Consistency of log-likelihood-based information criteria for selecting variables in high-dimensional canonical correlation analysis under nonnormality

Keisuke Fukui

Department of Mathematics Graduate School of Science,
Hiroshima University,
1-3-1 Kagamiyama, Higashi-Hiroshima Hiroshima 739-8626, Japan
E-mail: d126313@hiroshima-u.ac.jp

Abstract

The purpose of this paper is to clarify the conditions for consistency of the log-likelihood-based information criteria in canonical correlation analysis of q- and p-dimensional random vectors when the dimension p is large but does not exceed the sample size. Although the vector of observations is assumed to be normally distributed, we do not know whether the underlying distribution is actually normal. Therefore, conditions for consistency are evaluated in a high-dimensional asymptotic framework when the underlying distribution is not normal.

AMS 2010 subject classification: Primary 62H12; Secondary 62H20

Key words: AIC, assumption of normality, bias-corrected AIC, BIC, consistent AIC, high-dimensional asymptotic framework, HQC, nonnormality, redundant model selection, selection probability.

1 Introduction

Canonical correlation analysis (CCA) is a statistical method employed to investigate the relationships between a pair of q- and p-dimensional random vectors, $\boldsymbol{x} = (x_1, \ldots, x_q)'$ and $\boldsymbol{y} = (y_1, \ldots, y_p)'$, respectively. Introductions to CCA are provided in many textbooks for applied statistical analysis (see, e.g., Srivastava, 2002, chap. 14.7; Timm, 2002, chap. 8.7), and it has widespread applications in many fields (e.g., Doeswijk *et al.*, 2011; Khalil *et al.*, 2011; Vahedi, 2011; Sweeney *et al.*, 2013; Vilsaint *et al.*, 2013). Let $\boldsymbol{z} = (\boldsymbol{x}', \boldsymbol{y}')'$ be a $(p + q)$-dimensional vector with

$$E[\boldsymbol{z}] = \begin{pmatrix} \boldsymbol{\mu}_x \\ \boldsymbol{\mu}_y \end{pmatrix} = \boldsymbol{\mu}, \quad Cov[\boldsymbol{z}] = \begin{pmatrix} \boldsymbol{\Sigma}_{xx} & \boldsymbol{\Sigma}_{xy} \\ \boldsymbol{\Sigma}_{xy}' & \boldsymbol{\Sigma}_{yy} \end{pmatrix} = \boldsymbol{\Sigma},$$

where $\boldsymbol{\mu}_x$ and $\boldsymbol{\mu}_y$ are mean vectors of q- and p-dimensions, respectively; $\boldsymbol{\Sigma}_{xx}$ and $\boldsymbol{\Sigma}_{yy}$ are $q \times q$ and $p \times p$ covariance matrices of $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively; and $\boldsymbol{\Sigma}_{xy}$ is the $q \times p$ covariance matrix of $\boldsymbol{x}$ and $\boldsymbol{y}$. The square of the correlation between a pair of canonical correlation variables is obtained as the eigenvalue of $\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\Sigma}_{xy}'$ and the root of the k-th largest eigenvalue is called the k-th canonical correlation.

In an actual data analysis, it is important to remove the irrelevant variables for analysis. In CCA, the problem of removing irrelevant variables can be regarded as the selection of the redundancy model, and thus it has been widely investigated by many authors (e.g., McKay, 1977; Fujikoshi, 1982, 1985; Ogura, 2010). Suppose that j denotes a subset of $\omega = \{1, \dots, q\}$ containing q_j elements, and $\boldsymbol{x}_j$ denotes the q_j-dimensional vector consisting of the elements of $\boldsymbol{x}$ indexed by the elements of j, where q_A denotes the number of elements in a set of A, i.e., $q_A = \#(A)$. For example, if $j = \{1, 2, 4\}$, then $\boldsymbol{x}_j$ consists of the first, second, and fourth elements of $\boldsymbol{x}$. Without loss of generality, $\boldsymbol{x}$ can be divided into $\boldsymbol{x} = (\boldsymbol{x}_j', \boldsymbol{x}_{\bar{j}}')'$, where $\boldsymbol{x}_j$ and $\boldsymbol{x}_{\bar{j}}$ are q_j- and $q_{\bar{j}}$-dimensional vectors, respectively. Note that $\bar{A}$ denotes the compliment of the set A. Another expressions of $\boldsymbol{\mu}_x$, $\boldsymbol{\Sigma}_{xy}$ and $\boldsymbol{\Sigma}_{xx}$ corresponding to the divisions of $\boldsymbol{x}$ are

$$\boldsymbol{\mu}_x = \begin{pmatrix} \boldsymbol{\mu}_j \\ \boldsymbol{\mu}_{\bar{j}} \end{pmatrix}, \ \boldsymbol{\Sigma}_{xy} = \begin{pmatrix} \boldsymbol{\Sigma}_{jy} \\ \boldsymbol{\Sigma}_{\bar{j}y} \end{pmatrix}, \ \boldsymbol{\Sigma}_{xx} = \begin{pmatrix} \boldsymbol{\Sigma}_{jj} & \boldsymbol{\Sigma}_{j\bar{j}} \\ \boldsymbol{\Sigma}_{j\bar{j}}' & \boldsymbol{\Sigma}_{\bar{j}\bar{j}} \end{pmatrix}.$$

We are interested in whether the elements of $\boldsymbol{x}_{\bar{j}}$ are irrelevant variables in CCA. Let $\boldsymbol{z}_1, \dots, \boldsymbol{z}_n$ be n independent random vectors from $\boldsymbol{z}$, and let $\bar{\boldsymbol{z}}$ be the sample mean of $\boldsymbol{z}_1, \dots, \boldsymbol{z}_n$ given by $\bar{\boldsymbol{z}} = n^{-1}\sum_{i=1}^n \boldsymbol{z}_i$ and $\boldsymbol{S}$ be the usual unbiased estimator of $\boldsymbol{\Sigma}$ given by $\boldsymbol{S} = (n-1)^{-1}\sum_{i=1}^n(\boldsymbol{z}_i - \bar{\boldsymbol{z}})(\boldsymbol{z}_i - \bar{\boldsymbol{z}})'$, divided in the same way as we divided $\boldsymbol{\Sigma}$, as follows:

$$\boldsymbol{S} = \begin{pmatrix} \boldsymbol{S}_{xx} & \boldsymbol{S}_{xy} \\ \boldsymbol{S}_{xy}' & \boldsymbol{S}_{yy} \end{pmatrix} = \begin{pmatrix} \boldsymbol{S}_{jj} & \boldsymbol{S}_{j\bar{j}} & \boldsymbol{S}_{jy} \\ \boldsymbol{S}_{j\bar{j}}' & \boldsymbol{S}_{\bar{j}\bar{j}} & \boldsymbol{S}_{\bar{j}y} \\ \boldsymbol{S}_{jy}' & \boldsymbol{S}_{\bar{j}y}' & \boldsymbol{S}_{yy} \end{pmatrix}.$$

Suppose that $\boldsymbol{z}_1, \dots, \boldsymbol{z}_n \sim i.i.d.\ N_{p+q}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Following Fujikoshi (1985), the candidate model that $\boldsymbol{x}_{\bar{j}}$ is irrelevant is expressed as

$$\begin{aligned} M_j : (n-1)\boldsymbol{S} &\sim W_{p+q}(n-1, \boldsymbol{\Sigma}) \\ s.t.\ \mathrm{tr}(\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\Sigma}_{xy}') &= \mathrm{tr}(\boldsymbol{\Sigma}_{jj}^{-1}\boldsymbol{\Sigma}_{jy}\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\Sigma}_{jy}'). \end{aligned} \tag{1}$$

The candidate model is called the redundancy model. If the model M_j is selected as the best model, then we regard that $\boldsymbol{x}_{\bar{j}}$ is irrelevant. An estimator of $\boldsymbol{\Sigma}$ under model M_j in (1) is given by

$$\hat{\boldsymbol{\Sigma}}_j = \arg\min_{\boldsymbol{\Sigma}}\{F(\boldsymbol{S}, \boldsymbol{\Sigma})\ s.t.\ \mathrm{tr}(\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\Sigma}_{xy}') = \mathrm{tr}(\boldsymbol{\Sigma}_{jj}^{-1}\boldsymbol{\Sigma}_{jy}\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\Sigma}_{jy}')\}, \tag{2}$$

where $F(\boldsymbol{S}, \boldsymbol{\Sigma})$ is the Kullback-Leibler (KL) discrepancy function (see Kullback & Leibler, 1951) assessed by the Wishart density, and it is given by

$$F(\boldsymbol{S}, \boldsymbol{\Sigma}) = (n-1)\{\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{S}) - \log|\boldsymbol{\Sigma}^{-1}\boldsymbol{S}| - (p+q)\}, \tag{3}$$

except for the constant term. In the covariance structure analysis, the above discrepancy function is frequently called the maximum likelihood discrepancy function (see Jöreskog, 1967) or Stein's loss function (see James & Stein, 1961). From Fujikoshi and Kurata (2008) or Fujikoshi *et al.* (2010, chap. 11.5), we can see that an explicit form of $\hat{\boldsymbol{\Sigma}}_j$ in (2) is given by

$$\hat{\boldsymbol{\Sigma}}_j = \begin{pmatrix} \boldsymbol{S}_{jj} & \boldsymbol{S}_{j\bar{j}} & \boldsymbol{S}_{jy} \\ \boldsymbol{S}_{j\bar{j}}' & \boldsymbol{S}_{\bar{j}\bar{j}} & \boldsymbol{S}_{j\bar{j}}'\boldsymbol{S}_{jj}^{-1}\boldsymbol{S}_{jy} \\ \boldsymbol{S}_{jy}' & \boldsymbol{S}_{jy}'\boldsymbol{S}_{jj}^{-1}\boldsymbol{S}_{j\bar{j}} & \boldsymbol{S}_{yy} \end{pmatrix}. \tag{4}$$

Choosing the model by minimization of an information criterion is one of the primary selection methods. The most famous information criterion is Akaike's information criterion (AIC), which was proposed by Akaike (1973, 1974). Fujikoshi (1985) identified that the selection of the redundancy model in CCA is the selection of the covariance structure, and proposed using the AIC to select these structure for CCA. Many other information criteria have been proposed for CCA (see, e.g., Fujikoshi, 1985; Fujikoshi *et al.*, 2008; Hashiyama *et al.* 2011). The AIC is included in the family of log-likelihood-based information criteria (LLBICs); these are defined by adding a penalty term that expresses the complexity of the model for a negative twofold maximum log-likelihood. The family of LLBICs includes the bias-corrected AIC (AIC_c) proposed by Fujikoshi (1985), 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). The LLBIC for CCA is written as

$$\begin{aligned}\text{IC}_m(j) &= F(\boldsymbol{S}, \hat{\boldsymbol{\Sigma}}_j) + m(j) \\ &= (n-1)\log\frac{|\boldsymbol{S}_{yy\cdot j}|}{|\boldsymbol{S}_{yy\cdot x}|} + m(j),\end{aligned} \tag{5}$$

where $\boldsymbol{S}_{yy\cdot\ell} = \boldsymbol{S}_{yy} - \boldsymbol{S}'_{\ell y}\boldsymbol{S}_{\ell\ell}^{-1}\boldsymbol{S}_{\ell y}$ ($\ell = j, x$) and $m(j)$ is a positive penalty term that expresses the complexity of the model (1). The relations between LLBIC and most well-known information criteria are as follows:

$$\begin{aligned}\text{AIC} : m(j) &= p^2 + q^2 + p + q + 2pq_j, \\ \text{AIC}_\text{c} : m(j) &= (n-1)^2\left(\frac{p+q_j}{n-p-q_j-2} + \frac{q}{n-q-2} - \frac{q_j}{n-q_j-2} - \frac{p+q}{n-1}\right), \\ \text{BIC} : m(j) &= \left\{\frac{(p+q)(p+q+1)}{2} - p(q-q_j)\right\}\log n, \\ \text{CAIC} : m(j) &= \left\{\frac{(p+q)(p+q+1)}{2} - p(q-q_j)\right\}(1+\log n), \\ \text{HQC} : m(j) &= 2\left\{\frac{(p+q)(p+q+1)}{2} - p(q-q_j)\right\}\log\log n.\end{aligned} \tag{6}$$

When the asymptotic probability of an information criteria selecting the true model approaches 1, it is said to be *consistent*; this is one of its most important properties. In model selections, the true model is the candidate model with the set of true variables. The set of true variables is the smallest subset of variables which satisfies the condition in (1). In general, AIC is not consistent under the large-sample (LS) asymptotic framework in which only the sample size approaches ∞ (see e.g., Shibata, 1976; Nishii, 1984; Fujikoshi, 1982, 1985). When the AIC is used for model selection, its lack of consistency sometimes becomes a target for criticism, even though its purpose is not necessary to choose the true model.

