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 739-8526, 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, \( \mathbf{x} = (x_1, \ldots, x_q)' \) and \( \mathbf{y} = (y_1, \ldots, y_p)' \), 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 \( \mathbf{z} = (\mathbf{x}', \mathbf{y}')' \) be a \((p + q)\)-dimensional vector with

\[
E[\mathbf{z}] = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix} = \mu, \quad Cov[\mathbf{z}] = \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{pmatrix} = \Sigma,
\]

where \( \mu_x \) and \( \mu_y \) are mean vectors of \( q \)- and \( p \)-dimensions, respectively; \( \Sigma_{xx} \) and \( \Sigma_{yy} \) are \( q \times q \) and \( p \times p \) covariance matrices of \( \mathbf{x} \) and \( \mathbf{y} \), respectively; and \( \Sigma_{xy} \) is the \( q \times p \) covariance matrix of \( \mathbf{x} \) and \( \mathbf{y} \). The square of the correlation between a pair of canonical correlation variables is
obtained as the eigenvalue of $\Sigma^{-1}_{yy} \Sigma_{xy} \Sigma^{-1}_{yy} \Sigma'_{xy}$ 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 $\omega = \{1, \ldots, q\}$ containing $q_j$ elements, and $x_j$ denotes the $q_j$-dimensional vector consisting of the elements of $x$ indexed by the elements of $j$, where $q_A$ denotes the number of elements in a set $A$, i.e., $q_A = \#(A)$. For example, if $j = \{1, 2, 4\}$, then $x_j$ consists of the first, second, and fourth elements of $x$. Without loss of generality, $x$ can be divided into $x = (x'_j, x''_j)'$, where $x'_j$ and $x''_j$ are $q_j$- and $q_j$-dimensional vectors, respectively. Note that $\bar{A}$ denotes the compliment of the set $A$.

Another expressions of $\mu_x$, $\Sigma_{xy}$ and $\Sigma_{xx}$ corresponding to the divisions of $x$ are

$$
\mu_x = \begin{pmatrix} \mu_j \\ \mu_j \end{pmatrix}, \quad \Sigma_{xy} = \begin{pmatrix} \Sigma_{jj} \\ \Sigma_{jj} \end{pmatrix}, \quad \Sigma_{xx} = \begin{pmatrix} \Sigma_{jj} & \Sigma_{jj} \\ \Sigma_{jj} & \Sigma_{jj} \end{pmatrix}.
$$

We are interested in whether the elements of $x_j$ are irrelevant variables in CCA. Let $z_1, \ldots, z_n$ be $n$ independent random vectors from $x$, and let $\bar{z}$ be the sample mean of $x_1, \ldots, x_n$ given by $\bar{z} = n^{-1} \sum_{i=1}^n z_i$ and $S$ be the usual unbiased estimator of $\Sigma$ given by $S = (n-1)^{-1} \sum_{i=1}^n (z_i - \bar{z})(z_i - \bar{z})'$, divided in the same way as we divided $\Sigma$, as follows:

$$
S = \begin{pmatrix} S_{xx} & S_{xy} \\ S'_{xy} & S_{yy} \end{pmatrix} = \begin{pmatrix} S_{jj} & S_{jj} & S_{jj} \\ S'_{jj} & S'_{jj} & S'_{jj} \\ S_{jj} & S_{jj} & S_{jj} \end{pmatrix}.
$$

Suppose that $z_1, \ldots, z_n \sim i.i.d. \mathcal{N}_{p+q}(\mu, \Sigma)$. Following Fujikoshi (1985), the candidate model that $x_j$ is irrelevant is expressed as

$$
M_j : (n-1)S \sim W_{p+q}(n-1, \Sigma)
$$

s.t. $\text{tr}(\Sigma^{-1}_{xx} \Sigma_{xy} \Sigma^{-1}_{yy} \Sigma'_{xy}) = \text{tr}(\Sigma^{-1}_{jj} \Sigma_{jj} \Sigma^{-1}_{yy} \Sigma'_{jj})$. \hspace{1cm} (1)

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

$$
\hat{\Sigma}_j = \arg\min_{\Sigma} \{ F(S, \Sigma) \text{ s.t. } \text{tr}(\Sigma^{-1}_{xx} \Sigma_{xy} \Sigma^{-1}_{yy} \Sigma'_{xy}) = \text{tr}(\Sigma^{-1}_{jj} \Sigma_{jj} \Sigma^{-1}_{yy} \Sigma'_{jj}) \}, \hspace{1cm} (2)
$$

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

$$
F(S, \Sigma) = (n-1)\{ \text{tr}(\Sigma^{-1}S) - \log |\Sigma^{-1}S| - (p + q) \}, \hspace{1cm} (3)
$$

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 $\hat{\Sigma}_j$ in (2) is given by

$$
\hat{\Sigma}_j = \begin{pmatrix} S_{jj} & S'_{jj} & S_{jj} \\ S'_{jj} & S'_{jj} & S'_{jj} \\ S_{jj} & S_{jj} & S_{jj} \end{pmatrix}.
$$

2
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 (AICc) 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

$$IC_m(j) = F(S, \Sigma_j) + m(j)$$

where $$S_{yy}\ell = S_{yy} - S_{\ell y} S_{\ell y}^{-1} S_{\ell y} (\ell = j, x)$$ 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:

- AIC: $$m(j) = p^2 + q^2 + p + q + 2pq_j$$,
- AICc: $$m(j) = (n-1)^2 \left( \frac{p + q_j}{n-p-q_j} - 2 + \frac{q}{n-q-2} - \frac{q_j}{n-q_j-2} - \frac{p + q}{n-1} \right)$$,
- BIC: $$m(j) = \left\{ \frac{(p + q)(p + q + 1)}{2} - p(q - q_j) \right\} \log n$$,
- CAIC: $$m(j) = \left\{ \frac{(p + q)(p + q + 1)}{2} - p(q - q_j) \right\} (1 + \log n)$$,
- HQC: $$m(j) = 2 \left\{ \frac{(p + q)(p + q + 1)}{2} - p(q - q_j) \right\} \log \log n$$.

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 \( n \rightarrow \infty \) (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 \( n \rightarrow \infty \) 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 \( y \) is

\[
y = \mu_y + \Sigma_{y,y}^{-1}(x_j - \mu_j) + \Sigma_{yy,j}^{1/2} \varepsilon, \tag{7}
\]

where \( \varepsilon \) is a \( p \)-dimensional vector with \( E[\varepsilon] = 0_p \), \( \text{Cov}[\varepsilon] = I_p \), \( 0_p \) is a \( p \)-dimensional vector of zeros, \( x_j \) is a \( q_j \)-dimensional vector with \( E[x_j] = \mu_j \), \( \text{Cov}[x_j] = \Sigma_{j,j} \), 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 \( \Sigma \), 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 \( x \) 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 \( x \) 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 \( y_1, \ldots, y_n, x_1, \ldots, x_n \) and \( \varepsilon_1, \ldots, \varepsilon_n \) be \( n \) independent vectors from \( y, x \) and \( \varepsilon \), respectively. Then, the \( Y, X \) and \( \varepsilon \) are the \( n \times p \), \( n \times q \) and \( n \times p \) matrices given by

\[
Y = (I_n - J_n)(y_1, \ldots, y_n)',
X = (I_n - J_n)(x_1, \ldots, x_n)',
\varepsilon = (I_n - J_n)(\varepsilon_1, \ldots, \varepsilon_n)',
\]
where \( J_n = 1_n(1_n^t 1_n)^{-1} 1_n^t \) and \( 1_n \) is an \( n \)-dimensional vector of ones. Suppose that \( \mathbf{X}_j \) denotes the \( n \times q_j \) matrix consisting of the columns of \( \mathbf{X} \) indexed by the elements of \( j \). By using these matrices, the matrix form of the true model (7) is expressed as

\[
\mathbf{Y} = \mathbf{X}_j, \Sigma^{-1}_{j,y} \Sigma_{j,y} + \varepsilon \Sigma_{y,y}^{1/2}.
\] (8)

Henceforth, for simplicity, \( \mathbf{X}_j, X, q_j \), are represented as \( \mathbf{X}_r \) and \( q_r \), 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 \( \mathbf{Y} \) given \( \mathbf{X} \).

