TR-No. 13-11, Hiroshima Statistical Research Group, 1–27

Conditions for Consistency of a Log-Likelihood-Based Information Criterion in Normal Multivariate Linear Regression Models under the Violation of Normality Assumption

Hirokazu Yanagihara*

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

(Last Modified: July 30, 2013)

Abstract

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 a normality is assumed to the distribution of the candidate model, we frame the situation as 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 ∞. In both cases, our results are free of the influence of nonnormality in the true distribution.

Key words: 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.

E-mail address: yanagi@math.sci.hiroshima-u.ac.jp (Hirokazu Yanagihara)

1. Introduction

The multivariate linear regression model is one of 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 responses to a set of explanatory variables (see, e.g., Yoshimoto *et al*., 2005; Dien *et al*., 2006; Saxén & Sundell, 2006; Sárbu *et al*., 2008). Let $\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots .\boldsymbol{y}_n)'$ be an $n \times p$ matrix of p response variables, and let $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)'$ be an $n \times k$ matrix of nonstochastic centralized k 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 $\boldsymbol{X}$ 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 $\boldsymbol{X}$. We then consider the following multivariate linear regression model with k_j explanatory variables as the candidate model:

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

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) 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 affecting the response variables in a regression analysis, searching for the optimal subset j, i.e., variable selection, is essential. 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 (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 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.* (2012). 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 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 & 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 "$c_{n,p} \to c_0$".

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*. (2012) evaluated the consistency of various information criteria under the HD asymptotic framework, and pointed out that the AIC and $\mathrm{AIC_c}$ become consistent, but the BIC and CAIC sometimes become inconsistent.

Unfortunately, the results in previous works were obtained under the assumption that the distribution of the true model is a normal distribution. Although the normal distribution is assumed for the candidate model (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

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

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

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

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. As stated above, the consistency of an information criterion is assessed by the LS and HD asymptotic theories. It is common knowledge that the maximum log-likelihood of the model in (1) consists of the determinants of the maximum likelihood estimators (MLE) of the covariance matrix $\boldsymbol{\Sigma}_j$. 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 the MLE of $\boldsymbol{\Sigma}_j$ increases with an increase in the sample size. Yanagihara *et al*. (2012) avoided this difficulty by using a property of a random matrix distributed according to the Wishart distribution (see Fujikoshi *et al*., 2010, th. 3.2.4, p. 57). However, we cannot use this property because the normality of the true model is not assumed. Hence, it is necessary to consider a different idea, from Yanagihara *et al*. (2012), for assessing the consistency. In this paper, the moments of a specific random matrix and the distribution of the maximum eigenvalue of the estimator of the covariance matrix are used for assessing consistency. Under both the LS and HD asymptotic frameworks, the results we obtained indicate that the conditions for consistency are not influenced by nonnormality in 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 prepare several lemmas for assessing the consistency of an information criterion. In Sections 4, we obtain a necessary and sufficient condition to satisfy consistency under the LS asymptotic framework. In Section 5, we derive a sufficient condition to satisfy consistency under the HD asymptotic framework. In Section 6, we verify the adequacy of our claim by conducting numerical experiments. In Section 7, 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). First, we describe several classes of j that express subsets of $\boldsymbol{X}$ in the candidate model. Let $\mathcal{J}$ be a set of candidate models denoted by $\mathcal{J} = \{j_1, \ldots, j_K\}$, where K is the number of candidate models . We then separate $\mathcal{J}$ into two sets; one is the set of overspecified models for which the explanatory variables contain all the explanatory variables of the true model j_* (2), i.e., $\mathcal{J}_+ = \{j \in \mathcal{J} | j_* \subseteq j\}$, and the other is the set of underspecified models (those 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.,

$$j_+ = j \cup j_*. \tag{3}$$

Estimations for the unknown parameters $\boldsymbol{\mu}$, $\boldsymbol{\Theta}_j$, and $\boldsymbol{\Sigma}_j$ in the candidate model (1) are carried out by the maximum likelihood estimation, i.e., they are estimated by

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

where $\boldsymbol{P}_j$ and $\boldsymbol{J}_n$ are the projection matrices to the subspace spanned by the columns of $\boldsymbol{X}_j$ and $\boldsymbol{1}_n$, respectively, i.e., $\boldsymbol{P}_j = \boldsymbol{X}_j(\boldsymbol{X}_j'\boldsymbol{X}_j)^{-1}\boldsymbol{X}_j'$ and $\boldsymbol{J}_n = \boldsymbol{1}_n\boldsymbol{1}_n'/n$. In order to deal uniformly with 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

$$\mathrm{IC}_m(j) = n \log|\hat{\boldsymbol{\Sigma}}_j| + np(\log 2\pi + 1) + m(j), \tag{4}$$

where $m(j)$ is a positive penalty term that expresses the complexity of the model (1). An information criterion included in this family is specified by an individual penalty term $m(j)$. This family contains the AIC, $\mathrm{AIC_c}$, BIC, CAIC, and HQC as special cases:

$$m(j) = \begin{cases} 2p\{k_j + (p+3)/2\} & \text{(AIC)} \\ 2np\{k_j + (p+3)/2\}/(n - k_j - p - 2) & (\mathrm{AIC_c}) \\ p\{k_j + (p+3)/2\}\log n & \text{(BIC)} \\ p\{k_j + (p+3)/2\}(1 + \log n) & \text{(CAIC)} \\ 2p\{k_j + (p+3)/2\}\log\log n & \text{(HQC)} \end{cases}. \tag{5}$$

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). 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

$$\boldsymbol{\Sigma}_*^{-1/2}\boldsymbol{\Theta}_*'\boldsymbol{X}_*'(\boldsymbol{I}_n - \boldsymbol{P}_j)\boldsymbol{X}_*\boldsymbol{\Theta}_*\boldsymbol{\Sigma}_*^{-1/2} = \boldsymbol{\Gamma}_j\boldsymbol{\Gamma}_j', \tag{6}$$

where $\boldsymbol{\Gamma}_j$ is a $p \times \gamma_j$ ($\gamma_j \leq \min\{p, k_{j_* \cap j^c}\}$) matrix, and it has full column rank when p is large, i.e., $p \geq k_*$. Since $k_{j_* \cap j^c} \leq k_{j_+} - k_j$ holds for large p, $\gamma_j \leq k_{j_+} - k_j$ is satisfied for large p. It should be noted that $\boldsymbol{\Gamma}_j\boldsymbol{\Gamma}_j' = \boldsymbol{O}_{p,p}$ holds 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. $\lim_{n\to\infty} n^{-1}\boldsymbol{\Gamma}_j\boldsymbol{\Gamma}_j' = \boldsymbol{\Omega}_{j,0}$ exists and is positive semidefinite.

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

For which orders of $\boldsymbol{\Gamma}_j\boldsymbol{\Gamma}_j'$ and $\boldsymbol{\Gamma}_j'\boldsymbol{\Gamma}_j$ are adequate, see Yanagihara *et al*. (2012). 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.

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

$$\kappa_4^{(1)} = E[||\boldsymbol{\varepsilon}||^4] - p(p+2) = \sum_{a,b}^{p} \kappa_{aabb} + p(p+2), \tag{7}$$

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

$$\kappa_4^{(1)} = O(p^{1+s}) \text{ as } p \to \infty, \ s \in [0, 1]. \tag{8}$$

The positive constant s is changed by the distribution of $\boldsymbol{\varepsilon}$. For example, if $\varepsilon_1, \ldots, \varepsilon_p$ are independent random variables that are not distributed according to normal distributions, then $s = 0$. If $\boldsymbol{\varepsilon}$ is distributed according to an elliptical distribution other than the normal distribution (see, e.g., Fang *et al*., 1990), then $s = 1$. Hence, there is an additional assumption that can be regarded as a special case of assumption A2:

A2′. $\varepsilon_1, \ldots, \varepsilon_p$ are identically and independently distributed according to some distribution with $E[\varepsilon_1^4] < \infty$.

When the indicated assumptions hold, the following lemmas are satisfied (the proofs are given in Appendices A and B:

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

$$\boldsymbol{Q}_j' = (\boldsymbol{X}_j'\boldsymbol{X}_j)^{-1/2}\boldsymbol{X}_j = (\boldsymbol{q}_{j,1}, \ldots, \boldsymbol{q}_{j,n}), \quad \boldsymbol{q}_{j,i} = (q_{j,i1}, \ldots, q_{j,ik_j})'. \tag{9}$$

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

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

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

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

$$\boldsymbol{Z}_j = \boldsymbol{Q}_j'\boldsymbol{\mathcal{E}}, \tag{10}$$

where $\boldsymbol{Q}_j$ is given by (9). 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.

To ensure the asymptotic normality of $\boldsymbol{Z}_j$, Wakaki *et al*. (2002) assumed $\limsup_{n\to\infty} \|\boldsymbol{x}_i\|^4/n < \infty$, which is stronger than assumption A5.

