspect of selecting variables in this way is whether or not the chosen information criterion is consistent, i.e., whether the asymptotic probability of selecting the true model j_* approaches 1. The consistency properties of various information criteria for multivariate models have been reported, e.g., see Fujikoshi (1983, 1985) and Yanagihara *et al.* (2015). The property is determined by ordinary asymptotic theory, which is based on the following framework:

- Large-sample (LS) asymptotic framework: the sample size approaches ∞ under a fixed dimension p.

Under the LS asymptotic framework, it is a well-known fact that neither the AIC nor the $\mathrm{AIC_c}$ are consistent, but the BIC, CAIC, and HQC are consistent. Recently, there have been many investigations of statistical methods for high-dimensional data, in which p is large but still smaller than n (see, e.g., Fan

et al. (2008), Fujikoshi and Sakurai (2009)). It has been found that, for high-dimensional data, the following asymptotic framework is more suitable than the LS asymptotic framework (see, e.g., Fujikoshi *et al.* (2010)):

- High-dimensional (HD) asymptotic framework: the sample size and the dimension p 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 $c_{n,p} \to c_0$" as "$c_{n,p} \to c_0$". It should be emphasized that we assume that p always goes to ∞ in the HD asymptotic framework. Hence, $c_{n,p} \to 0$ is not a notation which means the LS asymptotic framework.

In this paper, the asymptotic theories based on the LS and HD asymptotic frameworks are named the LS and HD asymptotic theories, respectively. If an information criterion has the consistency property under the HD asymptotic framework, we will conclude that the information criterion is superior to one without the consistency property, for the purpose of selecting the true model from among the candidate models with high-dimensional response variables. Yanagihara *et al.* (2015) evaluated the consistency of various information criteria under the HD asymptotic framework, and pointed out that there are cases that the AIC and $\mathrm{AIC_c}$ are consistent, but the BIC and CAIC are inconsistent.

Unfortunately, the results in previous works were obtained under the assumption that the distribution of the true model, called the true distribution, is a normal distribution. Although the normal distribution is assumed for the candidate model (1.1), we are not able to determine whether this is actually correct. Hence, a natural assumption for the generating mechanism of $\boldsymbol{Y}$, i.e., the true model, is

$$\boldsymbol{Y} = \boldsymbol{1}_n \boldsymbol{\mu}_*' + \boldsymbol{X}_{j_*} \boldsymbol{\Theta}_* + \boldsymbol{\mathcal{E}} \boldsymbol{\Sigma}_*^{1/2}, \tag{1.2}$$

where $\boldsymbol{\mathcal{E}} = (\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n)'$ is an $n \times p$ matrix of error variables that are assumed to be

$$\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n \sim \text{i.i.d. } \boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_p)', \qquad E[\boldsymbol{\varepsilon}] = \boldsymbol{0}_p, \qquad \mathrm{Cov}[\boldsymbol{\varepsilon}] = \boldsymbol{I}_p.$$

Henceforth, for simplicity, we represent $\boldsymbol{X}_{j_*}$ and k_{j_*} as $\boldsymbol{X}_*$ and k_*, respectively.

The purpose of this paper is to determine which conditions are necessary so that, when the assumption of normality is violated, a log-likelihood-based information criterion satisfies the consistency property by the LS and HD asymptotic theories. It is common knowledge that the maximum log-likelihood of the model in (1.1) consists of the determinant of the maximum likelihood estimator (MLE) of the covariance matrix $\boldsymbol{\Sigma}_j$, which is denoted by $\hat{\boldsymbol{\Sigma}}_j$ whose detailed definition will be given in a later section. Hence, under the HD asymptotic framework, it is difficult to prove the convergence of the difference between the two negative twofold maximum log-likelihoods, because the dimension of $\hat{\boldsymbol{\Sigma}}_j$ increases with an increase in the sample size. Of course, convergence in probability of $\hat{\boldsymbol{\Sigma}}_j$ to $\boldsymbol{\Sigma}_*$ is not ensured under the HD asymptotic framework.

If $\boldsymbol{\varepsilon} \sim N_p(\boldsymbol{0}_p, \boldsymbol{I}_p)$ holds, $n\hat{\boldsymbol{\Sigma}}_j$ is distributed according to the central or non-central Wishart distribution with $n - k_j - 1$ degrees of freedom. Fujikoshi *et al.*

(2010, th. 3.2.4, p. 57), provided an available formula of the Wishart distribution for the HD asymptotic framework as

$$\frac{|\boldsymbol{V}_1|}{|\boldsymbol{V}_1+\boldsymbol{V}_2|}=\frac{|\boldsymbol{B}_1|}{|\boldsymbol{B}_1+\boldsymbol{B}_2|}, \tag{1.3}$$

where $\boldsymbol{V}_1$ and $\boldsymbol{V}_2$ are mutually independent and $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$ are also mutually independent random matrices, which are defined by

$$\begin{aligned}\boldsymbol{V}_1 &\sim W_p(n,\boldsymbol{I}_p), \qquad \boldsymbol{V}_2 \sim W_p(q,\boldsymbol{I}_p;\boldsymbol{M}'\boldsymbol{M}),\\ \boldsymbol{B}_1 &\sim W_q(n-p+q,\boldsymbol{I}_q), \qquad \boldsymbol{B}_2 \sim W_q(p,\boldsymbol{I}_q;\boldsymbol{M}\boldsymbol{M}').\end{aligned}$$

By applying this formula to $\hat{\boldsymbol{\Sigma}}_j$, we can evaluate the asymptotic behavior of a difference of maximum log-likelihood between two candidate models by using two random matrices whose dimensions do not increase with an increase in the sample size. By using this idea, Fujikoshi *et al.* (2014) and Yanagihara *et al.* (2015) derived the condition for consistency under the HD asymptotic framework. However, needless to say, we cannot use this idea because a specific distribution is not assumed to be the true distribution. Hence, it is necessary to consider a different idea for assessing the consistency.

In this paper, convergence in probability property for a linear combination of elements in a symmetric idempotent random matrix and the distribution of the maximum eigenvalue of the estimator of the covariance matrix are employed for assessing consistency. Under both the LS and HD asymptotic frameworks, the results we obtained show that the conditions for consistency do not depend on indicators measuring a discrepancy between the true distribution and the normal distribution, e.g., skewness, kurtosis and other higher-order cumulants of the true distribution.

This paper is organized as follows: In Section 2, we present the necessary notation and assumptions for an information criterion and a model. In Section 3, we obtain conditions to satisfy consistency under the LS and HD asymptotic frameworks. In Section 4, we compare with variable selections using AIC, $\mathrm{AIC_c}$, BIC, CAIC and HQC by conducting numerical experiments. In Section 5, we discuss our conclusions. Technical details are provided in the Appendix.

2. Notation and assumptions

In this section, we present the necessary notation and assumptions for assessing the consistency of an information criterion for the model j (1.1). 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 is the set of overspecified models of which the explanatory variables contain all the explanatory variables of the true model j_* (1.2), i.e., $\mathcal{J}_+=\{j\in\mathcal{J}\mid j_*\subseteq j\}$, and the other is the set of underspecified models that are not the overspecified models, i.e., $\mathcal{J}_-=\mathcal{J}_+^c\cap\mathcal{J}$, where A^c denotes the compliment of the set A. In particular, we express the minimum overspecified model including $j\in\mathcal{J}_-$ as j_+, i.e.,

$$j_+=j\cup j_*. \tag{2.1}$$

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

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

where $\boldsymbol{P}_j$ and $\boldsymbol{J}_n$ are the projection matrices to the subspace spanned by the columns of $\boldsymbol{X}_j$ and $\mathbf{1}_n$, respectively, i.e., $\boldsymbol{P}_j = \boldsymbol{X}_j(\boldsymbol{X}_j'\boldsymbol{X}_j)^{-1}\boldsymbol{X}_j'$ and $\boldsymbol{J}_n = \mathbf{1}_n\mathbf{1}_n'/n$. In order to permit unified handling for all the log-likelihood-based information criteria, we consider the family of criteria for which the value of the model j can be expressed as

$$\mathrm{IC}_m(j) = n\log|\hat{\boldsymbol{\Sigma}}_j| + np(\log 2\pi + 1) + m(j), \tag{2.2}$$

where $m(j)$ is a positive penalty term that expresses the complexity of the model (1.1). An information criterion included in this family is specified by an individual penalty term $m(j)$, and the AIC, $\mathrm{AIC_c}$, BIC, CAIC, and HQC are contained in this family as special cases:

$$m(j) = \begin{cases} 2p\{k_j + (p+3)/2\} & \text{(AIC)} \\ 2np\{k_j + (p+3)/2\}/(n - k_j - p - 2) & (\mathrm{AIC_c}) \\ p\{k_j + (p+3)/2\}\log n & \text{(BIC)} \\ p\{k_j + (p+3)/2\}(1 + \log n) & \text{(CAIC)} \\ 2p\{k_j + (p+3)/2\}\log\log n & \text{(HQC)} \end{cases}. \tag{2.3}$$

Although we will consider primarily the above five criteria, the family also includes information criteria for which the penalty terms are random variables, e.g., the modified AIC (MAIC) proposed by Fujikoshi and Satoh (1997), Takeuchi's information criterion (TIC) proposed by Takeuchi (1976), the extended information criterion (EIC) proposed by Ishiguro *et al.* (1997), the cross-validation (CV) criterion proposed by Stone (1974, 1977), and other bias-corrected AICs, such as those proposed by Fujikoshi *et al.* (2005), Yanagihara (2006), and Yanagihara *et al.* (2011) (for the details of those criteria, see Yanagihara *et al.* (2013)). The best subset of ω, which is chosen by minimizing $\mathrm{IC}_m(j)$, is written as

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

Let a $p \times p$ noncentrality matrix be denoted by

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

where $\boldsymbol{\Gamma}_j$ is a $p \times \gamma_j$ ($\gamma_j \le \min\{p, k_{j_*\cap j^c}\}$) matrix, and it has full column rank when p is large, i.e., $p \ge k_*$. Since $k_{j_*\cap j^c} \le k_{j_+} - k_j$ holds for large p, $\gamma_j \le k_{j_+} - k_j$ is satisfied for large p. It should be noted that $\boldsymbol{\Gamma}_j\boldsymbol{\Gamma}_j' = \boldsymbol{O}_{p,p}$ if and only if $j \in \mathcal{J}_-$, where $\boldsymbol{O}_{n,p}$ is an $n \times p$ matrix of zeros. Moreover, let $\|\boldsymbol{a}\|$ denote the Euclidean norm of the vector $\boldsymbol{a}$. Then, in order to assess the consistency of IC_m, the following assumptions are needed:

A1. The true model is included in the set of candidate models, i.e., $j_* \in \mathcal{J}$.
A2. $E[\|\boldsymbol{\varepsilon}\|^4]$ exists and has the order $O(p^2)$ as $p \to \infty$.
A3. For all $j \in \mathcal{J}_-$, $\lim_{n\to\infty} n^{-1}\boldsymbol{\Gamma}_j\boldsymbol{\Gamma}_j' = \boldsymbol{\Omega}_{j,0}$ exists.
A4. $\lim_{n\to\infty} n^{-1}\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{R}_0$ exists and is positive definite.
A5. $\sum_{i=1}^n \|\boldsymbol{x}_i\|^4 = o(n^2)$ as $n \to \infty$.
A6. For all $j \in \mathcal{J}_-$, $\lim_{c_{n,p}\to c_0}(np)^{-1}\boldsymbol{\Gamma}_j'\boldsymbol{\Gamma}_j = \boldsymbol{\Delta}_{j,0}$ exists and is positive definite.

For the adequacy of the orders of $\boldsymbol{\Gamma}_j\boldsymbol{\Gamma}_j'$ and $\boldsymbol{\Gamma}_j'\boldsymbol{\Gamma}_j$, see Yanagihara *et al.* (2015). For $\boldsymbol{R}$ in assumption A4, we write the 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.

Assumption A1 is a basic assumption for evaluating a consistency property of an information criterion, because the probability of selecting the true model always becomes 0 if assumption A1 does not hold. Although we do not assume the specific distribution to a distribution of $\boldsymbol{\varepsilon}$, we require a moment assumption of the true distribution as assumption A2. It is easy to see that assumption A2 holds if $\max_{a=1,\ldots,p} E[\varepsilon_a^4]$ is bounded. Assumptions A3, A4 and A5 are common assumptions for evaluating asymptotic distributions of estimators in linear regression models when an ordinary asymptotic theory, i.e., the LS asymptotic framework, is used. Especially, assumptions A4 and A5 are required for proving an asymptotic normality of $(\boldsymbol{X}_j'\boldsymbol{X}_j)^{-1/2}\boldsymbol{X}_j'\boldsymbol{\mathcal{E}}$ under nonnormality when the LS asymptotic framework is used (see Lemma A.2 in the Appendix A.1). To ensure the asymptotic normality of $(\boldsymbol{X}_j'\boldsymbol{X}_j)^{-1/2}\boldsymbol{X}_j'\boldsymbol{\mathcal{E}}$, Wakaki *et al.* (2002) assumed $\limsup_{n\to\infty} \|\boldsymbol{x}_i\|^4/n < \infty$, which is stronger than assumption A5. Assumption A6 plays a critical role in proving consistency under the HD asymptotic framework. In practice, it is strongly dependent on an increase velocity of the noncentrality matrix whether a consistency is satisfied or not.

If assumption A2 is satisfied, the multivariate kurtosis proposed by Mardia (1970) exists as

$$\kappa_4^{(1)} = E[\|\boldsymbol{\varepsilon}\|^4] - p(p+2) = \sum_{a,b}^{p} \kappa_{aabb} + p(p+2), \tag{2.5}$$

where the notation $\sum_{a_1,a_2,\ldots}^p$ means $\sum_{a_1=1}^p \sum_{a_2=1}^p \cdots$, and κ_{abcd} is the fourth-order multivariate cumulant of $\boldsymbol{\varepsilon}$, defined by

$$\kappa_{abcd} = E[\varepsilon_a\varepsilon_b\varepsilon_c\varepsilon_d] - \delta_{ab}\delta_{cd} - \delta_{ac}\delta_{bd} - \delta_{ad}\delta_{bc}.$$

Here δ_{ab} is the Kronecker delta, i.e., $\delta_{aa} = 1$ and $\delta_{ab} = 0$ for $a \neq b$. It is well known that $\kappa_4^{(1)} = 0$ when $\boldsymbol{\varepsilon} \sim N_p(\boldsymbol{0}_p, \boldsymbol{I}_p)$. In general, the order of $\kappa_4^{(1)}$ is such that

$$\kappa_4^{(1)} = O(p^s) \quad \text{as} \quad p \to \infty, \ \ s \in [0,2]. \tag{2.6}$$

The positive constant s is changed by the distribution of $\boldsymbol{\varepsilon}$. For example, if $\varepsilon_1,\ldots,\varepsilon_p$ are independent random variables from the normal distribution, then $s = 0$, if $\varepsilon_1,\ldots,\varepsilon_p$ are independent random variables that are not distributed

according to normal distributions, then $s = 1$. If $\boldsymbol{\varepsilon}$ is distributed according to an elliptical distribution other than the normal distribution (see, e.g., Fang *et al.* (1990)), then $s = 2$. In order to derive other conditions for consistency which are weaker than those obtained under assumption A2, we also prepare the following stronger distribution assumption than assumption A2:

A2′. $\varepsilon_1, \ldots, \varepsilon_p$ are identically and independently distributed according to some distribution with $E[\varepsilon_1^4] = O(1)$ as $p \to \infty$.

It is easy to see that $s = 1$ if assumption A2′ holds. Moreover, we can easily check that assumption A2 holds when assumption A2′ is satisfied.

Assumption 6 is strong because all the elements of $\boldsymbol{\Gamma}_j'\boldsymbol{\Gamma}_j$ are assumed to converge under the same orders. Let $\lambda_{\max}(\boldsymbol{A})$ denote the maximum eigenvalue of $\boldsymbol{A}$. Then, we prepare the following assumption which is a milder assumption than A6:

A7. For all $j \in \mathcal{J}_-$, $\lim_{c_{n,p} \to c_0} (np)^{-1} \lambda_{\max}(\boldsymbol{\Gamma}_j'\boldsymbol{\Gamma}_j) = \xi_{j,0}$ exists, and is positive.

By using assumption A7 instead of assumption A6, we can obtain conditions for consistency under a milder assumption, although the obtained conditions becomes more restrictive than those obtained under assumption A6.

3. Main results

In this section, we derive conditions to satisfy the consistency of IC_m. The conditions of specific criteria, AIC, $\mathrm{AIC_c}$, BIC, CAIC, and HQC in (2.3), are listed as in Table 1 which will be placed at the end of this section.

3.1. Conditions for consistency under the LS asymptotic framework

First, we study a consistency property of IC_m under the LS asymptotic framework, i.e., the ordinary asymptotic framework. The conditions such that IC_m is consistent under the LS asymptotic framework are derived in the following theorem (the proof is given in Appendix A.2):

THEOREM 1. *Suppose that assumptions* A1–A3 *hold. A variable selection using* IC_m *is consistent when* $n \to \infty$ *under a fixed* p *if the following conditions are satisfied simultaneously*:

C1-1. ${}^\forall j \in \mathcal{J}_+ \backslash \{j_*\}$, $\lim_{n\to\infty}\{m(j) - m(j_*)\} = \infty$.

C1-2. ${}^\forall j \in \mathcal{J}_-$, $\lim_{n\to\infty}\{m(j) - m(j_*)\}/n = 0$.

If one of the above two conditions is not satisfied, a variable selection using IC_m *is not consistent when* $n \to \infty$ *under a fixed* p.

The conditions in Theorem 1 are the same as those in Yanagihara *et al.* (2015) which were obtained under the assumption of normality. Hence, we can see that the conditions for consistency are free of the influence of nonnormality in the true distribution. Moreover, Theorem 1 points out a well-known fact that, when $n \to \infty$, the AIC and the $\mathrm{AIC_c}$ are not consistent in the selection of variables, but BIC, CAIC, and HQC are.

Although IC_m does not have the consistency property when $m(j) = O(1)$ as $n \to \infty$, the asymptotic probability of selecting the model j can be evaluated as

in the following corollary (the proof is given in Appendix A.3):

COROLLARY 1. *Suppose that assumptions* A1–A5 *hold, and the following condition holds*:

C1-3. $m(j) = O(1)$ *as* $n \to \infty$ *for all* $j \in \mathcal{J}_-$, *and* $\lim_{n\to\infty}\{m(j) - m(\ell)\} = m_0 p(k_j - k_\ell)$ *for all* $j, \ell \in \mathcal{J}_+$.