We now describe two classes of \( j \) that express subsets of \( \mathbf{X} \) 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 \( j_r \) in (8), that is, \( \mathcal{J}_+ = \{ j \in \mathcal{J} | j_r \subseteq j \} \) and the underspecified models, which are the models that are not overspecified model, that is, \( \mathcal{J}_- = \mathcal{J} - \mathcal{J}_+ \). In particular, we express the minimum overspecified model that includes \( j \in \mathcal{J}_- \) as \( j_+ \), and so

\[
j_+ = j \cup j_r.
\] (9)

By using \( \text{IC}_m \) in (5), the best subset of \( \omega \), which is chosen by minimizing \( \text{IC}_m \), is written as

\[
\hat{j}_m = \arg \min_{j \in \mathcal{J}} \text{IC}_m(j).
\]

Let a \( p \times p \) noncentrality matrix be denoted by

\[
\Gamma_j \mathbf{G}_j = \Sigma_{y,y}^{-1/2} \Sigma_{j,y} \Sigma_{j,y}^{-1} \mathbf{X}_r (I_n - \mathbf{P}_j) \mathbf{X}_r \Sigma_{j,y} \Sigma_{j,y}^{-1} \Sigma_{y,y}^{-1/2},
\] (10)

where \( \Gamma_j \) is a \( p \times \gamma_j \) matrix with \( \text{rank}(\Gamma_j) = \gamma_j \) and \( \mathbf{P}_j = \mathbf{X}_j (\mathbf{X}_j^t \mathbf{X}_j)^{-1} \mathbf{X}_j^t \). It should be noted that \( \Gamma_j \Gamma_j^t = O_{p,p} \) holds if and only if \( j \in \mathcal{J}_+ \), where \( O_{n,p} \) is an \( n \times p \) matrix of zeros. Moreover, for \( j \in \mathcal{J}_- \), we define

\[
\mathbf{A}_j = (I_n - \mathbf{P}_j) \mathbf{X}_r \Sigma_{j,y}^{-1} \Sigma_{j,y} \Sigma_{y,y}^{-1/2}.
\]

It is easy to see from the definition of the noncentrality matrix in (10) that \( \mathbf{A}_j^t \mathbf{A}_j = \Gamma_j \Gamma_j^t \).

By using a singular value decomposition, \( \mathbf{A}_j \) can be rewritten as

\[
\mathbf{A}_j = \mathbf{H}_j \mathbf{L}_j^{1/2} \mathbf{G}_j,
\] (11)

where \( \mathbf{H}_j = (h_{j,1}, \ldots, h_{j,\gamma_j}) \) and \( \mathbf{G}_j = (g_{j,1}, \ldots, g_{j,\gamma_j}) \) are \( n \times \gamma_j \) and \( \gamma_j \times \gamma_j \) matrices, that satisfy \( \mathbf{H}_j^t \mathbf{H}_j = I_{\gamma_j} \) and \( \mathbf{G}_j^t \mathbf{G}_j = I_{\gamma_j} \), respectively, and \( \mathbf{L}_j = \text{diag}(\alpha_{j,1}, \ldots, \alpha_{j,\gamma_j}) \) is a diagonal matrix of order \( \gamma_j \) whose diagonal elements \( \alpha_{j,k} \) are the squared singular values of \( \mathbf{A}_j \), which are assumed to be \( \alpha_{j,1} \geq \cdots \geq \alpha_{j,\gamma_j} \).

Furthermore, let \( ||\mathbf{a}|| \) denote the Euclidean norm of the vector \( \mathbf{a} \). Then, in order to assess the consistency of \( \text{IC}_m \), the following assumption are necessary:

A1. The true model is included in the set of candidate models, that is, \( j_r \in \mathcal{J} \).

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

A4. $\forall j \in J_-, \lim_{p \to \infty} p^{-1} \Sigma_{j,y}^{-1} \Sigma_{y,j}^{-1} = \Psi_j$ exists and

$$\text{tr}(\Sigma_{j,y}^{-1} \Sigma_{y,j}^{-1} \Psi_j) > 0.$$ 

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 $\varepsilon$ and $x$ 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, $X_j$ in $\Gamma_j \Gamma'_j$ is not random. However, in CCA, $X_j$ in $\Gamma_j \Gamma'_j$ 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

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

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

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

Here, $\delta_{ab}$ 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 $\varepsilon \sim N_p(0_p, I_p)$. In general, the order of $\kappa_{4}^{(1)}$ is

$$\kappa_{4}^{(1)} = O(p^s) \text{ as } p \to \infty, \quad s \in [0, 2]. \quad (13)$$

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

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

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. Furthermore, 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 AICc 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 \( c_b = \min\{1, \min_{j \in \mathcal{F}_-} 1/(2(q_s - q_j))\} \) and \( \mathcal{F}_- \) is a set of candidate models given by

\[
\mathcal{F}_- = \{j \in J | q_s - q_j > 0\}. \tag{15}
\]

Corollary 1 shows that, when \( c_{n,p} \to c_0 \), the AICc 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, AICc, BIC, CAIC, and HQC.

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Consistency</th>
<th>Conditions</th>
</tr>
</thead>
<tbody>
<tr>
<td>AIC</td>
<td>Conditionally holds</td>
<td>( c_0 \in (0, c_a) )</td>
</tr>
<tr>
<td>AICc &amp; HQC</td>
<td>Holds</td>
<td>- - - -</td>
</tr>
<tr>
<td>BIC &amp; CAIC</td>
<td>Conditionally holds</td>
<td>( c_0 \in (0, c_b) )</td>
</tr>
</tbody>
</table>

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, AICc, BIC, CAIC, and HQC were evaluated by Monte Carlo simulations with 10,000 iterations each.

Let \( \nu_1 = (\nu_{1,1}, \ldots, \nu_{1,p})' \sim N_p(0_p, I_p) \), \( \nu_2 = (\nu_{2,1}, \ldots, \nu_{2,q})' \sim N_q(0_q, I_q) \), \( \delta_1, \delta_2 \sim \chi_5^2 \), \( \omega_1, \ldots, \omega_2, \ldots, \omega_2 \sim i.i.d. \chi_5^2 \) and \( \omega_2, \ldots, \omega_2, \ldots, \omega_2, \ldots, \omega_2 \) be mutually independent random vectors and variables. Then, \( \varepsilon = (\varepsilon_1, \ldots, \varepsilon_p)' \) and \( x = (x_1, \ldots, x_q)' \) were generated from the following five distributions, as in Yanagihara (2013):

- Distribution 1 (the multivariate normal distribution).

\[ \varepsilon = \nu_1, \quad x = \nu_2. \]
• Distribution 2 (a scale mixture of the multivariate normal distribution).
  \[
  \varepsilon = \sqrt{\frac{\delta_1}{6}} \nu_1, \ x = \sqrt{\frac{\delta_2}{6}} \nu_2.
  \]