3. Preliminaries

In this section, we present some lemmas that we will use to derive the conditions for consistency of the penalty term $m(j)$ in $\mathrm{IC}_m(j)$ in (4). We first present two lemmas from basic probability theory (the proofs of these are given in Appendices C and D). In the next two lemmas, we do not specify the asymptotic framework because they are applicable to any asymptotic framework.

Lemma 3 *Let $h_{j,\ell}$ be some positive constant that depends on the models $j, \ell \in \mathcal{J}$. Then, we have*

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

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

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

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

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

$$\mathcal{D}(j,\ell) = \log\left(|\hat{\boldsymbol{\Sigma}}_j|/|\hat{\boldsymbol{\Sigma}}_\ell|\right). \tag{11}$$

Notice that

$$\mathrm{IC}_m(j) - \mathrm{IC}_m(\ell) = n\mathcal{D}(j,\ell) + m(j) - m(\ell). \tag{12}$$

From Lemma 3, 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 $\mathcal{D}(j, j_*)$ or a lower bound of $\mathcal{D}(j, j_*)$ divided by some constant.

Let $(\boldsymbol{A})_{ab}$ be the (a,b)th element of a matrix $\boldsymbol{A}$. Then, the following lemmas help us to prove the convergence in probability of $\mathcal{D}(j, j_*)$ or a lower bound of $\mathcal{D}(j, j_*)$ divided by some constant (the proofs of these lemmas are given in Appendices E, F, and G):

Lemma 5 *For any $n \times n$ symmetric matrix $\boldsymbol{A}$, let $\phi_1(\boldsymbol{A})$, $\phi_2(\boldsymbol{A})$, and $\phi_3(\boldsymbol{A})$ denote moments:*

$$\phi_1(\boldsymbol{A}) = E\left[\mathrm{tr}\left(\boldsymbol{\mathcal{E}}'\boldsymbol{A}\boldsymbol{\mathcal{E}}\right)\right], \quad \phi_2(\boldsymbol{A}) = E\left[\mathrm{tr}\left\{(\boldsymbol{\mathcal{E}}'\boldsymbol{A}\boldsymbol{\mathcal{E}})^2\right\}\right], \quad \phi_3(\boldsymbol{A}) = E\left[\mathrm{tr}(\boldsymbol{\mathcal{E}}'\boldsymbol{A}\boldsymbol{\mathcal{E}})^2\right].$$

Then, specific forms of $\phi_1(\boldsymbol{A})$, $\phi_2(\boldsymbol{A})$, and $\phi_3(\boldsymbol{A})$ are given as

$$\phi_1(\boldsymbol{A}) = p\mathrm{tr}(\boldsymbol{A}), \quad \phi_2(\boldsymbol{A}) = \kappa_4^{(1)}\sum_{a=1}^n\{(\boldsymbol{A})_{aa}\}^2 + p(p+1)\mathrm{tr}(\boldsymbol{A}^2) + p\mathrm{tr}(\boldsymbol{A})^2,$$
$$\phi_3(\boldsymbol{A}) = \kappa_4^{(1)}\sum_{a=1}^n\{(\boldsymbol{A})_{aa}\}^2 + p^2\mathrm{tr}(\boldsymbol{A})^2 + 2p\mathrm{tr}(\boldsymbol{A})^2,$$

where $\kappa_4^{(1)}$ is given by (7).

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

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

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

$$\boldsymbol{U} = (\boldsymbol{u}_1,\ldots,\boldsymbol{u}_n)' = (\boldsymbol{I}_n - \boldsymbol{J}_n)\boldsymbol{\mathcal{E}}, \quad \boldsymbol{W} = \boldsymbol{U}(\boldsymbol{U}'\boldsymbol{U})^{-1}\boldsymbol{U}', \tag{13}$$

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

$$||\alpha|| = ||\boldsymbol{\beta}|| = 1, \quad \boldsymbol{1}_n'\boldsymbol{\alpha} = \boldsymbol{1}_n'\boldsymbol{\beta} = 0, \quad \sum_{a=1}^n \alpha_a^2\beta_a^2 = o(1) \ \text{as} \ c_{n,p} \to c_0. \tag{14}$$

Then, we derive

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

Next, we show the decomposition of $\hat{\boldsymbol{\Sigma}}_j$ when $j \in \mathcal{J}_-$. Notice that

$$\begin{aligned}\boldsymbol{\Sigma}_*^{-1/2}\hat{\boldsymbol{\Sigma}}_j\boldsymbol{\Sigma}_*^{-1/2} = \frac{1}{n}\Big\{&\boldsymbol{\Gamma}_j\boldsymbol{\Gamma}_j' + \boldsymbol{\Sigma}_*^{-1/2}\boldsymbol{\Theta}_*'\boldsymbol{X}_*'(\boldsymbol{I}_n - \boldsymbol{P}_j)\boldsymbol{\mathcal{E}}\\ &+\boldsymbol{\mathcal{E}}'(\boldsymbol{I}_n - \boldsymbol{P}_j)\boldsymbol{X}_*\boldsymbol{\Theta}_*\boldsymbol{\Sigma}_*^{-1/2} + \boldsymbol{\mathcal{E}}'(\boldsymbol{I}_n - \boldsymbol{J}_n - \boldsymbol{P}_j)\boldsymbol{\mathcal{E}}\Big\},\end{aligned} \tag{15}$$

where $\boldsymbol{\Gamma}_j$ is given by (6). For $j \in \mathcal{J}_-$, we define an $n \times p$ matrix $\mathcal{A}_j$ as

$$\mathcal{A}_j = (\boldsymbol{I}_n - \boldsymbol{P}_j)\boldsymbol{X}_*\boldsymbol{\Theta}_*\boldsymbol{\Sigma}_*^{-1/2}. \tag{16}$$

It is easy to see from the definition of the noncentrality matrix in (6) that $\mathcal{A}_j'\mathcal{A}_j = \boldsymbol{\Gamma}_j\boldsymbol{\Gamma}_j'$. By using the singular value decomposition, $\mathcal{A}_j$ can be rewritten as

$$\mathcal{A}_j = \boldsymbol{H}_j\boldsymbol{L}_j^{1/2}\boldsymbol{G}_j', \tag{17}$$

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 $\mathcal{A}_j$. 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

$$\boldsymbol{\Gamma}_j'\boldsymbol{\Gamma}_j = \boldsymbol{C}_j\boldsymbol{L}_j\boldsymbol{C}_j'. \tag{18}$$

By using $\mathcal{A}_j$, equation (15) can be rewritten as

$$n\boldsymbol{\Sigma}_*^{-1/2}\hat{\boldsymbol{\Sigma}}_j\boldsymbol{\Sigma}_*^{-1/2} = \left(\boldsymbol{L}_j^{1/2}\boldsymbol{G}_j' + \boldsymbol{H}_j'\mathcal{E}\right)'\left(\boldsymbol{L}_j^{1/2}\boldsymbol{G}_j' + \boldsymbol{H}_j'\mathcal{E}\right) + \mathcal{E}'(\boldsymbol{I}_n - \boldsymbol{J}_n - \boldsymbol{P}_j - \boldsymbol{H}_j\boldsymbol{H}_j')\mathcal{E}. \tag{19}$$

Before concluding this section, we present the following lemma on $\boldsymbol{I}_n - \boldsymbol{J}_n - \boldsymbol{P}_j - \boldsymbol{H}_j\boldsymbol{H}_j'$ (the proof is given in Appendix H):

Lemma 8 *Let $\lambda_{\max}(\boldsymbol{A})$ denote the maximum eigenvalue of $\boldsymbol{A}$, and let $\boldsymbol{S}_j$ ($j \in \mathcal{J}_-$) be a $p\times p$ matrix defined by*

$$\boldsymbol{S}_j = \frac{1}{n}\mathcal{E}'(\boldsymbol{I}_n - \boldsymbol{J}_n - \boldsymbol{P}_j - \boldsymbol{H}_j\boldsymbol{H}_j')\mathcal{E}. \tag{20}$$

Then, we have

(*i*) *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* (3).