Recently, the consistencies of various information criteria have been reported for multivariate models under a high-dimensional (HD) asymptotic framework. A HD asymptotic framework is one in which the sample size and dimension p simultaneously approach ∞ under the condition that $c_{n,p} = p/n \to c_0 \in (0, 1]$ (for simplicity, we will write this as "$c_{n,p} \to c_0$"). Yanagihara *et al.* (2012) derived the conditions for consistency of the LLBIC for model selection in a multivariate linear regression model under the HD asymptotic framework, and they found that the AIC meets these conditions. Since, by definition, HD data have a large dimension p, evaluating the consistency of an information criterion under the HD asymptotic framework is more natural for HD data than evaluating it under the LS asymptotic framework.

The purpose of this paper is to clarify the conditions under which the LLBIC is consistent for model selection in CCA when the HD asymptotic framework is used. In previous works, many results were obtained under the assumption that the true distribution of the observation vector was the normal distribution (e.g., Shibata, 1976; Nishii, 1984; Yanagihara *et al.*, 2014; Fujikoshi *et al.*, 2014). However, we are not able to determine whether this assumption is actually correct. Hence, a natural assumption for the generating mechanism of the true model of $\boldsymbol{y}$ is

$$\boldsymbol{y} = \boldsymbol{\mu}_y + \boldsymbol{\Sigma}'_{j_*y}\boldsymbol{\Sigma}^{-1}_{j_*j_*}(\boldsymbol{x}_{j_*} - \boldsymbol{\mu}_{j_*}) + \boldsymbol{\Sigma}^{1/2}_{yy\cdot j_*}\boldsymbol{\varepsilon}, \tag{7}$$

where $\boldsymbol{\varepsilon}$ is a p-dimensional vector with $E[\boldsymbol{\varepsilon}] = \boldsymbol{0}_p$, $Cov[\boldsymbol{\varepsilon}] = \boldsymbol{I}_p$, $\boldsymbol{0}_p$ is a p-dimensional vector of zeros, $\boldsymbol{x}_{j_*}$ is a q_{j_*}-dimensional vector with $E[\boldsymbol{x}_{j_*}] = \boldsymbol{\mu}_{j_*}$, $Cov[\boldsymbol{x}_{j_*}] = \boldsymbol{\Sigma}_{j_*j_*}$ and j_* denotes the set of the true variables.

In deriving the conditions for consistency under the HD asymptotic framework, a primary problem is to prove the convergence in probability of the two log-determinants of estimators of $\boldsymbol{\Sigma}$, because the size of the matrix increases with an increase in the dimensions. Yanagihara *et al.* (2012, 2014) avoided this problem by using a property of a random matrix distributed according to the Wishart distribution (see Fujikoshi *et al.*, 2010, chap. 3.2.4, p. 57). In the present study, this method is unavailable, because the true distribution of the observations in (7) is nonnormal.

Yanagihara (2013) derived the conditions under the LLBIC is consistent in multivariate linear regression models with the assumption of a normal distribution when the HD asymptotic framework is used, even though the distribution on the true model is not normal. In Yanagihara (2013), the moments of a specific random matrix and the distribution of the maximum eigenvalue of the estimator of the covariance matrix were used for assessing consistency. In CCA, it is important to note that $\boldsymbol{x}$ is a random vector, which is different in the case of a multivariate linear regression model. Hence, the conditions for consistency in this study are derived under the assumption that $\boldsymbol{x}$ is a random vector.

This paper is organized as follows: In Section 2, we present the necessary notations and assumptions, and then we obtain sufficient conditions to ensure consistency under the HD asymptotic framework. In Section 3, we verify our claim by conducting numerical experiments. In Section 4, we discuss our conclusions. Technical details are provided in the Appendix.

2 Main result

In this section, we show the sufficient conditions for consistency of IC_m in (5). First, we present the necessary notations and assumptions for assessing the consistency of an information criterion for the model M_j in (1). Let $\boldsymbol{y}_1, \dots, \boldsymbol{y}_n$, $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ and $\boldsymbol{\varepsilon}_1, \dots, \boldsymbol{\varepsilon}_n$ be n independent vectors from $\boldsymbol{y}$, $\boldsymbol{x}$ and $\boldsymbol{\varepsilon}$, respectively. Then, the $\boldsymbol{Y}$, $\boldsymbol{X}$ and $\boldsymbol{\mathcal{E}}$ are the $n \times p$, $n \times q$ and $n \times p$ matrices given by

$$\begin{aligned}\boldsymbol{Y} &= (\boldsymbol{I}_n - \boldsymbol{J}_n)(\boldsymbol{y}_1, \dots, \boldsymbol{y}_n)', \\ \boldsymbol{X} &= (\boldsymbol{I}_n - \boldsymbol{J}_n)(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)', \\ \boldsymbol{\mathcal{E}} &= (\boldsymbol{I}_n - \boldsymbol{J}_n)(\varepsilon_1, \dots, \varepsilon_n)',\end{aligned}$$

where $\boldsymbol{J}_n = \mathbf{1}_n(\mathbf{1}_n'\mathbf{1}_n)^{-1}\mathbf{1}_n'$ and $\mathbf{1}_n$ is an n-dimensional vector of ones. Suppose that $\boldsymbol{X}_j$ denotes the $n \times q_j$ matrix consisting of the columns of $\boldsymbol{X}$ indexed by the elements of j. By using these matrices, the matrix form of the true model (7) is expressed as

$$\boldsymbol{Y} = \boldsymbol{X}_{j_*}\boldsymbol{\Sigma}_{j_*j_*}^{-1}\boldsymbol{\Sigma}_{j_*y} + \boldsymbol{\mathcal{E}}\boldsymbol{\Sigma}_{yy\cdot j_*}^{1/2}. \tag{8}$$

Henceforth, for simplicity, $\boldsymbol{X}_{j_*}$ and q_{j_*} are represented as $\boldsymbol{X}_*$ and q_*, respectively. From the above expression, it can be seen that we can regard the true model (8) as a multivariate linear model by considering the conditional distribution of $\boldsymbol{Y}$ given $\boldsymbol{X}$.

We now describe two classes of j that express subsets of $\boldsymbol{X}$ in the candidate model. Let $\mathcal{J}$ be the set of K candidate models denoted by $\mathcal{J} = \{j_1, \ldots, j_K\}$. We then separate $\mathcal{J}$ into two sets: the overspecified models, in which the set of variables contain all variables of the true model j_* in (8), that is, $\mathcal{J}_+ = \{j \in \mathcal{J} | j_* \subseteq j\}$ and the underspecified models, which are the models that are not overspecified model, that is, $\mathcal{J}_- = \bar{\mathcal{J}}_+ \cap \mathcal{J}$. In particular, we express the minimum overspecified model that includes $j \in \mathcal{J}_-$ as j_+, and so

$$j_+ = j \cup j_*. \tag{9}$$

By using IC_m in (5), the best subset of ω, which is chosen by minimizing IC_m, 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

$$\boldsymbol{\Gamma}_j\boldsymbol{\Gamma}_j' = \boldsymbol{\Sigma}_{yy\cdot j_*}^{-1/2}\boldsymbol{\Sigma}_{j_*y}'\boldsymbol{\Sigma}_{j_*j_*}^{-1}\boldsymbol{X}_*'(\boldsymbol{I}_n - \boldsymbol{P}_j)\boldsymbol{X}_*\boldsymbol{\Sigma}_{j_*j_*}^{-1}\boldsymbol{\Sigma}_{j_*y}\boldsymbol{\Sigma}_{yy\cdot j_*}^{-1/2}, \tag{10}$$

where $\boldsymbol{\Gamma}_j$ is a $p \times \gamma_j$ matrix with $\mathrm{rank}(\boldsymbol{\Gamma}_j) = \gamma_j$ and $\boldsymbol{P}_j = \boldsymbol{X}_j(\boldsymbol{X}_j'\boldsymbol{X}_j)^{-1}\boldsymbol{X}_j'$. 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, for $j \in \mathcal{J}_-$, we define

$$\boldsymbol{A}_j = (\boldsymbol{I}_n - \boldsymbol{P}_j)\boldsymbol{X}_*\boldsymbol{\Sigma}_{j_*j_*}^{-1}\boldsymbol{\Sigma}_{j_*y}\boldsymbol{\Sigma}_{yy\cdot j_*}^{-1/2}.$$

It is easy to see from the definition of the noncentrality matrix in (10) that $\boldsymbol{A}_j'\boldsymbol{A}_j = \boldsymbol{\Gamma}_j\boldsymbol{\Gamma}_j'$. By using a singular value decomposition, $\boldsymbol{A}_j$ can be rewritten as

$$\boldsymbol{A}_j = \boldsymbol{H}_j\boldsymbol{L}_j^{1/2}\boldsymbol{G}_j', \tag{11}$$

where $\boldsymbol{H}_j = (\boldsymbol{h}_{j,1}, \ldots, \boldsymbol{h}_{j,\gamma_j})$ and $\boldsymbol{G}_j = (\boldsymbol{g}_{j,1}, \ldots, \boldsymbol{g}_{j,\gamma_j})$ are $n \times \gamma_j$ and $\gamma_j \times \gamma_j$ matrices, that satisfy $\boldsymbol{H}_j'\boldsymbol{H}_j = \boldsymbol{I}_{\gamma_j}$ and $\boldsymbol{G}_j'\boldsymbol{G}_j = \boldsymbol{I}_{\gamma_j}$, respectively, and $\boldsymbol{L}_j = \mathrm{diag}(\alpha_{j,1}, \ldots, \alpha_{j,\gamma_j})$ is a diagonal matrix of order γ_j whose diagonal elements $\alpha_{j,k}$ are the squared singular values of $\boldsymbol{A}_j$, which are assumed to be $\alpha_{j,1} \geq \cdots \geq \alpha_{j,\gamma_j}$.

Furthermore, let $||\boldsymbol{a}||$ denote the Euclidean norm of the vector $\boldsymbol{a}$. Then, in order to assess the consistency of IC_m, the following assumption are necessary:

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

A2. $E[||\boldsymbol{\varepsilon}||^4]$ exists and has the order $O(p^2)$ as $p \to \infty$.

A3. $E[||\boldsymbol{x}||^4]$ exists.

A4. $^\forall j \in \mathcal{J}_-, \lim_{p\to\infty} p^{-1}\boldsymbol{\Sigma}_{j_*y}\boldsymbol{\Sigma}_{yy\cdot j_*}^{-1}\boldsymbol{\Sigma}_{j_*y}' = \boldsymbol{\Psi}_j$ exists and

$$\mathrm{tr}(\boldsymbol{\Sigma}_{j_*}^{-1}\boldsymbol{\Sigma}_{j_*j_*\cdot j}\boldsymbol{\Sigma}_{j_*}^{-1}\boldsymbol{\Psi}_j) > 0.$$

A1 is the basic assumption for evaluating the consistency of an information criterion, because the probability of selecting the true model becomes 0 if it does not hold. A2 and A3 are assumptions about the moments of the distribution of the true model, although $\boldsymbol{\varepsilon}$ and $\boldsymbol{x}$ are not assumed to represent a specific distribution. It is easy to see that A2 holds if $\max_{a=1,\ldots,p} E[\varepsilon_a^4]$ is bounded. A4 is used in assessing the noncentrality matrix. In the multivariate linear regression model, $\boldsymbol{X}_j$ in $\boldsymbol{\Gamma}_j\boldsymbol{\Gamma}_j'$ is not random. However in CCA, $\boldsymbol{X}_j$ in $\boldsymbol{\Gamma}_j\boldsymbol{\Gamma}_j'$ is random. Hence, a different assumption from the multivariate linear regression model is required in A4. If A2 is satisfied, the multivariate kurtosis proposed by Mardia (1970) exists as

$$\kappa_4^{(1)} = E[||\boldsymbol{\varepsilon}||^4] - p(p+2) = \sum_{a,b}^{p} \kappa_{aabb} + p(p+2), \tag{12}$$

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 as

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

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

$$\kappa_4^{(1)} = O(p^s) \text{ as } p \to \infty,\ \ s \in [0, 2]. \tag{13}$$

By using these notations and assumptions, we derived the following theorem for the sufficiency conditions for the consistency of the penalty term $m(j)$ (the proof was given in the Appendix A2).

Theorem 1 *Suppose that assumptions A1-A4 hold. Variable selection using IC_m is consistent when $c_{n,p} \to c_0$ if the following conditions are satisfied simultaneously:*

(C1) $^\forall j \in \mathcal{J}_+ \setminus \{j_*\},\ \lim_{c_{n,p}\to c_0}\{m(j) - m(j_*)\}/p > -c_0^{-1}(q_j - q_*)\log(1 - c_0)$.

(C2) $^\forall j \in \mathcal{J}_-,\ \lim_{c_{n,p}\to c_0}\{m(j) - m(j_*)\}/(n\log p) > -1/2$.