• Distribution 3 (a location-scale mixture of the multivariate normal distribution).
  \[
  \varepsilon = B_1^{-1/2} \left\{ 10 \left( \sqrt{\frac{\delta_1}{6}} - \eta \right) \mathbf{1}_p + \sqrt{\frac{\delta_1}{6}} \nu_1 \right\},
  \]
  \[
  x = B_2^{-1/2} \left\{ 10 \left( \sqrt{\frac{\delta_2}{6}} - \eta \right) \mathbf{1}_q + \sqrt{\frac{\delta_2}{6}} \nu_2 \right\},
  \]
  where \( \eta = 15\sqrt{\pi/16}, \ B_1 = I_p + 100(1 - \eta^2)\mathbf{1}_p\mathbf{1}_p', \) and \( B_2 = I_q + 100(1 - \eta^2)\mathbf{1}_q\mathbf{1}_q'. \)

• Distribution 4 (the independent \( t \)-distribution).
  \[
  \varepsilon_a = \sqrt{3} \nu_{1,a} \sqrt{\frac{1}{\delta \omega_{1,a}}}, \ x_a = \sqrt{3} \nu_{2,a} \sqrt{\frac{1}{\delta \omega_{2,a}}}.
  \]

• 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 \( \mu_y \) and \( \mu_{j*} \) were generated from \( U(-4, 4) \) and \( U(-3, 3) \), respectively, and \( j* = 3 \). Then, \( y \) was obtained from the true model (7). The structure of \( \Sigma \) was prepared for the following four cases (cases 1 and 2 are the same settings as in Fujikoshi, 2014):

Case 1.
  \[
  \Sigma = \begin{pmatrix}
  I_5 & R' \\
  R & I_p
  \end{pmatrix}, \ R = (R_1, O_{5,p-q})', \ R_1 = \text{diag}(\rho_1, \ldots, \rho_5),
  \]
  \[
  \rho_1 = 2\rho, \ \rho_2 = 3\rho/2, \ \rho_3 = \rho, \ \rho_4 = \rho_5 = 0, \ \rho = \sqrt{\frac{(4p/21)}{p + 1 + (4p/21)}}.
  \]

Case 2 (the structure of \( \Sigma \) is the same as in Case 1).
  \[
  \rho_1 = \tilde{\rho}, \ \rho_2 = 3\tilde{\rho}/4, \ \rho_3 = \tilde{\rho}/2, \ \rho_4 = \rho_5 = 0, \ \tilde{\rho} = \sqrt{\frac{p}{p + 1} \frac{(4p/21)}{1 + (4p/21)}}.
  \]

Case 3. \( \Sigma = \Phi \Phi' \), where \( \Phi \) is a \( (p + 5) \times (p + 5) \) matrix whose elements are distributed from \( U(0, 1/p + 5) \).
Case 4. $\Sigma = \Phi \Phi'$, where $\Phi$ 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 $\varepsilon$ and $x$ are from Distributions 1, 2, 3, 4, and 5, respectively, when using the AIC, the 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 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

<table>
<thead>
<tr>
<th>n</th>
<th>p</th>
<th>AIC</th>
<th>AICc</th>
<th>BIC</th>
<th>CAIC</th>
<th>HQC</th>
<th>AIC</th>
<th>AICc</th>
<th>BIC</th>
<th>CAIC</th>
<th>HQC</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>5</td>
<td>80.01</td>
<td>79.24</td>
<td>31.31</td>
<td>15.29</td>
<td>67.22</td>
<td>62.36</td>
<td>56.11</td>
<td>8.54</td>
<td>2.48</td>
<td>37.42</td>
</tr>
<tr>
<td>200</td>
<td>10</td>
<td>94.55</td>
<td>95.03</td>
<td>17.95</td>
<td>4.88</td>
<td>76.07</td>
<td>93.47</td>
<td>92.95</td>
<td>12.51</td>
<td>2.98</td>
<td>68.61</td>
</tr>
<tr>
<td>500</td>
<td>25</td>
<td>99.58</td>
<td>99.88</td>
<td>1.18</td>
<td>0.06</td>
<td>83.03</td>
<td>99.66</td>
<td>99.93</td>
<td>12.86</td>
<td>1.24</td>
<td>97.99</td>
</tr>
<tr>
<td>1000</td>
<td>50</td>
<td>99.99</td>
<td>100.00</td>
<td>0.00</td>
<td>0.00</td>
<td>85.92</td>
<td>100.00</td>
<td>100.00</td>
<td>6.25</td>
<td>0.13</td>
<td>99.99</td>
</tr>
<tr>
<td>c0 = 0.05 Case 3</td>
<td>n</td>
<td>p</td>
<td>AIC</td>
<td>AICc</td>
<td>BIC</td>
<td>CAIC</td>
<td>HQC</td>
<td>AIC</td>
<td>AICc</td>
<td>BIC</td>
<td>CAIC</td>
</tr>
<tr>
<td>----</td>
<td>---</td>
<td>-----</td>
<td>------</td>
<td>-----</td>
<td>------</td>
<td>-----</td>
<td>-----</td>
<td>------</td>
<td>-----</td>
<td>------</td>
<td>-----</td>
</tr>
<tr>
<td>100</td>
<td>5</td>
<td>88.12</td>
<td>94.09</td>
<td>80.83</td>
<td>63.32</td>
<td>94.94</td>
<td>85.77</td>
<td>92.28</td>
<td>68.05</td>
<td>47.49</td>
<td>90.37</td>
</tr>
<tr>
<td>200</td>
<td>10</td>
<td>96.08</td>
<td>98.70</td>
<td>96.04</td>
<td>84.35</td>
<td>99.82</td>
<td>95.67</td>
<td>98.70</td>
<td>86.36</td>
<td>64.22</td>
<td>99.62</td>
</tr>
<tr>
<td>500</td>
<td>25</td>
<td>99.68</td>
<td>99.92</td>
<td>99.99</td>
<td>98.41</td>
<td>100.00</td>
<td>99.61</td>
<td>99.88</td>
<td>99.12</td>
<td>89.53</td>
<td>100.00</td>
</tr>
<tr>
<td>1000</td>
<td>50</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>99.97</td>
<td>100.00</td>
<td>100.00</td>
<td>99.59</td>
<td>100.00</td>
</tr>
</tbody>
</table>

| c0 = 0.1 Case 1 | n  | p | AIC | AICc | BIC | CAIC | HQC | AIC | AICc | 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|

| c0 = 0.1 Case 3 | n  | p | AIC | AICc | BIC | CAIC | HQC | AIC | AICc | 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|

| c0 = 0.2 Case 1 | n  | p | AIC | AICc | BIC | CAIC | HQC | AIC | AICc | 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| 46.68| 0.00| 0.00| 0.00| 0.00|99.69|4.13|0.00|0.00|0.00|

| c0 = 0.2 Case 3 | n  | p | AIC | AICc | BIC | CAIC | HQC | AIC | AICc | 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|

| c0 = 0.3 Case 1 | n  | p | AIC | AICc | BIC | CAIC | HQC | AIC | AICc | 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|

| c0 = 0.3 Case 3 | n  | p | AIC | AICc | BIC | CAIC | HQC | AIC | AICc | 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|