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

(*iii*) *If assumption A2′ holds instead of assumption A2, the order of $\lambda_{\max}(\boldsymbol{S}_j)$ is changed to $O_p(1)$ from $O_p(p^{1/2})$.*

4. Conditions for Consistency under the LS Asymptotic Framework

In this section, we derive the conditions such that IC_m is consistent under the LS asymptotic framework, i.e., the ordinary asymptotic framework in which only n approaches ∞. 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 5 and 6 that

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

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

$$\left.\begin{array}{l} \dfrac{1}{n}\boldsymbol{\Sigma}_*^{-1/2}\boldsymbol{\Theta}_*'\boldsymbol{X}_*'(\boldsymbol{I}_n - \boldsymbol{P}_j)\mathcal{E} = O_p(n^{-1/2}) \\ \dfrac{1}{n}\mathcal{E}'(\boldsymbol{I}_n - \boldsymbol{J}_n - \boldsymbol{P}_j)\mathcal{E} = \boldsymbol{I}_p + O_p(n^{-1/2}) \end{array}\right. \text{ as } n \to \infty.$$

Using the above results and equation (15) yields

$$\boldsymbol{\Sigma}_*^{-1/2}\hat{\boldsymbol{\Sigma}}_j\boldsymbol{\Sigma}_*^{-1/2} \xrightarrow{p} \begin{cases} \boldsymbol{I}_p & ({}^\forall j \in \mathcal{J}_+) \\ \boldsymbol{I}_p + \boldsymbol{\Omega}_{j,0} & ({}^\forall j \in \mathcal{J}_-) \end{cases} \text{ as } n \to \infty. \tag{21}$$

The lower equation in (21) directly implies, for all $j \in \mathcal{J}_-$,

$$\mathcal{D}(j, j_*) \xrightarrow{p} \log|\boldsymbol{I}_p + \boldsymbol{\Omega}_{j,0}| \text{ as } n \to \infty, \tag{22}$$

where $\mathcal{D}(j, j_*)$ is given by (11) and $\boldsymbol{\Omega}_{j,0}$ is a limiting value of $\boldsymbol{\Gamma}_j\boldsymbol{\Gamma}_j'/n$, which is defined in assumption A3. Here, $\boldsymbol{\Gamma}_j\boldsymbol{\Gamma}_j'$ is the noncentrality matrix given by (6). On the other hand, for all $j \in \mathcal{J}_+\backslash\{j_*\}$, we have

$$\mathcal{D}(j, j_*) = -\log\left|\boldsymbol{I}_p + \mathcal{E}'(\boldsymbol{P}_j - \boldsymbol{P}_{j_*})\mathcal{E}\{\mathcal{E}'(\boldsymbol{I}_p - \boldsymbol{J}_n - \boldsymbol{P}_j)\mathcal{E}\}^{-1}\right|$$
$$= -\frac{1}{n}\mathrm{tr}(\boldsymbol{Z}_j'\boldsymbol{Z}_j - \boldsymbol{Z}_{j_*}'\boldsymbol{Z}_{j_*}) + o_p(n^{-1}) \text{ as } n \to \infty, \tag{23}$$

where $\boldsymbol{Z}_j$ is given by (10). Recall that $\boldsymbol{Z}_j = O_p(1)$ under assumption A2. From this result and equation (23), we derive, for all $j \in \mathcal{J}_+\backslash\{j_*\}$,

$$n\mathcal{D}(j, j_*) = O_p(1) \text{ as } n \to \infty. \tag{24}$$

Thus, Lemma 3 and equations (12), (22), and (24) lead us to the following theorem for the condition that IC_m is consistent:

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 the conditions in Yanagihara *et al*. (2012) 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. Suppose that 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 $j, \ell \in \mathcal{J}_+$.

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

$$\begin{aligned} P(\hat{j}_m = j) &= P(\cap_{\ell\in\mathcal{J}\backslash\{j\}}\{\mathrm{IC}_m(\ell) > \mathrm{IC}_m(j)\}) \\ &= P(\{\cap_{\ell\in\mathcal{J}_-\backslash\{j\}}\{\mathrm{IC}_m(\ell) > \mathrm{IC}_m(j)\}\} \cap \{\cap_{\ell\in\mathcal{J}_+\backslash\{j\}}\{\mathrm{IC}_m(\ell) > \mathrm{IC}_m(j)\}\}). \end{aligned} \quad (25)$$

The same way as was used in the calculation of (22) 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 the condition C1-3 that

$$\frac{1}{n}\{\mathrm{IC}_m(\ell_2) - \mathrm{IC}_m(\ell_1)\} \xrightarrow{p} \log|\boldsymbol{I}_p + \boldsymbol{\Omega}_{\ell_2,0}| > 0. \quad (26)$$

Equation (26) and Lemma 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 4 (ii), we have

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

Thus, from equation (25) and Lemma 4 (ii), we can see that

$$\lim_{n\to\infty} P(\hat{j}_m = j) = \begin{cases} 0 & (j \in \mathcal{J}_-) \\ \lim_{n\to\infty} P(\cap_{\ell\in\mathcal{J}_+\backslash\{j\}}\{\mathrm{IC}_m(\ell) > \mathrm{IC}_m(j)\}) & (j \in \mathcal{J}_+) \end{cases}. \quad (27)$$

On the other hand, by using equation (23), we have, for all $j, \ell \in \mathcal{J}_+$,

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

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

$$\mathrm{IC}_m(j) - \mathrm{IC}_m(\ell) \xrightarrow{p} -\mathrm{tr}(\boldsymbol{Z}_j'\boldsymbol{Z}_j - \boldsymbol{Z}_\ell'\boldsymbol{Z}_\ell) + m_0 p(k_j - k_\ell) \text{ as } n \to \infty. \quad (28)$$

Notice that $\mathrm{tr}(\boldsymbol{Z}_j'\boldsymbol{Z}_j) = \mathrm{vec}(\boldsymbol{Z}_j)'\mathrm{vec}(\boldsymbol{Z}_j)$ and $Cov[\mathrm{vec}(\boldsymbol{Z}_j), \mathrm{vec}(\boldsymbol{Z}_\ell)] = \boldsymbol{I}_p \otimes \boldsymbol{R}_{j,j,0}^{-1/2}\boldsymbol{R}_{j,\ell,0}\boldsymbol{R}_{\ell,\ell,0}^{-1/2}$, where $\boldsymbol{R}_{j,\ell,0}$ is the submatrix of $\boldsymbol{R}_0$, which is defined in assumption A4. Moreover, it follows from Lemma 2 that $\mathrm{vec}(\boldsymbol{Z}_j) \xrightarrow{d} N_{k_j p}(\boldsymbol{0}_{k_j p}, \boldsymbol{I}_{k_j p})$ as $n \to \infty$. Substituting equation (28) into equation (27) yields the following corollary:

Corollary 1 *Suppose that assumptions A1-A5 hold. When condition C1-3 holds, the asymptotic probability of selecting the model* j *by* IC_m *is*

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

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

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 $O(1)$ 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 $E[||\varepsilon||^8] < \infty$. Therefore, if $E[||\varepsilon||^8] < \infty$ holds, the asymptotic probabilities of selecting the model j by most bias-corrected AICs become the same as those in Corollary 1.

5. Conditions for Consistency under the HD Asymptotic Framework

In this section, we derive the conditions such that IC_m is consistent under the HD asymptotic framework, i.e., n and p approach ∞ simultaneously under the condition that $c_{n,p} \to c_0 \in [0, 1)$. Under the HD asymptotic framework, increasing the dimension of $\hat{\boldsymbol{\Sigma}}_j$ with an increase in the sample size n is a serious problem. Of course, convergence in probability of $\hat{\boldsymbol{\Sigma}}_j$ in (21) is not ensured. If $\varepsilon \sim N_p(\boldsymbol{0}_p, \boldsymbol{I}_p)$ holds, $n\hat{\boldsymbol{\Sigma}}_j$ is distributed according to the central or noncentral Wishart distribution with $n - k_j - 1$ degrees of freedom. From Fujikoshi *et al*. (2010), th. 3.2.4, p. 57, we can see that

$$\left|\frac{\boldsymbol{V}_1}{\boldsymbol{V}_1 + \boldsymbol{V}_2}\right| = \left|\frac{\boldsymbol{B}_1}{\boldsymbol{B}_1 + \boldsymbol{B}_2}\right|, \tag{29}$$

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

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

By applying this formula to $\hat{\boldsymbol{\Sigma}}_j$, we can evaluate the asymptotic behavior of $\mathcal{D}(j, j_*)$ by using two random matrices whose dimensions do not increase with an increase in the sample size. By using this idea, Yanagihara *et al*. (2012) derived the condition for consistency under the HD asymptotic framework. However, needless to say, we cannot use this idea in this paper, because the true distribution is not a normal distribution. Hence, it is necessary to use a different idea. We will employ the property of the convergence in probability of $\boldsymbol{W}$ in Lemma 7, and the distribution of $\lambda_{\max}(\boldsymbol{S}_j)$ in Lemma 8 to evaluate the asymptotic behavior, where $\boldsymbol{W}$ is given by (13).