Then, the asymptotic probability of selecting the model j *by* 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 $\boldsymbol{z}_j \sim N_{k_j p}(\boldsymbol{0}_{k_j p}, \boldsymbol{I}_{k_j p})$ *and* $\mathrm{Cov}[\boldsymbol{z}_j, \boldsymbol{z}_\ell] = \boldsymbol{I}_p \otimes \boldsymbol{R}_{j,j,0}^{-1/2}\boldsymbol{R}_{j,\ell,0}\boldsymbol{R}_{\ell,\ell,0}^{-1/2}$. *Especially, 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* IC_m *is simplified as*

$$\lim_{n\to\infty} P(\hat{j}_m = j_*) = \sum_{[k-k_*]} \prod_{i=1}^{k-k_*} \frac{1}{h_i!}\left\{\frac{F_{ip}(im_0p)}{i}\right\}^{h_i}, \tag{3.2}$$

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.,*

$$\begin{aligned}
&F_p(m_0p), \quad (k-k_*=1), \qquad \frac{1}{2}\{F_p(m_0p)^2 + F_{2p}(2m_0p)\}, \quad (k-k_*=2),\\
&\frac{1}{6}\{F_p(m_0p)^3 + 3F_p(m_0p)F_{2p}(2m_0p) + 2F_{3p}(3m_0p)\}, \qquad (k-k_*=3),\\
&\frac{1}{24}\{F_p(m_0p)^4 + 6F_p(m_0p)^2F_{2p}(2m_0p)\\
&\qquad +3F_{2p}(2m_0p)^2 + 8F_p(m_0p)F_{3p}(3m_0p) + 6F_{4p}(4m_0p)\}, \qquad (k-k_*=4).
\end{aligned}$$

From Corollary 1, we see that an asymptotic probability of selecting the model j can be obtained by applying Monte Carlo simulation to equation (3.1) which replaced $\boldsymbol{R}_0$ with $\boldsymbol{X}'\boldsymbol{X}/n$. Under a special case, we can derive the probability through distribution functions of chi-square distributions with several degrees of freedoms.

From Yanagihara *et al.* (2013), we see that the $m(j)$'s in the MAIC, TIC, EIC, CV criterion, and other bias-corrected AICs are $O_p(1)$ as $n \to \infty$ and $\lim_{n\to\infty}\{m(j) - m(\ell)\} = 2p(k_j - k_\ell)$, $^\forall j, \ell \in \mathcal{J}_+$ if $E[\|\boldsymbol{\varepsilon}\|^8] < \infty$. Therefore, if $E[\|\boldsymbol{\varepsilon}\|^8] < \infty$ holds, the asymptotic probabilities of selecting the model j by most bias-corrected AICs become the same as those in Corollary 1.

3.2. Conditions for consistency under the HD asymptotic framework

Next, we study a consistency property of IC_m under the HD asymptotic framework. The conditions such that IC_m is consistent under the HD asymptotic framework are derived in the following theorem (the proof is given in Appendix A.4):

THEOREM 2. *Suppose that assumptions* A1, A4 *and* A5 *hold, either assumption* A2 *or* A2′ *holds, and either assumption* A6 *or* A7 *holds. A variable selection using* IC_m *is consistent when* $c_{n,p} \to c_0$ *if the following conditions are satisfied simultaneously*:

C2-1. $^\forall j \in \mathcal{J}_+ \backslash \{j_*\}$, $\lim_{c_{n,p}\to c_0}\{m(j) - m(j_*)\}/p > -c_0^{-1}(k_j - k_*)\log(1 - c_0)$.

C2-2. $^\forall j \in \mathcal{J}_-$, $\lim_{c_{n,p}\to c_0}\{m(j) - m(j_*)\}/(n \log p) > -\beta_j$,

where β_j *is given by*

$$\beta_j = \begin{cases} \gamma_j/2 & (\textit{assumptions } \mathrm{A2} \textit{ and } \mathrm{A6} \textit{ are satisfied}) \\ \gamma_j & (\textit{assumptions } \mathrm{A2}' \textit{ and } \mathrm{A6} \textit{ are satisfied}) \\ 1/2 & (\textit{assumptions } \mathrm{A2} \textit{ and } \mathrm{A7} \textit{ are satisfied}) \\ 1 & (\textit{assumptions } \mathrm{A2}' \textit{ and } \mathrm{A7} \textit{ are satisfied}) \end{cases}. \tag{3.3}$$

When the sign "$>$" in condition C2-1 *becomes "$<$", a variable selection using* IC_m *is not consistent when* $(n, p) \to \infty$ *under* $c_{n,p} \to c_0$. *Under assumptions* A2′ *and* A6, *when the sign "$>$" in one of the two conditions* C2-1 *and* C2-2 *becomes "$<$", a variable selection using* IC_m *is not consistent when* $c_{n,p} \to c_0$.

From Theorem 2, we can see that the conditions for satisfying consistency are independent of indicators measuring a discrepancy between the true distribution and the normal distribution, e.g., skewness, kurtosis and other higher-order cumulants of the true distribution. In particular, when assumption A2′ is satisfied, the condition for consistency is the same as that in Yanagihara *et al.* (2015), which was obtained under the assumption that the normality assumption is correct. Hence, under assumption A2′, the conditions for satisfying consistency are free of the influence of nonnormality of the true distribution.

Although a sufficient condition for consistency has been derived, we still do not know which criteria satisfy the sufficient condition. Therefore, we clarify the condition for the consistency of specific criteria in (2.3). From the above results and Theorem 2, the consistency properties of specific criteria are clarified in the following corollary (the proof is given in Appendix A.5):

COROLLARY 2. *Suppose that assumptions* A1, A4 *and* A5 *hold, either assumption* A2 *or* A2′ *holds, and either assumption* A6 *or* A7 *holds.*

(i) *A variable selection using the* AIC *is consistent when* $c_{n,p} \to c_0$ *if* $c_0 \in [0, c_{\mathrm{a}})$ *holds, and it is not consistent if* $c_0 \in (c_{\mathrm{a}}, 1)$, *where* c_{a} (≈ 0.797) *is a constant satisfying*

$$\log(1 - c_{\mathrm{a}}) + 2c_{\mathrm{a}} = 0. \tag{3.4}$$

When $c_0 \in (c_{\mathrm{a}}, 1)$, $\lim_{c_{n,p} \to c_0} P(\hat{j}_m = j_{\mathrm{a}}) = 1$ *holds, where* j_{a} *is a candidate model defined by*

$$j_{\mathrm{a}} = \arg\max_{j \in \mathcal{J}_+} k_j. \tag{3.5}$$

(ii) *Variable selections using the* AIC_c *and* HQC *are consistent when* $c_{n,p} \to c_0$.

(iii) *Variable selections using the* BIC *and* CAIC *are consistent when* $c_{n,p} \to c_0$ *if* $c_0 \in [0, c_{\mathrm{b}})$ *holds, where* c_{b} *is a constant as*

$$c_{\mathrm{b}} = \begin{cases} c_{\mathrm{b},1}/2 & (\textit{assumptions } \mathrm{A2} \textit{ and } \mathrm{A6} \textit{ are satisfied}) \\ c_{\mathrm{b},1} & (\textit{assumptions } \mathrm{A2}' \textit{ and } \mathrm{A6} \textit{ are satisfied}) \\ c_{\mathrm{b},2}/2 & (\textit{assumptions } \mathrm{A2} \textit{ and } \mathrm{A7} \textit{ are satisfied}) \\ c_{\mathrm{b},2} & (\textit{assumptions } \mathrm{A2}' \textit{ and } \mathrm{A7} \textit{ are satisfied}) \end{cases}.$$

Here, $c_{\mathrm{b},1}$ *and* $c_{\mathrm{b},2}$ *are defined by*

$$\begin{aligned} c_{\mathrm{b},1} &= \min\left\{1, \min_{j \in \mathcal{S}_-} \gamma_j/(k_* - k_j)\right\}, \\ c_{\mathrm{b},2} &= \min\left\{1, \min_{j \in \mathcal{S}_-} 1/(k_* - k_j)\right\}, \end{aligned} \tag{3.6}$$

where $\mathcal{S}_-$ *is a set of candidate models given by*

$$\mathcal{S}_- = \{j \in \mathcal{J}_- \mid k_* - k_j > 0\}. \tag{3.7}$$

Under assumptions $\mathrm{A2}'$ *and* A6, *variable selections using the* BIC *and* CAIC *are not consistent when* $c_{n,p} \to c_0$ *if* $c_0 \in (c_{\mathrm{b},1}, 1)$ *holds. Then* $\lim_{c_{n,p} \to c_0} P(\hat{j}_m = j_{\mathrm{b}}) = 1$ *holds, where* j_{b} *is a candidate model defined by*

$$j_{\mathrm{b}} = \arg\min_{j \in \mathcal{S}_-} (\gamma_j + c_0 k_j). \tag{3.8}$$

Corollary 2 shows that, when $c_{n,p} \to c_0$, the $\mathrm{AIC_c}$ and HQC are always consistent in model selection whereas the AIC, BIC and CAIC are not always consistent. It is strongly dependent on the size of γ_j whether the BIC and CAIC have consistency properties although the condition for the AIC to satisfy consistency does not depend on the size of γ_j. This is a greater disadvantage of the BIC and CAIC compared to the AIC because nobody knows the size of γ_j.

4. Numerical Study

In this section, we compare the probabilities of selecting the true model with the AIC, $\mathrm{AIC_c}$, BIC, CAIC, and HQC in (2.3) by Monte Carlo simulations with 10,000 iterations. The ten candidate models $j_0 = \{\phi\}$ and $j_\alpha = \{1, \ldots, \alpha\}$ ($\alpha = 1, \ldots, 9$), with several different values of n and p, were prepared for Monte Carlo simulations. We independently generated $z_1, \ldots, z_n$ from $U(-1, 1)$. Using

Table 1. Summary of main results of specific criteria.

• Under the LS asymptotic framework

Criterion	Consistency
AIC & $\mathrm{AIC_c}$	Not hold
BIC, $\mathrm{AIC_c}$ & HQC	Hold

• Under the HD asymptotic framework

Criterion	Consistency	Assumption	Conditions	
			Consistency	Inconsistency
AIC	Conditionally hold		$c_0 \in [0, c_{\mathrm{a}})$	$c_0 \in (c_{\mathrm{a}}, 1)$
$\mathrm{AIC_c}$ & HQC	Hold		—	—
BIC & CAIC	Conditionally hold	A2 & A6	$c_0 \in [0, c_{\mathrm{b},1}/2)$	unknown
		A2′ & A6	$c_0 \in [0, c_{\mathrm{b},1})$	$c_0 \in (c_{\mathrm{b},1}, 1)$
		A2 & A7	$c_0 \in [0, c_{\mathrm{b},2}/2)$	unknown
		A2′ & A7	$c_0 \in [0, c_{\mathrm{b},2})$	unknown

Note 1) c_{a} is given by (3.4), and $c_{\mathrm{b},1}$ and $c_{\mathrm{b},2}$ are given by (3.6) (an inequality $c_{\mathrm{b},1} \geq c_{\mathrm{b},2}$ is always satisfied).

Note 2) If the condition for inconsistency holds, $\lim_{c_{n,p}\to c_0} P(\hat{j}_m = j_{\mathrm{a}}) = 1$ is satisfied in the case of AIC, and $\lim_{c_{n,p}\to c_0} P(\hat{j}_m = j_{\mathrm{b}}) = 1$ is satisfied in the cases of BIC and CAIC, where j_{a} and j_{b} are given by (3.5) and (3.8), respectively.

$z_1, \ldots, z_n$, we constructed an $n \times k$ matrix of explanatory variables $\boldsymbol{X}$, where the (a, b)th element was defined by $z_a^b - \bar{z}_b$ $(a = 1, \ldots, n;\ b = 1, \ldots, 9)$, where $\bar{z}_b = n^{-1}\sum_{a=1}^n z_a^b$. The true model was determined by $\boldsymbol{\mu}_* = \mathbf{1}_p$, $\boldsymbol{\Theta}_* = (1, 3, -4, 5)'\mathbf{1}_p'$, $j_* = \{1, 2, 3, 4\}$, and $\boldsymbol{\Sigma}_*$ in which the (i, j)th element was defined by $(0.8)^{|a-b|}$ $(a = 1, \ldots, p;\ b = 1, \ldots, p)$. Thus, j_α with $\alpha = 0, 1, 2, 3$ was the underspecified model, and j_α with $\alpha \geq 4$ was the overspecified model. Under such a true model, it is known that assumptions A1 and A3–A6 hold.

Let $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_p)' \sim N_p(\mathbf{0}_p, \boldsymbol{I}_p)$, $\delta \sim \chi_6^2$ and $\omega_1, \ldots, \omega_p \sim$ i.i.d. χ_5^2 be mutually independent random vectors and variables. Then, $\boldsymbol{\varepsilon}$ was generated from the following five distributions:

- Distribution 1 (multivariate normal distribution): $\boldsymbol{\varepsilon} = \boldsymbol{\nu}$.
- Distribution 2 (scale mixture of multivariate normal distribution): $\boldsymbol{\varepsilon} = \sqrt{\delta/6}\boldsymbol{\nu}$.
- Distribution 3 (scale and location mixtures of multivariate normal distribution): $\boldsymbol{\varepsilon} = \boldsymbol{\Psi}^{-1/2}\{10(\sqrt{\delta/6} - \eta)\mathbf{1}_p + \sqrt{\delta/6}\boldsymbol{\nu}\}$, where $\eta = 15\sqrt{\pi/3}/16$ and $\boldsymbol{\Psi} = \boldsymbol{I}_p + 100(1 - \eta^2)\mathbf{1}_p\mathbf{1}_p'$.
- Distribution 4 (independent t-distribution): $\varepsilon_a = \sqrt{3}\nu_a/\sqrt{5\omega_a}$.
- Distribution 5 (independent log-normal distribution): $\varepsilon_a = (\log \nu_a - \sqrt{e})/\sqrt{e(e-1)}$.

It is easy to see that distributions 1, 2 and 4 are symmetric, and distributions 3 and 5 are skewed. Although all the distributions satisfy assumption A2, only distributions 1, 4 and 5 satisfy assumption A2′.

In our numerical study, $\gamma_j = 1$ and $\max(k_* - k_j) = 4$ hold for all $j \in \mathcal{S}_-$. This implies that when $c_0 > 1/8$, the inequality $\gamma_j/2 > c_0(k_* - k_j)$ was not always

Table 2. Selection probabilities of the true model (%) in the case of distribution 1.

	Case 1						Case 2 ($c_0 = 0.02$)					
n	p	AIC	$\mathrm{AIC_c}$	BIC	CAIC	HQC	p	AIC	$\mathrm{AIC_c}$	BIC	CAIC	HQC
100	2	73.1	80.6	76.7	65.8	85.2	2	73.1	80.6	76.7	65.8	85.2
200	2	78.4	82.4	98.6	97.8	95.6	4	86.0	90.5	95.0	88.1	98.5
500	2	80.0	81.5	99.8	99.9	97.2	10	96.3	97.4	100.0	100.0	100.0
1000	2	80.1	80.9	99.9	100.0	97.6	20	99.4	99.6	100.0	100.0	100.0
∞	2	80.2	80.2	100.0	100.0	100.0	∞	100.0	100.0	100.0	100.0	100.0

	Case 3						Case 4 ($c_0 = 0.1$)					
n	p	AIC	$\mathrm{AIC_c}$	BIC	CAIC	HQC	p	AIC	$\mathrm{AIC_c}$	BIC	CAIC	HQC
100	10	86.4	73.0	5.2	0.3	55.8	10	86.6	73.5	5.4	0.3	55.7
200	10	95.5	98.2	67.8	37.9	98.4	20	98.7	99.8	17.9	0.8	96.4
500	10	96.2	97.4	100.0	100.0	100.0	50	100.0	100.0	99.0	69.8	100.0
1000	10	96.5	97.2	100.0	100.0	100.0	100	100.0	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	100.0	100.0

	Case 5						Case 6 ($c_0 = 0.3$)					
n	p	AIC	$\mathrm{AIC_c}$	BIC	CAIC	HQC	p	AIC	$\mathrm{AIC_c}$	BIC	CAIC	HQC
100	30	90.3	0.0	0.0	0.0	11.0	30	90.3	0.0	0.0	0.0	11.0
200	30	99.5	99.6	1.1	0.0	93.5	60	99.9	21.4	0.0	0.0	74.1
500	30	99.8	100.0	99.9	97.1	100.0	150	100.0	100.0	0.0	0.0	100.0
1000	30	99.8	99.9	100.0	100.0	100.0	300	100.0	100.0	0.0	0.0	100.0
∞	30	99.9	99.9	100.0	100.0	100.0	∞	100.0	100.0	0.0	0.0	100.0

	Case 7 ($c_0 = 0.0$)						Case 8 ($c_0 = 0.0$)					
n	p	AIC	$\mathrm{AIC_c}$	BIC	CAIC	HQC	p	AIC	$\mathrm{AIC_c}$	BIC	CAIC	HQC
100	30	90.3	0.0	0.0	0.0	11.0	30	90.3	0.0	0.0	0.0	11.0
200	32	99.6	99.5	0.4	0.0	93.0	40	99.7	97.5	0.0	0.0	88.9
500	35	99.9	100.0	99.8	94.1	100.0	50	100.0	100.0	99.2	70.1	100.0
1000	40	100.0	100.0	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	100.0	100.0	100.0

	Case 9 ($c_0 = 0.7$)						Case 10 ($c_0 = 0.85$)					
n	p	AIC	$\mathrm{AIC_c}$	BIC	CAIC	HQC	p	AIC	$\mathrm{AIC_c}$	BIC	CAIC	HQC
100	70	25.0	0.0	0.0	0.0	18.1	85	0.0	0.0	0.0	0.0	32.2
200	140	65.4	0.0	0.0	0.0	28.9	170	0.0	0.0	0.0	0.0	94.0
500	350	94.3	0.0	0.0	0.0	49.2	425	0.0	0.0	0.0	0.0	99.8
1000	700	99.4	0.0	0.0	0.0	65.3	850	0.0	0.0	0.0	0.0	100.0
∞	∞	100.0	100.0	0.0	0.0	100.0	∞	0.0	100.0	0.0	0.0	100.0

satisfied for all $j \in \mathcal{S}_-$. Thus, it is not clear whether the probability of selecting j_* by the BIC and CAIC converged to 1 as $c_{n,p} \to c_0 \in (1/8, 1)$ under distributions 2 and 3. On the other hand, when $c_0 > 1/4$, the inequality $\gamma_j > c_0(k_* - k_j)$ was not always satisfied for all $j \in \mathcal{S}_-$. Hence, the probability of selecting j_* by the BIC and CAIC did not converge to 1 as $c_{n,p} \to c_0 \in (1/4, 1)$ under distributions 1, 4 and 5.