10
Table 3. Selection probabilities of the true model (%) in the Case of Distribution 2

<table>
<thead>
<tr>
<th>c_0 = 0.05</th>
<th>Case 1</th>
<th>Case 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>p</td>
<td>AIC</td>
</tr>
<tr>
<td>100</td>
<td>5</td>
<td>70.88</td>
</tr>
<tr>
<td>200</td>
<td>10</td>
<td>83.11</td>
</tr>
<tr>
<td>500</td>
<td>25</td>
<td>90.40</td>
</tr>
<tr>
<td>1000</td>
<td>50</td>
<td>92.04</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>c_0 = 0.05</th>
<th>Case 3</th>
<th>Case 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>p</td>
<td>AIC</td>
</tr>
<tr>
<td>100</td>
<td>5</td>
<td>82.10</td>
</tr>
<tr>
<td>200</td>
<td>10</td>
<td>93.23</td>
</tr>
<tr>
<td>500</td>
<td>25</td>
<td>98.62</td>
</tr>
<tr>
<td>1000</td>
<td>50</td>
<td>99.49</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>c_0 = 0.1</th>
<th>Case 1</th>
<th>Case 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>p</td>
<td>AIC</td>
</tr>
<tr>
<td>100</td>
<td>10</td>
<td>64.80</td>
</tr>
<tr>
<td>200</td>
<td>20</td>
<td>72.02</td>
</tr>
<tr>
<td>500</td>
<td>50</td>
<td>75.68</td>
</tr>
<tr>
<td>1000</td>
<td>100</td>
<td>76.70</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>c_0 = 0.2</th>
<th>Case 1</th>
<th>Case 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>p</td>
<td>AIC</td>
</tr>
<tr>
<td>100</td>
<td>10</td>
<td>83.68</td>
</tr>
<tr>
<td>200</td>
<td>20</td>
<td>93.50</td>
</tr>
<tr>
<td>500</td>
<td>50</td>
<td>97.27</td>
</tr>
<tr>
<td>1000</td>
<td>100</td>
<td>98.00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>c_0 = 0.3</th>
<th>Case 1</th>
<th>Case 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>p</td>
<td>AIC</td>
</tr>
<tr>
<td>100</td>
<td>10</td>
<td>47.43</td>
</tr>
<tr>
<td>200</td>
<td>40</td>
<td>50.05</td>
</tr>
<tr>
<td>500</td>
<td>100</td>
<td>49.43</td>
</tr>
<tr>
<td>1000</td>
<td>200</td>
<td>49.49</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>c_0 = 0.3</th>
<th>Case 3</th>
<th>Case 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>p</td>
<td>AIC</td>
</tr>
<tr>
<td>100</td>
<td>10</td>
<td>79.91</td>
</tr>
<tr>
<td>200</td>
<td>40</td>
<td>87.60</td>
</tr>
<tr>
<td>500</td>
<td>100</td>
<td>93.24</td>
</tr>
<tr>
<td>1000</td>
<td>200</td>
<td>96.31</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>c_0 = 0.3</th>
<th>Case 3</th>
<th>Case 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>p</td>
<td>AIC</td>
</tr>
<tr>
<td>100</td>
<td>30</td>
<td>37.47</td>
</tr>
<tr>
<td>200</td>
<td>60</td>
<td>36.78</td>
</tr>
<tr>
<td>500</td>
<td>150</td>
<td>34.22</td>
</tr>
<tr>
<td>1000</td>
<td>300</td>
<td>34.70</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>c_0 = 0.3</th>
<th>Case 3</th>
<th>Case 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>p</td>
<td>AIC</td>
</tr>
<tr>
<td>100</td>
<td>30</td>
<td>73.63</td>
</tr>
<tr>
<td>200</td>
<td>60</td>
<td>82.74</td>
</tr>
<tr>
<td>500</td>
<td>150</td>
<td>89.72</td>
</tr>
<tr>
<td>1000</td>
<td>300</td>
<td>95.06</td>
</tr>
</tbody>
</table>
Table 4. Selection probabilities of the true model (%) in the Case of Distribution 3

<table>
<thead>
<tr>
<th>n</th>
<th>p</th>
<th>AIC</th>
<th>AICc</th>
<th>BIC</th>
<th>CAIC</th>
<th>HQC</th>
<th>AIC</th>
<th>AICc</th>
<th>BIC</th>
<th>CAIC</th>
<th>HQC</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>5</td>
<td>92.07</td>
<td>95.45</td>
<td>82.79</td>
<td>73.59</td>
<td>93.82</td>
<td>91.70</td>
<td>94.84</td>
<td>80.12</td>
<td>70.24</td>
<td>92.08</td>
</tr>
<tr>
<td>200</td>
<td>20</td>
<td>97.86</td>
<td>97.44</td>
<td>78.98</td>
<td>69.86</td>
<td>93.67</td>
<td>98.16</td>
<td>98.38</td>
<td>84.30</td>
<td>76.80</td>
<td>95.65</td>
</tr>
<tr>
<td>500</td>
<td>50</td>
<td>99.28</td>
<td>98.71</td>
<td>73.05</td>
<td>74.41</td>
<td>93.57</td>
<td>99.70</td>
<td>99.37</td>
<td>85.48</td>
<td>78.68</td>
<td>97.58</td>
</tr>
<tr>
<td>1000</td>
<td>100</td>
<td>99.50</td>
<td>98.79</td>
<td>67.48</td>
<td>57.84</td>
<td>92.03</td>
<td>98.88</td>
<td>99.69</td>
<td>83.60</td>
<td>77.61</td>
<td>97.11</td>
</tr>
<tr>
<td>100</td>
<td>10</td>
<td>94.17</td>
<td>99.37</td>
<td>99.96</td>
<td>99.82</td>
<td>99.79</td>
<td>94.12</td>
<td>99.53</td>
<td>99.95</td>
<td>99.89</td>
<td>99.84</td>
</tr>
<tr>
<td>200</td>
<td>20</td>
<td>98.79</td>
<td>99.96</td>
<td>99.99</td>
<td>99.99</td>
<td>100.00</td>
<td>98.87</td>
<td>99.89</td>
<td>99.99</td>
<td>99.99</td>
<td>100.00</td>
</tr>
<tr>
<td>500</td>
<td>50</td>
<td>99.99</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>99.99</td>
<td>100.00</td>
<td>100.00</td>
<td>99.99</td>
<td>100.00</td>
</tr>
<tr>
<td>1000</td>
<td>100</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
</tr>
<tr>
<td>100</td>
<td>0.1</td>
<td>94.17</td>
<td>99.37</td>
<td>99.96</td>
<td>99.82</td>
<td>99.79</td>
<td>94.12</td>
<td>99.53</td>
<td>99.95</td>
<td>99.89</td>
<td>99.84</td>
</tr>
<tr>
<td>200</td>
<td>20</td>
<td>98.79</td>
<td>99.96</td>
<td>99.99</td>
<td>99.99</td>
<td>100.00</td>
<td>98.87</td>
<td>99.89</td>
<td>99.99</td>
<td>99.99</td>
<td>100.00</td>
</tr>
<tr>
<td>500</td>
<td>50</td>
<td>99.99</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>99.99</td>
<td>100.00</td>
<td>100.00</td>
<td>99.99</td>
<td>100.00</td>
</tr>
<tr>
<td>1000</td>
<td>100</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
</tr>
<tr>
<td>100</td>
<td>0.2</td>
<td>94.17</td>
<td>99.37</td>
<td>99.96</td>
<td>99.82</td>
<td>99.79</td>
<td>94.12</td>
<td>99.53</td>
<td>99.95</td>
<td>99.89</td>
<td>99.84</td>
</tr>
<tr>
<td>200</td>
<td>20</td>
<td>98.79</td>
<td>99.96</td>
<td>99.99</td>
<td>99.99</td>
<td>100.00</td>
<td>98.87</td>
<td>99.89</td>
<td>99.99</td>
<td>99.99</td>
<td>100.00</td>
</tr>
<tr>
<td>500</td>
<td>50</td>
<td>99.99</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>99.99</td>
<td>100.00</td>
<td>100.00</td>
<td>99.99</td>
<td>100.00</td>
</tr>
<tr>
<td>1000</td>
<td>100</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
</tr>
<tr>
<td>100</td>
<td>0.3</td>
<td>94.17</td>
<td>99.37</td>
<td>99.96</td>
<td>99.82</td>
<td>99.79</td>
<td>94.12</td>
<td>99.53</td>
<td>99.95</td>
<td>99.89</td>
<td>99.84</td>
</tr>
<tr>
<td>200</td>
<td>20</td>
<td>98.79</td>
<td>99.96</td>
<td>99.99</td>
<td>99.99</td>
<td>100.00</td>
<td>98.87</td>
<td>99.89</td>
<td>99.99</td>
<td>99.99</td>
<td>100.00</td>
</tr>
<tr>
<td>500</td>
<td>50</td>
<td>99.99</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>99.99</td>
<td>100.00</td>
<td>100.00</td>
<td>99.99</td>
<td>100.00</td>
</tr>
<tr>
<td>1000</td>
<td>100</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
</tr>
</tbody>
</table>