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 (9). 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 1, it can be determined that $\boldsymbol{b}_{j,1}, \ldots, \boldsymbol{b}_{j,k_j}$ satisfy the conditions in (14) when assumptions A4 and A5 hold. Therefore, if assumptions A4 and A5 hold, we can derive

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

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:

$$\boldsymbol{Q}_j'\boldsymbol{W}\boldsymbol{Q}_j \xrightarrow{p} c_0\boldsymbol{I}_{k_j} \text{ as } c_{n,p} \to c_0. \tag{30}$$

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 (13). Then, by using equation (30) and the property of the determination (see, e.g., Harville, 1997, chap. 18, cor. 18.1.2), the following equation is satisfied for all $j \in \mathcal{J}_+\backslash\{j_*\}$:

$$\begin{aligned}\mathcal{D}(j, j_*) &= \log\frac{|\boldsymbol{\mathcal{E}}'(\boldsymbol{I}_n - \boldsymbol{J}_n - \boldsymbol{P}_j)\boldsymbol{\mathcal{E}}|}{|\boldsymbol{\mathcal{E}}'(\boldsymbol{I}_n - \boldsymbol{J}_n - \boldsymbol{P}_{j_*})\boldsymbol{\mathcal{E}}|} = \log\frac{|\boldsymbol{U}'(\boldsymbol{I}_n - \boldsymbol{P}_j)\boldsymbol{U}|}{|\boldsymbol{U}'(\boldsymbol{I}_n - \boldsymbol{P}_{j_*})\boldsymbol{U}|}\\ &= \log\frac{|\boldsymbol{I}_p - (\boldsymbol{U}'\boldsymbol{U})^{-1}\boldsymbol{U}'\boldsymbol{P}_j\boldsymbol{U}|}{|\boldsymbol{I}_p - (\boldsymbol{U}'\boldsymbol{U})^{-1}\boldsymbol{U}'\boldsymbol{P}_{j_*}\boldsymbol{U}|} = \log\frac{|\boldsymbol{I}_{k_j} - \boldsymbol{Q}_j'\boldsymbol{W}\boldsymbol{Q}_j|}{|\boldsymbol{I}_{k_*} - \boldsymbol{Q}_{j_*}'\boldsymbol{W}\boldsymbol{Q}_{j_*}|}\\ &\xrightarrow{p} (k_j - k_*)\log(1 - c_0) \text{ as } c_{n,p} \to c_0.\end{aligned} \tag{31}$$

It follows from equation (19) that for all $j \in \mathcal{J}_-$

$$\begin{aligned}\mathcal{D}(j, j_*) &= \log\frac{|\boldsymbol{\mathcal{E}}'(\boldsymbol{I}_n - \boldsymbol{J}_n - \boldsymbol{P}_j - \boldsymbol{H}_j\boldsymbol{H}_j')\boldsymbol{\mathcal{E}} + (\boldsymbol{L}_j^{1/2}\boldsymbol{G}_j' + \boldsymbol{H}_j'\boldsymbol{\mathcal{E}})'(\boldsymbol{L}_j^{1/2}\boldsymbol{G}_j' + \boldsymbol{H}_j'\boldsymbol{\mathcal{E}})|}{|\boldsymbol{\mathcal{E}}'(\boldsymbol{I}_n - \boldsymbol{J}_n - \boldsymbol{P}_{j_*})\boldsymbol{\mathcal{E}}|}\\ &= \log\frac{|\boldsymbol{I}_p + \boldsymbol{S}_j^{-1}(\boldsymbol{L}_j^{1/2}\boldsymbol{G}_j' + \boldsymbol{H}_j'\boldsymbol{\mathcal{E}})'(\boldsymbol{L}_j^{1/2}\boldsymbol{G}_j' + \boldsymbol{H}_j'\boldsymbol{\mathcal{E}})/n||\boldsymbol{U}'(\boldsymbol{I}_n - \boldsymbol{P}_j - \boldsymbol{H}_j\boldsymbol{H}_j')\boldsymbol{U}|}{|\boldsymbol{U}'(\boldsymbol{I}_n - \boldsymbol{P}_{j_*})\boldsymbol{U}|}\\ &= \log\frac{|\boldsymbol{I}_{\gamma_j} + (\boldsymbol{L}_j^{1/2}\boldsymbol{G}_j' + \boldsymbol{H}_j'\boldsymbol{\mathcal{E}})\boldsymbol{S}_j^{-1}(\boldsymbol{L}_j^{1/2}\boldsymbol{G}_j' + \boldsymbol{H}_j'\boldsymbol{\mathcal{E}})'/n||\boldsymbol{U}'(\boldsymbol{I}_n - \boldsymbol{P}_j - \boldsymbol{H}_j\boldsymbol{H}_j')\boldsymbol{U}|}{|\boldsymbol{U}'(\boldsymbol{I}_n - \boldsymbol{P}_{j_*})\boldsymbol{U}|}\\ &\geq \log\left|\lambda_{\max}(\boldsymbol{S}_j)\boldsymbol{I}_{\gamma_j} + \boldsymbol{C}_j(\boldsymbol{L}_j^{1/2}\boldsymbol{G}_j' + \boldsymbol{H}_j'\boldsymbol{\mathcal{E}})(\boldsymbol{L}_j^{1/2}\boldsymbol{G}_j' + \boldsymbol{H}_j'\boldsymbol{\mathcal{E}})'\boldsymbol{C}_j'/n\right|\\ &\quad + \log\frac{|\boldsymbol{U}'(\boldsymbol{I}_n - \boldsymbol{P}_j - \boldsymbol{H}_j\boldsymbol{H}_j')\boldsymbol{U}|}{|\boldsymbol{U}'(\boldsymbol{I}_n - \boldsymbol{P}_{j_*})\boldsymbol{U}|} - \gamma_j\log\lambda_{\max}(\boldsymbol{S}_j)\\ &= \mathcal{D}_1(j, j_*) + \mathcal{D}_2(j, j_*) + \mathcal{D}_3(j, j_*),\end{aligned} \tag{32}$$

where $\boldsymbol{H}_j$, $\boldsymbol{L}_j$, and $\boldsymbol{G}_j$ are given in (17); $\boldsymbol{C}_j$ is given by (18); and $\boldsymbol{S}_j$ is given by (20).

We first evaluate the asymptotic behavior of $\mathcal{D}_1(j, j_*)$ in (32). 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 5 and 6 that

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

where $\kappa_4^{(1)}$ is given by (7), and s is some positive constant given by (8). These equations imply that $\boldsymbol{C}_j\boldsymbol{H}_j'\boldsymbol{\mathcal{E}}\boldsymbol{\mathcal{E}}'\boldsymbol{H}_j\boldsymbol{C}_j' = p\boldsymbol{I}_{\gamma_j} + O_p(p^{(1+s)/2}) = O_p(p)$ as $c_{n,p} \to c_0$. Moreover, from Hölder's inequality, we have

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

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 $c_{n,p} \to c_0$. Additionally, it follows from equation (ii) in Lemma 8 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

$$\left|\frac{1}{p}\left\{\lambda_{\max}(\boldsymbol{S}_j)\boldsymbol{I}_{\gamma_j} + \frac{1}{n}\boldsymbol{C}_j(\boldsymbol{L}_j^{1/2}\boldsymbol{G}_j' + \boldsymbol{H}_j'\boldsymbol{\mathcal{E}})(\boldsymbol{L}_j^{1/2}\boldsymbol{G}_j' + \boldsymbol{H}_j'\boldsymbol{\mathcal{E}})'\boldsymbol{C}_j'\right\}\right| \xrightarrow{p} |\boldsymbol{\Delta}_{j,0}| \text{ as } c_{n,p} \to c_0,$$

where $\boldsymbol{\Delta}_{j,0}$ is a limiting value of $\boldsymbol{\Gamma}_j'\boldsymbol{\Gamma}_j/(np)$, which is defined in assumption A6. Notice that

$$\mathcal{D}_1(j, j_*) = \log\left[p^{\gamma_j}\left|\left\{\lambda_{\max}(\boldsymbol{S}_j)\boldsymbol{I}_{\gamma_j} + \boldsymbol{C}_j(\boldsymbol{L}_j^{1/2}\boldsymbol{G}_j' + \boldsymbol{H}_j'\boldsymbol{\mathcal{E}})(\boldsymbol{L}_j^{1/2}\boldsymbol{G}_j' + \boldsymbol{H}_j'\boldsymbol{\mathcal{E}})'\boldsymbol{C}_j'/n\right\}/p\right|\right].$$

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

$$\frac{1}{\log p}\mathcal{D}_1(j, j_*) \xrightarrow{p} \gamma_j \text{ as } c_{n,p} \to c_0. \tag{33}$$

Next, we evaluate the asymptotic behavior of $\mathcal{D}_2(j, j_*)$ in (32). From equation (30) 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 equation (i) in Lemma 8, we can see that

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

It follows from equation (i) in Lemma 8 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 (3). Thus, we also have

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

The above upper and lower bounds of $\mathcal{D}_2(j, j_*)$ imply that

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

Finally, we evaluate the asymptotic behavior of $\mathcal{D}_3(j, j_*)$ in (32). The asymptotic behavior of this term depends on whether we assume A2 or A2′. Let $I(x > a)$ be an indicator function, i.e., $I(x > a) = 1$ if $x > a$ and $I(x > a) = 0$ if $x \le a$. Notice that