Table 3. Selection probabilities of the true model (%) in the case of distribution 2.

	Case 1						Case 2 ($c_0 = 0.02$)					
n	p	AIC	$\mathrm{AIC_c}$	BIC	CAIC	HQC	p	AIC	$\mathrm{AIC_c}$	BIC	CAIC	HQC
100	2	73.5	80.7	76.4	65.9	84.4	2	73.5	80.7	76.4	65.9	84.4
200	2	78.2	82.3	98.6	97.8	95.1	4	86.9	91.0	95.1	88.1	98.3
500	2	79.9	81.5	99.8	99.9	97.0	10	96.6	97.7	100.0	99.9	100.0
1000	2	80.0	80.7	99.9	100.0	97.5	20	99.3	99.6	100.0	100.0	100.0
∞	2	80.2	80.2	100.0	100.0	100.0	∞	100.0	100.0	100.0	100.0	100.0

	Case 3						Case 4 ($c_0 = 0.1$)					
n	p	AIC	$\mathrm{AIC_c}$	BIC	CAIC	HQC	p	AIC	$\mathrm{AIC_c}$	BIC	CAIC	HQC
100	10	86.7	75.3	6.8	0.5	59.8	10	86.7	75.3	6.8	0.5	59.8
200	10	95.1	98.2	69.4	40.4	98.5	20	98.7	99.9	23.7	1.9	96.8
500	10	96.2	97.4	100.0	99.9	100.0	50	100.0	100.0	99.2	76.5	100.0
1000	10	96.5	97.1	100.0	100.0	100.0	100	100.0	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	100.0	100.0

	Case 5						Case 6 ($c_0 = 0.3$)					
n	p	AIC	$\mathrm{AIC_c}$	BIC	CAIC	HQC	p	AIC	$\mathrm{AIC_c}$	BIC	CAIC	HQC
100	30	92.4	0.0	0.0	0.0	18.4	30	92.4	0.0	0.0	0.0	18.4
200	30	99.5	99.7	2.5	0.0	94.5	60	99.8	40.8	0.0	0.0	86.5
500	30	99.8	100.0	100.0	97.5	100.0	150	100.0	100.0	0.0	0.0	100.0
1000	30	99.9	100.0	100.0	100.0	100.0	300	100.0	100.0	0.0	0.0	100.0
∞	30	99.9	99.9	100.0	100.0	100.0	∞	100.0	100.0	—	—	100.0

	Case 7 ($c_0 = 0.0$)						Case 8 ($c_0 = 0.0$)					
n	p	AIC	$\mathrm{AIC_c}$	BIC	CAIC	HQC	p	AIC	$\mathrm{AIC_c}$	BIC	CAIC	HQC
100	30	92.4	0.0	0.0	0.0	18.4	30	92.4	0.0	0.0	0.0	18.4
200	32	99.5	99.7	1.2	0.0	94.8	40	99.6	98.4	0.0	0.0	92.8
500	35	99.9	100.0	99.9	95.3	100.0	50	99.9	100.0	99.2	76.6	100.0
1000	40	100.0	100.0	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	100.0	100.0	100.0

	Case 9 ($c_0 = 0.7$)						Case 10 ($c_0 = 0.85$)					
n	p	AIC	$\mathrm{AIC_c}$	BIC	CAIC	HQC	p	AIC	$\mathrm{AIC_c}$	BIC	CAIC	HQC
100	70	22.9	0.0	0.0	0.0	36.0	85	0.0	0.0	0.0	0.0	33.8
200	140	63.5	0.0	0.0	0.0	62.1	170	0.0	0.0	0.0	0.0	98.5
500	350	93.9	0.0	0.0	0.0	86.3	425	0.0	0.0	0.0	0.0	100.0
1000	700	99.5	0.0	0.0	0.0	97.7	850	0.0	0.0	0.0	0.0	100.0
∞	∞	100.0	100.0	—	—	100.0	∞	0.0	100.0	—	—	100.0

Tables 2, 3, 4, 5 and 6 show the probability of selecting the true model by the AIC, $\mathrm{AIC_c}$, BIC, CAIC, and HQC when the distributions of $\boldsymbol{\varepsilon}$ are 1, 2, 3, 4 and 5, respectively. For $n = \infty$ or $p = \infty$, we list the theoretical values obtained from Theorems 1 and 2. Especially the theoretical values of AIC and $\mathrm{AIC_c}$ in Cases 1, 3 and 5 are calculated by formula (3.2) in Corollary 1. The symbol "—" indicates that the theoretical value is not clear. It should be em-

Table 4. Selection probabilities of the true model (%) in the case of distribution 3.

	Case 1						Case 2 ($c_0 = 0.02$)					
n	p	AIC	$\mathrm{AIC_c}$	BIC	CAIC	HQC	p	AIC	$\mathrm{AIC_c}$	BIC	CAIC	HQC
100	2	73.5	80.5	77.0	66.4	85.1	2	73.5	80.5	77.0	66.4	85.1
200	2	78.7	82.7	98.4	97.6	95.3	4	86.6	90.5	94.9	88.9	98.3
500	2	79.5	81.1	99.8	99.9	96.7	10	96.0	97.3	100.0	100.0	100.0
1000	2	79.5	80.4	99.9	100.0	97.8	20	99.4	99.7	100.0	100.0	100.0
∞	2	80.6	80.6	100.0	100.0	100.0	∞	100.0	100.0	100.0	100.0	100.0

	Case 3						Case 4 ($c_0 = 0.1$)					
n	p	AIC	$\mathrm{AIC_c}$	BIC	CAIC	HQC	p	AIC	$\mathrm{AIC_c}$	BIC	CAIC	HQC
100	10	86.3	75.9	6.3	0.5	59.8	10	86.3	75.9	6.3	0.5	59.8
200	10	95.1	98.4	69.3	39.3	98.4	20	98.6	99.9	23.3	1.7	97.1
500	10	96.4	97.5	100.0	100.0	100.0	50	100.0	100.0	99.5	77.9	100.0
1000	10	96.6	97.0	100.0	100.0	100.0	100	100.0	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	100.0	100.0

	Case 5						Case 6 ($c_0 = 0.3$)					
n	p	AIC	$\mathrm{AIC_c}$	BIC	CAIC	HQC	p	AIC	$\mathrm{AIC_c}$	BIC	CAIC	HQC
100	30	91.3	0.0	0.0	0.0	16.8	30	91.3	0.0	0.0	0.0	16.8
200	30	99.6	99.8	2.0	0.0	94.8	60	99.8	35.1	0.0	0.0	85.4
500	30	99.9	100.0	99.9	97.5	100.0	150	100.0	100.0	0.0	0.0	100.0
1000	30	99.8	99.9	100.0	100.0	100.0	300	100.0	100.0	0.0	0.0	100.0
∞	30	99.9	99.9	100.0	100.0	100.0	∞	100.0	100.0	—	—	100.0

	Case 7 ($c_0 = 0.0$)						Case 8 ($c_0 = 0.0$)					
n	p	AIC	$\mathrm{AIC_c}$	BIC	CAIC	HQC	p	AIC	$\mathrm{AIC_c}$	BIC	CAIC	HQC
100	30	91.3	0.0	0.0	0.0	16.8	30	91.3	0.0	0.0	0.0	16.8
200	32	99.5	99.7	0.9	0.0	94.7	40	99.6	98.6	0.0	0.0	92.9
500	35	99.9	100.0	99.9	95.3	100.0	50	100.0	100.0	99.4	77.2	100.0
1000	40	100.0	100.0	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	100.0	100.0	100.0

	Case 9 ($c_0 = 0.7$)						Case 10 ($c_0 = 0.85$)					
n	p	AIC	$\mathrm{AIC_c}$	BIC	CAIC	HQC	p	AIC	$\mathrm{AIC_c}$	BIC	CAIC	HQC
100	70	23.2	0.0	0.0	0.0	26.2	85	0.0	0.0	0.0	0.0	31.7
200	140	63.1	0.0	0.0	0.0	43.8	170	0.0	0.0	0.0	0.0	96.6
500	350	93.7	0.0	0.0	0.0	68.7	425	0.0	0.0	0.0	0.0	99.9
1000	700	99.4	0.0	0.0	0.0	87.8	850	0.0	0.0	0.0	0.0	100.0
∞	∞	100.0	100.0	—	—	100.0	∞	0.0	100.0	—	—	100.0

phasized that the probability of selecting j_* by the AIC did not converge to 1 in Case 10. From the tables, we can see that in the cases of the AIC, $\mathrm{AIC_c}$, and HQC, the greater the dimension and sample size, the greater the probabilities. Compared with the results obtained from the AIC, $\mathrm{AIC_c}$, and HQC, the probabilities for the $\mathrm{AIC_c}$ and HQC tended to be higher than those for the AIC when n was not small. In the cases of the BIC and CAIC, the greater the

Table 5. Selection probabilities of the true model (%) in the case of distribution 4.

	Case 1						Case 2 ($c_0 = 0.02$)					
n	p	AIC	$\mathrm{AIC_c}$	BIC	CAIC	HQC	p	AIC	$\mathrm{AIC_c}$	BIC	CAIC	HQC
100	2	73.6	80.9	77.9	68.8	85.0	2	72.8	80.8	77.9	67.9	85.2
200	2	78.3	82.0	98.4	97.6	95.0	4	85.9	90.4	95.0	88.7	98.3
500	2	80.1	81.7	99.8	100.0	97.0	10	96.2	97.3	100.0	100.0	100.0
1000	2	79.3	80.2	99.9	99.9	97.6	20	99.4	99.6	100.0	100.0	100.0
∞	2	80.6	80.6	100.0	100.0	100.0	∞	100.0	100.0	100.0	100.0	100.0

	Case 3						Case 4 ($c_0 = 0.1$)					
n	p	AIC	$\mathrm{AIC_c}$	BIC	CAIC	HQC	p	AIC	$\mathrm{AIC_c}$	BIC	CAIC	HQC
100	10	86.6	75.0	6.4	0.4	60.1	10	87.2	76.1	6.2	0.4	59.3
200	10	95.3	98.3	69.4	40.2	98.6	20	98.9	99.8	21.2	1.2	96.6
500	10	96.1	97.5	100.0	100.0	100.0	50	100.0	100.0	99.1	71.5	100.0
1000	10	96.5	97.1	100.0	100.0	100.0	100	100.0	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	100.0	100.0

	Case 5						Case 6 ($c_0 = 0.3$)					
n	p	AIC	$\mathrm{AIC_c}$	BIC	CAIC	HQC	p	AIC	$\mathrm{AIC_c}$	BIC	CAIC	HQC
100	30	91.3	0.0	0.0	0.0	13.6	30	91.2	0.0	0.0	0.0	13.6
200	30	99.4	99.6	1.4	0.0	94.4	60	99.8	25.0	0.0	0.0	77.9
500	30	99.7	100.0	99.9	97.3	100.0	150	100.0	100.0	0.0	0.0	100.0
1000	30	99.9	99.9	100.0	100.0	100.0	300	100.0	100.0	0.0	0.0	100.0
∞	30	99.9	99.9	100.0	100.0	100.0	∞	100.0	100.0	0.0	0.0	100.0

	Case 7 ($c_0 = 0.0$)						Case 8 ($c_0 = 0.0$)					
n	p	AIC	$\mathrm{AIC_c}$	BIC	CAIC	HQC	p	AIC	$\mathrm{AIC_c}$	BIC	CAIC	HQC
100	30	90.9	0.0	0.0	0.0	12.6	30	90.7	0.0	0.0	0.0	13.0
200	32	99.5	99.6	0.5	0.0	93.4	40	99.6	97.7	0.0	0.0	90.8
500	35	99.9	100.0	99.9	94.5	100.0	50	100.0	100.0	99.1	71.7	100.0
1000	40	100.0	100.0	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	100.0	100.0	100.0

	Case 9 ($c_0 = 0.7$)						Case 10 ($c_0 = 0.85$)					
n	p	AIC	$\mathrm{AIC_c}$	BIC	CAIC	HQC	p	AIC	$\mathrm{AIC_c}$	BIC	CAIC	HQC
100	70	24.2	0.0	0.0	0.0	20.6	85	0.0	0.0	0.0	0.0	32.2
200	140	66.2	0.0	0.0	0.0	31.0	170	0.0	0.0	0.0	0.0	94.6
500	350	94.4	0.0	0.0	0.0	50.2	425	0.0	0.0	0.0	0.0	99.8
1000	700	99.5	0.0	0.0	0.0	65.0	850	0.0	0.0	0.0	0.0	100.0
∞	∞	100.0	100.0	0.0	0.0	100.0	∞	0.0	100.0	0.0	0.0	100.0

dimension and sample size were, the higher the selection probabilities became, with the exception of Cases 6, 9 and 10. This was because there is a possibility that variable selections using the BIC and the CAIC are not consistent in Cases 6, 9 and 10. Indeed, BIC and CAIC do not have consistency under distributions 1, 4 and 5. 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

Table 6. Selection probabilities of the true model (%) in the case of distribution 5.

	Case 1						Case 2 ($c_0 = 0.02$)					
n	p	AIC	$\mathrm{AIC_c}$	BIC	CAIC	HQC	p	AIC	$\mathrm{AIC_c}$	BIC	CAIC	HQC
100	2	74.7	81.7	83.2	76.9	86.4	2	74.4	81.3	83.4	77.1	86.1
200	2	79.5	83.4	98.1	97.6	94.5	4	86.2	90.5	96.6	93.0	98.2
500	2	79.9	81.4	99.7	99.9	96.7	10	96.0	97.3	100.0	100.0	100.0
1000	2	79.2	80.0	99.8	99.9	97.4	20	99.4	99.6	100.0	100.0	100.0
∞	2	80.6	80.6	100.0	100.0	100.0	∞	100.0	100.0	100.0	100.0	100.0

	Case 3						Case 4 ($c_0 = 0.1$)					
n	p	AIC	$\mathrm{AIC_c}$	BIC	CAIC	HQC	p	AIC	$\mathrm{AIC_c}$	BIC	CAIC	HQC
100	10	89.4	88.3	26.2	5.5	79.4	10	89.0	88.2	25.5	5.4	78.7
200	10	94.5	97.9	83.9	62.8	99.4	20	98.3	99.8	49.5	11.1	99.2
500	10	96.4	97.5	100.0	100.0	100.0	50	100.0	100.0	99.9	90.2	100.0
1000	10	96.4	97.1	100.0	100.0	100.0	100	100.0	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	100.0	100.0

	Case 5						Case 6 ($c_0 = 0.3$)					
n	p	AIC	$\mathrm{AIC_c}$	BIC	CAIC	HQC	p	AIC	$\mathrm{AIC_c}$	BIC	CAIC	HQC
100	30	95.1	0.5	0.0	0.0	39.7	30	95.1	0.4	0.0	0.0	39.6
200	30	99.4	100.0	11.8	0.1	98.7	60	99.8	61.2	0.0	0.0	95.2
500	30	99.8	100.0	100.0	99.3	100.0	150	100.0	100.0	0.0	0.0	100.0
1000	30	99.9	100.0	100.0	100.0	100.0	300	100.0	100.0	0.0	0.0	100.0
∞	30	99.9	99.9	100.0	100.0	100.0	∞	100.0	100.0	0.0	0.0	100.0

	Case 7 ($c_0 = 0.0$)						Case 8 ($c_0 = 0.0$)					
n	p	AIC	$\mathrm{AIC_c}$	BIC	CAIC	HQC	p	AIC	$\mathrm{AIC_c}$	BIC	CAIC	HQC
100	30	94.9	0.5	0.0	0.0	40.1	30	95.0	0.3	0.0	0.0	39.4
200	32	99.4	100.0	7.9	0.0	98.5	40	99.5	99.6	0.5	0.0	98.0
500	35	99.9	100.0	100.0	98.5	100.0	50	100.0	100.0	99.9	90.7	100.0
1000	40	100.0	100.0	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	100.0	100.0	100.0

	Case 9 ($c_0 = 0.7$)						Case 10 ($c_0 = 0.85$)					
n	p	AIC	$\mathrm{AIC_c}$	BIC	CAIC	HQC	p	AIC	$\mathrm{AIC_c}$	BIC	CAIC	HQC
100	70	23.5	0.0	0.0	0.0	39.0	85	0.0	0.0	0.0	0.0	33.5
200	140	64.7	0.0	0.0	0.0	51.3	170	0.0	0.0	0.0	0.0	97.3
500	350	94.7	0.0	0.0	0.0	64.9	425	0.0	0.0	0.0	0.0	99.9
1000	700	99.2	0.0	0.0	0.0	77.0	850	0.0	0.0	0.0	0.0	100.0
∞	∞	100.0	100.0	0.0	0.0	100.0	∞	0.0	100.0	0.0	0.0	100.0

and the CAIC were consistent in variable selection, these probabilities became high as n and p increased. In Cases 9 and 10, although $\mathrm{AIC_c}$ was consistent in variable selection, the selection probabilities of the $\mathrm{AIC_c}$ were always 0. This is because $h(n, c_0, j) = \{m(j) - m(j_*)\}/(n \log p)$ for $\mathrm{AIC_c}$ were negative and did not converge to 0 even under huge n, e.g., $h(100000, 0.7, j_0) = -3.626$ and $h(100000, 0.8, j_0) = -15.320$. Moreover, we could not find notable differences

between the simulation results obtained from normal and nonnormal distributions. This indicates that, for variable selection even under the HD asymptotic framework, the effect of the violation of the normality assumption is not large.

5. Conclusion and discussion

In this paper, we derived the conditions to satisfy the consistency property of a log-likelihood-based information criterion in (2.2) for selecting variables in multivariate linear regression models with the normality assumption, but for which normality may be violated in the true model. The information criteria considered in this paper were defined by adding a positive penalty term to the negative twofold maximum log-likelihood; hence, the family of information criteria that we considered included as special cases the AIC, $\mathrm{AIC_c}$, BIC, CAIC, and HQC. The consistency property was studied under the LS and HD asymptotic theories. In both cases, the conditions obtained were independent of indicators measuring a discrepancy between the true distribution and the normal distribution, e.g., skewness, kurtosis and other higher-order cumulants of the true distribution. Under the LS asymptotic framework, we obtained the conditions for consistency and inconsistency, which were equivalent to that derived under the normality assumption. Under the HD asymptotic framework, the sufficient condition for consistency was obtained. The condition was slightly more constrictive than that derived under the normality assumption. But with a strong assumption for the true distribution, i.e., all the elements of $\boldsymbol{\varepsilon}$ are independent, the condition coincided with that derived under the normality assumption.