We can see from Theorem 1 that the conditions for consistency are similar to those in the multivariate regression model derived by Yanagihara and colleagues (Yanagihara *et al.*, 2012; Yanagihara, 2013). This is because the CCA can be regarded as an extension of the multivariate regression model. Futhermore, the conditions for consistency in Theorem 1 is also similar to those in Yanagihara *et al.* (2014), which is derived for a CCA when a normal distribution is assumed to the true model. This indicates that the conditions for consistency are free of the influence of nonnormality in the distribution of the true model.

Using Theorem 1, the conditions for consistency of specific criteria can be clarified by the following corollary (the proof is given in the Appendix A3):

Corollary 1 *Suppose that assumptions A1-A4 are satisfied. Then we have*

1. *A model selection using the AIC is consistent when $c_{n,p} \to c_0$ if $c_0 \in (0, c_a]$ holds, where $c_a (\approx 0.797)$ is a constant satisfying*

$$\log(1 - c_a) + 2c_a = 0. \tag{14}$$

2. *Model selections using the $\mathrm{AIC_c}$ and HQC are consistent when $c_{n,p} \to c_0$.*

3. *Model selections using the BIC and CAIC are consistent when $c_{n,p} \to c_0$ if $c_0 \in (0, c_b/2]$ holds, where $c_b = \min\{1, \min_{j \in \mathcal{F}_-} 1/\{2(q_* - q_j)\}\}$ and $\mathcal{F}_-$ is a set of candidate models given by*

$$\mathcal{F}_- = \{j \in \mathcal{J} | q_* - q_j > 0\}. \tag{15}$$

Corollary 1 shows that, when $c_{n,p} \to c_0$, the $\mathrm{AIC_c}$ and HQC are always consistent in model selection, whereas the AIC, BIC, and CAIC are not always consistent. The consistency of the BIC and CAIC is strongly dependent on values of parameters in the true model, but this is not true for the AIC. This sets the BIC and CAIC at a great disadvantage compared to the AIC, because the real values of parameters in the true model is unknowable. Table 1 lists the conditions required for consistency for each of the following criteria: AIC, $\mathrm{AIC_c}$, BIC, CAIC, and HQC.

TABLE 1: Conditions for consistency

Criterion	Consistency	Conditions
AIC	Conditionally holds	$c_0 \in [0, c_a)$
$\mathrm{AIC_c}$ & HQC	Holds	- - - -
BIC & CAIC	Conditionally holds	$c_0 \in [0, c_b)$

Note) c_a and c_b are given in COROLLARY 1.

3 Numerical Study

In this section, we conduct numerical studies to examine the validity of our claim. The probabilities of selecting the true model by the AIC, $\mathrm{AIC_c}$, BIC, CAIC, and HQC were evaluated by Monte Carlo simulations with 10,000 iterations each.

Let $\boldsymbol{\nu}_1 = (\nu_{1,1}, \dots, \nu_{1,p})' \sim N_p(\mathbf{0}_p, \boldsymbol{I}_p)$, $\boldsymbol{\nu}_2 = (\nu_{2,1}, \dots, \nu_{2,q})' \sim N_q(\mathbf{0}_q, \boldsymbol{I}_q)$, $\delta_1, \delta_2 \sim \chi^2_6$, $\omega_{1,1}, \dots, \omega_{2,p} \sim i.i.d. \chi^2_5$ and $\omega_{2,1}, \dots, \omega_{2,q} \sim i.i.d. \chi^2_5$ be mutually independent random vectors and variables. Then, $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_p)'$ and $\boldsymbol{x} = (x_1, \dots, x_q)'$ were generated from the following five distributions, as in Yanagihara (2013):

- Distribution 1 (the multivariate normal distribution).

$$\boldsymbol{\varepsilon} = \boldsymbol{\nu}_1, \ \boldsymbol{x} = \boldsymbol{\nu}_2.$$

- Distribution 2 (a scale mixture of the multivariate normal distribution).

$$\boldsymbol{\varepsilon} = \sqrt{\frac{\delta_1}{6}}\boldsymbol{\nu}_1, \ \boldsymbol{x} = \sqrt{\frac{\delta_2}{6}}\boldsymbol{\nu}_2.$$

- Distribution 3 (a location-scale mixture of the multivariate normal distribution).

$$\boldsymbol{\varepsilon} = \boldsymbol{B}_1^{-1/2}\left\{10\left(\sqrt{\frac{\delta_1}{6}} - \eta\right)\mathbf{1}_p + \sqrt{\frac{\delta_1}{6}}\boldsymbol{\nu}_1\right\},$$
$$\boldsymbol{x} = \boldsymbol{B}_2^{-1/2}\left\{10\left(\sqrt{\frac{\delta_2}{6}} - \eta\right)\mathbf{1}_q + \sqrt{\frac{\delta_2}{6}}\boldsymbol{\nu}_2\right\},$$

where $\eta = 15\sqrt{\pi/3}/16$, $\boldsymbol{B}_1 = \boldsymbol{I}_p + 100(1-\eta^2)\mathbf{1}_p\mathbf{1}_p'$, and $\boldsymbol{B}_2 = \boldsymbol{I}_q + 100(1-\eta^2)\mathbf{1}_q\mathbf{1}_q'$.

- Distribution 4 (the independent t-distribution).

$$\varepsilon_a = \frac{\sqrt{3}\nu_{1,a}}{\sqrt{5\omega_{1,a}}}, \ x_a = \frac{\sqrt{3}\nu_{2,a}}{\sqrt{5\omega_{2,a}}}.$$

- Distribution 5 (the independent log-normal distribution).

$$\varepsilon_a = \frac{\log \nu_{1,a} - \sqrt{e}}{\sqrt{e(e-1)}}, \ x_a = \frac{\log \nu_{2,a} - \sqrt{e}}{\sqrt{e(e-1)}}.$$

It is easy to see that distributions 1, 2, and 4 are symmetric, and distributions 3 and 5 are skewed.

The mean vectors $\boldsymbol{\mu}_y$ and $\boldsymbol{\mu}_{j_*}$ were generated from $U(-4, 4)$ and $U(-3, 3)$, respectively, and $j_* = 3$. Then, $\boldsymbol{y}$ was obtained from the true model (7). The structure of $\boldsymbol{\Sigma}$ was prepared for the following four cases (cases 1 and 2 are the same settings as in Fujikoshi, 2014):

Case 1.

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{I}_5 & \boldsymbol{R}' \\ \boldsymbol{R} & \boldsymbol{I}_p \end{pmatrix}, \ \boldsymbol{R} = (\boldsymbol{R}_1, \boldsymbol{O}_{5,p-q})', \ \boldsymbol{R}_1 = \mathrm{diag}(\rho_1, \ldots, \rho_5),$$

$$\rho_1 = 2\rho, \ \rho_2 = 3\rho/2, \ \rho_3 = \rho, \ \rho_4 = \rho_5 = 0, \ \rho = \sqrt{\frac{(4p/21)}{p+1+(4p/21)}}.$$

Case 2 (the structure of $\boldsymbol{\Sigma}$ is the same as in Case 1).

$$\rho_1 = \tilde{\rho}, \ \rho_2 = 3\tilde{\rho}/4, \ \rho_3 = \tilde{\rho}/2, \ \rho_4 = \rho_5 = 0, \ \tilde{\rho} = \sqrt{\frac{p}{p+1}}\sqrt{\frac{(4p/21)}{1+(4p/21)}}.$$

Case 3. $\boldsymbol{\Sigma} = \boldsymbol{\Phi}\boldsymbol{\Phi}'$, where $\boldsymbol{\Phi}$ is a $(p+5) \times (p+5)$ matrix whose elements are distributed from $U(0, 1/p+5)$.

Case 4. $\boldsymbol{\Sigma} = \boldsymbol{\Phi}\boldsymbol{\Phi}'$, where$\boldsymbol{\Phi}$ is a $(p+8) \times (p+8)$ matrix whose elements are distributed from $U(0, 1/p+8)$.

In these settings, data are generated under the following combinations of n and p:

- $c_0 = 0.05$: $(n, p) = (100, 5), (200, 10), (500, 25), (1000, 50)$.
- $c_0 = 0.1$: $(n, p) = (100, 10), (200, 20), (500, 50), (1000, 100)$.
- $c_0 = 0.2$: $(n, p) = (100, 20), (200, 40), (500, 100), (1000, 200)$.
- $c_0 = 0.3$: $(n, p) = (100, 30), (200, 60), (500, 150), (1000, 300)$.

Tables 2 through 6 show the selection probability (i.e., the probability of selecting the true model) when $\boldsymbol{\varepsilon}$ and $\boldsymbol{x}$ are from Distributions 1, 2, 3, 4, and 5, respectively, when using the AIC, the $\mathrm{AIC_c}$, the BIC, the CAIC, and the HQC. From these tables, we can see that the selection probability of the AIC tends to increase in most settings when p and n were large. The $\mathrm{AIC_c}$ and HQC had the same tendency as that of the AIC, that is, when n and p were large, their selection probabilities tended to increase. On the other hand, the selection probabilities of the BIC and CAIC decreased for larger values of n and p. Moreover, it was worth noting that the selection probabilities of the BIC and CAIC depend on the distribution settings, this may be because the conditions for consistency of the BIC and CAIC have a strong dependence on the values of parameters in the true model. We repeated the simulations for several models and obtained similar results, and these validated our claim.

Table 2. Selection probabilities of the true model (%) in the Case of Distribution 1

$c_0 = 0.05$ Case 1 / Case 2

n	p	AIC	AIC$_\mathrm{c}$	BIC	CAIC	HQC	AIC	AIC$_\mathrm{c}$	BIC	CAIC	HQC
100	5	80.01	79.24	31.31	15.29	67.22	62.36	56.11	8.54	2.48	37.42
200	10	94.55	95.03	17.95	4.88	76.07	93.47	92.95	12.51	2.98	68.61
500	25	99.58	99.88	1.18	0.06	83.03	99.66	99.93	12.86	1.24	97.99
1000	50	99.99	100.00	0.00	0.00	85.92	100.00	100.00	6.25	0.13	99.99

$c_0 = 0.05$ Case 3 / Case 4

n	p	AIC	AIC$_\mathrm{c}$	BIC	CAIC	HQC	AIC	AIC$_\mathrm{c}$	BIC	CAIC	HQC
100	5	88.12	94.09	80.83	63.32	94.94	85.77	92.28	68.05	47.49	90.37
200	10	96.08	98.70	96.04	84.35	99.82	95.67	98.70	86.36	64.22	99.62
500	25	99.68	99.92	99.99	98.41	100.00	99.61	99.88	99.12	89.53	100.00
1000	50	100.00	100.00	100.00	100.00	100.00	99.97	100.00	100.00	99.59	100.00

$c_0 = 0.1$ Case 1 / Case 2

n	p	AIC	AIC$_\mathrm{c}$	BIC	CAIC	HQC	AIC	AIC$_\mathrm{c}$	BIC	CAIC	HQC
100	10	70.89	49.01	2.14	0.15	31.16	65.76	42.52	1.24	0.10	24.92
200	20	86.25	62.95	0.01	0.00	17.14	93.81	78.96	0.22	0.01	32.36
500	50	97.74	81.43	0.00	0.00	2.19	100.00	99.43	0.00	0.00	36.62
1000	100	99.76	92.53	0.00	0.00	0.03	100.00	100.00	0.00	0.00	30.78

$c_0 = 0.1$ Case 3 / Case 4

n	p	AIC	AIC$_\mathrm{c}$	BIC	CAIC	HQC	AIC	AIC$_\mathrm{c}$	BIC	CAIC	HQC
100	10	93.28	95.23	41.77	13.54	88.98	91.66	89.10	24.65	5.28	79.32
200	20	98.98	99.88	40.28	7.35	98.35	99.03	99.62	17.78	1.30	94.04
500	50	99.98	100.00	32.00	1.57	100.00	100.00	100.00	9.86	0.01	99.97
1000	100	100.00	100.00	27.28	0.14	100.00	100.00	100.00	4.61	0.00	100.00

$c_0 = 0.2$ Case 1 / Case 2

n	p	AIC	AIC$_\mathrm{c}$	BIC	CAIC	HQC	AIC	AIC$_\mathrm{c}$	BIC	CAIC	HQC
100	20	43.70	2.00	0.00	0.00	2.94	54.98	4.17	0.01	0.00	5.62
200	40	46.18	0.70	0.00	0.00	0.02	76.68	6.28	0.00	0.00	1.21
500	100	46.50	0.05	0.00	0.00	0.00	96.04	6.35	0.00	0.00	0.00
1000	200	45.68	0.00	0.00	0.00	0.00	99.69	4.13	0.00	0.00	0.00