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

It follows from equation (ii) in Lemma 8 that $\lambda_{\max}(\boldsymbol{S}_j)I(\lambda_{\max}(\boldsymbol{S}_j) \ge p^{1/2})/p^{1/2}$ is $O_p(1)$ as $c_{n,p} \to c_0$ and is larger than or equal to 1 when assumption A2 holds. This implies that

$$\frac{1}{\log p}\underline{\mathcal{D}}_3(j, j_*) \xrightarrow{p} -\frac{1}{2}\gamma_j \text{ as } c_{n,p} \to c_0. \tag{35}$$

On the other hand, if assumption A2′ holds instead of assumption A2, it follows from equation (iii) in Lemma 8 that $\log \lambda_{\max}(\boldsymbol{S}_j) = O_p(1)$ as $c_{n,p} \to c_0$. This implies that

$$\frac{1}{\log p}\mathcal{D}_3(j, j_*) \xrightarrow{p} 0 \text{ as } c_{n,p} \to c_0. \tag{36}$$

Combining (32), (34), (33), (35), and (36) yields

$$\frac{1}{\log p}\mathcal{D}(j, j_*) \geq \begin{cases} \{\mathcal{D}_1(j, j_*) + \mathcal{D}_2(j, j_*) + \underline{\mathcal{D}}_3(j, j_*)\}/\log p \xrightarrow{p} \gamma_j/2 & \text{(when A2 holds)} \\ \{\mathcal{D}_1(j, j_*) + \mathcal{D}_2(j, j_*) + \mathcal{D}_3(j, j_*)\}/\log p \xrightarrow{p} \gamma_j & \text{(when A2′ holds)} \end{cases}, \tag{37}$$

as $c_{n,p} \to c_0$. From the results (31) and (37), equation (12), and equation (ii) in Lemma 3, the following theorem is derived:

Theorem 2 *Suppose that assumptions A1, A2, and A4–A6 hold. A variable selection using IC_m is consistent when $(n, p) \to \infty$ under $c_{n,p} \to c_0$ 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. $\forall j \in \mathcal{J}_-, \lim_{c_{n,p}\to c_0}\{m(j) - m(j_*)\}/(n\log p) > -\gamma_j/2.$

If assumption A2′ is satisfied instead of A2, condition C2-2 is relaxed as

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

It should be kept in mind that $\lim_{c\to 0} c^{-1}\log(1 - c) = -1$, and $c^{-1}\log(1 - c)$ is a monotonically decreasing function in $0 \leq c < 1$. From Theorem 2, we can see that the conditions for satisfying consistency are free of the influence of nonnormality in the true distribution. In particular, when assumption A2′ is satisfied instead of assumption A2, the sufficient condition for consistency is the same as that in Yanagihara *et al*. (2012), which was obtained under the assumption that the normality assumption is correct.

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 (5). First, we consider the AIC and AIC_c. Notice that $m(j) - m(j_*)$ in the AIC_c can be expanded as

$$m(j) - m(j_*) = \frac{(k_j - k_*)(2 - c_{n,p})p}{(1 - c_{n,p})^2} + O(pn^{-1}) \text{ as } c_{n,p} \to c_0. \tag{38}$$

Hence, the differences between the penalty terms of the AICs and the AIC_cs 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_c. Furthermore, it follows from equality (38) that

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

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_c always satisfy the condition C2-1, and those in the AIC satisfy the condition C2-1 if $c_0 \in [0, c_\text{a})$, where c_a (≈ 0.797) is a constant satisfying

$$\log(1-c_\text{a}) + 2c_\text{a} = 0. \tag{39}$$

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

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

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

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

Let $\mathcal{S}_-$ be a set defined by

$$\mathcal{S}_- = \{j \in \mathcal{J}_- | k_* - k_j > 0\}. \tag{40}$$

When $j \in \mathcal{S}_-^\text{c} \cap \mathcal{J}_-$, condition C2-2 is satisfied because $c_0(k_j-k_*) \geq 0$ holds. When $j \in \mathcal{S}_-$, condition C2-2 is satisfied if $c_0 < \gamma_j/\{2(k_*-k_j)\}$ holds for all $j \in \mathcal{S}_-$. 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, the consistency properties of specific criteria are clarified in the following corollary:

Corollary 2 *Suppose that assumptions A1, A2, and A4–A6 are satisfied.*

(*i*) *A variable selection using the AIC is consistent if* $c_0 \in [0, c_\text{a})$ *holds, and it is not consistent if* $c_0 \in (c_\text{a}, 1)$ *holds, where* c_a *is given by* (39).

(ii) *Variable selections using the AIC$_c$ and HQC are consistent.*

(iii) *Variable selections using the BIC and CAIC are consistent if $c_0 \in [0, c_b)$ holds, where $c_b = \min\{1, \min_{j \in \mathcal{S}_-} \gamma_j / \{2(k_* - k_j)\}\}$ and $\mathcal{S}_-$ is given by (40). If assumption A2′ is satisfied instead of A2, the condition $c_0 \in [0, c_b)$ is relaxed as $c_0 \in [0, c'_b)$, where $c'_b = \min\{1, \min_{j \in \mathcal{S}_-} \gamma_j / (k_* - k_j)\}$.*

Corollary 2 shows that, when $c_{n,p} \to c_0$, the AIC, AIC$_c$, and HQC are consistent in model selection if $c_0 \in [0, c_a)$ for the AIC, and if $c_0 \in [0, 1)$ for the AIC$_c$ and HQC. Therefore, the ranges of values for (n, p) that satisfy consistency are wider for the AIC$_c$ and HQC than that for the AIC. Moreover, Corollary 2 indicates that the BIC and the CAIC are not always consistent in variable selection when $c_{n,p} \to c_0$. Since $c_0 < 1$ and $k_{j_+} - k_j > k_* - k_j$ for all $j \in \mathcal{S}_-$, $\gamma_j > c_0(k_* - k_j)$ is satisfied if $\gamma_j = k_{j_+} - k_j$ holds. In contrast, if $c_0 = 0$, then $\gamma_j > c_0(k_* - k_j)$ is satisfied. Therefore, we can see that variable selections using the BIC and the CAIC are consistent as $c_{n,p} \to c_0$ if $\gamma_j = k_{j_+} - k_j$ and $c_0 \in (0, 1/2)$ hold, or $c_{n,p}$ converges to 0. However, if the previous condition does not hold, we cannot determine if variable selections using the BIC and the CAIC are consistent as $c_{n,p} \to c_0$.

6. Numerical Study

In this section, we numerically examine the validity of our claim. The probability of selecting the true model by the AIC, AIC$_c$, BIC, CAIC, and HQC in (5) was evaluated by Monte Carlo simulations with 10,000 iterations. The ten candidate models $j_\alpha = \{1, \ldots, \alpha\}$ $(\alpha = 1, \ldots, k)$, 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 $z_1, \ldots, z_n$, we constructed an $n \times k$ matrix of explanatory variables $\boldsymbol{X}$, where the (a, b)th element was defined by z_a^{b-1} $(a = 1, \ldots, n; b = 1, \ldots, k)$. The true model was determined by $\boldsymbol{\Theta}_* = (1, 1, 3, -4, 5)' \boldsymbol{1}'_p$, $j_* = \{1, 2, 3, 4, 5\}$, 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 $\alpha = 1, \ldots, 4$ was the underspecified model, and j_α with $\alpha \geq 5$ was the overspecified model.

Let $\boldsymbol{\nu} \sim N_p(\boldsymbol{0}_p, \boldsymbol{I}_p)$ and $\delta \sim \chi_6^2$ be a mutually independent random vector and variable. Then, $\boldsymbol{\varepsilon}$ was generated from the following three 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): $\boldsymbol{\varepsilon} = \boldsymbol{\Psi}^{-1/2}\{10(\sqrt{\delta/6} - \eta)\boldsymbol{1}_p + \sqrt{\delta/6}\boldsymbol{\nu}\}$, where $\eta = 15\sqrt{\pi/3}/16$ and $\boldsymbol{\Psi} = \boldsymbol{I}_p + 100(1 - \eta^2)\boldsymbol{1}_p\boldsymbol{1}'_p$.

It is easy to see that distributions 1 and 2 are symmetric, and distribution 3 is skewed.

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 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 $c_{n,p} \to c_0 \in (1/8, 1)$.