Under the HD asymptotic framework, when normality is assumed for the true distribution, we can assess the asymptotic behavior of the difference between the two negative twofold maximum log-likelihoods by two random matrices whose dimensions do not increase with an increase in the sample size, after applying the formula in (1.3) to $\hat{\boldsymbol{\Sigma}}_j$, which is the same method used in Yanagihara *et al.* (2015) and Fujikoshi *et al.* (2014). However, we cannot use this because our setting assumes that the normality assumption is violated. Hence, we employed the convergence in probability property for a linear combination of elements in a symmetric idempotent random matrix and the distribution of the maximum eigenvalue of the estimator of the covariance matrix, to evaluate the asymptotic behavior.

If we assume the existence of $E[\|\boldsymbol{\varepsilon}\|^6]$, and that $E[\|\boldsymbol{\varepsilon}\|^6] = O(p^3)$ as $p \to \infty$, equation (ii) in Lemma A.8 is changed to $\lambda_{\max}(\boldsymbol{S}_j) = O_p(p^{1/3})$. This directly implies that β_j in condition C2-2 is relaxed to $2\gamma_j/3$. If we assume the existence of $E[\|\boldsymbol{\varepsilon}\|^{2r}]$, and that $E[\|\boldsymbol{\varepsilon}\|^{2r}] = O(p^r)$ as $p \to \infty$ for all $r \geq 1$, β_j in condition C2-2 may be relaxed to γ_j, which is equivalent to the condition obtained from the normality assumption.

Appendix

A.1. Lemmas to prove theorems and corollaries

In this section, we describe several lemmas for deriving main results. In Lemmas A.3 and A.4, we do not specify the asymptotic framework because these

are applicable to any asymptotic framework. Proofs of all the lemmas are given in Appendix A.6.

LEMMA A.1. *Let* $\boldsymbol{Q}_j$ *be an* $n \times k_j$ *matrix defined by*

$$\boldsymbol{Q}_j' = (\boldsymbol{X}_j'\boldsymbol{X}_j)^{-1/2}\boldsymbol{X}_j' = (\boldsymbol{q}_{j,1}, \ldots, \boldsymbol{q}_{j,n}), \qquad \boldsymbol{q}_{j,i} = (q_{j,i1}, \ldots, q_{j,ik_j})'. \tag{A.1}$$

Suppose that assumptions A4 *and* A5 *are satisfied. Then, we have*

$$\sum_{i=1}^{n} |q_{j,ia}q_{j,ib}q_{j,ic}q_{j,id}| = o(1) \qquad \text{as} \qquad n \to \infty,$$

where a, b, c, d *are arbitrary positive integers not larger than* k_j.

LEMMA A.2. *Let* $\boldsymbol{Z}_j$ *be a* $k_j \times p$ *matrix defined by*

$$\boldsymbol{Z}_j = \boldsymbol{Q}_j'\boldsymbol{\mathcal{E}}, \tag{A.2}$$

where $\boldsymbol{Q}_j$ *is given by* (A.1). *Suppose that assumptions* A2, A4, *and* A5 *are satisfied. Then,* $\boldsymbol{Z}_j \xrightarrow{d} N_{k_j \times p}(\boldsymbol{O}_{k_j,p}, \boldsymbol{I}_{k_j p})$ *as* $n \to \infty$ *holds.*

LEMMA A.3. *Let* IC_m *denote a log-likelihood-based information criterion defined by* (2.2), *and* $h_{j,\ell}$ *be some positive constant that depends on the models* $j, \ell \in \mathcal{J}$. *Then, we have*

(i) $j, \ell \in \mathcal{J}$, $j \neq \ell$, $\frac{1}{h_{j,\ell}}\{\mathrm{IC}_m(j) - \mathrm{IC}_m(\ell)\} \geq T_{j,\ell} \xrightarrow{p} \tau_{j,\ell} > 0 \Rightarrow P(\mathrm{IC}_m(j) < \mathrm{IC}_m(\ell)) \to 0$ *and* $P(\mathrm{IC}_m(j) > \mathrm{IC}_m(\ell)) \to 1$,

(ii) $^{\forall}\ell \in \mathcal{J}\backslash\{j\}$, $\frac{1}{h_{\ell,j}}\{\mathrm{IC}_m(\ell) - \mathrm{IC}_m(j)\} \geq T_{\ell,j} \xrightarrow{p} \tau_{\ell,j} > 0 \Rightarrow P(\hat{j}_m = j) \to 1$,

(iii) $^{\exists}\ell_0 \in \mathcal{J}\backslash\{j\}$ *s.t.* $\frac{1}{h_{j,\ell_0}}\{\mathrm{IC}_m(j) - \mathrm{IC}_m(\ell_0)\} \geq T_{j,\ell_0} \xrightarrow{p} \tau_{j,\ell_0} > 0 \Rightarrow P(\hat{j}_m = j) \to 0$.

LEMMA A.4. *Let* A *and* B *be events. Then, the following equations are satisfied*:

(i) $P(B) \to 0 \Rightarrow P(A \cap B) \to 0$,

(ii) $P(B) \to 1 \Rightarrow \lim P(A \cap B) = \lim P(A)$.

LEMMA A.5. *For any* $n \times n$ *symmetric matrix* $\boldsymbol{A}$, *let* $\phi_1(\boldsymbol{A})$ *and* $\phi_2(\boldsymbol{A})$ *denote moments*:

$$\phi_1(\boldsymbol{A}) = E[\mathrm{tr}(\boldsymbol{\mathcal{E}}'\boldsymbol{A}\boldsymbol{\mathcal{E}})], \qquad \phi_2(\boldsymbol{A}) = E[\mathrm{tr}\{(\boldsymbol{\mathcal{E}}'\boldsymbol{A}\boldsymbol{\mathcal{E}})^2\}].$$

Then, specific forms of $\phi_1(\boldsymbol{A})$ *and* $\phi_2(\boldsymbol{A})$ *are given as*

$$\phi_1(\boldsymbol{A}) = p\,\mathrm{tr}(\boldsymbol{A}), \qquad \phi_2(\boldsymbol{A}) = \kappa_4^{(1)}\sum_{a=1}^{n}\{(\boldsymbol{A})_{aa}\}^2 + p(p+1)\,\mathrm{tr}(\boldsymbol{A}^2) + p\,\mathrm{tr}(\boldsymbol{A})^2,$$

where $\kappa_4^{(1)}$ *is given by* (2.5), *and* $(\boldsymbol{A})_{ab}$ *is the* (a,b)*th element of a matrix* $\boldsymbol{A}$.

LEMMA A.6. *For any $n \times n$ symmetric idempotent matrix $\boldsymbol{A}$, we have*

$$\sum_{a=1}^{n} \{(\boldsymbol{A})_{aa}\}^2 = O(\mathrm{tr}(\boldsymbol{A})) \quad \text{as} \quad \mathrm{tr}(\boldsymbol{A}) \to \infty,$$

where $(\boldsymbol{A})_{ab}$ is the (a,b)th element of a matrix $\boldsymbol{A}$.

LEMMA A.7. *Let $\boldsymbol{U}$ and $\boldsymbol{W}$ be $n \times p$ and $n \times n$ random matrices, respectively, defined by*

$$\boldsymbol{U} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n)' = (\boldsymbol{I}_n - \boldsymbol{J}_n)\boldsymbol{\mathcal{E}}, \qquad \boldsymbol{W} = \boldsymbol{U}(\boldsymbol{U}'\boldsymbol{U})^{-1}\boldsymbol{U}', \tag{A.3}$$

and let $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)'$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_n)'$ be n-dimensional vectors satisfying

$$\begin{aligned} &\|\boldsymbol{\alpha}\| = \|\boldsymbol{\beta}\| = 1, \qquad \boldsymbol{1}_n'\boldsymbol{\alpha} = \boldsymbol{1}_n'\boldsymbol{\beta} = 0, \\ &\sum_{a=1}^{n} \alpha_a^2 \beta_a^2 = o(1) \quad \text{as} \quad c_{n,p} \to c_0. \end{aligned} \tag{A.4}$$

Then, we derive

$$\boldsymbol{\alpha}'\boldsymbol{W}\boldsymbol{\beta} \xrightarrow{p} c_0 \boldsymbol{\alpha}'\boldsymbol{\beta} \qquad \text{as} \qquad c_{n,p} \to c_0.$$

LEMMA A.8. *Let $\boldsymbol{\mathcal{A}}_j$ $(j \in \mathcal{J}_-)$ be an $n \times p$ matrix defined by*

$$\boldsymbol{\mathcal{A}}_j = (\boldsymbol{I}_n - \boldsymbol{P}_j)\boldsymbol{X}_*\boldsymbol{\Theta}_*\boldsymbol{\Sigma}_*^{-1/2} = \boldsymbol{H}_j\boldsymbol{L}_j^{1/2}\boldsymbol{G}_j', \tag{A.5}$$

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 $\boldsymbol{\mathcal{A}}_j$, and let $\boldsymbol{S}_j$ $(j \in \mathcal{J}_-)$ be a $p \times p$ matrix defined by

$$\boldsymbol{S}_j = \frac{1}{n}\boldsymbol{\mathcal{E}}'(\boldsymbol{I}_n - \boldsymbol{J}_n - \boldsymbol{P}_j - \boldsymbol{H}_j\boldsymbol{H}_j')\boldsymbol{\mathcal{E}}. \tag{A.6}$$

Then, we have

(i) $^\forall j \in \mathcal{J}_-$, $n\boldsymbol{\Sigma}_*^{-1/2}\hat{\boldsymbol{\Sigma}}_j\boldsymbol{\Sigma}_*^{-1/2} = (\boldsymbol{L}_j^{1/2}\boldsymbol{G}_j' + \boldsymbol{H}_j'\boldsymbol{\mathcal{E}})'(\boldsymbol{L}_j^{1/2}\boldsymbol{G}_j' + \boldsymbol{H}_j'\boldsymbol{\mathcal{E}}) + n\boldsymbol{S}_j$.

(ii) *The $n \times n$ matrix $\boldsymbol{I}_n - \boldsymbol{J}_n - \boldsymbol{P}_j - \boldsymbol{H}_j\boldsymbol{H}_j'$ is idempotent, and $\boldsymbol{P}_{j_+}(\boldsymbol{P}_j - \boldsymbol{H}_j\boldsymbol{H}_j') = \boldsymbol{P}_j - \boldsymbol{H}_j\boldsymbol{H}_j'$ holds, where j_+ is given by* (2.1).

(iii) *If assumption* A2 *holds, $\lambda_{\max}(\boldsymbol{S}_j) = O_p(p^{1/2})$ as $c_{n,p} \to c_0$ is satisfied.*

A.2. Proof of Theorem 1

Let $\mathcal{D}(j,\ell)$ $(j,\ell \in \mathcal{J})$ be the difference between two negative twofold maximum log-likelihoods divided by n, such that

$$\mathcal{D}(j,\ell) = \log(|\hat{\boldsymbol{\Sigma}}_j|/|\hat{\boldsymbol{\Sigma}}_\ell|). \tag{A.7}$$

Notice that

$$\mathrm{IC}_m(j) - \mathrm{IC}_m(\ell) = n\mathcal{D}(j,\ell) + m(j) - m(\ell). \tag{A.8}$$

From Lemma A.3 (ii) and (iii), we see that, to obtain the conditions of $m(j)$ such that $\mathrm{IC}_m(j)$ is consistent, we only have to show the convergence in probability of $\mathcal{D}(j, j_*)$ divided by some constant.

Notice that

$$\begin{aligned}\boldsymbol{\Sigma}_*^{-1/2}\hat{\boldsymbol{\Sigma}}_j\boldsymbol{\Sigma}_*^{-1/2} = \frac{1}{n}\{&\boldsymbol{\Gamma}_j\boldsymbol{\Gamma}_j' + \boldsymbol{\Sigma}_*^{-1/2}\boldsymbol{\Theta}_*'\boldsymbol{X}_*'(\boldsymbol{I}_n - \boldsymbol{P}_j)\boldsymbol{\mathcal{E}} \\ &+ \boldsymbol{\mathcal{E}}'(\boldsymbol{I}_n - \boldsymbol{P}_j)\boldsymbol{X}_*\boldsymbol{\Theta}_*\boldsymbol{\Sigma}_*^{-1/2} + \boldsymbol{\mathcal{E}}'(\boldsymbol{I}_n - \boldsymbol{J}_n - \boldsymbol{P}_j)\boldsymbol{\mathcal{E}}\},\end{aligned} \tag{A.9}$$

where $\boldsymbol{\Gamma}_j$ is given by (2.4). Let $\mathrm{vec}(\boldsymbol{A})$ denote 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). Suppose that assumptions A2 and A3 are satisfied. It follows from Lemmas A.5 and A.6 that

$$\begin{aligned}&\mathrm{tr}\{\mathrm{Cov}[\mathrm{vec}(\boldsymbol{\Sigma}_*^{-1/2}\boldsymbol{\Theta}_*'\boldsymbol{X}_*'(\boldsymbol{I}_n - \boldsymbol{P}_j)\boldsymbol{\mathcal{E}})]\} = \phi_1(\boldsymbol{\Gamma}_j\boldsymbol{\Gamma}_j') = \mathrm{tr}(\boldsymbol{\Gamma}_j\boldsymbol{\Gamma}_j')p = O(n), \\ &\mathrm{tr}\{\mathrm{Cov}[\mathrm{vec}(\boldsymbol{\mathcal{E}}'(\boldsymbol{I}_n - \boldsymbol{J}_n - \boldsymbol{P}_j)\boldsymbol{\mathcal{E}})]\} \\ &\quad = \phi_2(\boldsymbol{I}_n - \boldsymbol{J}_n - \boldsymbol{P}_j) - p\,\mathrm{tr}\{(\boldsymbol{I}_n - \boldsymbol{J}_n - \boldsymbol{P}_j)^2\} \\ &\quad = \kappa_4^{(1)}\sum_{a=1}^{n}\{(\boldsymbol{I}_n - \boldsymbol{J}_n - \boldsymbol{P}_j)_{aa}\}^2 + p(p+1)(n - k_j - 1) = O(n),\end{aligned}$$

as $n \to \infty$. These equations imply that

$$\begin{aligned}&\frac{1}{n}\boldsymbol{\Sigma}_*^{-1/2}\boldsymbol{\Theta}_*'\boldsymbol{X}_*'(\boldsymbol{I}_n - \boldsymbol{P}_j)\boldsymbol{\mathcal{E}} = O_p(n^{-1/2}), \\ &\frac{1}{n}\boldsymbol{\mathcal{E}}'(\boldsymbol{I}_n - \boldsymbol{J}_n - \boldsymbol{P}_j)\boldsymbol{\mathcal{E}} = \boldsymbol{I}_p + O_p(n^{-1/2}),\end{aligned}$$

as $n \to \infty$. Using the above results and equation (A.9) yields

$$\boldsymbol{\Sigma}_*^{-1/2}\hat{\boldsymbol{\Sigma}}_j\boldsymbol{\Sigma}_*^{-1/2} \xrightarrow{p} \begin{cases}\boldsymbol{I}_p & ({}^\forall j \in \mathcal{J}_+) \\ \boldsymbol{I}_p + \boldsymbol{\Omega}_{j,0} & ({}^\forall j \in \mathcal{J}_-)\end{cases} \quad \text{as} \quad n \to \infty. \tag{A.10}$$

The lower equation in (A.10) directly implies, for all $j \in \mathcal{J}_-$,

$$\mathcal{D}(j, j_*) \xrightarrow{p} \log|\boldsymbol{I}_p + \boldsymbol{\Omega}_{j,0}| \quad \text{as} \quad n \to \infty, \tag{A.11}$$

where $\mathcal{D}(j, j_*)$ is given by (A.7) and $\boldsymbol{\Omega}_{j,0}$ is a limiting value of $\boldsymbol{\Gamma}_j\boldsymbol{\Gamma}_j'/n$, which is defined in assumption A3. On the other hand, for all $j \in \mathcal{J}_+\backslash\{j_*\}$, we have

$$\begin{aligned}\mathcal{D}(j, j_*) &= -\log|\boldsymbol{I}_p + \boldsymbol{\mathcal{E}}'(\boldsymbol{P}_j - \boldsymbol{P}_{j_*})\boldsymbol{\mathcal{E}}\{\boldsymbol{\mathcal{E}}'(\boldsymbol{I}_n - \boldsymbol{J}_n - \boldsymbol{P}_j)\boldsymbol{\mathcal{E}}\}^{-1}| \\ &= -\frac{1}{n}\,\mathrm{tr}(\boldsymbol{Z}_j'\boldsymbol{Z}_j - \boldsymbol{Z}_{j_*}'\boldsymbol{Z}_{j_*}) + o_p(n^{-1}) \quad \text{as} \quad n \to \infty,\end{aligned} \tag{A.12}$$

where $\boldsymbol{Z}_j$ is given by (A.2). Recall that $\boldsymbol{Z}_j = O_p(1)$ under assumption A2. From this result and equation (A.12), we derive, for all $j \in \mathcal{J}_+ \backslash \{j_*\}$,

$$n\mathcal{D}(j, j_*) = O_p(1) \quad \text{as} \quad n \to \infty. \tag{A.13}$$

Thus, Lemma A.3 (ii) and (iii), and equations (A.8), (A.11), and (A.13) lead us to the conditions in Theorem 1.