$c_0 = 0.2$ Case 3 / Case 4

n	p	AIC	AIC$_\mathrm{c}$	BIC	CAIC	HQC	AIC	AIC$_\mathrm{c}$	BIC	CAIC	HQC
100	20	94.18	49.12	0.85	0.00	53.71	90.08	30.18	0.12	0.00	35.98
200	40	99.76	83.58	0.00	0.00	57.81	99.52	67.62	0.00	0.00	37.10
500	100	100.00	99.96	0.00	0.00	78.03	100.00	99.49	0.00	0.00	52.33
1000	200	100.00	100.00	0.00	0.00	99.81	100.00	100.00	0.00	0.00	97.96

$c_0 = 0.3$ Case 1 / Case 2

n	p	AIC	AIC$_\mathrm{c}$	BIC	CAIC	HQC	AIC	AIC$_\mathrm{c}$	BIC	CAIC	HQC
100	30	27.92	0.00	0.00	0.00	0.10	43.50	0.00	0.00	0.00	0.97
200	60	21.75	0.00	0.00	0.00	0.00	54.80	0.00	0.00	0.00	0.02
500	150	11.36	0.00	0.00	0.00	0.00	68.94	0.00	0.00	0.00	0.00
1000	300	4.13	0.00	0.00	0.00	0.00	80.42	0.00	0.00	0.00	0.00

$c_0 = 0.3$ Case 3 / Case 4

n	p	AIC	AIC$_\mathrm{c}$	BIC	CAIC	HQC	AIC	AIC$_\mathrm{c}$	BIC	CAIC	HQC
100	30	89.60	0.29	0.00	0.00	17.18	85.65	0.07	0.00	0.00	9.41
200	60	99.34	1.09	0.00	0.00	8.98	98.66	0.13	0.00	0.00	3.17
500	150	100.00	11.88	0.00	0.00	5.06	100.00	3.14	0.00	0.00	0.74
1000	300	100.00	97.41	0.00	0.00	50.09	100.00	93.84	0.00	0.00	33.20

Table 3. Selection probabilities of the true model (%) in the Case of Distribution 2

$c_0 = 0.05$ Case 1 / Case 2

n	p	AIC	AIC$_c$	BIC	CAIC	HQC	AIC	AIC$_c$	BIC	CAIC	HQC
100	5	70.88	70.46	41.68	30.97	62.29	59.53	55.40	23.14	15.84	43.97
200	10	83.11	81.24	36.61	27.35	64.22	81.02	78.56	33.51	24.30	60.22
500	25	90.40	87.54	29.00	21.80	62.23	94.25	92.27	39.50	30.74	72.06
1000	50	92.04	89.24	23.62	17.80	59.04	96.50	95.29	40.56	32.51	75.13

$c_0 = 0.05$ Case 3 / Case 4

n	p	AIC	AIC$_c$	BIC	CAIC	HQC	AIC	AIC$_c$	BIC	CAIC	HQC
100	5	82.10	85.17	66.30	56.92	81.89	78.86	81.42	60.29	49.99	77.19
200	10	93.23	94.36	71.94	63.35	88.57	92.09	92.88	65.70	55.36	85.87
500	25	98.62	98.46	75.12	67.20	92.92	98.03	97.71	68.73	60.07	89.75
1000	50	99.49	99.30	76.25	68.69	94.38	99.34	99.04	71.52	64.30	93.01

$c_0 = 0.1$ Case 1 / Case 2

n	p	AIC	AIC$_c$	BIC	CAIC	HQC	AIC	AIC$_c$	BIC	CAIC	HQC
100	10	64.80	51.08	17.11	10.27	40.08	61.25	47.16	15.50	9.31	36.22
200	20	72.02	58.14	12.12	6.83	36.35	77.47	64.79	16.47	10.28	42.66
500	50	75.68	61.46	7.04	4.19	29.85	86.82	76.86	15.27	10.13	46.72
1000	100	76.70	63.06	5.69	3.49	26.86	89.08	80.67	13.46	8.82	46.74

$c_0 = 0.1$ Case 3 / Case 4

n	p	AIC	AIC$_c$	BIC	CAIC	HQC	AIC	AIC$_c$	BIC	CAIC	HQC
100	10	83.68	79.58	47.70	35.06	72.52	80.40	73.00	39.56	27.40	65.88
200	20	93.50	89.31	46.82	34.92	76.79	91.20	85.29	40.31	29.12	71.40
500	50	97.27	94.75	46.91	36.65	80.37	96.54	93.27	41.18	31.03	76.67
1000	100	98.00	96.28	47.83	38.02	83.41	98.04	95.66	41.87	32.54	80.15

$c_0 = 0.2$ Case 1 / Case 2

n	p	AIC	AIC$_c$	BIC	CAIC	HQC	AIC	AIC$_c$	BIC	CAIC	HQC
100	20	47.43	14.72	3.47	1.45	16.31	55.08	20.38	5.48	2.42	22.46
200	40	50.05	17.70	1.64	0.77	12.28	64.78	29.86	4.12	1.81	21.77
500	100	49.43	18.32	0.80	0.42	7.86	69.83	35.59	2.71	1.34	18.57
1000	200	49.49	18.56	0.42	0.22	6.33	71.86	38.38	1.71	0.83	16.91

$c_0 = 0.2$ Case 3 / Case 4

n	p	AIC	AIC$_c$	BIC	CAIC	HQC	AIC	AIC$_c$	BIC	CAIC	HQC
100	20	79.91	50.57	20.96	10.53	52.44	75.03	42.11	15.57	7.22	44.63
200	40	87.60	62.47	17.41	8.76	52.42	84.93	56.40	13.64	6.58	46.10
500	100	93.24	75.42	15.87	8.48	56.02	91.56	70.49	12.39	6.55	50.00
1000	200	96.31	83.78	17.75	10.84	63.24	95.63	81.52	15.48	8.92	60.03

$c_0 = 0.3$ Case 1 / Case 2

n	p	AIC	AIC$_c$	BIC	CAIC	HQC	AIC	AIC$_c$	BIC	CAIC	HQC
100	30	37.47	2.05	0.85	0.24	7.48	48.51	4.57	1.96	0.61	13.55
200	60	36.78	2.81	0.43	0.17	4.76	54.14	6.72	1.17	0.37	10.51
500	150	34.22	2.75	0.13	0.05	2.43	57.06	8.62	0.52	0.17	7.66
1000	300	34.70	2.99	0.04	0.02	1.80	56.92	9.41	0.17	0.06	5.96

$c_0 = 0.3$ Case 3 / Case 4

n	p	AIC	AIC$_c$	BIC	CAIC	HQC	AIC	AIC$_c$	BIC	CAIC	HQC
100	30	73.63	16.62	7.94	2.32	35.67	70.68	13.42	6.12	1.74	31.38
200	60	82.74	27.83	6.00	2.17	36.05	78.83	23.01	4.81	1.73	30.72
500	150	89.72	41.14	4.98	2.02	38.36	88.05	38.43	4.29	1.76	35.95
1000	300	95.06	59.40	7.02	3.09	50.27	94.51	57.81	5.91	2.89	48.35

Table 4. Selection probabilities of the true model (%) in the Case of Distribution 3

$c_0 = 0.05$ Case 1 / Case 2

n	p	AIC	AIC$_c$	BIC	CAIC	HQC	AIC	AIC$_c$	BIC	CAIC	HQC
100	5	88.85	94.65	94.12	90.68	96.36	87.31	92.30	85.31	79.04	92.31
200	10	95.66	97.94	94.29	91.23	98.37	95.57	97.81	93.27	89.64	97.97
500	25	99.50	99.67	93.22	90.07	98.52	99.60	99.80	96.22	94.04	99.22
1000	50	99.86	99.87	91.82	88.26	98.46	99.91	99.92	96.38	94.76	99.50

$c_0 = 0.05$ Case 3 / Case 4

n	p	AIC	AIC$_c$	BIC	CAIC	HQC	AIC	AIC$_c$	BIC	CAIC	HQC
100	5	88.98	95.67	99.80	99.78	98.83	87.48	96.53	99.85	99.76	98.61
200	10	95.95	98.56	100.00	100.00	99.93	95.66	98.51	100.00	99.99	99.94
500	25	99.68	99.96	100.00	100.00	100.00	99.63	99.88	100.00	99.99	100.00
1000	50	99.98	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00

$c_0 = 0.1$ Case 1 / Case 2

n	p	AIC	AIC$_c$	BIC	CAIC	HQC	AIC	AIC$_c$	BIC	CAIC	HQC
100	10	92.07	95.45	82.79	73.59	93.32	91.70	94.84	80.12	70.24	92.08
200	20	97.86	97.44	78.98	69.86	93.67	98.16	98.38	84.30	76.80	95.65
500	50	99.28	98.71	73.05	64.11	93.57	99.70	99.37	85.48	78.68	97.58
1000	100	99.50	98.79	67.48	57.84	92.03	99.88	99.69	83.60	77.61	97.11

$c_0 = 0.1$ Case 3 / Case 4

n	p	AIC	AIC$_c$	BIC	CAIC	HQC	AIC	AIC$_c$	BIC	CAIC	HQC
100	10	94.17	99.37	99.96	99.82	99.79	94.12	99.53	99.95	99.89	99.84
200	20	98.79	99.96	99.99	99.99	100.00	98.87	99.89	99.99	99.99	100.00
500	50	99.99	100.00	100.00	100.00	100.00	100.00	100.00	100.00	99.99	100.00
1000	100	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00

$c_0 = 0.2$ Case 1 / Case 2

n	p	AIC	AIC$_c$	BIC	CAIC	HQC	AIC	AIC$_c$	BIC	CAIC	HQC
100	20	92.60	79.33	52.46	36.37	80.81	94.40	84.37	61.07	44.56	85.66
200	40	96.22	84.23	43.23	28.85	78.19	98.19	91.50	59.73	44.81	87.59
500	100	97.15	87.72	31.62	19.98	74.35	99.02	95.05	52.37	38.88	88.36
1000	200	97.43	87.97	23.28	14.81	70.63	99.24	95.81	44.88	32.47	86.65

$c_0 = 0.2$ Case 3 / Case 4

n	p	AIC	AIC$_c$	BIC	CAIC	HQC	AIC	AIC$_c$	BIC	CAIC	HQC
100	20	97.54	100.00	99.92	99.63	99.98	97.36	99.97	99.90	99.66	99.97
200	40	99.78	100.00	100.00	99.93	100.00	99.81	100.00	99.96	99.88	100.00
500	100	100.00	100.00	100.00	99.98	100.00	100.00	100.00	100.00	100.00	100.00
1000	200	100.00	100.00	100.00	99.99	100.00	100.00	100.00	100.00	99.99	100.00

$c_0 = 0.3$ Case 1 / Case 2

n	p	AIC	AIC$_c$	BIC	CAIC	HQC	AIC	AIC$_c$	BIC	CAIC	HQC
100	30	89.84	44.40	29.49	14.31	67.31	93.39	57.59	43.18	24.00	78.32
200	60	93.25	51.09	19.56	9.03	60.82	97.12	69.56	35.49	20.54	77.67
500	150	94.29	55.46	9.90	4.47	52.73	98.08	76.25	23.36	13.05	74.07
1000	300	94.62	55.92	5.34	2.13	46.42	98.34	77.47	16.54	8.23	69.95

$c_0 = 0.3$ Case 3 / Case 4

n	p	AIC	AIC$_c$	BIC	CAIC	HQC	AIC	AIC$_c$	BIC	CAIC	HQC
100	30	97.85	99.93	99.77	98.69	100.00	97.81	99.90	99.72	98.79	99.98
200	60	99.90	100.00	99.92	99.49	100.00	99.89	99.99	99.92	99.47	99.99
500	150	100.00	100.00	99.98	99.83	100.00	100.00	100.00	99.97	99.83	100.00
1000	300	100.00	100.00	99.99	99.93	100.00	100.00	100.00	99.99	99.96	100.00

Table 5. Selection probabilities of the true model (%) in the Case of Distribution 4