Tables 1, 2, and 3 show the probability of selecting the true model by the AIC, AIC$_c$, BIC, CAIC, and HQC when the distributions of $\boldsymbol{\varepsilon}$ are 1, 2, and 3, respectively. For $n = \infty$ or $p = \infty$, we list the

Table 1. Selection Probabilities of the True Model (%) in the Case of Distribution 1

	Case 1						Case 2 ($c_0 = 0.01$)					
n	p	AIC	AIC$_c$	BIC	CAIC	HQC	p	AIC	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	AIC$_c$	BIC	CAIC	HQC	p	AIC	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	AIC$_c$	BIC	CAIC	HQC	p	AIC	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	AIC$_c$	BIC	CAIC	HQC	p	AIC	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

theoretical values obtained from Theorems 1 and 2. In particular, by using the result in Yanagihara *et al*. (2012), we can obtain the theoretical values of the asymptotic selection probabilities of the true model by the BIC and CAIC if the distribution of ε is normal, even for Case 6. The symbol "—" indicates that the theoretical value is not clear. From the tables, we can see that in the cases of the AIC, AIC$_c$, and HQC, the greater the dimension and sample size, the greater the probabilities. Compared with the results obtained from the AIC, AIC$_c$, and HQC, the probabilities for the 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 dimension and sample size were, the higher the selection probabilities became, with the exception of Case 6. This was because there is a possibility that variable selections using the BIC and the CAIC are not consistent in Case 6. Additionally, when n was small and p was large, the selection probabilities of the BIC and the CAIC were both very low. However, if the BIC and the CAIC were consistent in variable selection, these probabilities became high as n and p increased. Moreover, we 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 violation of the normality assumption is not large.

Table 2. Selection Probabilities of the True Model (%) in the Case of Distribution 2

	Case 1						Case 2 (c_0 = 0.01)					
n	p	AIC	AIC$_c$	BIC	CAIC	HQC	p	AIC	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	AIC$_c$	BIC	CAIC	HQC	p	AIC	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	AIC$_c$	BIC	CAIC	HQC	p	AIC	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	AIC$_c$	BIC	CAIC	HQC	p	AIC	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

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

7. Conclusion and Discussion

In this paper, we derived the conditions to satisfy the consistency property of a log-likelihood-based information criterion in (4) for selecting variables in the multivariate linear regression models with the normality assumption, but for which normality is 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, 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 free from the influence of nonnormality in the true distribution. Under the LS asymptotic framework, we ob-

Table 3. Selection Probabilities of the True Model (%) in the Case of Distribution 3

	Case 1						Case 2 ($c_0 = 0.01$)					
n	p	AIC	AIC_c	BIC	CAIC	HQC	p	AIC	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	AIC_c	BIC	CAIC	HQC	p	AIC	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	AIC_c	BIC	CAIC	HQC	p	AIC	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	AIC_c	BIC	CAIC	HQC	p	AIC	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

tained the necessary and sufficient condition for consistency, which was equivalent to that derived under the normality assumption. Under the HD asymptotic framework, the sufficient condition for consistency was obtained. The condition was slightly stronger 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 $\mathcal{D}(j, j_*)$ by two random matrices whose dimensions do not increase with an increase in the sample size, after applying the formula in (29) to $\hat{\boldsymbol{\Sigma}}_j$, which is the same method used in Yanagihara *et al*. (2012). However, we cannot use this because our setting assumes that the normality assumption is violated. Hence, we employed the convergence in probability of $\boldsymbol{W}$ in Lemma 7, and the distribution of $\lambda_{\max}(\boldsymbol{S}_j)$ in Lemma 8, 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 (i) in Lemma 8 is changed to $\lambda_{\max}(\boldsymbol{S}_j) = O_p(p^{1/3})$. This directly implies that condition C2-2 is relaxed to $\lim_{c_{n,p}\to c_0}\{m(j) - m(j_*)\}/(n \log p) < -2\gamma_j/3$. If we assume the existence of $E[\|\boldsymbol{\varepsilon}\|^{2r}]$, and that $E[\|\varepsilon\|^{2r}] = O(p^r)$ as $p \to \infty$ for all $r \geq 1$, condition C2-2 may be relaxed to $\lim_{c_{n,p} \to c_0} \{m(j) - m(j_*)\}/(n \log p) < -\gamma_j$, which is equivalent to the condition obtained from the normality assumption.

Acknowledgment The author thanks Prof. Hirofumi Wakaki and Prof. Yasunori Fujikoshi, Hiroshima University, for their helpful comments on the assumptions necessary to satisfy consistency. 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. In *2nd International Symposium on Information Theory* (eds. B. N. Petrov & F. Csáki), pp. 267–281. Akadémiai Kiadó, Budapest.

Akaike, H. (1974): A new look at the statistical model identification. *Institute of Electrical and Electronics Engineers. Transactions on 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. *The Annals of Probability* **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. *Journal of Bioscience and Bioengineering* **102**, 34–40.

Fan, J., Fan, Y. and Lv, J. (2008): High dimensional covariance matrix estimation using a factor model. *Journal of 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 Mathematical Journal* **13**, 203–214.

Fujikoshi, Y. (1985): Selection of variables in two-group discriminant analysis by error rate and Akaike's information criteria. *Journal of Multivariate Analysis* **17**, 27–37.

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

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

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. *American Journal of Mathematical and Management Sciences* **25**, 221–258.

Hannan, E. J. and Quinn, B. G. (1979): The determination of the order of an autoregression. *Journal of the Royal Statistical Society*, *Series* **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. *Annals of the Institute of Statistical Mathematics* **49**, 411–434.

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

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. *Journal of Environmental Radioactivity* **87**, 62–76.

Schwarz, G. (1978): Estimating the dimension of a model. *The Annals of Statistics* **6**, 461–464.

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

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

Stone, M. (1974): Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society*,

Series **B 36**, 111–147.
Stone, M. (1977): An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society, Series* **B 39**, 44–47.
Takeuchi, K. (1976): Distribution of information statistics and criteria for adequacy of models. *Mathematical Science* **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 Mathematical Journal* **32**, 17–50.
Yanagihara, H. (2006): Corrected version of *AIC* for selecting multivariate normal linear regression models in a general nonnormal case. *Journal of Multivariate Analysis* **97**, 1070–1089.
Yanagihara, H., Kamo, K. and Tonda, T. (2011): Second-order bias-corrected AIC in multivariate normal linear models under nonnormality. *The Canadian Journal of Statistics* **39**, 126–146.
Yanagihara, H., Wakaki, H. and Fujikoshi, Y. (2012): A consistency property of the AIC for multivariate linear models when the dimension and the sample size are large. *TR 12-08, Statistical Research Group, Hiroshima University*, Hiroshima.
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. *Journal of Japanese Forestry Society* **87**, 504–512 (in Japanese).

Appendix

A. Proof of Lemma 1

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

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

The above equation implies that

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

Moreover, assumption A3 indicates $\lambda_{\min}(\boldsymbol{X}'\boldsymbol{X}) = O(n)$. Hence, by combining this equation, equation (A.1), and assumption A4, we have proved Lemma 1.

B. Proof of Lemma 2

In order to prove Lemma 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 (9). It is clear that $\boldsymbol{\nu}_{j,1}, \ldots, \boldsymbol{\nu}_{j,n}$ are independent, and $E[\boldsymbol{\nu}_{j,i}] = \boldsymbol{0}_{pk_j}$, $Cov[\boldsymbol{\nu}_{j,i}] = \boldsymbol{I}_p \otimes \boldsymbol{q}_{j,i}\boldsymbol{q}_{j,i}'$, 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 $\sum_{i=1}^{n} Cov[\boldsymbol{\nu}_{j,i}] = \boldsymbol{I}_{pk_j}$ hold, where $\boldsymbol{Q}_j$ is given by (9). Then, for all $\epsilon > 0$, we derive

$$E\left[||\boldsymbol{\nu}_{j,i}||^2 I(||\boldsymbol{\nu}_{j,i}|| > \epsilon)\right]^2 \leq E\left[||\boldsymbol{\nu}_{j,i}||^4\right] E\left[I(||\boldsymbol{\nu}_{j,i}|| > \epsilon)^2\right] \\ = E\left[||\boldsymbol{\nu}_{j,i}||^4\right] P(||\boldsymbol{\nu}_{j,i}|| > \epsilon) \leq \frac{1}{\epsilon^4} E\left[||\boldsymbol{\nu}_{j,i}||^4\right]^2.$$

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 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

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

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.