A.3. Proof of Corollary 1

Notice that the probability that a model j is selected by IC_m is

$$\begin{aligned} P(\hat{j}_m = j) &= P(\cap_{\ell \in \mathcal{J} \backslash \{j\}} \{\mathrm{IC}_m(\ell) > \mathrm{IC}_m(j)\}) \\ &= P(\{\cap_{\ell \in \mathcal{J}_- \backslash \{j\}} \{\mathrm{IC}_m(\ell) > \mathrm{IC}_m(j)\}\} \\ &\qquad \cap \{\cap_{\ell \in \mathcal{J}_+ \backslash \{j\}} \{\mathrm{IC}_m(\ell) > \mathrm{IC}_m(j)\}\}). \end{aligned} \tag{A.14}$$

The same way as was previously used in the calculation of (A.11) yields $\mathcal{D}(\ell_2, \ell_1) \xrightarrow{p} \log |\boldsymbol{I}_p + \boldsymbol{\Omega}_{\ell_2,0}|$ as $n \to \infty$ for all $\ell_1 \in \mathcal{J}_+ \backslash \{j\}$ and $\ell_2 \in \mathcal{J}_- \backslash \{j\}$. It follows from this result and condition C1-3 that

$$\frac{1}{n}\{\mathrm{IC}_m(\ell_2) - \mathrm{IC}_m(\ell_1)\} \xrightarrow{p} \log |\boldsymbol{I}_p + \boldsymbol{\Omega}_{\ell_2,0}| > 0. \tag{A.15}$$

Equation (A.15) and Lemma A.3 (iii) imply that $\lim_{n\to\infty} P(\hat{j}_m = j) = 0$ holds for all $j \in \mathcal{J}_-$, and they also imply that

$$\lim_{n\to\infty} P(\mathrm{IC}_m(\ell_2) > \mathrm{IC}_m(\ell_1)) = 1.$$

Using the above equation and Lemma A.4 (ii), we have

$$\lim_{n\to\infty} P(\cap_{\ell \in \mathcal{J}_- \backslash \{j\}} \{\mathrm{IC}_m(\ell) > \mathrm{IC}_m(j)\}) = 1, \qquad (\forall j \in \mathcal{J}_+).$$

Thus, from equation (A.14) and Lemma A.4 (ii), we can see that

$$\begin{aligned} &\lim_{n\to\infty} P(\hat{j}_m = j) \\ &\quad = \begin{cases} 0 & (j \in \mathcal{J}_-) \\ \lim_{n\to\infty} P(\cap_{\ell \in \mathcal{J}_+ \backslash \{j\}} \{\mathrm{IC}_m(\ell) > \mathrm{IC}_m(j)\}) & (j \in \mathcal{J}_+) \end{cases}. \end{aligned} \tag{A.16}$$

On the other hand, by using equation (A.12), we have, for all $j, \ell \in \mathcal{J}_+$,

$$n\mathcal{D}(j, \ell) = n\{\mathcal{D}(j, j_*) - \mathcal{D}(\ell, j_*)\} = -\operatorname{tr}(\boldsymbol{Z}_j'\boldsymbol{Z}_j - \boldsymbol{Z}_\ell'\boldsymbol{Z}_\ell) + o_p(1) \quad \text{as} \quad n \to \infty.$$

This equation and $\lim_{n\to\infty}\{m(j) - m(\ell)\} = m_0 p(k_j - k_\ell)$ for all $j, \ell \in \mathcal{J}_+$ imply that

$$\begin{aligned} &\mathrm{IC}_m(j) - \mathrm{IC}_m(\ell) \\ &\quad \xrightarrow{p} -\operatorname{tr}(\boldsymbol{Z}_j'\boldsymbol{Z}_j - \boldsymbol{Z}_\ell'\boldsymbol{Z}_\ell) + m_0 p(k_j - k_\ell) \quad \text{as} \quad n \to \infty. \end{aligned} \tag{A.17}$$

Notice that $\mathrm{tr}(\boldsymbol{Z}_j'\boldsymbol{Z}_j) = \mathrm{vec}(\boldsymbol{Z}_j)'\mathrm{vec}(\boldsymbol{Z}_j)$ and

$$\begin{aligned}\mathrm{Cov}[\mathrm{vec}(\boldsymbol{Z}_j), \mathrm{vec}(\boldsymbol{Z}_\ell)] &= \boldsymbol{I}_p \otimes (\boldsymbol{X}_j'\boldsymbol{X}_j)^{-1/2}\boldsymbol{X}_j'\boldsymbol{X}_\ell(\boldsymbol{X}_\ell'\boldsymbol{X}_\ell)^{-1/2} \\ &\to \boldsymbol{I}_p \otimes \boldsymbol{R}_{j,j,0}^{-1/2}\boldsymbol{R}_{j,\ell,0}\boldsymbol{R}_{\ell,\ell,0}^{-1/2} \quad \text{as} \quad n \to \infty,\end{aligned}$$

where $\boldsymbol{R}_{j,\ell,0}$ is the submatrix of $\boldsymbol{R}_0$, which is defined in assumption A4. Moreover, it follows from Lemma A.2 that $\mathrm{vec}(\boldsymbol{Z}_j) \xrightarrow{d} N_{k_jp}(\boldsymbol{0}_{k_jp}, \boldsymbol{I}_{k_jp})$ as $n \to \infty$. Substituting equation (A.17) into equation (A.16) yields equation (3.1) in Corollary 1.

Let $\boldsymbol{z}_j$ be the same random vector as in (3.1). 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

$$P\left(\max_{\alpha=1,\ldots,k-k_*} \sum_{i=1}^{\alpha}(w_i - m_0p) < 0\right), \tag{A.18}$$

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.18) is explicitly evaluated as (3.2).

A.4. Proof of Theorem 2

Just like the proof of Theorem 1, to obtain the conditions of $m(j)$ such that $\mathrm{IC}_m(j)$ is consistent, we only have to show the convergence in probability of $\mathcal{D}(j, j_*)$ in (A.7), or lower bounds of $\mathcal{D}(j, j_*)$ divided by some constant.

Let us give another expression of $\boldsymbol{Q}_j$ as $\boldsymbol{Q}_j = (\boldsymbol{b}_{j,1}, \ldots, \boldsymbol{b}_{j,k_j})$, where $\boldsymbol{b}_{j,a} = (q_{j,1a}, \ldots, q_{j,na})'$ and $\boldsymbol{Q}_j$ is given by (A.1). Then, it is clear that $\boldsymbol{b}_{j,a}'\boldsymbol{b}_{j,b} = \delta_{ab}$, because $\boldsymbol{Q}_j'\boldsymbol{Q}_j = \boldsymbol{I}_{k_j}$ holds. Moreover, $\boldsymbol{Q}_j'\boldsymbol{1}_n = \boldsymbol{0}_{k_j}$ holds because $\boldsymbol{X}_j$ is centralized. From these equations and Lemma A.1, it can be determined that $\boldsymbol{b}_{j,1}, \ldots, \boldsymbol{b}_{j,k_j}$ satisfy the conditions in (A.4) when assumptions A4 and A5 hold. Therefore, if assumptions A4 and A5 hold, it follows from Lemma A.7 that

$$\boldsymbol{b}_{j,a}'\boldsymbol{W}\boldsymbol{b}_{j,b} \xrightarrow{p} c_0\delta_{ab} \quad \text{as} \quad c_{n,p} \to c_0,$$

where $\boldsymbol{W}$ is given by (A.3). Since $\boldsymbol{b}_{j,a}'\boldsymbol{W}\boldsymbol{b}_{j,b}$ is the (a,b)th element of $\boldsymbol{Q}_j'\boldsymbol{W}\boldsymbol{Q}_j$, the following equation is satisfied if assumptions A4 and A5 hold:

$$\boldsymbol{Q}_j'\boldsymbol{W}\boldsymbol{Q}_j \xrightarrow{p} c_0\boldsymbol{I}_{k_j} \quad \text{as} \quad c_{n,p} \to c_0. \tag{A.19}$$

First, we show convergence in probability of $\mathcal{D}(j, j_*)$ when $j \in \mathcal{J}_+$. Notice that $\boldsymbol{P}_j\boldsymbol{\mathcal{E}} = \boldsymbol{P}_j\boldsymbol{U}$ holds for all $j \in \mathcal{J}$ because $\boldsymbol{X}_j$ is centralized, where $\boldsymbol{U}$ is given by (A.3). Then, by using equation (A.19) and the property of the determination

(see, e.g., Harville (1997), chap. 18, cor. 18.1.2), the following equation is satisfied for all $j \in \mathcal{J}_+ \backslash \{j_*\}$ under assumptions A1, A4 and A5:

$$\begin{aligned}
\mathcal{D}(j, j_*) &= \log \frac{|\boldsymbol{\mathcal{E}}'(\boldsymbol{I}_n - \boldsymbol{J}_n - \boldsymbol{P}_j)\boldsymbol{\mathcal{E}}|}{|\boldsymbol{\mathcal{E}}'(\boldsymbol{I}_n - \boldsymbol{J}_n - \boldsymbol{P}_{j_*})\boldsymbol{\mathcal{E}}|} = \log \frac{|\boldsymbol{U}'(\boldsymbol{I}_n - \boldsymbol{P}_j)\boldsymbol{U}|}{|\boldsymbol{U}'(\boldsymbol{I}_n - \boldsymbol{P}_{j_*})\boldsymbol{U}|} \\
&= \log \frac{|\boldsymbol{I}_p - (\boldsymbol{U}'\boldsymbol{U})^{-1}\boldsymbol{U}'\boldsymbol{P}_j\boldsymbol{U}|}{|\boldsymbol{I}_p - (\boldsymbol{U}'\boldsymbol{U})^{-1}\boldsymbol{U}'\boldsymbol{P}_{j_*}\boldsymbol{U}|} = \log \frac{|\boldsymbol{I}_{k_j} - \boldsymbol{Q}_j'\boldsymbol{W}\boldsymbol{Q}_j|}{|\boldsymbol{I}_{k_*} - \boldsymbol{Q}_{j_*}'\boldsymbol{W}\boldsymbol{Q}_{j_*}|} \\
&\xrightarrow{p} (k_j - k_*)\log(1 - c_0) \quad \text{as} \quad c_{n,p} \to c_0.
\end{aligned} \tag{A.20}$$

Next, we show convergence in probability of a lower bound of $\mathcal{D}(j, j_*)/\log p$ when $j \in \mathcal{J}_-$. The asymptotic behavior of this term depends on whether we assume A6 or A7. At the beginning, we assume A6. Let $\boldsymbol{C}_j$ be a $\gamma_j \times \gamma_j$ orthogonal matrix that diagonalizes $\boldsymbol{\Gamma}_j'\boldsymbol{\Gamma}_j$ to $\boldsymbol{L}_j$, and hence $\boldsymbol{\Gamma}_j'\boldsymbol{\Gamma}_j = \boldsymbol{C}_j\boldsymbol{L}_j\boldsymbol{C}_j'$. It follows from Lemma A.8 (i) that for all $j \in \mathcal{J}_-$

$$\begin{aligned}
\frac{|\hat{\boldsymbol{\Sigma}}_j|}{|\hat{\boldsymbol{\Sigma}}_{j_*}|} &= \frac{|n\boldsymbol{S}_j + (\boldsymbol{L}_j^{1/2}\boldsymbol{G}_j' + \boldsymbol{H}_j'\boldsymbol{\mathcal{E}})'(\boldsymbol{L}_j^{1/2}\boldsymbol{G}_j' + \boldsymbol{H}_j'\boldsymbol{\mathcal{E}})|}{|\boldsymbol{\mathcal{E}}'(\boldsymbol{I}_n - \boldsymbol{J}_n - \boldsymbol{P}_{j_*})\boldsymbol{\mathcal{E}}|} \\
&= \frac{|\boldsymbol{I}_p + \boldsymbol{S}_j^{-1}(\boldsymbol{L}_j^{1/2}\boldsymbol{G}_j' + \boldsymbol{H}_j'\boldsymbol{\mathcal{E}})'(\boldsymbol{L}_j^{1/2}\boldsymbol{G}_j' + \boldsymbol{H}_j'\boldsymbol{\mathcal{E}})/n||n\boldsymbol{S}_j|}{|\boldsymbol{U}'(\boldsymbol{I}_n - \boldsymbol{P}_{j_*})\boldsymbol{U}|} \\
&= \frac{|\boldsymbol{I}_{\gamma_j} + (\boldsymbol{L}_j^{1/2}\boldsymbol{G}_j' + \boldsymbol{H}_j'\boldsymbol{\mathcal{E}})\boldsymbol{S}_j^{-1}(\boldsymbol{L}_j^{1/2}\boldsymbol{G}_j' + \boldsymbol{H}_j'\boldsymbol{\mathcal{E}})'/n||n\boldsymbol{S}_j|}{|\boldsymbol{U}'(\boldsymbol{I}_n - \boldsymbol{P}_{j_*})\boldsymbol{U}|},
\end{aligned} \tag{A.21}$$

where $\boldsymbol{H}_j$, $\boldsymbol{L}_j$, and $\boldsymbol{G}_j$ are given in (A.5); and $\boldsymbol{S}_j$ is given by (A.6). Hence, we have

$$\frac{|\hat{\boldsymbol{\Sigma}}_j|}{|\hat{\boldsymbol{\Sigma}}_{j_*}|} \geq \frac{|\lambda_{\max}(\boldsymbol{S}_j)\boldsymbol{I}_{\gamma_j} + p\hat{\boldsymbol{\Delta}}_j||n\boldsymbol{S}_j|}{\lambda_{\max}(\boldsymbol{S}_j)^{\gamma_j}|\boldsymbol{U}'(\boldsymbol{I}_n - \boldsymbol{P}_{j_*})\boldsymbol{U}|},$$

where $\hat{\boldsymbol{\Delta}}_j$ is a $\gamma_j \times \gamma_j$ matrix defined by

$$\hat{\boldsymbol{\Delta}}_j = \frac{1}{np}\boldsymbol{C}_j(\boldsymbol{L}_j^{1/2}\boldsymbol{G}_j' + \boldsymbol{H}_j'\boldsymbol{\mathcal{E}})(\boldsymbol{L}_j^{1/2}\boldsymbol{G}_j' + \boldsymbol{H}_j'\boldsymbol{\mathcal{E}})'\boldsymbol{C}_j'.$$

This implies that

$$\begin{aligned}
\mathcal{D}(j, j_*) &\geq \log|\lambda_{\max}(\boldsymbol{S}_j)\boldsymbol{I}_{\gamma_j} + p\hat{\boldsymbol{\Delta}}_j| \\
&\quad + \log\frac{|n\boldsymbol{S}_j|}{|\boldsymbol{U}'(\boldsymbol{I}_n - \boldsymbol{P}_{j_*})\boldsymbol{U}|} - \gamma_j\log\lambda_{\max}(\boldsymbol{S}_j) \\
&= \mathcal{D}\mathrm{l}_{1,1}(j) + \mathcal{D}\mathrm{l}_2(j) + \mathcal{D}\mathrm{l}_3(j).
\end{aligned} \tag{A.22}$$

We first evaluate the asymptotic behavior of $\mathcal{D}\mathrm{l}_{1,1}(j)$ in (A.22). Recall that $\boldsymbol{C}_j\boldsymbol{L}_j\boldsymbol{C}_j' = \boldsymbol{\Gamma}_j'\boldsymbol{\Gamma}_j = O(np)$ as $c_{n,p} \to c_0$. It is easy to see that $E[\boldsymbol{C}_j\boldsymbol{H}_j'\boldsymbol{\mathcal{E}}\boldsymbol{\mathcal{E}}'\boldsymbol{H}_j\boldsymbol{C}_j'] = p\boldsymbol{I}_{\gamma_j}$. Furthermore, it follows from Lemmas A.5 and A.6 that

$$\begin{aligned}
&E[\mathrm{tr}\{(\boldsymbol{C}_j\boldsymbol{H}_j'\boldsymbol{\mathcal{E}}\boldsymbol{\mathcal{E}}'\boldsymbol{H}_j\boldsymbol{C}_j' - p\boldsymbol{I}_{\gamma_j})'(\boldsymbol{C}_j\boldsymbol{H}_j'\boldsymbol{\mathcal{E}}\boldsymbol{\mathcal{E}}'\boldsymbol{H}_j\boldsymbol{C}_j' - p\boldsymbol{I}_{\gamma_j})\}] \\
&\quad = \phi_2(\boldsymbol{H}_j\boldsymbol{H}_j') - p^2\gamma_j \\
&\quad = \kappa_4^{(1)}\sum_{a=1}^{n}\{(\boldsymbol{H}_j\boldsymbol{H}_j')_{aa}\}^2 + p\gamma_j(\gamma_j + 1) = O(\max\{p, p^s\}) \quad \text{as} \quad c_{n,p} \to c_0,
\end{aligned}$$

where $\kappa_4^{(1)}$ is given by (2.5), and s is some positive constant given by (2.6). These equations imply that $\boldsymbol{C}_j\boldsymbol{H}_j'\boldsymbol{\mathcal{E}}\boldsymbol{\mathcal{E}}'\boldsymbol{H}_j\boldsymbol{C}_j' = p\boldsymbol{I}_{\gamma_j} + O_p(\max\{p^{1/2}, p^{s/2}\}) = O_p(p)$ as $c_{n,p} \to c_0$. Moreover, from Hölder's inequality, we have

$$\begin{aligned}
\mathrm{tr}(\boldsymbol{C}_j\boldsymbol{L}_j^{1/2}\boldsymbol{G}_j'\boldsymbol{\mathcal{E}}'\boldsymbol{H}_j\boldsymbol{C}_j')^2 &= \{\mathrm{vec}(\boldsymbol{G}_j\boldsymbol{L}_j^{1/2}\boldsymbol{C}_j')'\mathrm{vec}(\boldsymbol{\mathcal{E}}'\boldsymbol{H}_j\boldsymbol{C}_j')\}^2 \\
&\leq \|\mathrm{vec}(\boldsymbol{G}_j\boldsymbol{L}_j^{1/2}\boldsymbol{C}_j')\|^2\|\mathrm{vec}(\boldsymbol{\mathcal{E}}'\boldsymbol{H}_j\boldsymbol{C}_j')\|^2 \\
&= \mathrm{tr}(\boldsymbol{\Gamma}_j'\boldsymbol{\Gamma}_j)\,\mathrm{tr}(\boldsymbol{C}_j\boldsymbol{H}_j'\boldsymbol{\mathcal{E}}\boldsymbol{\mathcal{E}}'\boldsymbol{H}_j\boldsymbol{C}_j') \\
&= O_p(np^2) \quad \text{as} \quad c_{n,p} \to c_0.
\end{aligned}$$

This implies that $\boldsymbol{C}_j\boldsymbol{L}_j^{1/2}\boldsymbol{G}_j'\boldsymbol{\mathcal{E}}'\boldsymbol{H}_j\boldsymbol{C}_j' = O_p(n^{1/2}p)$ as $c_{n,p} \to c_0$. Therefore, we have

$$\hat{\boldsymbol{\Delta}}_j \xrightarrow{p} \boldsymbol{\Delta}_{j,0} \quad \text{as} \quad c_{n,p} \to c_0,$$

where $\boldsymbol{\Delta}_{j,0}$ is a limiting value of $\boldsymbol{\Gamma}_j'\boldsymbol{\Gamma}_j/(np)$, which is defined in assumption A6. It follows from Lemma A.8 (iii) that $\lambda_{\max}(\boldsymbol{S}_j)\boldsymbol{I}_{\gamma_j} = O_p(p^{1/2})$ as $c_{n,p} \to c_0$ if assumption A2 holds. By using these equations, we derive

$$\left|\frac{1}{p}\{\lambda_{\max}(\boldsymbol{S}_j)\boldsymbol{I}_{\gamma_j} + p\hat{\boldsymbol{\Delta}}_j\}\right| \xrightarrow{p} |\boldsymbol{\Delta}_{j,0}| \quad \text{as} \quad c_{n,p} \to c_0. \tag{A.23}$$