$c_0 = 0.05$ Case 1							Case 2				
n	p	AIC	$\mathrm{AIC_c}$	BIC	CAIC	HQC	AIC	$\mathrm{AIC_c}$	BIC	CAIC	HQC
100	5	78.20	77.81	32.68	16.90	66.21	62.22	55.45	9.93	3.10	38.18
200	10	92.82	93.13	20.00	6.68	74.06	91.95	91.39	15.13	4.56	68.45
500	25	99.54	99.71	2.64	0.28	80.76	99.55	99.87	17.48	3.16	96.38
1000	50	99.99	99.98	0.10	0.03	84.23	99.98	100.00	10.39	0.93	99.72
$c_0 = 0.05$ Case 3							Case 4				
n	p	AIC	$\mathrm{AIC_c}$	BIC	CAIC	HQC	AIC	$\mathrm{AIC_c}$	BIC	CAIC	HQC
100	5	87.88	93.62	79.14	62.57	94.41	86.54	92.39	68.39	48.09	90.36
200	10	95.49	98.14	94.70	82.68	99.87	95.16	98.45	85.14	64.50	99.45
500	25	99.63	99.90	99.89	98.06	100.00	99.68	99.94	98.36	87.81	100.00
1000	50	100.00	100.00	100.00	99.99	100.00	100.00	100.00	100.00	99.20	100.00
$c_0 = 0.1$ Case 1							Case 2				
n	p	AIC	$\mathrm{AIC_c}$	BIC	CAIC	HQC	AIC	$\mathrm{AIC_c}$	BIC	CAIC	HQC
100	10	69.39	48.63	3.18	0.34	32.22	65.19	43.49	2.12	0.24	27.32
200	20	84.31	62.58	0.12	0.02	19.98	91.63	77.31	0.62	0.04	34.25
500	50	96.47	79.43	0.02	0.00	3.66	99.85	98.55	0.06	0.02	38.67
1000	100	99.44	90.44	0.00	0.00	0.16	100.00	99.96	0.00	0.00	33.97
$c_0 = 0.1$ Case 3							Case 4				
n	p	AIC	$\mathrm{AIC_c}$	BIC	CAIC	HQC	AIC	$\mathrm{AIC_c}$	BIC	CAIC	HQC
100	10	92.95	94.50	43.66	16.67	88.52	91.42	88.14	28.99	7.68	79.38
200	20	98.68	99.81	41.77	10.51	97.82	98.82	99.54	20.37	2.26	93.68
500	50	99.98	100.00	34.09	3.09	100.00	99.98	100.00	12.52	0.23	99.94
1000	100	100.00	100.00	30.53	0.78	100.00	100.00	100.00	7.42	0.15	100.00
$c_0 = 0.2$ Case 1							Case 2				
n	p	AIC	$\mathrm{AIC_c}$	BIC	CAIC	HQC	AIC	$\mathrm{AIC_c}$	BIC	CAIC	HQC
100	20	44.40	2.83	0.01	0.00	3.79	55.29	6.02	0.03	0.00	7.41
200	40	46.23	1.21	0.00	0.00	0.17	74.94	9.11	0.00	0.00	2.40
500	100	46.74	0.21	0.00	0.00	0.00	93.21	8.66	0.00	0.00	0.09
1000	200	46.50	0.03	0.00	0.00	0.00	98.87	6.62	0.00	0.00	0.01
$c_0 = 0.2$ Case 3							Case 4				
n	p	AIC	$\mathrm{AIC_c}$	BIC	CAIC	HQC	AIC	$\mathrm{AIC_c}$	BIC	CAIC	HQC
100	20	93.63	50.57	1.16	0.01	54.88	89.19	32.12	0.34	0.00	38.31
200	40	99.61	82.54	0.01	0.00	57.02	99.44	66.86	0.01	0.00	38.29
500	100	100.00	99.93	0.00	0.00	77.16	100.00	99.37	0.00	0.00	52.97
1000	200	100.00	100.00	0.00	0.00	99.42	100.00	100.00	0.00	0.00	96.68
$c_0 = 0.3$ Case 1							Case 2				
n	p	AIC	$\mathrm{AIC_c}$	BIC	CAIC	HQC	AIC	$\mathrm{AIC_c}$	BIC	CAIC	HQC
100	30	28.13	0.01	0.00	0.00	0.24	44.62	0.02	0.00	0.00	1.77
200	60	23.64	0.00	0.00	0.00	0.00	54.96	0.00	0.00	0.00	0.08
500	150	13.48	0.00	0.00	0.00	0.00	67.55	0.01	0.00	0.00	0.01
1000	300	5.65	0.00	0.00	0.00	0.00	78.08	0.01	0.00	0.00	0.00
$c_0 = 0.3$ Case 3							Case 4				
n	p	AIC	$\mathrm{AIC_c}$	BIC	CAIC	HQC	AIC	$\mathrm{AIC_c}$	BIC	CAIC	HQC
100	30	88.89	0.41	0.01	0.00	19.33	84.12	0.08	0.00	98.79	99.98
200	60	99.23	1.56	0.00	0.00	11.25	98.24	0.50	0.00	99.47	99.99
500	150	100.00	13.89	0.00	0.00	6.66	100.00	4.92	0.00	99.83	100.00
1000	300	100.00	96.20	0.00	0.00	51.00	100.00	91.87	0.00	99.96	100.00

Table 6. Selection probabilities of the true model (%) in the Case of Distribution 5

$c_0 = 0.05$ Case 1							Case 2				
n	p	AIC	AIC$_c$	BIC	CAIC	HQC	AIC	AIC$_c$	BIC	CAIC	HQC
100	5	68.55	66.94	37.54	26.26	58.66	68.77	67.46	36.44	25.73	58.13
200	10	82.13	81.29	30.08	18.65	65.52	81.80	81.24	29.76	18.67	64.76
500	25	94.76	94.71	14.85	7.22	69.75	94.34	94.61	15.47	7.38	69.62
1000	50	98.55	98.62	5.12	1.94	70.50	98.48	98.41	5.21	1.88	70.23

$c_0 = 0.05$ Case 3							Case 4				
n	p	AIC	AIC$_c$	BIC	CAIC	HQC	AIC	AIC$_c$	BIC	CAIC	HQC
100	5	83.07	87.90	74.46	62.14	87.55	79.29	85.69	66.74	52.36	83.95
200	10	90.71	94.44	88.08	75.94	97.71	89.81	94.00	79.21	63.46	96.66
500	25	97.14	98.33	98.15	91.54	99.89	97.03	98.38	93.55	78.81	99.82
1000	50	99.31	99.61	99.87	98.01	99.97	99.07	99.52	99.20	93.10	99.95

$c_0 = 0.1$ Case 1							Case 2				
n	p	AIC	AIC$_c$	BIC	CAIC	HQC	AIC	AIC$_c$	BIC	CAIC	HQC
100	10	59.32	47.56	12.74	6.17	36.77	59.04	47.53	12.47	5.91	36.65
200	20	70.33	56.89	4.31	1.67	29.71	70.95	56.51	4.21	1.69	29.34
500	50	85.63	68.94	0.56	0.20	16.92	84.84	67.64	0.55	0.19	16.28
1000	100	93.21	77.33	0.08	0.02	7.46	93.14	76.98	0.06	0.03	7.27

$c_0 = 0.1$ Case 3							Case 4				
n	p	AIC	AIC$_c$	BIC	CAIC	HQC	AIC	AIC$_c$	BIC	CAIC	HQC
100	10	85.68	87.01	49.37	28.97	81.53	83.01	82.24	38.34	19.34	75.53
200	20	94.55	97.36	47.50	23.41	92.70	94.19	96.21	32.56	12.82	87.38
500	50	98.98	99.67	41.80	15.27	99.54	99.00	99.73	26.01	6.66	98.53
1000	100	99.79	99.88	39.29	10.02	100.00	99.84	99.93	21.94	3.76	99.98

$c_0 = 0.2$ Case 1							Case 2				
n	p	AIC	AIC$_c$	BIC	CAIC	HQC	AIC	AIC$_c$	BIC	CAIC	HQC
100	20	42.87	11.39	1.35	0.41	12.64	42.95	11.00	1.36	0.37	12.15
200	40	46.97	9.62	0.18	0.04	4.73	47.13	9.12	0.11	0.04	4.43
500	100	48.63	5.15	0.00	0.00	0.69	47.48	5.17	0.00	0.00	0.64
1000	200	48.18	2.37	0.00	0.00	0.11	48.73	2.19	0.00	0.00	0.11

$c_0 = 0.2$ Case 3							Case 4				
n	p	AIC	AIC$_c$	BIC	CAIC	HQC	AIC	AIC$_c$	BIC	CAIC	HQC
100	20	84.44	53.91	9.39	1.37	56.51	80.68	41.22	5.14	0.69	45.79
200	40	96.33	76.28	1.95	0.23	58.61	94.87	65.24	0.77	0.10	44.32
500	100	99.73	97.82	0.15	0.01	68.54	99.69	94.99	0.04	0.00	53.72
1000	200	99.93	100.00	0.01	0.00	93.01	99.95	100.00	0.07	0.00	85.80

$c_0 = 0.3$ Case 1							Case 2				
n	p	AIC	AIC$_c$	BIC	CAIC	HQC	AIC	AIC$_c$	BIC	CAIC	HQC
100	30	32.52	0.91	0.26	0.02	4.55	44.24	2.30	0.70	0.19	9.44
200	60	31.53	0.32	0.01	0.01	0.80	53.04	1.71	0.06	0.00	3.75
500	150	24.33	0.11	0.00	0.00	0.09	60.55	0.82	0.00	0.00	0.57
1000	300	17.70	0.05	0.00	0.00	0.00	66.13	0.30	0.00	0.00	0.10

$c_0 = 0.3$ Case 3							Case 4				
n	p	AIC	AIC$_c$	BIC	CAIC	HQC	AIC	AIC$_c$	BIC	CAIC	HQC
100	30	79.72	5.54	0.72	0.05	30.74	77.07	3.48	0.43	0.02	24.33
200	60	95.30	10.57	0.04	0.00	23.13	93.24	5.70	0.01	0.00	14.91
500	150	99.86	27.58	0.01	0.00	19.93	99.79	16.97	0.00	0.00	11.77
1000	300	100.00	84.93	0.00	0.00	50.09	99.98	80.00	0.00	0.00	44.04

4 Conclusion and Discussion

In this paper, we derived the conditions that the LLBIC in (6) is consistent in selecting the best model for a CCA, when the normality assumption to the true model is violated. The information criteria considered in this paper are defined by adding a positive penalty term to the negative twofold maximum log-likelihood, hence, the family of information criteria that we considered includes as special cases the AIC, $\mathrm{AIC_c}$, BIC, CAIC, and HQC. If we define consistency by meaning that the probability of selecting the true model approaches 1, then, in general, under the LS asymptotic framework, neither the AIC nor the $\mathrm{AIC_c}$ are consistent, but the BIC, CAIC, and HQC are. In this paper, we derived the conditions for consistency under the HD asymptotic framework. Understanding the asymptotic behavior of the difference between the two negative twofold maximum log-likelihoods are important because the dimension of the maximum log-likelihood increases with an increase in the sample size. If a normal distribution is assumed to the true model, it is possible to use a method that uses the properties of Wishart distribution (see Yanagihara *et al.*, 2012; Fujikoshi *et al.*, 2014). However, we cannot use this method in this paper, because we considered a case in which the normality assumption is violated for the true model. Hence, to evaluate the asymptotic behavior, we considered the convergence in probability for a linear combination of elements in a symmetric idempotent random matrix and the distribution of the maximum eigenvalues of the estimators of the covariance matrix. A basic idea for evaluating consistency is the same as in Yanagihara (2013). However, in Yanagihara (2013), $\boldsymbol{x}$ was not a random vector. Hence, we extended Yanagihara's method to the case that $\boldsymbol{x}$ is a random vector.

The results of our analysis and simulations confirmed that the AIC and $\mathrm{AIC_c}$ are consistent, and in some cases, the BIC is not consistent. These results are similar to those obtained for a multivariate regression model proposed by Yanagihara and colleagues (Yanagihara *et al.*, 2012, 2014; Yanagihara 2013).

Appendix

A1. Lemmas for Proving Theorems and Corollaries

In this section, we prepare some lemmas that we will use to derive the conditions for consistency of the penalty term $m(j)$ in IC_m in (5). We first present Lemma 1, which addresses the expectation of a moment (the proof was given in Yanagihara, 2013).

Lemma 1 *For any $n \times n$ symmetric matrix $\boldsymbol{A}$,*

$$E\left[\mathrm{tr}\left\{(\boldsymbol{\mathcal{E}}'\boldsymbol{A}\boldsymbol{\mathcal{E}})^2\right\}\right] = \kappa_4^{(1)} \sum_{a=1}^{n} \{(\boldsymbol{A})_{aa}\}^2 + p(p+1)\mathrm{tr}(\boldsymbol{A}^2) + p\,\mathrm{tr}(\boldsymbol{A})^2,$$

where $\kappa_4^{(1)}$ is given by (12)*, and $(\boldsymbol{A})_{ab}$ is the (a,b)th element of $\boldsymbol{A}$.*

Next, we present Lemma 2, which is the key lemma for deriving the conditions for consistency. In this study, we derived the conditions necessary for achieving Lemma 2 (the proof was given in Yanagihara, 2013).

Lemma 2 *Let $b_{j,\ell}$ be some positive constant that depends on the models, $j, \ell \in \mathcal{J}$. Then, we have*

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

Lemmas 3, 4, and 5 were used for evaluating the asymptotic behavior of each term (the proofs are given in Appendices A4, A5 and A6).