C. Proof of Lemma 3

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

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

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

$$P(|T_{j,\ell} - \tau_{j,\ell}| > \tau_{j,\ell}) = P(\{T_{j,\ell} > 2\tau_{j,\ell}\} \cup \{T_{j,\ell} < 0\}) \\ \geq P(T_{j,\ell} < 0) \geq P(\mathrm{IC}_m(j) - \mathrm{IC}_m(\ell) < 0). \tag{C.2}$$

Since equation (C.1) holds for all $\epsilon > 0$, the first probability in (C.2) converges to 0. This indicates that $P(\mathrm{IC}_m(j) < \mathrm{IC}_m(\ell)) \to 0$. Furthermore, it is common knowledge that equation (C.1) is equivalent to

$$P(|T_{j,\ell} - \tau_{j,\ell}| \leq \epsilon) \to 1, \ ^\forall\epsilon > 0. \tag{C.3}$$

By the same method as in the calculation of (C.2), we derive

$$P(|T_{j,\ell} - \tau_{j,\ell}| \leq \tau_{j,\ell}/2) \leq P(|T_{j,\ell} - \tau_{j,\ell}| < \tau_{j,\ell}) = P(\{T_{j,\ell} > 0\} \cap \{T_{j,\ell} < 2\tau_{j,\ell}\}) \\ \leq P(T_{j,\ell} > 0) \leq P(\mathrm{IC}_m(j) - \mathrm{IC}_m(\ell) > 0). \tag{C.4}$$

Since equation (C.3) holds for all $\epsilon > 0$, the first probability in (C.4) 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

$$P(\hat{j}_m = j) = 1 - P(\hat{j}_m \neq j) = 1 - P(\cup_{\ell\in\mathcal{J}\backslash\{j\}}\{\mathrm{IC}_m(\ell) < \mathrm{IC}_m(j)\}) \\ \geq 1 - \sum_{\ell\in\mathcal{J}\backslash\{j\}} P(\mathrm{IC}_m(\ell) < \mathrm{IC}_m(j)). \tag{C.5}$$

Since $T_{\ell,j} \xrightarrow{p} \tau_{\ell,j} > 0$ holds for all $\ell \in \mathcal{J}\backslash\{j\}$, we can see from Lemma 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 (C.5), 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 (C.1), (C.2), and (C.5), we obtain

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

Consequently, equation (iii) is proved.

D. Proof of Lemma 4

First, we prove equation (i) in Lemma 4. Notice that $P(A \cap B) \le \min\{P(A), P(B)\}$. It follows from this equation and the assumption $P(B) \to 0$ that $\min\{P(A), P(B)\} \to 0$. Hence, equation (i) is proved. Next, we show equation (ii) in Lemma 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 4 is proved.

E. Proof of Lemma 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 $\varepsilon_1, \dots, \varepsilon_n$ are identically and independently distributed, $E[\varepsilon_a \varepsilon_a'] = \boldsymbol{I}_p$, and $E[\|\varepsilon_a\|^4] = \kappa_4^{(1)} + p(p+2)$, where $\kappa_4^{(1)}$ is given by (7). These equations imply that

$$\begin{aligned} E\left[\mathrm{tr}\left\{(\boldsymbol{\mathcal{E}}' \boldsymbol{A} \boldsymbol{\mathcal{E}})^2\right\}\right] &= \sum_{a,b,c,d}^{n} (\boldsymbol{A})_{ad} (\boldsymbol{A})_{bc} E[\varepsilon_a' \varepsilon_b \varepsilon_c' \varepsilon_d] \\ &= \sum_{a=1}^{n} \{(\boldsymbol{A})_{aa}\}^2 E\left[(\varepsilon_a' \varepsilon_a)^2\right] + \sum_{a \ne b}^{n} \left[\{(\boldsymbol{A})_{aa}\}\{(\boldsymbol{A})_{bb}\} E\left[(\varepsilon_a' \varepsilon_b)^2\right]\right. \\ &\quad \left. + \{(\boldsymbol{A})_{ab}\}^2 \left\{E[\varepsilon_a' \varepsilon_a \varepsilon_b' \varepsilon_b] + E[(\varepsilon_a' \varepsilon_b)^2]\right\}\right] \\ &= \kappa_4^{(1)} \sum_{a=1}^{n} \{(\boldsymbol{A})_{aa}\}^2 + p(p+1)\mathrm{tr}(\boldsymbol{A}^2) + p\,\mathrm{tr}(\boldsymbol{A})^2, \end{aligned}$$

and

$$\begin{aligned} E\left[\mathrm{tr}(\boldsymbol{\mathcal{E}}' \boldsymbol{A} \boldsymbol{\mathcal{E}})^2\right] &= \sum_{a,b,c,d}^{n} (\boldsymbol{A})_{ab} (\boldsymbol{A})_{cd} E[\varepsilon_a' \varepsilon_b \varepsilon_c' \varepsilon_d] \\ &= \sum_{a=1}^{n} \{(\boldsymbol{A})_{aa}\}^2 E\left[(\varepsilon_a' \varepsilon_a)^2\right] \\ &\quad + \sum_{a \ne b}^{n} \left[\{(\boldsymbol{A})_{aa}\}\{(\boldsymbol{A})_{bb}\} E\left[\varepsilon_a' \varepsilon_a \varepsilon_b' \varepsilon_b\right] + 2\{(\boldsymbol{A})_{ab}\}^2 E[(\varepsilon_a' \varepsilon_b)^2]\right] \\ &= \kappa_4^{(1)} \sum_{a=1}^{n} \{(\boldsymbol{A})_{aa}\}^2 + p^2 \mathrm{tr}(\boldsymbol{A})^2 + p\,\mathrm{tr}(\boldsymbol{A}^2). \end{aligned}$$

Consequently, Lemma 5 is proved.

F. Proof of Lemma 6

Notice that

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

Hence, Lemma 6 is proved.

G. Proof of Lemma 7

Let w_{ab} be the (a,b)th element of $\boldsymbol{W}$, and let $\bar{\varepsilon}$ be the sample mean of $\varepsilon_1,\ldots,\varepsilon_n$, i.e., $\bar{\varepsilon} = \sum_{i=1}^{n}\varepsilon_i$, where $\boldsymbol{W}$ is given by (13). It follows from $w_{ab} = \boldsymbol{u}_a'(\boldsymbol{U}'\boldsymbol{U})^{-1}\boldsymbol{u}_b$ and $\boldsymbol{u}_a = \varepsilon_a - \bar{\varepsilon}$ 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 (13). Recall that $\boldsymbol{W}$ is a symmetric idempotent matrix and $\boldsymbol{W}\boldsymbol{1}_n = \boldsymbol{0}_n$ holds. These imply that

$$0 \le w_{aa} \le 1, \quad |w_{ab}| \le \sqrt{w_{aa}w_{bb}} \le 1 \quad (a = 1,\ldots,n; b = 1,\ldots,n; a \ne b), \tag{G.1}$$

and

$$\begin{aligned}
&p = \mathrm{tr}(\boldsymbol{W}) = \sum_{a=1}^{n} w_{aa}, \quad p = \mathrm{tr}(\boldsymbol{W}^2) = \sum_{a=1}^{n} w_{aa}^2 + \sum_{a\ne b}^{n} w_{ab}^2,\\
&p^2 = \mathrm{tr}(\boldsymbol{W})^2 = \sum_{a=1}^{n} w_{aa}^2 + \sum_{a\ne b}^{n} w_{aa}w_{bb}, \quad 0 = \boldsymbol{1}_n'\boldsymbol{W}\boldsymbol{1}_n = \sum_{a=1}^{n} w_{aa} + \sum_{a\ne b}^{n} w_{ab},\\
&0 = \boldsymbol{1}'\boldsymbol{W}^2\boldsymbol{1}_n = \sum_{a=1}^{n} w_{aa}^2 + \sum_{a\ne b}^{n}(2w_{aa}w_{ab} + w_{ab}^2) + \sum_{a\ne b\ne c}^{n} w_{ab}w_{ac},\\
&0 = \mathrm{tr}(\boldsymbol{W})\boldsymbol{1}_n'\boldsymbol{W}\boldsymbol{1}_n = \sum_{a=1}^{n} w_{aa}^2 + \sum_{a\ne b}^{n}(2w_{aa}w_{ab} + w_{aa}w_{bb}) + \sum_{a\ne b\ne c}^{n} w_{aa}w_{bc},\\
&0 = (\boldsymbol{1}_n'\boldsymbol{W}\boldsymbol{1}_n)^2 = \sum_{a=1}^{n} w_{aa}^2 + \sum_{a\ne b}^{n}(w_{aa}w_{bb} + 2w_{ab}^2 + 4w_{aa}w_{ab})\\
&\qquad + 2\sum_{a\ne b\ne c}^{n}(w_{aa}w_{bc} + 2w_{ab}w_{ac}) + \sum_{a\ne b\ne c\ne d}^{n} w_{ab}w_{cd},
\end{aligned} \tag{G.2}$$

where the notation $\sum_{a_1\ne a_2\ne\cdots}^{n}$ means $\sum_{a_1=1}^{n}\sum_{a_2=1,a_2\ne a_1}^{n}\cdots$. 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 the equations in (G.2), we derive for $a \ne b \ne c \ne d$

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

It follows from equation (G.1) that

$$
w_{ab}^2 = O_p(1) \text{ as } c_{n,p} \to c_0. \tag{G.4}
$$

Hölder's inequality implies that

$$
|E[w_{aa}w_{ab}]| \le E[|w_{aa}w_{ab}|] \le \sqrt{E[w_{aa}^2]E[w_{ab}^2]}. \tag{G.5}
$$