AIC: Akaike Information Criterion; AICc: Corrected Akaike Information Criterion; BIC: Bayesian Information Criterion; CAIC: Consistent Akaike Information Criterion; HQC: Hannan-Quinn Information Criterion; P: Probability.
<table>
<thead>
<tr>
<th>n</th>
<th>p</th>
<th>AIC</th>
<th>AICc</th>
<th>BIC</th>
<th>CAIC</th>
<th>HQC</th>
<th>AIC</th>
<th>AICc</th>
<th>BIC</th>
<th>CAIC</th>
<th>HQC</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>5</td>
<td>78.20</td>
<td>77.81</td>
<td>32.68</td>
<td>16.90</td>
<td>66.21</td>
<td>62.22</td>
<td>55.45</td>
<td>9.93</td>
<td>3.10</td>
<td>38.18</td>
</tr>
<tr>
<td>200</td>
<td>10</td>
<td>92.82</td>
<td>93.13</td>
<td>20.00</td>
<td>6.68</td>
<td>74.06</td>
<td>91.95</td>
<td>91.39</td>
<td>15.13</td>
<td>4.56</td>
<td>68.45</td>
</tr>
<tr>
<td>500</td>
<td>25</td>
<td>99.54</td>
<td>99.71</td>
<td>2.64</td>
<td>0.28</td>
<td>80.76</td>
<td>95.55</td>
<td>99.87</td>
<td>17.48</td>
<td>3.16</td>
<td>96.38</td>
</tr>
<tr>
<td>1000</td>
<td>50</td>
<td>99.99</td>
<td>99.98</td>
<td>0.10</td>
<td>0.03</td>
<td>84.23</td>
<td>99.98</td>
<td>100.00</td>
<td>10.39</td>
<td>0.93</td>
<td>99.72</td>
</tr>
<tr>
<td>100</td>
<td>5</td>
<td>87.88</td>
<td>93.62</td>
<td>79.14</td>
<td>62.57</td>
<td>94.41</td>
<td>86.54</td>
<td>92.39</td>
<td>68.39</td>
<td>48.09</td>
<td>90.36</td>
</tr>
<tr>
<td>200</td>
<td>10</td>
<td>95.49</td>
<td>98.14</td>
<td>94.70</td>
<td>82.68</td>
<td>99.87</td>
<td>95.16</td>
<td>98.45</td>
<td>85.14</td>
<td>64.50</td>
<td>99.45</td>
</tr>
<tr>
<td>500</td>
<td>25</td>
<td>99.63</td>
<td>99.90</td>
<td>99.89</td>
<td>98.06</td>
<td>100.00</td>
<td>99.68</td>
<td>99.94</td>
<td>98.36</td>
<td>87.81</td>
<td>100.00</td>
</tr>
<tr>
<td>1000</td>
<td>50</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>99.99</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
</tr>
<tr>
<td>100</td>
<td>10</td>
<td>69.39</td>
<td>48.63</td>
<td>3.18</td>
<td>0.34</td>
<td>32.22</td>
<td>65.19</td>
<td>43.49</td>
<td>2.12</td>
<td>0.24</td>
<td>27.32</td>
</tr>
<tr>
<td>200</td>
<td>20</td>
<td>84.31</td>
<td>62.58</td>
<td>0.12</td>
<td>0.02</td>
<td>19.98</td>
<td>91.63</td>
<td>77.31</td>
<td>0.62</td>
<td>0.04</td>
<td>34.25</td>
</tr>
<tr>
<td>500</td>
<td>50</td>
<td>96.47</td>
<td>79.43</td>
<td>0.02</td>
<td>0.00</td>
<td>3.66</td>
<td>99.85</td>
<td>98.55</td>
<td>0.06</td>
<td>0.02</td>
<td>38.67</td>
</tr>
<tr>
<td>1000</td>
<td>100</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>99.99</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>99.20</td>
<td>100.00</td>
</tr>
<tr>
<td>100</td>
<td>10</td>
<td>92.95</td>
<td>94.50</td>
<td>43.66</td>
<td>16.67</td>
<td>88.52</td>
<td>91.42</td>
<td>88.14</td>
<td>28.99</td>
<td>7.68</td>
<td>79.38</td>
</tr>
<tr>
<td>200</td>
<td>20</td>
<td>98.68</td>
<td>99.81</td>
<td>41.77</td>
<td>10.51</td>
<td>97.82</td>
<td>98.82</td>
<td>99.54</td>
<td>20.37</td>
<td>2.26</td>
<td>93.68</td>
</tr>
<tr>
<td>500</td>
<td>50</td>
<td>99.98</td>
<td>100.00</td>
<td>34.09</td>
<td>3.09</td>
<td>100.00</td>
<td>99.98</td>
<td>100.00</td>
<td>12.52</td>
<td>0.23</td>
<td>99.94</td>
</tr>
<tr>
<td>1000</td>
<td>100</td>
<td>100.00</td>
<td>100.00</td>
<td>30.53</td>
<td>0.78</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>7.42</td>
<td>0.15</td>
<td>100.00</td>
</tr>
<tr>
<td>100</td>
<td>20</td>
<td>44.40</td>
<td>2.83</td>
<td>0.01</td>
<td>0.00</td>
<td>3.79</td>
<td>55.29</td>
<td>6.02</td>
<td>0.03</td>
<td>0.00</td>
<td>7.41</td>
</tr>
<tr>
<td>200</td>
<td>40</td>
<td>46.23</td>
<td>1.21</td>
<td>0.00</td>
<td>0.00</td>
<td>0.17</td>
<td>74.94</td>
<td>9.11</td>
<td>0.00</td>
<td>0.00</td>
<td>2.40</td>
</tr>
<tr>
<td>500</td>
<td>100</td>
<td>46.74</td>
<td>0.21</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>93.21</td>
<td>8.66</td>
<td>0.00</td>
<td>0.00</td>
<td>0.09</td>
</tr>
<tr>
<td>1000</td>
<td>200</td>
<td>46.50</td>
<td>0.03</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>98.87</td>
<td>6.62</td>
<td>0.00</td>
<td>0.00</td>
<td>0.01</td>
</tr>
<tr>
<td>100</td>
<td>30</td>
<td>28.13</td>
<td>0.01</td>
<td>0.00</td>
<td>0.00</td>
<td>0.24</td>
<td>44.62</td>
<td>0.02</td>
<td>0.00</td>
<td>0.00</td>
<td>1.77</td>
</tr>
<tr>
<td>200</td>
<td>60</td>
<td>23.64</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>54.96</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.08</td>
</tr>
<tr>
<td>500</td>
<td>150</td>
<td>13.48</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>67.55</td>
<td>0.01</td>
<td>0.00</td>
<td>0.00</td>
<td>0.01</td>
</tr>
<tr>
<td>1000</td>
<td>300</td>
<td>5.65</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>78.08</td>
<td>0.01</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>100</td>
<td>30</td>
<td>88.89</td>
<td>0.41</td>
<td>0.01</td>
<td>0.00</td>
<td>19.33</td>
<td>84.12</td>
<td>0.08</td>
<td>0.00</td>
<td>98.79</td>
<td>99.98</td>
</tr>
<tr>
<td>200</td>
<td>60</td>
<td>99.23</td>
<td>1.56</td>
<td>0.00</td>
<td>0.00</td>
<td>11.25</td>
<td>98.24</td>
<td>0.50</td>
<td>0.00</td>
<td>99.47</td>
<td>99.99</td>
</tr>
<tr>
<td>500</td>
<td>150</td>
<td>13.89</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>6.66</td>
<td>100.00</td>
<td>4.92</td>
<td>0.00</td>
<td>99.83</td>
<td>100.00</td>
</tr>
<tr>
<td>1000</td>
<td>300</td>
<td>96.20</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>51.00</td>
<td>100.00</td>
<td>91.87</td>
<td>0.00</td>
<td>99.96</td>
<td>100.00</td>
</tr>
</tbody>
</table>
Table 6. Selection probabilities of the true model (%) in the Case of Distribution 5