Lemma 3 *Let $\boldsymbol{W}$ be an $n \times n$ random matrix, defined by $\boldsymbol{W} = \boldsymbol{\mathcal{E}}(\boldsymbol{\mathcal{E}}'\boldsymbol{\mathcal{E}})^{-1}\boldsymbol{\mathcal{E}}'$. Then, for any $\ell \in \mathcal{J}$, we obtain*

$$\frac{1}{n-1}\boldsymbol{X}_\ell'\boldsymbol{W}\boldsymbol{X}_\ell \xrightarrow{p} c_0\boldsymbol{\Sigma}_{\ell\ell}.$$

Lemma 4 *Let $\lambda_{\max}(\boldsymbol{A})$ denote the maximum eigenvalue of $\boldsymbol{A}$, and let $\boldsymbol{V}_j$ be a $p \times p$ matrix defined by*

$$\boldsymbol{V}_j = \frac{1}{n}\boldsymbol{\mathcal{E}}'(\boldsymbol{I}_n - \boldsymbol{P}_j - \boldsymbol{H}_j\boldsymbol{H}_j')\boldsymbol{\mathcal{E}},$$

where $\boldsymbol{P}_j$ and $\boldsymbol{H}_j$ are given by (10) *and* (11)*, respectively. If assumption A2 holds, $\lambda_{\max}(\boldsymbol{V}_j) = O_p(p^{1/2})$ is satisfied.*

Lemma 5 *If assumptions A2 and A4 hold, $\alpha_{j,1} = O_p(np)$ is satisfied, and $\liminf_{c_{n,p}\to c_0} \alpha_{j,1}/(np) > 0$, where $\alpha_{j,1}$ is the maximum diagonal element of $\boldsymbol{L}_j$ given by* (11).

A2. Proof of Theorem 1

Let $D(j,\ell)$ ($j, \ell \in \mathcal{J}$) be the difference between two negative twofold maximum log-likelihoods divided by $(n-1)$, such that

$$D(j,\ell) = \log\frac{|\boldsymbol{S}_{yy\cdot j}|}{|\boldsymbol{S}_{yy\cdot \ell}|}.$$

Note that

$$\mathrm{IC}_m(j) - \mathrm{IC}_m(j_*) = (n-1)D(j,j_*) + m(j) - m(j_*).$$

From Lemma 2, we see that to obtain the conditions on $m(j)$ such that $\mathrm{IC}_m(j)$ is consistent, we only have to show the convergence in probability of $D(j,j_*)$ or a lower bound on $D(j,j_*)$ divided by some constant.

First, we show the convergence in probability of $D(j,j_*)$ when $j \in \mathcal{J}_+$. Note that $\boldsymbol{P}_j\boldsymbol{Y} = \boldsymbol{P}_j\boldsymbol{\mathcal{E}}$ holds for all j, since $\boldsymbol{X}_*$ is centralized. From the property of the determinant (see, e.g., Harville, 1997, chap. 18, cor. 18.1.2), the following equation are satisfied for all $j \in \mathcal{J}_+ \setminus \{j_*\}$ under the given assumptions:

$$\begin{aligned} D(j,j_*) &= \log\frac{|\boldsymbol{Y}'(\boldsymbol{I}_n - \boldsymbol{P}_j)\boldsymbol{Y}|}{|\boldsymbol{Y}'(\boldsymbol{I}_n - \boldsymbol{P}_{j_*})\boldsymbol{Y}|} = \log\frac{|\boldsymbol{\mathcal{E}}'(\boldsymbol{I}_n - \boldsymbol{P}_j)\boldsymbol{\mathcal{E}}|}{|\boldsymbol{\mathcal{E}}'(\boldsymbol{I}_n - \boldsymbol{P}_{j_*})\boldsymbol{\mathcal{E}}|} \\ &= \log\frac{|\boldsymbol{I}_n - (\boldsymbol{\mathcal{E}}'\boldsymbol{\mathcal{E}})^{-1}\boldsymbol{\mathcal{E}}'\boldsymbol{P}_j\boldsymbol{\mathcal{E}}|}{|\boldsymbol{I}_n - (\boldsymbol{\mathcal{E}}'\boldsymbol{\mathcal{E}})^{-1}\boldsymbol{\mathcal{E}}'\boldsymbol{P}_{j_*}\boldsymbol{\mathcal{E}}|} \\ &= \log\frac{|\boldsymbol{X}_j'\boldsymbol{X}_j - \boldsymbol{X}_j'\boldsymbol{W}\boldsymbol{X}_j|\,|\boldsymbol{X}_*'\boldsymbol{X}_*|}{|\boldsymbol{X}_*'\boldsymbol{X}_* - \boldsymbol{X}_*'\boldsymbol{W}\boldsymbol{X}_*|\,|\boldsymbol{X}_j'\boldsymbol{X}_j|}. \end{aligned}$$

Hence, by using Lemma 3 and $(n-1)^{-1}\boldsymbol{X}_\ell'\boldsymbol{X}_\ell \xrightarrow{p} \boldsymbol{\Sigma}_{\ell\ell}$ for all $\ell \in \mathcal{J}$, we obtain

$$D(j, j_*) \xrightarrow{p} (q_j - q_{j*})\log(1 - c_0). \tag{A1}$$

Next, we show the convergence in probability of a lower bound on $D(j, j_*)/\log p$ when $j \in \mathcal{J}_-$. It follows that for all $j \in \mathcal{J}_-$,

$$\begin{aligned}
D(j, j_*) &= \log \frac{\left|(\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{V}_j\right|}{|\boldsymbol{\mathcal{E}}'(\boldsymbol{I}_n - \boldsymbol{P}_{j_*})\boldsymbol{\mathcal{E}}|} \\
&= \log \left| \boldsymbol{I}_p + \frac{\sum_{a=1}^{\gamma_j} \boldsymbol{V}_j^{-1}(\sqrt{\alpha_{j,a}}\boldsymbol{g}_{j,a} + \boldsymbol{\mathcal{E}}'\boldsymbol{h}_{j,a})(\sqrt{\alpha_{j,a}}\boldsymbol{g}_{j,a} + \boldsymbol{\mathcal{E}}'\boldsymbol{h}_{j,a})'}{n} \right| \\
&\quad + \log \frac{|n\boldsymbol{V}_j|}{|\boldsymbol{\mathcal{E}}'(\boldsymbol{I}_n - \boldsymbol{P}_{j_*})\boldsymbol{\mathcal{E}}|} \\
&\geq \log \left| \boldsymbol{I}_p + \frac{\boldsymbol{V}_j^{-1}(\sqrt{\alpha_{j,1}}\boldsymbol{g}_{j,1} + \boldsymbol{\mathcal{E}}'\boldsymbol{h}_{j,1})(\sqrt{\alpha_{j,1}}\boldsymbol{g}_{j,1} + \boldsymbol{\mathcal{E}}'\boldsymbol{h}_{j,1})'}{n} \right| \\
&\quad + \log \frac{|n\boldsymbol{V}_j|}{|\boldsymbol{\mathcal{E}}'(\boldsymbol{I}_n - \boldsymbol{P}_{j_*})\boldsymbol{\mathcal{E}}|} \\
&= \log \left\{ 1 + \frac{(\sqrt{\alpha_{j,1}}\boldsymbol{g}_{j,1} + \boldsymbol{\mathcal{E}}'\boldsymbol{h}_{j,1})'\boldsymbol{V}_j^{-1}(\sqrt{\alpha_{j,1}}\boldsymbol{g}_{j,1} + \boldsymbol{\mathcal{E}}'\boldsymbol{h}_{j,1})}{n} \right\} \\
&\quad + \log \frac{|n\boldsymbol{V}_j|}{|\boldsymbol{\mathcal{E}}'(\boldsymbol{I}_n - \boldsymbol{P}_{j_*})\boldsymbol{\mathcal{E}}|} \\
&\geq \log \left\{ \lambda_{\max}(\boldsymbol{V}_j) + \frac{(\sqrt{\alpha_{j,1}}\boldsymbol{g}_{j,1} + \boldsymbol{\mathcal{E}}'\boldsymbol{h}_{j,1})'(\sqrt{\alpha_{j,1}}\boldsymbol{g}_{j,1} + \boldsymbol{\mathcal{E}}'\boldsymbol{h}_{j,1})}{n} \right\} \\
&\quad + \log \frac{|n\boldsymbol{V}_j|}{|\boldsymbol{\mathcal{E}}'(\boldsymbol{I}_n - \boldsymbol{P}_{j_*})\boldsymbol{\mathcal{E}}|} - \log \lambda_{\max}(\boldsymbol{V}_j) \\
&= D_1(j) + D_2(j) + D_3(j),
\end{aligned} \tag{A2}$$

where

$$\begin{aligned}
D_1(j) &= \log\left\{\lambda_{\max}(\boldsymbol{V}_j) + p\xi_j\right\}, \\
D_2(j) &= \log \frac{|n\boldsymbol{V}_j|}{|\boldsymbol{\mathcal{E}}'(\boldsymbol{I}_n - \boldsymbol{P}_{j_*})\boldsymbol{\mathcal{E}}|}, \\
D_3(j) &= -\log \lambda_{\max}(\boldsymbol{V}_j),
\end{aligned}$$

and $\xi_j = (\sqrt{\alpha_{j,1}}\boldsymbol{g}_{j,1} + \boldsymbol{\mathcal{E}}'\boldsymbol{h}_{j,1})'(\sqrt{\alpha_{j,1}}\boldsymbol{g}_{j,1} + \boldsymbol{\mathcal{E}}'\boldsymbol{h}_{j,1})/(np)$.

First, we evaluate the asymptotic behavior of $D_1(j)$ in (A2). From the equation $\boldsymbol{h}_{j,l}'\boldsymbol{h}_{j,1} = 1$, it is easy to see that

$$E[\boldsymbol{h}_{j,1}'\boldsymbol{\mathcal{E}}\boldsymbol{\mathcal{E}}'\boldsymbol{h}_{j,1}] = p.$$

Moreover, it follows from Lemma 1 that

$$\begin{aligned}
E[(\boldsymbol{h}_{j,1}'\boldsymbol{\mathcal{E}}\boldsymbol{\mathcal{E}}'\boldsymbol{h}_{j,1} - p)^2] &= \kappa_4^{(1)} \sum_{a=1}^{n} \{(\boldsymbol{h}_{j,1}\boldsymbol{h}_{j,1}')_{aa}\}^2 + 2p \\
&= O(\max\{p, p^s\}),
\end{aligned}$$

where $\kappa_4^{(1)}$ is given by (12), and s is a positive constant given by (13). Hence, we have

$$\boldsymbol{h}_{j,1}'\boldsymbol{\mathcal{E}}\boldsymbol{\mathcal{E}}'\boldsymbol{h}_{j,1} = p + O_p(\max\{p^{1/2}, p^{s/2}\}) = O_p(p). \tag{A3}$$

Moreover, note that $\boldsymbol{g}_{j,1}\boldsymbol{g}_{j,1}'$ is an idempotent matrix,

$$\begin{aligned}\left(\sqrt{\alpha_{j,1}}\boldsymbol{g}_{j,1}\boldsymbol{\mathcal{E}}'\boldsymbol{h}_{j,1}\right)^2 &= \alpha_{j,1}\boldsymbol{h}_{j,1}'\boldsymbol{\mathcal{E}}\boldsymbol{g}_{j,1}\boldsymbol{g}_{j,1}'\boldsymbol{\mathcal{E}}'\boldsymbol{h}_{j,1} \\ &\le \alpha_{j,1}\boldsymbol{h}_{j,1}'\boldsymbol{\mathcal{E}}\boldsymbol{\mathcal{E}}'\boldsymbol{h}_{j,1} \\ &= O_p(np^2).\end{aligned}$$

This implies that

$$\sqrt{\alpha_{j,1}}\boldsymbol{g}_{j,1}\boldsymbol{\mathcal{E}}'\boldsymbol{h}_{j,1} = O_p(n^{1/2}p). \tag{A4}$$

From Lemma 5, (A3), and (A4), we have

$$\xi_j = O_p(1). \tag{A5}$$

By using (A5) and Lemma 4, we obtain

$$\begin{aligned}\frac{1}{\log p}D_1(j) &= \frac{1}{\log p}\log\left\{\lambda_{\max}(\boldsymbol{V}_j) + p\xi_j\right\} \\ &= \frac{1}{\log p}\log\left\{\frac{1}{p}\lambda_{\max}(\boldsymbol{V}_j) + \xi_j\right\} + 1 \\ &\xrightarrow{p} 1.\end{aligned} \tag{A6}$$