Notice that

$$\mathcal{D}\mathrm{l}_{1,1}(j) = \log\left[p^{\gamma_j}\left|\frac{1}{p}\{\lambda_{\max}(\boldsymbol{S}_j)\boldsymbol{I}_{\gamma_j} + p\hat{\boldsymbol{\Delta}}_j\}\right|\right].$$

It follows from the above results and the positive definiteness of $\boldsymbol{\Delta}_{j,0}$ that

$$\frac{1}{\log p}\mathcal{D}\mathrm{l}_{1,1}(j) \xrightarrow{p} \gamma_j \quad \text{as} \quad c_{n,p} \to c_0. \tag{A.24}$$

Next, we evaluate the asymptotic behavior of $\mathcal{D}\mathrm{l}_2(j)$ in (A.22). From equation (A.19) and the result $(\boldsymbol{I}_n - \boldsymbol{P}_j - \boldsymbol{H}_j\boldsymbol{H}_j')(\boldsymbol{I}_n - \boldsymbol{P}_j) = \boldsymbol{I}_n - \boldsymbol{P}_j - \boldsymbol{H}_j\boldsymbol{H}_j'$, obtained from Lemma A.8 (ii), we can see that

$$\begin{aligned}
\mathcal{D}\mathrm{l}_2(j) &\leq \log\frac{|\boldsymbol{U}'(\boldsymbol{I}_n - \boldsymbol{P}_j)\boldsymbol{U}|}{|\boldsymbol{U}'(\boldsymbol{I}_n - \boldsymbol{P}_{j_*})\boldsymbol{U}|} = \log\frac{|\boldsymbol{I}_{k_j} - \boldsymbol{Q}_j'\boldsymbol{W}\boldsymbol{Q}_j|}{|\boldsymbol{I}_{k_{j_*}} - \boldsymbol{Q}_{j_*}'\boldsymbol{W}\boldsymbol{Q}_{j_*}|} \\
&\xrightarrow{p} (k_j - k_*)\log(1 - c_0) \quad \text{as} \quad c_{n,p} \to c_0.
\end{aligned}$$

It follows from Lemma A.8 (ii) that $(\boldsymbol{I}_n - \boldsymbol{P}_{j_+})(\boldsymbol{I}_n - \boldsymbol{P}_j - \boldsymbol{H}_j\boldsymbol{H}_j') = \boldsymbol{I}_n - \boldsymbol{P}_{j_+}$, where j_+ is given by (2.1). Thus, we also have

$$\begin{aligned}
\mathcal{D}\mathrm{l}_2(j) &\geq \log\frac{|\boldsymbol{U}'(\boldsymbol{I}_n - \boldsymbol{P}_{j_+})\boldsymbol{U}|}{|\boldsymbol{U}'(\boldsymbol{I}_n - \boldsymbol{P}_{j_*})\boldsymbol{U}|} = \log\frac{|\boldsymbol{I}_{k_{j_+}} - \boldsymbol{Q}_{j_+}'\boldsymbol{W}\boldsymbol{Q}_{j_+}|}{|\boldsymbol{I}_{k_{j_*}} - \boldsymbol{Q}_{j_*}'\boldsymbol{W}\boldsymbol{Q}_{j_*}|} \\
&\xrightarrow{p} (k_{j_+} - k_*)\log(1 - c_0) \quad \text{as} \quad c_{n,p} \to c_0.
\end{aligned}$$

The above upper and lower bounds of $\mathcal{D}\mathrm{l}_2(j)$ imply that

$$\frac{1}{\log p}\mathcal{D}\mathrm{l}_2(j) \xrightarrow{p} 0 \quad \text{as} \quad c_{n,p} \to c_0. \tag{A.25}$$

Finally, we evaluate the asymptotic behavior of $\mathcal{D}\mathrm{l}_3(j)$ in (A.22). The asymptotic behavior of this term depends on whether we assume A2 or A2$'$. Since $-\log x \le -x+1$ for all $x \ge 0$, we have

$$\begin{aligned}\mathcal{D}\mathrm{l}_3(j) &= -\frac{1}{2}\gamma_j \log p - \gamma_j \log \frac{\lambda_{\max}(\boldsymbol{S}_j)}{\sqrt{p}} \\ &\ge -\frac{1}{2}\gamma_j \log p - \gamma_j \left\{ \frac{\lambda_{\max}(\boldsymbol{S}_j)}{\sqrt{p}} - 1 \right\} = \mathcal{D}\mathrm{l}_{3,1}(j).\end{aligned}$$

It follows from Lemma A.8 (iii) that $\lambda_{\max}(\boldsymbol{S}_j)/p^{1/2}$ is $O_p(1)$ as $c_{n,p} \to c_0$ when assumption A2 holds. This implies that

$$\frac{1}{\log p}\mathcal{D}\mathrm{l}_3(j) \xrightarrow{p} -\frac{1}{2}\gamma_j \quad \text{as} \quad c_{n,p} \to c_0. \tag{A.26}$$

On the other hand, another lower bound of $\mathcal{D}\mathrm{l}_3(j)$ is derived as

$$\mathcal{D}\mathrm{l}_3(j) \ge -\gamma_j\{\lambda_{\max}(\boldsymbol{S}_j) - 1\} \ge -\gamma_j\{\lambda_{\max}(\boldsymbol{\mathcal{E}}'\boldsymbol{\mathcal{E}}/n) - 1\} = \mathcal{D}\mathrm{l}_{3,2}(j).$$

If assumption A2$'$ holds instead of assumption A2, it follows from Bai and Yin (1993) that $\lambda_{\max}(\boldsymbol{\mathcal{E}}'\boldsymbol{\mathcal{E}}/n) \xrightarrow{\text{a.s.}} (1+c_0^{1/2})^2$ as $c_{n,p} \to c_0$. This implies that

$$\frac{1}{\log p}\mathcal{D}\mathrm{l}_{3,2}(j) \xrightarrow{p} 0 \quad \text{as} \quad c_{n,p} \to c_0. \tag{A.27}$$

Combining (A.22), (A.25), (A.24), (A.26), and (A.27) yields

$$\frac{1}{\log p}\mathcal{D}(j,j_*) \ge \begin{cases} \{\mathcal{D}\mathrm{l}_{1,1}(j) + \mathcal{D}\mathrm{l}_2(j) + \mathcal{D}\mathrm{l}_{3,1}(j)\}/\log p \xrightarrow{p} \gamma_j/2 \\ \qquad\qquad\qquad\qquad\qquad\qquad \text{(when A2 holds)} \\ \{\mathcal{D}\mathrm{l}_{1,1}(j) + \mathcal{D}\mathrm{l}_2(j) + \mathcal{D}\mathrm{l}_{3,2}(j)\}/\log p \xrightarrow{p} \gamma_j \\ \qquad\qquad\qquad\qquad\qquad\qquad \text{(when A2}' \text{ holds)} \end{cases}, \tag{A.28}$$

as $c_{n,p} \to c_0$.

Secondly, we assume A7 instead of A6. Let $\alpha_{j,a}$ be the ath largest eigenvalue of $\boldsymbol{\Gamma}_j'\boldsymbol{\Gamma}_j$. Without loss of generality, we express that $\boldsymbol{L}_j = \mathrm{diag}(\alpha_{j,1},\ldots,\alpha_{j,\gamma_j})$, $\boldsymbol{G}_j = (\boldsymbol{g}_{j,1},\ldots,\boldsymbol{g}_{j,\gamma_j})$ and $\boldsymbol{H}_j = (\boldsymbol{h}_{j,1},\ldots,\boldsymbol{h}_{j,\gamma_j})$. By using these expressions, $n\boldsymbol{\Sigma}_*^{-1/2}\hat{\boldsymbol{\Sigma}}_j\boldsymbol{\Sigma}_*^{-1/2}$ can be rewritten as

$$n\boldsymbol{\Sigma}_*^{-1/2}\hat{\boldsymbol{\Sigma}}_j\boldsymbol{\Sigma}_*^{-1/2} = n\boldsymbol{S}_j + \sum_{a=1}^{\gamma_j}(\sqrt{\alpha_{j,a}}\boldsymbol{g}_{j,a} + \boldsymbol{\mathcal{E}}'\boldsymbol{h}_{j,a})(\sqrt{\alpha_{j,a}}\boldsymbol{g}_{j,a} + \boldsymbol{\mathcal{E}}'\boldsymbol{h}_{j,a})'.$$

It follows from the above expression, and equations (A.21) and (A.22) that

$$\begin{aligned}\mathcal{D}(j,j_*) &\ge \frac{|n\boldsymbol{S}_j + (\sqrt{\alpha_{j,1}}\boldsymbol{g}_{j,1} + \boldsymbol{\mathcal{E}}'\boldsymbol{h}_{j,1})(\sqrt{\alpha_{j,1}}\boldsymbol{g}_{j,1} + \boldsymbol{\mathcal{E}}'\boldsymbol{h}_{j,1})'|}{|\boldsymbol{\mathcal{E}}'(\boldsymbol{I}_n - \boldsymbol{J}_n - \boldsymbol{P}_{j_*})\boldsymbol{\mathcal{E}}|} \\ &= \frac{\{1 + (\sqrt{\alpha_{j,1}}\boldsymbol{g}_{j,1} + \boldsymbol{\mathcal{E}}'\boldsymbol{h}_{j,1})'\boldsymbol{S}_j^{-1}(\sqrt{\alpha_{j,1}}\boldsymbol{g}_{j,1} + \boldsymbol{\mathcal{E}}'\boldsymbol{h}_{j,1})/n\}|n\boldsymbol{S}_j|}{|\boldsymbol{\mathcal{E}}'(\boldsymbol{I}_n - \boldsymbol{J}_n - \boldsymbol{P}_{j_*})\boldsymbol{\mathcal{E}}|} \\ &\ge \mathcal{D}\mathrm{l}_{1,2}(j) + \mathcal{D}\mathrm{l}_2(j) + \frac{1}{\gamma_j}\mathcal{D}\mathrm{l}_3(j),\end{aligned} \tag{A.29}$$

where $\mathcal{D}\mathrm{l}_2(j)$ and $\mathcal{D}\mathrm{l}_3(j)$ are given by (A.22), and $\mathcal{D}\mathrm{l}_{1,2}(j)$ is defined by

$$\mathcal{D}\mathrm{l}_{1,2}(j) = \log\{\lambda_{\max}(\boldsymbol{S}_j) + p\hat{\xi}_j\}.$$

Here, $\hat{\xi}_j$ is defined by

$$\hat{\xi}_j = \frac{1}{np}(\sqrt{\alpha_{j,1}}\boldsymbol{g}_{j,1} + \boldsymbol{\mathcal{E}}'\boldsymbol{h}_{j,1})'(\sqrt{\alpha_{j,1}}\boldsymbol{g}_{j,1} + \boldsymbol{\mathcal{E}}'\boldsymbol{h}_{j,1}).$$

By using the same method as in the proof of equation (A.23), $\hat{\xi}_j \xrightarrow{p} \xi_{j,0}$ as $c_{n,p} \to c_0$ is derived under assumption A7, where $\xi_{j,0}$ is a limiting value of $\lambda_{\max}(\boldsymbol{\Gamma}_j'\boldsymbol{\Gamma}_j)/(np)$, which is defined in assumption A7. This imples that

$$\text{(A.30)} \quad \frac{1}{\log p}\mathcal{D}\mathrm{l}_{1,2}(j) = \frac{1}{\log p}\log\left\{\frac{1}{p}\lambda_{\max}(\boldsymbol{S}_j) + \hat{\xi}_j\right\} + 1 \xrightarrow{p} 1 \quad \text{as} \quad c_{n,p} \to c_0.$$

Combining (A.28), (A.29) and (A.30) yields

$$\text{(A.31)} \quad \frac{1}{\log p}\mathcal{D}(j, j_*) \geq \begin{cases} \{\mathcal{D}\mathrm{l}_{1,2}(j) + \mathcal{D}\mathrm{l}_2(j) + \mathcal{D}\mathrm{l}_{3,1}(j)/\gamma_j\}/\log p \xrightarrow{p} 1/2 \\ \qquad\qquad\qquad\qquad\qquad\qquad \text{(when A2 holds)} \\ \{\mathcal{D}\mathrm{l}_{1,2}(j) + \mathcal{D}\mathrm{l}_2(j) + \mathcal{D}\mathrm{l}_{3,2}(j)/\gamma_j\}/\log p \xrightarrow{p} 1 \\ \qquad\qquad\qquad\qquad\qquad\qquad \text{(when A2}' \text{ holds)} \end{cases},$$

as $c_{n,p} \to c_0$.

Thirdly, in order to obtain the conditions of $m(j)$ such that $\mathrm{IC}_m(j)$ is inconsistent, we show the convergence in probability of the upper bound of $\mathcal{D}(j, j_*)/\log p$ when $j \in \mathcal{J}_-$ under assumptions A2′ and A6. Notice that $n\boldsymbol{S}_j = \boldsymbol{U}'\boldsymbol{U} - \boldsymbol{\mathcal{E}}'(\boldsymbol{P}_j + \boldsymbol{H}_j\boldsymbol{H}_j')\boldsymbol{\mathcal{E}}$, where $\boldsymbol{U}$ is given by (A.3). Hence, we derive

$$\begin{aligned} \text{(A.32)} \quad & |n\boldsymbol{S}_j + (\boldsymbol{L}_j^{1/2}\boldsymbol{G}_j' + \boldsymbol{H}_j'\boldsymbol{\mathcal{E}})'(\boldsymbol{L}_j^{1/2}\boldsymbol{G}_j' + \boldsymbol{H}_j'\boldsymbol{\mathcal{E}})| \\ & \leq |\boldsymbol{U}'\boldsymbol{U} + (\boldsymbol{L}_j^{1/2}\boldsymbol{G}_j' + \boldsymbol{H}_j'\boldsymbol{\mathcal{E}})'(\boldsymbol{L}_j^{1/2}\boldsymbol{G}_j' + \boldsymbol{H}_j'\boldsymbol{\mathcal{E}})| \\ & = |\boldsymbol{I}_{\gamma_j} + (\boldsymbol{L}_j^{1/2}\boldsymbol{G}_j' + \boldsymbol{H}_j'\boldsymbol{\mathcal{E}})(\boldsymbol{U}'\boldsymbol{U}/n)^{-1}(\boldsymbol{L}_j^{1/2}\boldsymbol{G}_j' + \boldsymbol{H}_j'\boldsymbol{\mathcal{E}})'/n| \\ & \quad \cdot |\boldsymbol{U}'\boldsymbol{U}/n|. \end{aligned}$$

It follows from Jensen's inequality that

$$\begin{aligned} & |\boldsymbol{I}_{\gamma_j} + (\boldsymbol{L}_j^{1/2}\boldsymbol{G}_j' + \boldsymbol{H}_j'\boldsymbol{\mathcal{E}})(\boldsymbol{U}'\boldsymbol{U}/n)^{-1}(\boldsymbol{L}_j^{1/2}\boldsymbol{G}_j' + \boldsymbol{H}_j'\boldsymbol{\mathcal{E}})'/n|^{1/\gamma_j} \\ & \quad \leq \frac{1}{\gamma_j}\operatorname{tr}\{\boldsymbol{I}_{\gamma_j} + (\boldsymbol{L}_j^{1/2}\boldsymbol{G}_j' + \boldsymbol{H}_j'\boldsymbol{\mathcal{E}})(\boldsymbol{U}'\boldsymbol{U}/n)^{-1}(\boldsymbol{L}_j^{1/2}\boldsymbol{G}_j' + \boldsymbol{H}_j'\boldsymbol{\mathcal{E}})'/n\} \\ & \quad = 1 + \frac{1}{n\gamma_j}\operatorname{tr}\{(\boldsymbol{L}_j^{1/2}\boldsymbol{G}_j' + \boldsymbol{H}_j'\boldsymbol{\mathcal{E}})(\boldsymbol{U}'\boldsymbol{U}/n)^{-1}(\boldsymbol{L}_j^{1/2}\boldsymbol{G}_j' + \boldsymbol{H}_j'\boldsymbol{\mathcal{E}})'\} \\ & \quad \leq 1 + \frac{p}{\gamma_j\lambda_{\min}(\boldsymbol{U}'\boldsymbol{U}/n)}\operatorname{tr}(\hat{\boldsymbol{\Delta}}_j) = \mathcal{R}(j), \end{aligned}$$

where $\lambda_{\min}(\boldsymbol{A})$ denotes the smallest eigenvalue of a matrix $\boldsymbol{A}$, and $\hat{\boldsymbol{\Delta}}_j$ is given by (A.4). By using the above result, and equations (A.21) and (A.32), we obtain

$$\mathcal{D}(j, j_*) \leq \gamma_j \log \mathcal{R}(j) + \log \frac{|\boldsymbol{U}'\boldsymbol{U}|}{|\boldsymbol{U}'(\boldsymbol{I}_n - \boldsymbol{P}_{j_*})\boldsymbol{U}|} = \mathcal{D}\mathrm{u}_1(j) + \mathcal{D}\mathrm{u}_2(j). \tag{A.33}$$

Using the same method as in the proof of equation (A.25) yields

$$\frac{1}{\log p}\mathcal{D}\mathrm{u}_2(j) \xrightarrow{p} 0 \quad \text{as} \quad c_{n,p} \to c_0. \tag{A.34}$$

It follows form Xiao and Zhou (2010) that $\lambda_{\min}(\boldsymbol{U}'\boldsymbol{U}/n) \xrightarrow{\text{a.s.}} (1-c_0^{1/2})^2$ as $c_{n,p} \to c_0$ when assumption A2′ holds. By using this result and the same method as in the proof of (A.24), we derive $\mathcal{R}(j)/p \xrightarrow{p} \mathrm{tr}(\boldsymbol{\Delta}_{j,0})/\{\gamma_j(1-c_0^{1/2})^2\}$ as $c_{n,p} \to c_0$. This implies that

$$\frac{1}{\log p}\mathcal{D}\mathrm{u}_1(j) = \frac{1}{\log p}\gamma_j \log\{\mathcal{R}(j)/p\} + \gamma_j \xrightarrow{p} \gamma_j \quad \text{as} \quad c_{n,p} \to c_0. \tag{A.35}$$

From the results (A.33), (A.34) and equation (A.35), under assumption A2′, we have

$$\frac{1}{\log p}\mathcal{D}(j, j_*) \leq \frac{1}{\log p}\{\mathcal{D}\mathrm{u}_1(j) + \mathcal{D}\mathrm{u}_2(j)\} \xrightarrow{p} \gamma_j \quad \text{as} \quad c_{n,p} \to c_0. \tag{A.36}$$

Consequently, the results (A.20), (A.28), (A.31) and (A.36), equation (A.8), and Lemma A.3 (ii) and (iii) lead us to the conditions in Theorem 2.