<table>
<thead>
<tr>
<th>$c_0$</th>
<th>Case 1</th>
<th>Case 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$n$</td>
<td>$p$</td>
</tr>
<tr>
<td>0.05</td>
<td>100</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>50</td>
</tr>
<tr>
<td>0.1</td>
<td>100</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>50</td>
</tr>
<tr>
<td>0.2</td>
<td>100</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>100</td>
</tr>
<tr>
<td>0.3</td>
<td>100</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>100</td>
</tr>
<tr>
<td>0.4</td>
<td>100</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>100</td>
</tr>
</tbody>
</table>

Note: $c_0$ represents the significance level of the model selection criteria; $n$ and $p$ denote the sample size and the number of parameters, respectively; AIC, AICc, BIC, CAIC, and HQC are the Akaike Information Criterion, the corrected Akaike Information Criterion, the Bayesian Information Criterion, the Consistent Akaike Information Criterion, and the Hannan-Quinn Information Criterion, respectively.
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, AICc, 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 AICc 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), \( \mathbf{x} \) was not a random vector. Hence, we extended Yanagihara’s method to the case that \( \mathbf{x} \) is a random vector.

The results of our analysis and simulations confirmed that the AIC and AICc 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 \( \text{IC}_m \) 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 \( \mathbf{A} \),

\[
E \left[ \text{tr} \left\{ (\mathbf{E}' \mathbf{A} \mathbf{E})^2 \right\} \right] = \kappa_4^{(1)} \sum_{a=1}^{n} ((\mathbf{A})_{aa})^2 + p(p+1)\text{tr}(\mathbf{A}^2) + p\text{tr}(\mathbf{A})^2;
\]

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

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
\[ \forall \ell \in \mathcal{J} \setminus \{j\}, \quad \frac{1}{b_{j,\ell}} \{ \text{IC}_m(\ell) - \text{IC}_m(j) \} \geq T_{j,\ell} \Rightarrow \tau_{j,\ell} > 0 \Rightarrow P(\hat{j}_m = j) \to 1. \]

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 $W$ be an $n \times n$ random matrix, defined by $W = \mathcal{E}(\mathcal{E}')^{-1}\mathcal{E}'$. Then, for any $\ell \in \mathcal{J}$, we obtain
\[ \frac{1}{n-1} X'_{\ell}W X_{\ell} \overset{p}{\to} c_0 \Sigma_{\ell \ell}. \]

Lemma 4 Let $\lambda_{\max}(A)$ denote the maximum eigenvalue of $A$, and let $V_j$ be a $p \times p$ matrix defined by
\[ V_j = \frac{1}{n} \mathcal{E}'(I_n - P_j - H_j H_j') \mathcal{E}, \]
where $P_j$ and $H_j$ are given by (10) and (11), respectively. If assumption A2 holds, $\lambda_{\max}(V_j) = O_p(n^{1/2})$ is satisfied.

Lemma 5 If assumptions A2 and A4 hold, $\alpha_{j,1} = O_p(np)$ is satisfied, and $\lim \inf_{n \to \infty} \alpha_{j,1}/(np) > 0$, where $\alpha_{j,1}$ is the maximum diagonal element of $L_j$ 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
\[ D(j, \ell) = \log \frac{|S_{y_jy_j}|}{|S_{y_jy_\ell}|}. \]

Note that
\[ \text{IC}_m(j) - \text{IC}_m(j_*) = (n - 1)D(j, j_*) + m(j) - m(j_*). \]

From Lemma 2, we see that to obtain the conditions on $m(j)$ such that $\text{IC}_m(j)$ 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 $P_j Y = P_j \mathcal{E}$ holds for all $j$, since $X_*$ 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:
\[ D(j, j_*) = \log \left| \frac{Y' (I_n - P_j) Y}{Y' (I_n - P_{j_*}) Y} \right| = \log \left| \frac{\mathcal{E}' (I_n - P_j) \mathcal{E}}{\mathcal{E}' (I_n - P_{j_*}) \mathcal{E}} \right| \\
= \log \left| \frac{I_n - (\mathcal{E}' \mathcal{E})^{-1} \mathcal{E}' P_j \mathcal{E}}{I_n - (\mathcal{E}' \mathcal{E})^{-1} \mathcal{E}' P_{j_*} \mathcal{E}} \right| \\
= \log \left| \frac{X_j' X_j - X_{j_*}' W X_{j_*}}{X_j' X_j - X_{j_*}' W X_{j_*}} \right|. \]
Moreover, it follows from Lemma 1 that it is easy to see that

\[
D(j, j_s) \overset{p}{\to} (q_j - q_{j_s}) \log(1 - c_0).
\]  

(A1)

Next, we show the convergence in probability of a lower bound on \( D(j, j_s) / \log p \) when \( j \in J_- \). It follows that for all \( j \in J_- \),

\[
D(j, j_s) = \log \left| (L_j^{1/2} G_j' + H_j' \mathcal{E})(L_j^{1/2} G_j' + H_j' \mathcal{E}) + nV_j \right| \frac{\mathcal{E}'(I_n - P_{j_s})\mathcal{E}}{n} \\
= \log \left| I_p + \sum_{a=1}^{G_j} V_j^{-1}(\sqrt{\alpha_{j,a}} g_{j,a} + \mathcal{E}h_{j,a})(\sqrt{\alpha_{j,a}} g_{j,a} + \mathcal{E}'h_{j,a}) \right| nV_j \\
+ \log \left| \mathcal{E}'(I_n - P_{j_s})\mathcal{E} \right| \\
\geq \log \left| I_p + \frac{V_j^{-1}(\sqrt{\alpha_{j,1}} g_{j,1} + \mathcal{E}h_{j,1})(\sqrt{\alpha_{j,1}} g_{j,1} + \mathcal{E}'h_{j,1})}{n} \right| nV_j \\
+ \log \left| \mathcal{E}'(I_n - P_{j_s})\mathcal{E} \right| \\
= \log \left\{ 1 + \frac{(\sqrt{\alpha_{j,1}} g_{j,1} + \mathcal{E}h_{j,1})'(\sqrt{\alpha_{j,1}} g_{j,1} + \mathcal{E}'h_{j,1})}{n} \right\} \\
+ \log \left| \mathcal{E}'(I_n - P_{j_s})\mathcal{E} \right| - \log \lambda_{\text{max}}(V_j) \\
= D_1(j) + D_2(j) + D_3(j),
\]  

(A2)

where

\[
D_1(j) = \log \left\{ \lambda_{\text{max}}(V_j) + p \xi_j \right\}, \\
D_2(j) = \log \left| \mathcal{E}'(I_n - P_{j_s})\mathcal{E} \right|, \\
D_3(j) = -\log \lambda_{\text{max}}(V_j)
\]

and \( \xi_j = (\sqrt{\alpha_{j,1}} g_{j,1} + \mathcal{E}h_{j,1})'(\sqrt{\alpha_{j,1}} g_{j,1} + \mathcal{E}'h_{j,1})/(np) \).