Next, we evaluate the asymptotic behavior of $D_2(j)$ in (A2). From Lemma 3 and the result $(\boldsymbol{I}_n - \boldsymbol{P}_j - \boldsymbol{H}_j\boldsymbol{H}_j')(\boldsymbol{I}_n - \boldsymbol{P}_j) = \boldsymbol{I}_n - \boldsymbol{P}_j - \boldsymbol{H}_j\boldsymbol{H}_j'$, we can see that

$$\begin{aligned}D_2(j) &\le \log\frac{|\boldsymbol{\mathcal{E}}'(\boldsymbol{I}_n - \boldsymbol{P}_j)\boldsymbol{\mathcal{E}}|}{|\boldsymbol{\mathcal{E}}'(\boldsymbol{I}_n - \boldsymbol{P}_{j_*})\boldsymbol{\mathcal{E}}|} \\ &= \log\frac{\left|\boldsymbol{X}_j'\boldsymbol{X}_j - \boldsymbol{X}_j'\boldsymbol{W}\boldsymbol{X}_j\right|\,|\boldsymbol{X}_*'\boldsymbol{X}_*|}{|\boldsymbol{X}_*'\boldsymbol{X}_* - \boldsymbol{X}_*'\boldsymbol{W}\boldsymbol{X}_*|\,\left|\boldsymbol{X}_j'\boldsymbol{X}_j\right|} \\ &\xrightarrow{p} (k_j - k_{j_*})\log(1 - c_0),\end{aligned}$$

where $\boldsymbol{W}$ is given in Lemma 3. It follows 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 (9). Thus, we also have

$$\begin{aligned}D_2(j) &\ge \log\frac{\left|\boldsymbol{\mathcal{E}}'(\boldsymbol{I}_n - \boldsymbol{P}_{j_+})\boldsymbol{\mathcal{E}}\right|}{|\boldsymbol{\mathcal{E}}'(\boldsymbol{I}_n - \boldsymbol{P}_{j_*})\boldsymbol{\mathcal{E}}|} \\ &= \log\frac{\left|\boldsymbol{X}_{j+}'\boldsymbol{X}_{j+} - \boldsymbol{X}_{j+}'\boldsymbol{W}\boldsymbol{X}_{j+}\right|\,|\boldsymbol{X}_*'\boldsymbol{X}_*|}{|\boldsymbol{X}_*'\boldsymbol{X}_* - \boldsymbol{X}_*'\boldsymbol{W}\boldsymbol{X}_*|\,\left|\boldsymbol{X}_{j+}'\boldsymbol{X}_{j+}\right|} \\ &\xrightarrow{p} (k_{j_+} - k_{j_*})\log(1 - c_0).\end{aligned}$$

The above upper and lower bounds on $D_2(j)$ imply that

$$\frac{1}{\log p}D_2(j) \xrightarrow{p} 0. \tag{A7}$$

Finally, we evaluate the asymptotic behavior of $D_3(j)$ in (A2). Since $-\log x \leq -x+1$ for any $x \geq 0$, we have

$$\begin{aligned} D_3(j) &= \frac{1}{2}\log p - \log \frac{\lambda_{\max}(\boldsymbol{V}_j)}{\sqrt{p}} \\ &\geq \frac{1}{2}\log p - \left\{ \frac{\lambda_{\max}(\boldsymbol{V}_j)}{\sqrt{p}} - 1 \right\} = D_{3,1}(j). \end{aligned}$$

It follows from Lemma 4 that

$$\frac{1}{\log p} D_{3,1}(j) \xrightarrow{p} \frac{1}{2}. \tag{A8}$$

Consequently, combining (A2), (A6), (A7), and (A8) yields,

$$\begin{aligned} \frac{1}{\log p} \log D(j, j_*) &= \frac{1}{\log p} \{ D_1(j) + D_2(j) + D_3(j) \} \\ &\geq \frac{1}{\log p} \{ D_1(j) + D_2(j) + D_{3,1}(j) \} \\ &\xrightarrow{p} \frac{1}{2}. \end{aligned} \tag{A9}$$

As a result, from Lemma 2, (A1), and (A9), we can obtain the conditions given in Theorem 1.

A3. Proof of Corollary 1

First, we consider the AIC and $\mathrm{AIC_c}$. According to an expansion of $m(j) - m(j_*)$ in the $\mathrm{AIC_c}$, the differences between the penalty terms of the $\mathrm{AIC_c}$s are

$$\begin{aligned} & m(j) - m(j_*) \\ &= \frac{(q_j - q_*)(2 - c_{n,p})p}{(1 - c_{n,p})^2} \left(1 + \frac{q_j + q_* - 2}{n} \right) \left(1 - \frac{1}{n} \right)^2 + O\left(pn^{-1} \right). \end{aligned} \tag{A10}$$

Moreover, the differences between the penalty terms of the AICs are

$$\frac{1}{n \log p} \{ m(j) - m(j_*) \} = \frac{2 c_{n,p}}{\log p} (q_j - q_{j_*}).$$

Hence, the convergence of the differences between the penalty terms of the AICs and those of the $\mathrm{AIC_c}$s is

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

This indicates that the condition C2 holds for both the AIC and the $\mathrm{AIC_c}$. Furthermore, it follows from equation (A10) that

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

Since $c^{-1}\log(1-c) + (1-c)^{-1} + (1-c)^{-2}$ is a monotonically increasing function when $0 \leq c < 1$, it follows that $c_0^{-1} \log(1 - c_0) + (1 - c_0)^{-1} + (1 - c_0)^{-2} > 0$ holds. That is, the penalty terms in the $\mathrm{AIC_c}$ always satisfy the condition C1 when $j \in \mathcal{J} \setminus \{j_*\}$, and those in the AIC satisfy the condition C1 if $c_0 \in [0, c_a)$, where c_a is given by (14).

Next, we consider the BIC and the CAIC. When $j \in \mathcal{J}_+ \setminus \{j_*\}$, the difference between the penalty term of the BIC and that of the CAIC is

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

Thus, the condition C1 holds. Moreover, it is easy to obtain

$$\frac{1}{n \log p} \{m(j) - m(j_*)\} = \begin{cases} c_{n,p}(q_j - q_{j_*}) \left(\dfrac{-\log c_{n,p}}{\log p} + 1 \right) & \text{(BIC)} \\[2ex] c_{n,p}(q_j - q_{j_*}) \left(\dfrac{1 - \log c_{n,p}}{\log p} + 1 \right) & \text{(CAIC)} \end{cases}.$$

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

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

When $j \in \bar{S}_- \cap \mathcal{J}_-$, condition C2 is satisfied because $c_0(qj - q_*) \geq 0$ holds, where S_- is given by (15). When $j \in S_-$, then for all $j \in S_-$, condition C2 is satisfied if $c_0 < 1/\{2(q_* - q_j)\}$ holds.

Finally, the HQC is considered. When $j \in \mathcal{J}_+ \setminus \{j_*\}$, the difference between the penalty terms of the HQCs is

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

Thus, the condition C1 holds. Moreover, it is easy to see that

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

From this equation, we obtain

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

Hence, condition C2 holds. From the above results and Theorem 1, Corollary 1 is proved.

A4. Proof of Lemma 3

For any $\ell \in \mathcal{J}$, let $\boldsymbol{X}_\ell = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{q_\ell})$, let $\boldsymbol{x}_k = (x_{1k}, \ldots, x_{nk})'$, and let w_{ab} be the (a, b)th element of $\boldsymbol{W}$. Then, $\boldsymbol{x}_s' \boldsymbol{W} \boldsymbol{x}_t$, which is the (s, t)th element of $\boldsymbol{X}_\ell' \boldsymbol{W} \boldsymbol{X}_\ell$, is expressed as

$$\boldsymbol{x}_s' \boldsymbol{W} \boldsymbol{x}_t = \sum_{a=1}^{n} x_{as} x_{at} w_{aa} + \sum_{a \neq b}^{n} x_{as} x_{bt} w_{ab}. \tag{A11}$$

Moreover, we can calculate

$$
\begin{aligned}
\left(\boldsymbol{x}_s'\boldsymbol{W}\boldsymbol{x}_t\right)^2 &= \sum_{a=1}^{n} x_{as}^2 x_{at}^2 w_{aa}^2 + \sum_{a\neq b\neq c\neq d}^{n} x_{as}x_{bs}x_{ct}x_{dt}w_{ab}w_{cd} \\
&\quad + \sum_{a\neq b}^{n} \left\{ x_{as}x_{bs}x_{at}x_{bt}\left(w_{aa}w_{bb} + w_{ab}^2\right) + x_{as}^2 x_{bt}^2 w_{ab}^2 + 2(x_{as}^2 x_{at}x_{bt} \right. \\
&\quad \left. + x_{as}x_{bs}x_{at}^2) w_{aa}w_{ab} \right\} + \sum_{a\neq b\neq c}^{n} \left\{ 2x_{as}x_{bs}x_{at}x_{ct} + (x_{as}^2 x_{bt}x_{ct} \right. \\
&\quad \left. + 2x_{as}x_{bs}x_{at}x_{ct} + x_{bs}x_{cs}x_{at}^2) w_{ab}w_{ac} \right\}, \qquad \text{(A12)}
\end{aligned}
$$

where the notation $\sum_{a_1\neq a_2\neq\cdots}^{n}$ means $\sum_{a_1=1}^{n}\sum_{a_2=1,a_2\neq a_1}^{n}\cdots$. Notice that $\boldsymbol{X}'\boldsymbol{1}_n = \boldsymbol{0}_q$ and so

$$
\begin{aligned}
&\sum_{a,b}^{n} x_{as}x_{bt} = \sum_{a=1}^{n} x_{as} = \sum_{a=1}^{n} x_{at} = 0, \quad \sum_{a\neq b}^{n} x_{as}x_{bt} = -\boldsymbol{x}_s'\boldsymbol{x}_t, \\
&\sum_{a\neq b}^{n} x_{as}x_{bs}x_{at}x_{bt} = (\boldsymbol{x}_s'\boldsymbol{x}_t)^2 - \sum_{a=1}^{n} x_{as}^2 x_{at}^2, \quad \sum_{a\neq b}^{n} x_{as}^2 x_{bt}^2 = \boldsymbol{x}_s'\boldsymbol{x}_s\boldsymbol{x}_t'\boldsymbol{x}_t - \sum_{a=1}^{n} x_{as}^2 x_{at}^2, \\
&\sum_{a\neq b} x_{as}^2 x_{at}x_{bt} = \sum_{a\neq b} x_{as}x_{bs}x_{at}^2 = -\sum_{a=1}^{n} x_{as}^2 x_{at}^2, \qquad \text{(A13)} \\
&\sum_{a\neq b\neq c}^{n} x_{as}x_{bs}x_{at}x_{ct} = \sum_{a\neq b\neq c}^{n} x_{as}^2 x_{bt}x_{ct} = \sum_{a\neq b\neq c}^{n} x_{bs}x_{cs}x_{at}^2 \\
&\quad = \sum_{a=1}^{n} x_{as}^2 x_{at}^2 + \sum_{a\neq b}^{n} x_{at}x_{bt}x_{as}x_{bs}.
\end{aligned}
$$

Note that $\boldsymbol{x}_s'\boldsymbol{x}_t$ is the (s,t)th element of $\boldsymbol{X}_\ell'\boldsymbol{X}_\ell$, and $(n-1)^{-1}\boldsymbol{X}_\ell'\boldsymbol{X}_\ell \xrightarrow{p} \boldsymbol{\Sigma}_{\ell\ell}$. Here, since $\boldsymbol{W}$ is a symmetric idempotent matrix and $\boldsymbol{W}\boldsymbol{1}_n = \boldsymbol{0}_n$ holds, we obtain the following equations:

$$
0 \leq w_{aa} \leq |w_{ab}| \leq \sqrt{w_{aa}w_{bb}} \leq 1 \quad (a = 1,\ldots,n; b = 1,\ldots,n; a\neq b), \qquad \text{(A14)}
$$

and

$$
\begin{aligned}
&\mathrm{tr}(\boldsymbol{W}) = \sum_{a=1}^{n} w_{aa} = p, \quad \mathrm{tr}(\boldsymbol{W}^2) = \sum_{a=1}^{n} w_{aa}^2 + \sum_{a\neq b}^{n} w_{ab}^2 = p, \\
&\mathrm{tr}(\boldsymbol{W})^2 = \sum_{a=1}^{n} w_{aa}^2 + \sum_{a\neq b}^{n} w_{aa}w_{bb} = p^2, \quad \mathbf{1}_n'\boldsymbol{W}\mathbf{1}_n = \sum_{a=1}^{n} w_{aa} + \sum_{a\neq b}^{n} w_{ab} = 0, \\
&\mathbf{1}_n'\boldsymbol{W}^2\mathbf{1}_n = \sum_{a=1}^{n} w_{aa}^2 + \sum_{a\neq b}^{n}(2w_{aa}w_{ab} + w_{ab}^2) + \sum_{a\neq b\neq c} w_{ab}w_{ac} = 0, \\
&\mathrm{tr}(\boldsymbol{W})\mathbf{1}_n'\boldsymbol{W}\mathbf{1}_n = \sum_{a=1}^{n} w_{aa}^2 + \sum_{a\neq b}^{n}(2w_{aa}w_{ab} + w_{aa}w_{bb}) + \sum_{a\neq b\neq c} w_{aa}w_{bc} = 0, \\
&\left(\mathbf{1}_n'\boldsymbol{W}\mathbf{1}_n\right)^2 = \sum_{a=1}^{n} w_{aa}^2 + \sum_{a\neq b}^{n}(w_{aa}w_{ab} + 2w_{ab}^2 + 4w_{aa}w_{ab}) \\
&\qquad + 2\sum_{a\neq b\neq c}(w_{aa}w_{bc} + 2w_{ab}w_{ac}) + \sum_{a\neq b\neq c\neq d} w_{ab}w_{cd} = 0.
\end{aligned}
\tag{A15}
$$