A.5. Proof of Corollary 2

First, we consider the AIC and AIC$_\mathrm{c}$. Notice that $m(j) - m(j_*)$ in the AIC$_\mathrm{c}$ can be expanded as

$$m(j) - m(j_*) = \frac{(k_j - k_*)(2 - c_{n,p})p}{(1 - c_{n,p})^2} + O(pn^{-1}) \quad \text{as} \quad c_{n,p} \to c_0. \tag{A.37}$$

Hence, the differences between the penalty terms of the AICs and the AIC$_\mathrm{c}$s converge as

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

This indicates that condition C2-2 holds for the AIC and AIC$_\mathrm{c}$. Furthermore, it follows from equality (A.37) that

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

Notice that, in $0 \leq c < 1$, $c^{-1}\log(1-c)+2$ is a monotonically decreasing function, and $c^{-1}\log(1-c) + (1-c)^{-1} + (1-c)^{-2}$ is a monotonically increasing function. Hence, when $j \in \mathcal{J}\backslash\{j_*\}$, the penalty terms in the AIC$_\mathrm{c}$ always satisfy condition

C2-1, and those in the AIC satisfy condition C2-1 if $c_0 \in [0, c_{\mathrm{a}})$, where c_{a} is given by (3.4). Next, we consider the BIC and CAIC. When $j \in \mathcal{J}_+ \backslash \{j_*\}$, the difference between the penalty terms of the BIC and the CAIC is

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

Thus, condition C2-1 holds. Moreover, it is easy to obtain

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

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 C2-2 is satisfied because $c_0(k_j - k_*) \geq 0$ holds, where $\mathcal{S}_-$ is given by (3.7). When $j \in \mathcal{S}_-$, condition C2-2 is satisfied if $c_0 < \beta_j / (k_* - k_j)$ holds for all $j \in \mathcal{S}_-$, where β_j is given by (3.3). Finally, the case of HQC is considered. When $j \in \mathcal{J}_+ \backslash \{j_*\}$, the difference between the penalty terms of the HQCs is

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

Moreover, it is easy to derive

$$\frac{1}{n \log p} \{m(j) - m(j_*)\} = 2(k_j - k_*) c_{n,p} \left\{ \frac{\log \log p}{\log p} + \frac{\log(1 - \log c_{n,p} / \log p)}{\log p} \right\}.$$

This implies that

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

Thus, conditions C2-1 and C2-2 hold. From the above results and Theorem 2, Corollary 2 is proved.

A.6. Proofs of lemmas in Appendix A.1

A.6.1. Proof of Lemma A.1

Let $\boldsymbol{X}_j = (\boldsymbol{x}_{j,1}, \ldots, \boldsymbol{x}_{j,n})'$. Notice that $\|\boldsymbol{x}_{j,i}\| \leq \|\boldsymbol{x}_i\|$ and $\lambda_{\min}(\boldsymbol{X}'\boldsymbol{X}) \leq \lambda_{\min}(\boldsymbol{X}_j'\boldsymbol{X}_j)$ hold because $\boldsymbol{X}_j$ is a submatrix of $\boldsymbol{X}$. Hence, for any integer a not larger than k_j, we have

$$|q_{j,ia}| \leq \|\boldsymbol{q}_{j,i}\| = \|\boldsymbol{x}_{j,i}'(\boldsymbol{X}_j'\boldsymbol{X}_j)^{-1}\boldsymbol{x}_{j,i}\| \leq \frac{\|\boldsymbol{x}_{j,i}\|}{\lambda_{\min}(\boldsymbol{X}_j'\boldsymbol{X}_j)^{1/2}} \leq \frac{\|\boldsymbol{x}_i\|}{\lambda_{\min}(\boldsymbol{X}'\boldsymbol{X})^{1/2}},$$

where $\lambda_{\min}(\boldsymbol{A})$ denote the smallest eigenvalue of a matrix $\boldsymbol{A}$. The above equation implies that

$$\sum_{i=1}^n |q_{j,ia} q_{j,ib} q_{j,ic} q_{j,id}| \leq \sum_{i=1}^n |q_{j,ia}||q_{j,ib}||q_{j,ic}||q_{j,id}| \leq \frac{\sum_{i=1}^n \|\boldsymbol{x}_i\|^4}{\lambda_{\min}(\boldsymbol{X}'\boldsymbol{X})^2}. \tag{A.38}$$

Moreover, assumption A4 indicates that

$$\frac{n}{\lambda_{\min}(\boldsymbol{X}'\boldsymbol{X})} = \frac{1}{\lambda_{\min}(\boldsymbol{X}'\boldsymbol{X}/n)} < \mathrm{tr}\{(\boldsymbol{X}'\boldsymbol{X}/n)^{-1}\} \to \mathrm{tr}(\boldsymbol{R}_0^{-1}) \quad \text{as} \quad n \to \infty,$$

where $\boldsymbol{R}_0$ is a limiting value of $\boldsymbol{X}'\boldsymbol{X}/n$, which is defined in assumption A4. Hence, by combining this equation, equation (A.38), and assumption A5, we have proved Lemma A.1.

A.6.2. Proof of Lemma A.2

In order to prove Lemma A.2, we have only to show that Lindeberg's condition (see, e.g., Serfling (2001), th. B, p. 30) is satisfied. Let $\boldsymbol{\nu}_{j,i} = (\boldsymbol{I}_p \otimes \boldsymbol{q}_{j,i})\boldsymbol{\varepsilon}_i$, where $\boldsymbol{q}_{j,i}$ is given by (A.1). It is clear that $\boldsymbol{\nu}_{j,1}, \ldots, \boldsymbol{\nu}_{j,n}$ are independent, and $E[\boldsymbol{\nu}_{j,i}] = \boldsymbol{0}_{pk_j}$, $\mathrm{Cov}[\boldsymbol{\nu}_{j,i}] = \boldsymbol{I}_p \otimes \boldsymbol{q}_{j,i}\boldsymbol{q}_{j,i}'$, and $E[\|\boldsymbol{\nu}_{j,i}\|^2] = p\boldsymbol{q}_{j,i}'\boldsymbol{q}_{j,i}$. Besides these, $\mathrm{vec}(\boldsymbol{Q}_j'\boldsymbol{\mathcal{E}}) = \sum_{i=1}^n \boldsymbol{\nu}_{j,i}$ and $\sum_{i=1}^n \mathrm{Cov}[\boldsymbol{\nu}_{j,i}] = \boldsymbol{I}_{pk_j}$ hold, where $\boldsymbol{Q}_j$ is given by (A.1). Let $I(x > a)$ be an indicator function, i.e., $I(x > a) = 1$ if $x > a$ and $I(x > a) = 0$ if $x \le a$. Then, for all $\epsilon > 0$, we derive

$$\begin{aligned} E[\|\boldsymbol{\nu}_{j,i}\|^2 I(\|\boldsymbol{\nu}_{j,i}\| > \epsilon)]^2 &\le E[\|\boldsymbol{\nu}_{j,i}\|^4]E[I(\|\boldsymbol{\nu}_{j,i}\| > \epsilon)^2] \\ &= E[\|\boldsymbol{\nu}_{j,i}\|^4]P(\|\boldsymbol{\nu}_{j,i}\| > \epsilon) \le \frac{1}{\epsilon^4}E[\|\boldsymbol{\nu}_{j,i}\|^4]^2. \end{aligned}$$

Since assumption A2 holds, $E[\|\boldsymbol{\nu}_{j,i}\|^4]$ exists and becomes $(\boldsymbol{q}_{j,i}'\boldsymbol{q}_{j,i})^2\{\kappa_4^{(1)} + p(p+2)\}$. Moreover, it follows from Lemma A.1 that $\sum_{i=1}^n (\boldsymbol{q}_{j,i}'\boldsymbol{q}_{j,i})^2 = o(1)$ as $n \to \infty$ holds, because we assume assumptions A4 and A5. Hence, we have

$$\sum_{i=1}^n E[\|\boldsymbol{\nu}_{j,i}\|^2 I(\|\boldsymbol{\nu}_{j,i}\| > \epsilon)] \le \frac{1}{\epsilon^2}\{\kappa_4^{(1)} + p(p+2)\}\sum_{i=1}^n (\boldsymbol{q}_{j,i}'\boldsymbol{q}_{j,i})^2 \to 0 \quad \text{as} \quad n \to \infty.$$

This means that Lindeberg's condition, i.e., $\lim_{n\to\infty}\sum_{i=1}^n E[\|\boldsymbol{\nu}_{j,i}\|^2 I(\|\boldsymbol{\nu}_{j,i}\| > \epsilon)] = 0$, is satisfied.

A.6.3. Proof of Lemma A.3

First, we show the proof of equation (i) in Lemma A.3. If $T_{j,\ell} \xrightarrow{p} \tau_{j,\ell} > 0$ holds, then

$$P(|T_{j,\ell} - \tau_{j,\ell}| > \epsilon) \to 0, \quad {}^\forall\epsilon > 0. \tag{A.39}$$

Recall that $\{\mathrm{IC}_m(j) - \mathrm{IC}_m(\ell)\}/h_{j,\ell} \ge T_{j,\ell}$ holds. Thus, the following equation is satisfied:

$$\begin{aligned} P(|T_{j,\ell} - \tau_{j,\ell}| > \tau_{j,\ell}) &= P(\{T_{j,\ell} > 2\tau_{j,\ell}\} \cup \{T_{j,\ell} < 0\}) \\ &\ge P(T_{j,\ell} < 0) \ge P(\mathrm{IC}_m(j) - \mathrm{IC}_m(\ell) < 0). \end{aligned} \tag{A.40}$$

Since equation (A.39) holds for all $\epsilon > 0$, the first probability in (A.40) converges to 0. This indicates that $P(\mathrm{IC}_m(j) < \mathrm{IC}_m(\ell)) \to 0$. Furthermore, it is common knowledge that equation (A.39) is equivalent to

$$P(|T_{j,\ell} - \tau_{j,\ell}| \le \epsilon) \to 1, \quad {}^\forall\epsilon > 0. \tag{A.41}$$

By the same method as in the calculation of (A.40), we derive

$$\begin{aligned}(\text{A.42})\qquad P(|T_{j,\ell}-\tau_{j,\ell}|\le \tau_{j,\ell}/2) &\le P(|T_{j,\ell}-\tau_{j,\ell}|<\tau_{j,\ell})\\ &= P(\{T_{j,\ell}>0\}\cap\{T_{j,\ell}<2\tau_{j,\ell}\})\\ &\le P(T_{j,\ell}>0)\le P(\mathrm{IC}_m(j)-\mathrm{IC}_m(\ell)>0).\end{aligned}$$

Since equation (A.41) holds for all $\epsilon>0$, the first probability in (A.42) converges to 1. This indicates that $P(\mathrm{IC}_m(j)>\mathrm{IC}_m(\ell))\to 1$.

Next, we show the proof of equations (ii) and (iii). From basic probability theory, we obtain

$$\begin{aligned}(\text{A.43})\qquad P(\hat{j}_m=j) &= 1-P(\hat{j}_m\ne j)=1-P(\cup_{\ell\in\mathcal{J}\backslash\{j\}}\{\mathrm{IC}_m(\ell)<\mathrm{IC}_m(j)\})\\ &\ge 1-\sum_{\ell\in\mathcal{J}\backslash\{j\}}P(\mathrm{IC}_m(\ell)<\mathrm{IC}_m(j)).\end{aligned}$$

Since $T_{\ell,j}\xrightarrow{p}\tau_{\ell,j}>0$ holds for all $\ell\in\mathcal{J}\backslash\{j\}$, we can see from Lemma A.3 (i) that $P(\mathrm{IC}_m(\ell)<\mathrm{IC}_m(j))\to 0$ for all $\ell\in\mathcal{J}\backslash\{j\}$. By using this result and equation (A.43), we prove equation (ii). Suppose that $^{\exists}\ell_0\in\mathcal{J}\backslash\{j\}$ *s.t.* $T_{j,\ell_0}\xrightarrow{p}\tau_{j,\ell_0}>0$. Then, by using the same method as that by which we calculated (A.39), (A.40), and (A.43), we obtain

$$\begin{aligned}P(\hat{j}_m=j) &= P(\cap_{\ell\in\mathcal{J}\backslash\{j\}}\{\mathrm{IC}_m(j)<\mathrm{IC}_m(\ell)\})\le P(\mathrm{IC}_m(j)<\mathrm{IC}_m(\ell_0))\\ &\le P(T_{j,\ell_0}<0)\le P(|T_{j,\ell_0}-\tau_{j,\ell_0}|>\tau_{j,\ell_0})\to 0.\end{aligned}$$

Consequently, equation (iii) is proved.

A.6.4. Proof of Lemma A.4

First, we prove equation (i) in Lemma A.4. Notice that $P(A\cap B)\le\min\{P(A),P(B)\}$. It follows from this equation and the assumption of $P(B)$ that $\min\{P(A),P(B)\}\to 0$. Hence, equation (i) is proved. Next, we show equation (ii) in Lemma A.4. Since $P(B)\to 1$ holds, $P(B^c)\to 0$ holds. It follows from this result and equation (i) that $P(A\cap B^c)\to 0$. Notice that $A=A\cap(B\cup B^c)=(A\cap B)\cup(A\cap B^c)$, and $A\cap B$ and $A\cap B^c$ are mutually exclusive events. Hence, we have

$$P(A)=P((A\cap B)\cup(A\cap B^c))=P(A\cap B)+P(A\cap B^c).$$

Recall that $P(A\cap B^c)\to 0$. Therefore, equation (ii) in Lemma A.4 is proved.

A.6.5. Proof of Lemma A.5

It is easy to obtain that $E[\boldsymbol{\mathcal{E}}'\boldsymbol{A}\boldsymbol{\mathcal{E}}]=\mathrm{tr}(\boldsymbol{A})\boldsymbol{I}_p$. Recall that $\boldsymbol{\varepsilon}_1,\ldots,\boldsymbol{\varepsilon}_n$ are identically and independently distributed, $E[\boldsymbol{\varepsilon}_a\boldsymbol{\varepsilon}_a']=\boldsymbol{I}_p$, and $E[\|\boldsymbol{\varepsilon}_a\|^4]=\kappa_4^{(1)}+p(p+2)$, where $\kappa_4^{(1)}$ is given by (2.5). These equations imply that

$$E[\mathrm{tr}\{(\boldsymbol{\mathcal{E}}'\boldsymbol{A}\boldsymbol{\mathcal{E}})^2\}]=\sum_{a,b,c,d}^{n}(\boldsymbol{A})_{ad}(\boldsymbol{A})_{bc}E[\boldsymbol{\varepsilon}_a'\boldsymbol{\varepsilon}_b\boldsymbol{\varepsilon}_c'\boldsymbol{\varepsilon}_d]$$

$$
\begin{aligned}
&= \sum_{a=1}^{n} \{(\boldsymbol{A})_{aa}\}^2 E[(\boldsymbol{\varepsilon}_a' \boldsymbol{\varepsilon}_a)^2] \\
&\quad + \sum_{a \neq b}^{n} [\{(\boldsymbol{A})_{aa}\}\{(\boldsymbol{A})_{bb}\} E[(\boldsymbol{\varepsilon}_a' \boldsymbol{\varepsilon}_b)^2] \\
&\qquad + \{(\boldsymbol{A})_{ab}\}^2 \{E[\boldsymbol{\varepsilon}_a' \boldsymbol{\varepsilon}_a \boldsymbol{\varepsilon}_b' \boldsymbol{\varepsilon}_b] + E[(\boldsymbol{\varepsilon}_a' \boldsymbol{\varepsilon}_b)^2]\}] \\
&= \kappa_4^{(1)} \sum_{a=1}^{n} \{(\boldsymbol{A})_{aa}\}^2 + p(p+1)\operatorname{tr}(\boldsymbol{A}^2) + p \operatorname{tr}(\boldsymbol{A})^2.
\end{aligned}
$$

Consequently, Lemma A.5 is proved.

A.6.6. Proof of Lemma A.6

Notice that

$$
\sum_{a=1}^{n} \{(\boldsymbol{A})_{aa}\}^2 \leq \sum_{a,b}^{n} \{(\boldsymbol{A})_{ab}\}^2 = \sum_{a=1}^{n} (\boldsymbol{A})_{aa} = \operatorname{tr}(\boldsymbol{A}).
$$

Hence, Lemma A.6 is proved.