Combining equations (G.3), (G.4), and (G.5) yields

$$
\begin{aligned}
&E[w_{aa}] = c_{n,p}, && E[w_{ab}] = O(n^{-1}), && E[w_{aa}^2] = O(1), \\
&E[w_{aa}w_{bb}] = c_{n,p}^2 + O(n^{-1}), && E[w_{ab}^2] = O(n^{-1}), && E[w_{aa}w_{ab}] = O(n^{-1/2}), \\
&E[w_{aa}w_{bc}] = O(n^{-3/2}), && E[w_{ab}w_{ac}] = O(n^{-3/2}), && E[w_{ab}w_{cd}] = O(n^{-5/2}),
\end{aligned} \tag{G.6}
$$

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

$$
\begin{aligned}
\boldsymbol{\alpha}'\boldsymbol{W}\boldsymbol{\beta} &= \sum_{a=1}^{n} \alpha_a\beta_a w_{aa} + \sum_{a\neq b}^{n} \alpha_a\beta_b w_{ab}, \\
(\boldsymbol{\alpha}'\boldsymbol{W}\boldsymbol{\beta})^2 &= \sum_{a=1}^{n} \alpha_a^2\beta_a^2 w_{aa}^2 + \sum_{a\neq b\neq c\neq d}^{n} \alpha_a\alpha_c\beta_b\beta_d w_{ab}w_{cd} \\
&\quad + \sum_{a\neq b}^{n} \left\{\alpha_a\alpha_b\beta_a\beta_b(w_{aa}w_{bb} + w_{ab}^2) + 2\left(\alpha_a^2\beta_a\beta_b + \alpha_a\alpha_b\beta_a^2\right) w_{aa}w_{ab} + \alpha_a^2\beta_b^2 w_{ab}^2\right\} \\
&\quad + \sum_{a\neq b\neq c}^{n} \left\{2\alpha_a\alpha_b\beta_a\beta_c w_{aa}w_{bc} + \left(\alpha_a^2\beta_b\beta_c + 2\alpha_a\alpha_b\beta_a\beta_c + \alpha_b\alpha_c\beta_a^2\right) w_{ab}w_{ac}\right\}.
\end{aligned}
$$

It follows from conditions $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ in (14) that

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

as $c_{n,p} \to c_0$. Consequently, by using the above results and the expectations in (G.6), we derive

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

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

H. Proof of Lemma 8

First, we prove equation (i) in Lemma 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 (3). 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 $\boldsymbol{1}_n' \boldsymbol{\mathcal{A}}_j = \boldsymbol{0}_p'$, where $\boldsymbol{\mathcal{A}}_j$ is given by (16). Hence, we have

$$\begin{aligned}
\boldsymbol{P}_j \boldsymbol{\mathcal{A}}_j = \boldsymbol{O}_{n,p} &\Leftrightarrow \boldsymbol{P}_j \boldsymbol{H}_j \boldsymbol{L}_j^{1/2} \boldsymbol{G}_j' = \boldsymbol{O}_{n,p} \Leftrightarrow \boldsymbol{P}_j \boldsymbol{H}_j \boldsymbol{L}_j^{1/2} \boldsymbol{G}_j' \boldsymbol{G}_j \boldsymbol{L}_j^{-1/2} \boldsymbol{H}_j' = \boldsymbol{O}_{n,n} \\
&\Leftrightarrow \boldsymbol{P}_j \boldsymbol{H}_j \boldsymbol{H}_j' = \boldsymbol{O}_{n,n}, \\
(\boldsymbol{I}_n - \boldsymbol{P}_{j_+})\boldsymbol{\mathcal{A}}_j = \boldsymbol{O}_{n,p} &\Leftrightarrow (\boldsymbol{I}_n - \boldsymbol{P}_{j_+})\boldsymbol{H}_j \boldsymbol{L}_j^{1/2} \boldsymbol{G}_j' = \boldsymbol{O}_{n,p} \\
&\Leftrightarrow (\boldsymbol{I}_n - \boldsymbol{P}_{j_+})\boldsymbol{H}_j \boldsymbol{L}_j^{1/2} \boldsymbol{G}_j' \boldsymbol{G}_j \boldsymbol{L}_j^{-1/2} \boldsymbol{H}_j' = \boldsymbol{O}_{n,n} \\
&\Leftrightarrow (\boldsymbol{I}_n - \boldsymbol{P}_{j_+})\boldsymbol{H}_j \boldsymbol{H}_j' = \boldsymbol{O}_{n,n} \Leftrightarrow \boldsymbol{P}_{j_+} \boldsymbol{H}_j \boldsymbol{H}_j' = \boldsymbol{H}_j \boldsymbol{H}_j', \\
\boldsymbol{1}_n' \boldsymbol{\mathcal{A}}_j = \boldsymbol{0}_p' &\Rightarrow \boldsymbol{J}_n \boldsymbol{\mathcal{A}}_j = \boldsymbol{0}_{n,p} \Leftrightarrow \boldsymbol{J}_n \boldsymbol{H}_j \boldsymbol{L}_j^{1/2} \boldsymbol{G}_j' = \boldsymbol{0}_{n,p} \\
&\Leftrightarrow \boldsymbol{J}_n \boldsymbol{H}_j \boldsymbol{L}_j^{1/2} \boldsymbol{G}_j' \boldsymbol{G}_j \boldsymbol{L}_j^{-1/2} \boldsymbol{H}_j' = \boldsymbol{O}_{n,n} \Leftrightarrow \boldsymbol{J}_n \boldsymbol{H}_j \boldsymbol{H}_j' = \boldsymbol{O}_{n,n},
\end{aligned}$$

where $\boldsymbol{H}_j$, $\boldsymbol{L}_j$, and $\boldsymbol{G}_j$ were given in (17). Hence, equation (i) in Lemma 8 is proved.

Next, we prove equation (ii) in Lemma 8. It follows from elementary linear algebra that

$$\lambda_{\max}(\boldsymbol{S}_j) \geq \mathrm{tr}(\boldsymbol{S}_j)/p, \quad \lambda_{\max}(\boldsymbol{S}_j) \leq \sqrt{\mathrm{tr}(\boldsymbol{S}_j^2)},$$

where $\boldsymbol{S}_j$ is given by (20). From Lemma 6 and equation (i) in Lemma 8, we can see that

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

The above result and Lemma 5 imply that

$$\begin{aligned}
Var[\mathrm{tr}(\boldsymbol{S}_j)] &= \frac{1}{n^2 p^2} \left\{ \phi_3(\boldsymbol{I}_n - \boldsymbol{J}_n - \boldsymbol{P}_j - \boldsymbol{H}_j \boldsymbol{H}_j') - \phi_1(\boldsymbol{I}_n - \boldsymbol{J}_n - \boldsymbol{P}_j - \boldsymbol{H}_j \boldsymbol{H}_j')^2 \right\} \\
&= O(n^{-1} p^{s-1}) \text{ as } c_{n,p} \to c_0, \\
E\left[\mathrm{tr}(\boldsymbol{S}_j^2) \right] &= \frac{1}{n^2} \phi_2(\boldsymbol{I}_n - \boldsymbol{J}_n - \boldsymbol{P}_j - \boldsymbol{H}_j \boldsymbol{H}_j') = O(p) \text{ as } c_{n,p} \to c_0,
\end{aligned}$$

where s is some positive constant given by (8). The variance of $\mathrm{tr}(\boldsymbol{S}_j)$ leads us to the equation $\mathrm{tr}(\boldsymbol{S}_j)/p \stackrel{p}{\to} E[\mathrm{tr}(\boldsymbol{S}_j)/p] = 1 - (k_j + \gamma_j + 1)/n \to 1$ as $c_{n,p} \to c_0$. Moreover, the expectation of $\mathrm{tr}(\boldsymbol{S}_j^2)$ leads us to the equation $\mathrm{tr}(\boldsymbol{S}_j^2)^{1/2} = O_p(p^{1/2})$ as $c_{n,p} \to c_0$. Hence, equation (ii) in Lemma 8 is proved.

Finally, we prove equation (iii) in Lemma 8. Suppose that assumption A2′ holds instead of assumption A2. Then, it follows from Bai and Yin (1993) that $\lambda_{\max}(\boldsymbol{\mathcal{E}}' \boldsymbol{\mathcal{E}}/n) \stackrel{a.s.}{\to} (1 + c_0^{1/2})^2$ as $c_{n,p} \to c_0$.

Since $\lambda_{\max}(\boldsymbol{S}_j) \leq \lambda_{\max}(\boldsymbol{\mathcal{E}}'\boldsymbol{\mathcal{E}}/n)$ is satisfied without assumption A2′, we have

$$\limsup_{c_{n,p} \to c_0} \lambda_{\max}(\boldsymbol{S}_j) = (1 + \sqrt{c_0})^2.$$

These indicates that equation (iii) in Lemma 8 is proved.