Since w_{aa} $(a = 1, \dots, n)$ are identically distributed, and w_{ab} $(a = 1, \dots, n; b = a+1, \dots, n)$ are also identically distributed, from the equations in (A15), and for $a \neq b \neq c \neq d$, we obtain

$$
\begin{aligned}
&p = nE[w_{aa}], \\
&p = nE[w_{aa}^2] + n(n-1)E[w_{ab}^2], \\
&p^2 = nE[w_{aa}^2] + n(n-1)E[w_{aa}w_{bb}], \\
&0 = nE[w_{aa}] + n(n-1)E[w_{ab}], \\
&0 = nE[w_{aa}^2] + n(n-1)\left(2E[w_{aa}w_{ab}] + E[w_{ab^2}]\right) + n(n-1)(n-2)E[w_{ab}w_{ac}], \\
&0 = nE[w_{aa}^2] + n(n-1)\left(2E[w_{aa}w_{ab}] + E[w_{aa}w_{bb}]\right) + n(n-1)(n-2)E[w_{aa}w_{bc}], \\
&0 = nE[w_{aa}^2] + n(n-1)\left(E[w_{aa}w_{bb}] + 2E[w_{ab^2}] + 4E[w_{aa}w_{ab}]\right) \\
&\qquad + 2n(n-1)(n-2)\left(E[w_{aa}w_{bc}] + 2E[w_{ab}w_{ac}]\right) \\
&\qquad + n(n-1)(n-2)(n-3)E[w_{ab}w_{cd}].
\end{aligned}
\tag{A16}
$$

It follows from equation (A14) that $E[w_{aa}^2] \leq 1$. Combining this result and equation (A16) yields

$$
\begin{aligned}
&E[w_{aa}] = c_{n,p}, && E[w_{ab}] = O(n^{-1}), \\
&E[w_{aa}^2] = O(1), && E[w_{aa}w_{bb}] = c_{n,p}^2 + O(n^{-1}), \\
&E[w_{ab}^2] = O(n^{-1}), && E[w_{aa}w_{ab}] = O(n^{-1}), \\
&E[w_{aa}w_{bc}] = O(n^{-1}), && E[w_{ab}w_{ac}] = O(n^{-2}), \\
&E[w_{ab}w_{cd}] = O(n^{-2}),
\end{aligned}
\tag{A17}
$$

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

Let σ_{st} be the (s,t)th element of $\boldsymbol{\Sigma}_{\ell\ell}$. Then, by using (A11), (A12), (A13), and (A17), we have

$$\frac{1}{n-1}E\left[\boldsymbol{x}_s'\boldsymbol{W}\boldsymbol{x}_t\right] \to c_0\sigma_{st}, \quad \frac{1}{(n-1)^2}E\left[\left(\boldsymbol{x}_s'\boldsymbol{W}\boldsymbol{x}_t\right)^2\right] \to c_0^2\sigma_{st}^2.$$

The above equations directly imply that $(n-1)^{-1}Var\left[\boldsymbol{x}_s'\boldsymbol{W}\boldsymbol{x}_t\right] \to 0$ as $c_{n,p} \to 0$. Hence, the (s,t)th element of $\boldsymbol{X}_\ell'\boldsymbol{W}\boldsymbol{X}_\ell$ converges, as follows:

$$\frac{1}{n-1}\boldsymbol{x}_s'\boldsymbol{W}\boldsymbol{x}_t \xrightarrow{p} c_0\sigma_{st}.$$

Therefore, Lemma 3 is proved.

A5. Proof of Lemma 4

It follows from elementary linear algebra that

$$\lambda_{\max}(\boldsymbol{V}_j) \le \lambda_{\max}\left(\frac{1}{n}\boldsymbol{\mathcal{E}}'\boldsymbol{\mathcal{E}}\right) \le \sqrt{\frac{1}{n^2}\mathrm{tr}\left\{\left(\boldsymbol{\mathcal{E}}'\boldsymbol{\mathcal{E}}\right)^2\right\}}.$$

From Lemma 1, we can see that

$$E\left[\frac{1}{n^2}\mathrm{tr}\left\{\left(\boldsymbol{\mathcal{E}}'\boldsymbol{\mathcal{E}}\right)^2\right\}\right] = \frac{1}{n}\kappa_4^{(1)} + \frac{1}{n}p(p+1) + p = O(p).$$

The above equation and Jensen's inequality lead us to the equation

$$E\left[\sqrt{\frac{1}{n^2}\mathrm{tr}\left\{\left(\boldsymbol{\mathcal{E}}'\boldsymbol{\mathcal{E}}\right)^2\right\}}\right] \le \sqrt{E\left[\frac{1}{n^2}\mathrm{tr}\left\{\left(\boldsymbol{\mathcal{E}}'\boldsymbol{\mathcal{E}}\right)^2\right\}\right]} = O(p^{1/2}).$$

This directly implies that $n^{-1}[\mathrm{tr}\{(\boldsymbol{\mathcal{E}}'\boldsymbol{\mathcal{E}})^2\}]^{1/2} = O_p(p^{1/2})$. Hence, Lemma 4 is proved.

A6. Proof of Lemma 5

It follows from elementary linear algebra that

$$\begin{aligned}
\frac{1}{np}\alpha_{j,1} = \frac{1}{np}\lambda_{\max}(\boldsymbol{L}_j) &\le \frac{1}{np}\mathrm{tr}(\boldsymbol{L}_j) \\
&= \frac{1}{np}\mathrm{tr}\left(\boldsymbol{\Gamma}_j\boldsymbol{\Gamma}_j'\right) \\
&= \frac{1}{np}\mathrm{tr}\left\{\boldsymbol{X}_*'(\boldsymbol{I}_n - \boldsymbol{P}_j)\boldsymbol{X}_*\boldsymbol{\Sigma}_{j_*j_*}^{-1}\boldsymbol{\Sigma}_{j_*y}\boldsymbol{\Sigma}_{yy\cdot j_*}^{-1}\boldsymbol{\Sigma}_{j_*y}'\boldsymbol{\Sigma}_{j_*j_*}^{-1}\right\} \\
&\le \frac{1}{np}\mathrm{tr}\left\{\boldsymbol{X}_*'\boldsymbol{X}_*\boldsymbol{\Sigma}_{j_*j_*}^{-1}\boldsymbol{\Sigma}_{j_*y}\boldsymbol{\Sigma}_{yy\cdot j_*}^{-1}\boldsymbol{\Sigma}_{j_*y}'\boldsymbol{\Sigma}_{j_*j_*}^{-1}\right\} \\
&\xrightarrow{p} \mathrm{tr}\left(\boldsymbol{\Psi}_j\boldsymbol{\Sigma}_{j_*j_*}^{-1}\right).
\end{aligned}$$

From the above equations and assumptions A2 and A4, we have

$$\alpha_{j,1} = O_p(np).$$

Moreover, it also follows from elementary linear algebra that

$$
\begin{aligned}
\frac{1}{np}\alpha_{j,1} = \frac{1}{np}\lambda_{\max}(\boldsymbol{L}_j) \geq \frac{1}{\gamma_j np}\mathrm{tr}\,(\boldsymbol{L}_j) & \\
&= \frac{1}{\gamma_j np}\mathrm{tr}\left\{\boldsymbol{X}_*'(\boldsymbol{I}_n - \boldsymbol{P}_j)\boldsymbol{X}_*\boldsymbol{\Sigma}_{j_*j_*}^{-1}\boldsymbol{\Sigma}_{j_*y}\boldsymbol{\Sigma}_{yy\cdot j_*}^{-1}\boldsymbol{\Sigma}_{j_*y}'\boldsymbol{\Sigma}_{j_*j_*}^{-1}\right\} \\
&\xrightarrow{p} \mathrm{tr}\left\{\boldsymbol{\Sigma}_{j_*j_*}^{-1}\boldsymbol{\Sigma}_{j_*j_*\cdot j}\boldsymbol{\Sigma}_{j_*j_*}^{-1}\boldsymbol{\Psi}_j\right\}.
\end{aligned}
$$

Hence, with assumption A4, this implies that

$$
\liminf_{c_{n,p}\to c_0}\frac{1}{np}\alpha_{j,1} > 0.
$$

Consequently, Lemma 5 is proved.

References

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

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

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

[4] Doeswijk, T. G., Hageman, J. A., Westerhuis, J. A., Tikunov, Y., Bovy, A. & van Eeuwijk, F. A. (2011). Canonical correlation analysis of multiple sensory directed metabolomics data blocks reveals corresponding parts between data blocks. Chemometr. Intell. Lab., **107**, 371–376.

[5] Fujikoshi, Y. (1982). A test for additional information in canonical correlation analysis. Ann. Inst. Statist. Math., **34**, 523–530.

[6] Fujikoshi, Y. (1985). Selection of variables in discriminant analysis and canonical correlation analysis. In Multivariate Analysis VI (Ed. P. R. Krishnaiah), 219–236, North-Holland, Amsterdam.

[7] Fujikoshi, Y. (2014). High-dimensional properties of AIC and Cp for estimation of dimensionality in multivariate models. TR 14-02, Statistical Research Group, Hiroshima University, Hiroshima.

[8] Fujikoshi, Y. & Kurata, H. (2008). Information criterion for some conditional independence structures. In New Trends in Psychometrics (Eds. K. Shigemasu, A. Okada, T. Imaizumi & T. Hoshino), 69–78, Universal Academy Press, Tokyo.

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

[10] Fujikoshi, Y., Sakurai, T., Kanda, S. & Sugiyama, T. (2008). Bootstrap information criterion for selection of variables in canonical correlation analysis. J. Inst. Sci. Engi., Chuo Univ., **14**, 31–49 (in Japanese).

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

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

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

[14] Hashiyama, Y., Yanagihara, H. & Fujikoshi, Y. (2011). Jackknife bias correction of the AIC for selecting variables in canonical correlation analysis under model misspecification. Linear Algebra Appl., **455**, 82–106.

[15] James, W. & Stein, C. (1961). Estimation with quadratic loss. In Proc. 4th Berkeley Sympos. Math. Statist. and Prob., **1**, 361–379.

[16] Jöreskog, K. G. (1967). Some contributions to maximum likelihood factor analysis. Psychometrika, **32**, 443–482.

[17] Khalil, B., Ouarda, T. B. M. J. & St-Hilaire, A. (2011). Estimation of water quality characteristics at ungauged sites using artificial neural networks and canonical correlation analysis. J. Hydrol., **405**, 277–287.

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

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

[20] McKay, R. J. (1977). Variable selection in multivariate regression: an application of simultaneous test procedures. J. Roy. Statist. Soc., Ser. **B 39**, 371–380.

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

[22] Ogura, T. (2010). A variable selection method in principal canonical correlation analysis. Comput. Statist. Data Anal., **54**, 1117–1123.

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

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

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

[26] Sweeney, K. T., McLoone, S. F. & Ward, T. E. (2013). The use of ensemble empirical mode decomposition with canonical correlation analysis as a novel artifact removal technique. IEEE Trans. Biomed. Eng., **60**, 97–105.

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

[28] Vahedi, S. (2011). Canonical correlation analysis of procrastination, learning strategies and statistics anxiety among Iranian female college students. Procedia Soc. Behav. Sci., **30**, 1620–1624.

[29] Vilsaint, C. L., Aiyer, S. M., Wilson, M. N., Shaw, D. S. & Dishion, T. J. (2013). The ecology of early childhood risk: A canonical correlation analysis of children's adjustment, family, and community context in a high-risk sample. J. Prim. Prev., **34**, 261–277.

[30] Yanagihara, H. (2013). Conditions for consistency of a log-likelihood-based information criterion in normal multivariate linear regression models under the violation of normality assumption. TR 13-11, Statistical Research Group, Hiroshima University, Hiroshima.

[31] Yanagihara, H., Hashiyama, Y. & Fujikoshi, Y. (2014). High-Dimensional asymptotic behaviors of differences between the log-determinants of two Wishart matrices. TR 14-10, Statistical Research Group, Hiroshima University, Hiroshima.

[32] Yanagihara, H., Wakaki, H. & 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.