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

\[
E[h_{j,1} \mathcal{E} \mathcal{E}'h_{j,1}] = p.
\]

Moreover, it follows from Lemma 1 that

\[
E[(h_{j,1} \mathcal{E} \mathcal{E}'h_{j,1} - p)^2] = \kappa_4^{(1)} \sum_{a=1}^{G_j} (h_{j,a} h_{j,1}^{'})^2 \geq 2p \\
= O(\max\{p, p^*\}),
\]

17
where $\kappa_4^{(1)}$ is given by (12), and $s$ is a positive constant given by (13). Hence, we have
\[ h_j' \mathcal{E} \mathcal{E}' h_{j,1} = p + O_p(\max\{p^{1/2}, p^{s/2}\}) = O_p(p). \tag{A3} \]
Moreover, note that $g_{j,1}g_{j,1}'$ is an idempotent matrix,
\[ \left( \sqrt{\alpha_j h_{j,1} g_{j,1}' \mathcal{E} \mathcal{E}' h_{j,1}} \right)^2 = \alpha_j h_{j,1} g_{j,1}' \mathcal{E} \mathcal{E}' h_{j,1} \leq \alpha_j h_{j,1}' \mathcal{E} \mathcal{E}' h_{j,1} = O_p(np^2). \]
This implies that
\[ \sqrt{\alpha_j g_{j,1}' \mathcal{E} h_{j,1}} = O_p(n^{1/2}p). \tag{A4} \]
From Lemma 5, (A3), and (A4), we have
\[ \xi_j = O_p(1). \tag{A5} \]
By using (A5) and Lemma 4, we obtain
\[ \frac{1}{\log p} D_1(j) = \frac{1}{\log p} \log \left\{ \lambda_{\max}(V_j) + p\xi_j \right\} = \frac{1}{\log p} \log \left\{ \frac{1}{p} \lambda_{\max}(V_j) + \xi_j \right\} + 1 \overset{P}{\to} 1. \tag{A6} \]
Next, we evaluate the asymptotic behavior of $D_2(j)$ in (A2). From Lemma 3 and the result $(I_n - P_j - H_jH_j') (I_n - P_j) = I_n - P_j - H_jH_j'$, we can see that
\[ D_2(j) \leq \log \frac{|\mathcal{E}'(I_n - P_j) \mathcal{E}|}{|\mathcal{E}'(I_n - P_{j+}) \mathcal{E}|} = \log \frac{|X_j'X_j - X_j'WX_j||X_j'X_*|}{|X_*'X_* - X_*'WX_*||X_j'X_j|} \overset{P}{\to} (k_j - k_{j+}) \log(1 - c_0), \]
where $W$ is given in Lemma 3. It follows that $(I_n - P_{j+}) (I_n - P_j - H_jH_j') = I_n - P_{j+}$, where $j_+$ is given by (9). Thus, we also have
\[ D_2(j) \geq \log \frac{|\mathcal{E}'(I_n - P_{j+}) \mathcal{E}|}{|\mathcal{E}'(I_n - P_{j+}) \mathcal{E}|} = \log \frac{|X_{j+}'X_{j+} - X_{j+}'WX_{j+}||X_{j+}'X_*|}{|X_*'X_* - X_*'WX_*||X_{j+}'X_{j+}|} \overset{P}{\to} (k_{j+} - k_j) \log(1 - c_0). \]
The above upper and lower bounds on $D_2(j)$ imply that
\[ \frac{1}{\log p} D_2(j) \overset{P}{\to} 0. \tag{A7} \]
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

$$D_3(j) = \frac{1}{2} \log p - \log \left( \frac{\lambda_{\text{max}}(V_j)}{\sqrt{p}} \right) \geq \frac{1}{2} \log p - \left\{ \frac{\lambda_{\text{max}}(V_j)}{\sqrt{p}} - 1 \right\} = D_{3,1}(j).$$

It follows from Lemma 4 that

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

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

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

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 $\text{AIC}_c$. According to an expansion of $m(j) - m(j_*)$ in the $\text{AIC}_c$, the differences between the penalty terms of the $\text{AIC}_c$s are

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

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

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

Hence, the convergence of the differences between the penalty terms of the AICs and those of the $\text{AIC}_c$s is

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

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

$$\lim_{c_{n,p} \to 0} \frac{1}{p} \{m(j) - m(j_*)\} = \begin{cases} 2(q_j - q_*) & \text{(AIC)} \\ (q_j - q_*) \{(1 - c_0)^{-1} + (1 - c_0)^{-2}\} & \text{(AIC}_c) \end{cases}.$$

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 AICc 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

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

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

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

When \( j \in \mathcal{S}_- \cap \mathcal{J}_- \), condition C2 is satisfied because \( c_0(q_j - q_{j_*}) \geq 0 \) holds, where \( \mathcal{S}_- \) is given by (15). When \( j \in \mathcal{S}_- \), then for all \( j \in \mathcal{S}_- \), 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 \( X_\ell = (x_1, \ldots, x_{q_\ell}) \), let \( x_k = (x_{1k}, \ldots, x_{nk})' \), and let \( w_{ab} \) be the \((a, b)\)th element of \( W \). Then, \( x_\ell' W x_\ell \), which is the \((s, t)\)th element of \( X_\ell' W X_\ell \), is expressed as

\[
x_\ell' W x_\ell = \sum_{a=1}^{n} x_{as} x_{at} w_{aa} + \sum_{a \neq b}^{n} x_{as} x_{bt} w_{ab}.
\]  

(A11)
Moreover, we can calculate

\[ (x_s'Wx_t)^2 = \sum_{a=1}^{n} x_{as}^2 w_{aa} + \sum_{a \neq b \neq c \neq d} x_{as}x_{bs}x_{ct}w_{ab}w_{cd} \]

\[ + \sum_{a \neq b} \left\{ x_{as}x_{bs}x_{at}x_bl (w_{aa}w_{bb} + w_{ab}^2) + x_{as}x_{bt}w_{ab}^2 + 2(x_{as}x_{at}x_{bt}) \right\} \]

\[ + x_{as}x_{bs}x_{at}w_{aa}w_{ab} + \sum_{a \neq b \neq c} \left\{ 2x_{as}x_{bs}x_{ct} + (x_{as}x_{at}x_{ct} \right\} \]

\[ + 2x_{as}x_{bs}x_{at}x_{ct} + x_{bs}x_{cs}x_{at}w_{ab}w_{ac} \}, \quad (A12) \]

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

\[ \sum_{a,b}^{n} x_{as}x_{bt} = \sum_{a=1}^{n} x_{as} = \sum_{a=1}^{n} x_{at} = 0, \quad \sum_{a \neq b}^{n} x_{as}x_{bt} = -x_s'x_t, \]

\[ \sum_{a \neq b}^{n} x_{as}x_{bs}x_{at}x_bl = (x_s'x_t)^2 - \sum_{a=1}^{n} x_{as}^2 x_{at}, \quad \sum_{a \neq b}^{n} x_{as}^2 x_{bt} = x_s'x_s'x_t - \sum_{a=1}^{n} x_{as}^2 x_{at}, \]

\[ \sum_{a \neq b}^{n} x_{as}^2 x_{at}x_bl = \sum_{a \neq b}^{n} x_{as}x_{bs}x_{at}^2 = -\sum_{a=1}^{n} x_{as}^2 x_{at}, \]

\[ \sum_{a \neq b \neq c}^{n} x_{as}x_{bs}x_{at}x_{ct} = \sum_{a \neq b \neq c}^{n} x_{as}x_{bt}x_{ct} = \sum_{a \neq b \neq c}^{n} x_{bs}x_{cs}x_{at}^2 \]

\[ = \sum_{a=1}^{n} x_{as}^2 x_{at}^2 + \sum_{a \neq b}^{n} x_{at}x_{bt}x_{as}x_{bs}, \quad (A13) \]