A.6.7. Proof of Lemma A.7

Let w_{ab} be the (a,b)th element of $\boldsymbol{W}$, and let $\bar{\boldsymbol{\varepsilon}}$ be the sample mean of $\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n$, i.e., $\bar{\boldsymbol{\varepsilon}} = \sum_{i=1}^{n} \boldsymbol{\varepsilon}_i$, where $\boldsymbol{W}$ is given by (A.3). It follows from $w_{ab} = \boldsymbol{u}_a'(\boldsymbol{U}'\boldsymbol{U})^{-1}\boldsymbol{u}_b$ and $\boldsymbol{u}_a = \boldsymbol{\varepsilon}_a - \bar{\boldsymbol{\varepsilon}}$ that the diagonal elements of $\boldsymbol{W}$ are identically distributed and the upper (or lower) off-diagonal elements of $\boldsymbol{W}$ are also identically distributed, where $\boldsymbol{U}$ is given by (A.3). Let $\boldsymbol{W}_d = \operatorname{diag}(w_{11}, \ldots, w_{nn})$. Recall that $\boldsymbol{W}$ is a symmetric idempotent matrix and $\boldsymbol{W}\boldsymbol{1}_n = \boldsymbol{0}_n$ holds. These imply that

$$
\begin{aligned}
&0 \leq w_{aa} \leq 1, \\
&|w_{ab}| \leq \sqrt{w_{aa} w_{bb}} \leq 1 \qquad (a = 1, \ldots, n; b = 1, \ldots, n; a \neq b),
\end{aligned} \tag{A.44}
$$

and

$$
\begin{aligned}
&p = \operatorname{tr}(\boldsymbol{W}) = \sum_{a=1}^{n} w_{aa}, \qquad p = \operatorname{tr}(\boldsymbol{W}^2) = \sum_{a=1}^{n} w_{aa}^2 + \sum_{a \neq b}^{n} w_{ab}^2, \\
&p^2 = \operatorname{tr}(\boldsymbol{W})^2 = \sum_{a=1}^{n} w_{aa}^2 + \sum_{a \neq b}^{n} w_{aa} w_{bb}, \\
&0 = \boldsymbol{1}_n' \boldsymbol{W} \boldsymbol{1}_n = \sum_{a=1}^{n} w_{aa} + \sum_{a \neq b}^{n} w_{ab}, \\
&0 = \boldsymbol{1}_n' \boldsymbol{W}_d \boldsymbol{W} \boldsymbol{1}_n = \sum_{a=1}^{n} w_{aa}^2 + \sum_{a \neq b}^{n} w_{aa} w_{ab}, \\
&0 = \boldsymbol{1}_n' \boldsymbol{W}^2 \boldsymbol{1}_n = \sum_{a=1}^{n} w_{aa}^2 + \sum_{a \neq b}^{n} (2 w_{aa} w_{ab} + w_{ab}^2) + \sum_{a \neq b \neq c}^{n} w_{ab} w_{ac}, \\
&0 = \operatorname{tr}(\boldsymbol{W}) \boldsymbol{1}_n' \boldsymbol{W} \boldsymbol{1}_n = \sum_{a=1}^{n} w_{aa}^2 + \sum_{a \neq b}^{n} (2 w_{aa} w_{ab} + w_{aa} w_{bb})
\end{aligned} \tag{A.45}
$$

$$+ \sum_{a\neq b\neq c}^{n} w_{aa}w_{bc},$$

$$0 = (\mathbf{1}_n' \boldsymbol{W} \mathbf{1}_n)^2 = \sum_{a=1}^{n} w_{aa}^2 + \sum_{a\neq b}^{n}(w_{aa}w_{bb} + 2w_{ab}^2 + 4w_{aa}w_{ab})$$
$$+ 2\sum_{a\neq b\neq c}^{n}(w_{aa}w_{bc} + 2w_{ab}w_{ac}) + \sum_{a\neq b\neq c\neq d}^{n} w_{ab}w_{cd},$$

where the notation $\sum_{a_1\neq a_2\neq\cdots}^{n}$ means $\sum_{a_1=1}^{n}\sum_{a_2=1,a_2\neq a_1}^{n}\cdots$. Since w_{aa} $(a = 1,\ldots,n)$ are identically distributed and w_{ab} $(a = 1,\ldots,n; b = a+1,\ldots,n)$ are also identically distributed, from equations in (A.45), we derive for $a \neq b \neq c \neq d$

$$\begin{aligned}
&p = nE[w_{aa}], \qquad p = nE[w_{aa}^2] + n(n-1)E[w_{ab}^2],\\
&p^2 = nE[w_{aa}^2] + n(n-1)E[w_{aa}w_{bb}], \qquad 0 = nE[w_{aa}] + n(n-1)E[w_{ab}],\\
&0 = nE[w_{aa}^2] + n(n-1)E[w_{aa}w_{ab}],\\
&0 = nE[w_{aa}^2] + n(n-1)(2E[w_{aa}w_{ab}] + E[w_{ab}^2])\\
&\qquad + n(n-1)(n-2)E[w_{ab}w_{ac}],\\
&0 = nE[w_{aa}^2] + n(n-1)(2E[w_{aa}w_{ab}] + E[w_{aa}w_{bb}])\\
&\qquad + n(n-1)(n-2)E[w_{aa}w_{bc}],\\
&0 = nE[w_{aa}^2] + n(n-1)(E[w_{aa}w_{bb}] + 2E[w_{ab}^2] + 4E[w_{aa}w_{ab}])\\
&\qquad + 2n(n-1)(n-2)(E[w_{aa}w_{bc}] + 2E[w_{ab}w_{ac}])\\
&\qquad + n(n-1)(n-2)(n-3)E[w_{ab}w_{cd}].
\end{aligned} \tag{A.46}$$

It follows from equation (A.44) that $E[w_{aa}^2] \leq 1$. Combining this result and equation (A.46) yields

$$\begin{aligned}
&E[w_{aa}] = c_{n,p}, && E[w_{ab}] = O(n^{-1}), && E[w_{aa}^2] = O(1),\\
&E[w_{aa}w_{bb}] = c_{n,p}^2 + O(n^{-1}), && E[w_{ab}^2] = O(n^{-1}),\\
&E[w_{aa}w_{ab}] = O(n^{-1}), && E[w_{aa}w_{bc}] = O(n^{-1}),\\
&E[w_{ab}w_{ac}] = O(n^{-2}), && E[w_{ab}w_{cd}] = O(n^{-2}),
\end{aligned} \tag{A.47}$$

as $c_{n,p} \to c_0$, where a, b, c, d are arbitrary positive integers not larger than n and $a \neq b \neq c \neq d$.

Notice that

$$\boldsymbol{\alpha}' \boldsymbol{W} \boldsymbol{\beta} = \sum_{a=1}^{n} \alpha_a\beta_a w_{aa} + \sum_{a\neq b}^{n} \alpha_a\beta_b w_{ab},$$

and

$$(\boldsymbol{\alpha}' \boldsymbol{W} \boldsymbol{\beta})^2 = \sum_{a=1}^{n} \alpha_a^2\beta_a^2 w_{aa}^2 + \sum_{a\neq b\neq c\neq d}^{n} \alpha_a\alpha_c\beta_b\beta_d w_{ab}w_{cd}$$
$$+ \sum_{a\neq b}^{n}\{\alpha_a\alpha_b\beta_a\beta_b(w_{aa}w_{bb} + w_{ab}^2)$$

$$
\begin{aligned}
&\qquad + 2(\alpha_a^2\beta_a\beta_b + \alpha_a\alpha_b\beta_a^2)w_{aa}w_{ab} + \alpha_a^2\beta_b^2 w_{ab}^2\} \\
&+ \sum_{a\neq b\neq c}^{n} \{2\alpha_a\alpha_b\beta_a\beta_c w_{aa}w_{bc} \\
&\qquad + (\alpha_a^2\beta_b\beta_c + 2\alpha_a\alpha_b\beta_a\beta_c + \alpha_b\alpha_c\beta_a^2)w_{ab}w_{ac}\}.
\end{aligned}
$$

It follows from conditions $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ in (A.4) that

$$
\begin{aligned}
&\sum_{a\neq b}^{n}\alpha_a\beta_b = -\boldsymbol{\alpha}'\boldsymbol{\beta}, \qquad \sum_{a=1}^{n}\alpha_a^2\beta_a^2 = o(1), \qquad \sum_{a\neq b}^{n}\alpha_a\alpha_b\beta_a\beta_b = (\boldsymbol{\alpha}'\boldsymbol{\beta})^2 + o(1), \\
&\sum_{a\neq b}^{n}(\alpha_a^2\beta_a\beta_b + \alpha_a\alpha_b\beta_a^2) = o(1), \qquad \sum_{a\neq b}^{n}\alpha_a^2\beta_b^2 = 1 + o(1), \\
&\sum_{a\neq b\neq c}^{n}\alpha_a\alpha_b\beta_a\beta_c = O(1), \qquad \sum_{a\neq b\neq c}^{n}(\alpha_a^2\beta_b\beta_c + 2\alpha_a\alpha_b\beta_a\beta_c + \alpha_b\alpha_c\beta_a^2) = O(1), \\
&\sum_{a\neq b\neq c\neq d}^{n}\alpha_a\alpha_c\beta_b\beta_d = O(1),
\end{aligned}
$$

as $c_{n,p} \to c_0$. Consequently, by using the above results and the expectations in (A.47), we derive

$$
E[\boldsymbol{\alpha}'\boldsymbol{W}\boldsymbol{\beta}] \to c_0\boldsymbol{\alpha}'\boldsymbol{\beta}, \quad E[(\boldsymbol{\alpha}'\boldsymbol{W}\boldsymbol{\beta})^2] \to c_0^2(\boldsymbol{\alpha}'\boldsymbol{\beta})^2 \quad \text{as} \quad c_{n,p} \to c_0.
$$

The above equations directly imply that $\mathrm{Var}[\boldsymbol{\alpha}'\boldsymbol{W}\boldsymbol{\beta}] \to 0$ as $c_{n,p} \to c_0$. Hence, Lemma A.7 is proved.

A.6.8. Proof of Lemma A.8

First, we prove equation (i) in Lemma A.8. Recall that $\boldsymbol{\Gamma}_j\boldsymbol{\Gamma}_j' = \boldsymbol{G}_j\boldsymbol{L}_j\boldsymbol{G}_j'$, where $\boldsymbol{\Gamma}_j$ is given by (2.4), and $\boldsymbol{L}_j$, and $\boldsymbol{G}_j$ were given in (A.5). By using $\boldsymbol{\mathcal{A}}_j$ in (A.5) and $\boldsymbol{S}_j$ in (A.6), equation (A.9) can be rewritten as

$$
\begin{aligned}
n\boldsymbol{\Sigma}_*^{-1/2}\hat{\boldsymbol{\Sigma}}_j\boldsymbol{\Sigma}_*^{-1/2} &= \boldsymbol{\Gamma}_j\boldsymbol{\Gamma}_j' + \boldsymbol{\mathcal{A}}_j'\boldsymbol{\mathcal{E}} + \boldsymbol{\mathcal{E}}'\boldsymbol{\mathcal{A}}_j + \boldsymbol{\mathcal{E}}'\boldsymbol{H}_j\boldsymbol{H}_j'\boldsymbol{\mathcal{E}} + n\boldsymbol{S}_j \\
&= \boldsymbol{G}_j\boldsymbol{L}_j\boldsymbol{G}_j' + \boldsymbol{G}_j\boldsymbol{L}_j^{1/2}\boldsymbol{H}_j'\boldsymbol{\mathcal{E}} + \boldsymbol{\mathcal{E}}'\boldsymbol{H}_j\boldsymbol{L}_j^{1/2}\boldsymbol{G}_j' + \boldsymbol{\mathcal{E}}'\boldsymbol{H}_j\boldsymbol{H}_j'\boldsymbol{\mathcal{E}} + n\boldsymbol{S}_j \\
&= (\boldsymbol{L}_j^{1/2}\boldsymbol{G}_j' + \boldsymbol{H}_j'\boldsymbol{\mathcal{E}})'(\boldsymbol{L}_j^{1/2}\boldsymbol{G}_j' + \boldsymbol{H}_j'\boldsymbol{\mathcal{E}}) + n\boldsymbol{S}_j,
\end{aligned}
$$

where $\boldsymbol{H}_j$ is given by (A.5). Hence, equation (i) in Lemma A.8 is proved.

Next, we prove equation (ii) in Lemma A.8. Notice that $\boldsymbol{P}_j(\boldsymbol{I}_n - \boldsymbol{P}_j) = \boldsymbol{O}_{n,n}$, $(\boldsymbol{I}_n - \boldsymbol{P}_{j_+})(\boldsymbol{I}_n - \boldsymbol{P}_j)\boldsymbol{X}_* = \boldsymbol{O}_{n,k_*}$, and $\boldsymbol{1}_n'(\boldsymbol{I}_n - \boldsymbol{P}_j)\boldsymbol{X}_* = \boldsymbol{0}_{k_*}'$, where j_+ is given by (2.1). These imply that $\boldsymbol{P}_j\boldsymbol{\mathcal{A}}_j = \boldsymbol{O}_{n,p}$, $(\boldsymbol{I}_n - \boldsymbol{P}_{j_+})\boldsymbol{\mathcal{A}}_j = \boldsymbol{O}_{n,p}$, and $\boldsymbol{1}_n'\boldsymbol{\mathcal{A}}_j = \boldsymbol{0}_p'$. Hence, we have

$$
\begin{aligned}
&\boldsymbol{P}_j\boldsymbol{\mathcal{A}}_j = \boldsymbol{O}_{n,p} \Leftrightarrow \boldsymbol{P}_j\boldsymbol{H}_j\boldsymbol{L}_j^{1/2}\boldsymbol{G}_j' = \boldsymbol{O}_{n,p} \Leftrightarrow \boldsymbol{P}_j\boldsymbol{H}_j\boldsymbol{L}_j^{1/2}\boldsymbol{G}_j'\boldsymbol{G}_j\boldsymbol{L}_j^{-1/2}\boldsymbol{H}_j' = \boldsymbol{O}_{n,n} \\
&\qquad\qquad \Leftrightarrow \boldsymbol{P}_j\boldsymbol{H}_j\boldsymbol{H}_j' = \boldsymbol{O}_{n,n}, \\
&(\boldsymbol{I}_n - \boldsymbol{P}_{j_+})\boldsymbol{\mathcal{A}}_j = \boldsymbol{O}_{n,p} \Leftrightarrow (\boldsymbol{I}_n - \boldsymbol{P}_{j_+})\boldsymbol{H}_j\boldsymbol{L}_j^{1/2}\boldsymbol{G}_j' = \boldsymbol{O}_{n,p}
\end{aligned}
$$

$$\Leftrightarrow (\boldsymbol{I}_n - \boldsymbol{P}_{j+})\boldsymbol{H}_j\boldsymbol{L}_j^{1/2}\boldsymbol{G}_j'\boldsymbol{G}_j\boldsymbol{L}_j^{-1/2}\boldsymbol{H}_j' = \boldsymbol{O}_{n,n}$$
$$\Leftrightarrow (\boldsymbol{I}_n - \boldsymbol{P}_{j+})\boldsymbol{H}_j\boldsymbol{H}_j' = \boldsymbol{O}_{n,n} \Leftrightarrow \boldsymbol{P}_{j+}\boldsymbol{H}_j\boldsymbol{H}_j' = \boldsymbol{H}_j\boldsymbol{H}_j',$$

$$\boldsymbol{1}_n'\boldsymbol{\mathcal{A}}_j = \boldsymbol{0}_p' \Leftrightarrow \boldsymbol{J}_n\boldsymbol{\mathcal{A}}_j = \boldsymbol{O}_{n,p} \Leftrightarrow \boldsymbol{J}_n\boldsymbol{H}_j\boldsymbol{L}_j^{1/2}\boldsymbol{G}_j' = \boldsymbol{O}_{n,p}$$
$$\Leftrightarrow \boldsymbol{J}_n\boldsymbol{H}_j\boldsymbol{L}_j^{1/2}\boldsymbol{G}_j'\boldsymbol{G}_j\boldsymbol{L}_j^{-1/2}\boldsymbol{H}_j' = \boldsymbol{O}_{n,n} \Leftrightarrow \boldsymbol{J}_n\boldsymbol{H}_j\boldsymbol{H}_j' = \boldsymbol{O}_{n,n}.$$

Hence, equation (ii) in Lemma A.8 is proved.

Finally, we prove equation (iii) in Lemma A.8. It follows from elementary linear algebra that

$$\lambda_{\max}(\boldsymbol{S}_j) \leq \sqrt{\mathrm{tr}(\boldsymbol{S}_j^2)}.$$

From Lemma A.6 and equation (ii) in Lemma A.8, we can see that

$$\sum_{a=1}^{n}\{(\boldsymbol{I}_n - \boldsymbol{J}_n - \boldsymbol{P}_j - \boldsymbol{H}_j\boldsymbol{H}_j')_{aa}\}^2 = O(n) \quad \text{as} \quad c_{n,p} \to c_0.$$

The above result and Lemma A.5 imply that

$$E[\mathrm{tr}(\boldsymbol{S}_j^2)] = \frac{1}{n^2}\phi_2(\boldsymbol{I}_n - \boldsymbol{J}_n - \boldsymbol{P}_j - \boldsymbol{H}_j\boldsymbol{H}_j') = O(p) \quad \text{as} \quad c_{n,p} \to c_0.$$

The expectation of $\mathrm{tr}(\boldsymbol{S}_j^2)$ and Jensen's inequality lead us to the equation

$$E[\mathrm{tr}(\boldsymbol{S}_j^2)^{1/2}] \leq \sqrt{E[\mathrm{tr}(\boldsymbol{S}_j^2)]} = O(p^{1/2}) \quad \text{as} \quad c_{n,p} \to c_0.$$

This directly implies that $\mathrm{tr}(\boldsymbol{S}_j^2)^{1/2} = O_p(p^{1/2})$ as $c_{n,p} \to c_0$. Hence, equation (iii) in Lemma A.8 is proved.

Acknowledgements

The author thanks Prof. Hirofumi Wakaki and Prof. Yasunori Fujikoshi, Hiroshima University, for their helpful comments on the assumptions necessary to satisfy consistency. Furthermore, the author also wishes to thank the referees, the associate editor and the editor for their helpful suggestions. This research was partially supported by the Ministry of Education, Science, Sports, and Culture, and a Grant-in-Aid for Challenging Exploratory Research, #25540012, 2013–2015.

References

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

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

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

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

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

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.

Fan, J., Fan, Y. and Lv, J. (2008). High dimensional covariance matrix estimation using a factor model, *J. Econometrics*, **147**, 186–197.

Fang, K. T., Kotz, S. and Ng, K. W. (1990). *Symmetric Multivariate and Related Distributions*, Chapman & Hall/CRC, London.

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

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.

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

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

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

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

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.

Hannan, E. J. and Quinn, B. G. (1979). The determination of the order of an autoregression, *J. Roy. Statist. Soc. Ser. B*, **41**, 190–195.

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

Ishiguro, M., Sakamoto, Y. and Kitagawa, G. (1997). Bootstrapping log likelihood and EIC, an extension of AIC, *Ann. Inst. Statist. Math.*, **49**, 411–434.

Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications, *Biometrika*, **57**, 519–530.

Obozinski, G., Wainwright, M. J. and Jordan, M. I. (2008). High-dimensional union support recovery in multivariate regression, *Advances in Neural Information Processing Systems* **21** (eds. D. Koller, D. Schuurmans, Y. Bengio and L. Bottou), MIT Press, Cambridge, MA.

Obozinski, G., Wainwright, M. J. and Jordan, M. I. (2011). Support union recovery in high-dimensional multivariate regression, *Ann. Statist.*, **39**, 1–47.

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.

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.

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

Serfling, R. J. (2001). *Approximation Theorems of Mathematical Statistics*, (Paperback ed.), John Wiley & Sons, Inc.

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

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

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions, *J. Roy. Statist. Soc. Ser. B*, **36**, 111–147.

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion, *J. Roy. Statist. Soc. Ser. B*, **39**, 44–47.

Takeuchi, K. (1976). Distribution of information statistics and criteria for adequacy of models, *Math. Sci.*, **153**, 12–18 (in Japanese).

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

Wakaki, H., Yanagihara, H. and Fujikoshi, Y. (2002). Asymptotic expansions of the null distributions of test statistics for multivariate linear hypothesis under nonnormality, *Hiroshima Math. J.*, **32**, 17–50.

Xiao, H. and Zhou, W. (2010). Almost sure limit of the smallest eigenvalue of some sample correlation matrices, *J. Theoret. Probab.*, **23**, 1–20.

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.

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.

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

Yanagihara, H., Kamo, K., Imori, S. and Yamamura, M. (2013). A study on the bias-correction effect of the AIC for selecting variables in normal multivariate linear regression models under model misspecification, *TR 13-08*, *Statistical Research Group*, *Hiroshima University*, Hiroshima.

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).