Note that \( x_s'x_t \) is the \((s, t)\)th element of \( X_s'X_t \), and \((n-1)^{-1}X_s'X_t \stackrel{p}{\approx} \Sigma_{\ell \ell} \). Here, since \( W \) is a symmetric idempotent matrix and \( W1_n = 0_n \) holds, we obtain the following equations:

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

\[ \text{tr}(W) = \sum_{a=1}^{n} w_{aa} = p, \quad \text{tr}(W^2) = \sum_{a=1}^{n} w_{aa}^2 + \sum_{a \neq b}^{n} w_{ab}^2 = p, \]

\[ \text{tr}(W)^2 = \sum_{a=1}^{n} w_{aa}^2 + \sum_{a \neq b}^{n} w_{aa} w_{bb} = p^2, \quad 1_n' W 1_n = \sum_{a=1}^{n} w_{aa} + \sum_{a \neq b}^{n} w_{ab} = 0, \]

\[ 1_n' W^2 1_n = \sum_{a=1}^{n} w_{aa}^2 + \sum_{a \neq b}^{n} (2w_{aa} w_{ab} + w_{ab}^2) + \sum_{a \neq b \neq c} w_{ab} w_{ac} = 0, \quad (A15) \]

\[ \text{tr}(W) 1_n' W 1_n = \sum_{a=1}^{n} w_{aa} + \sum_{a \neq b}^{n} (2w_{aa} w_{ab} + w_{aa} w_{bb}) + \sum_{a \neq b \neq c} w_{aa} w_{bc} = 0, \]

\[ (1_n' W 1_n)^2 = \sum_{a=1}^{n} w_{aa}^2 + \sum_{a \neq b}^{n} (w_{aa} w_{ab} + 2w_{ab}^2 + 4w_{aa} w_{ab}) \]

\[ + 2 \sum_{a \neq b \neq c} (w_{aa} w_{bc} + 2w_{ab} w_{ac}) + \sum_{a \neq b \neq c \neq d} w_{ab} w_{cd} = 0. \]

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 (A15), and for \( a \neq b \neq c \neq d \), we obtain

\[ p = n E[w_{aa}], \]

\[ p = n E[w_{aa}^2] + n(n-1) E[w_{ab}^2], \]

\[ p^2 = n E[w_{aa}^2] + n(n-1) E[w_{aa} w_{bb}], \]

\[ 0 = n E[w_{aa}] + n(n-1) E[w_{ab}], \]

\[ 0 = n E[w_{aa}^2] + n(n-1) (2E[w_{aa} w_{ab}] + E[w_{ab}^2]) + n(n-1)(n-2) E[w_{ab} w_{ac}], \]

\[ 0 = n E[w_{aa}^2] + n(n-1) (2E[w_{aa} w_{ab}] + E[w_{aa} w_{bb}]) + n(n-1)(n-2) E[w_{aa} w_{bc}], \]

\[ 0 = n E[w_{aa}^2] + n(n-1) (E[w_{aa} w_{bb}] + 2E[w_{ab}^2] + 4E[w_{aa} w_{ab}]) \]

\[ + 2n(n-1)(n-2) (E[w_{aa} w_{bc}] + 2E[w_{ab} w_{ac}]) \]

\[ + n(n-1)(n-2)(n-3) E[w_{ab} w_{cd}] . \]

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

\[ E[w_{aa}] = c_{n,p}, \quad E[w_{ab}] = O(n^{-1}), \]

\[ E[w_{aa}^2] = O(1), \quad E[w_{aa} w_{bb}] = c_{n,p}^2 + O(n^{-1}), \]

\[ E[w_{ab}^2] = O(n^{-1}), \quad E[w_{aa} w_{ab}] = O(n^{-1}), \]

\[ E[w_{aa} w_{bc}] = O(n^{-1}), \quad E[w_{ab} w_{ac}] = O(n^{-2}), \]

\[ E[w_{ab} w_{cd}] = O(n^{-2}), \quad (A17) \]

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 $\sigma_{st}$ be the $(s,t)$th element of $\Sigma_{\ell\ell}$. Then, by using (A11), (A12), (A13), and (A17), we have
\[
\frac{1}{n-1} E[x'_s W x_t] \rightarrow c_0 \sigma_{st}, \quad \frac{1}{(n-1)^2} E[(x'_s W x_t)^2] \rightarrow c_0^2 \sigma_{st}^2.
\]
The above equations directly imply that $(n-1)^{-1} \operatorname{Var}[x'_s W x_t] \rightarrow 0$ as $c_{n,p} \rightarrow 0$. Hence, the $(s,t)$th element of $X'_s W X_\ell$ converges, as follows:
\[
\frac{1}{n-1} x'_s W x_t \overset{p}{\rightarrow} c_0 \sigma_{st}.
\]
Therefore, Lemma 3 is proved.

**A5. Proof of Lemma 4**

It follows from elementary linear algebra that
\[
\lambda_{\text{max}}(V_j) \leq \lambda_{\text{max}} \left( \frac{1}{n} \mathcal{E}' \mathcal{E} \right) \leq \sqrt{\frac{1}{n^2} \text{tr} \left\{ (\mathcal{E}' \mathcal{E})^2 \right\}}.
\]
From Lemma 1, we can see that
\[
E \left[ \frac{1}{n^2} \text{tr} \left\{ (\mathcal{E}' \mathcal{E})^2 \right\} \right] = \frac{1}{n} \kappa_4^{(1)} + \frac{1}{n} p(1 + p) + p = O(p).
\]
The above equation and Jensen’s inequality lead us to the equation
\[
E \left[ \sqrt{\frac{1}{n^2} \text{tr} \left\{ (\mathcal{E}' \mathcal{E})^2 \right\}} \right] \leq \sqrt{E \left[ \frac{1}{n^2} \text{tr} \left\{ (\mathcal{E}' \mathcal{E})^2 \right\} \right]} = O(p^{1/2}).
\]
This directly implies that $n^{-1} \operatorname{tr}\{((\mathcal{E}' \mathcal{E})^2)^{1/2}\} = O_p(p^{1/2})$. Hence, Lemma 4 is proved.

**A6. Proof of Lemma 5**

It follows from elementary linear algebra that
\[
\frac{1}{np} \alpha_{j,1} = \frac{1}{np} \lambda_{\text{max}}(L_j) \leq \frac{1}{np} \text{tr}(L_j)
\]
\[
= \frac{1}{np} \text{tr}(\Gamma_j' \Gamma_j)
\]
\[
= \frac{1}{np} \text{tr} \left\{ X'_j (I_n - P_j) X_j \Sigma_{jj,ss}^{-1} \Sigma_{ss,yy}^{-1} \Sigma_{yy,ss}^{-1} \Sigma_{ss,yy}^{-1} \right\}
\]
\[
\leq \frac{1}{np} \text{tr} \left\{ X'_j X_j \Sigma_{jj,ss}^{-1} \Sigma_{ss,yy}^{-1} \Sigma_{yy,ss}^{-1} \Sigma_{ss,yy}^{-1} \right\}
\]
\[
\overset{p}{\rightarrow} \text{tr} \left( \Psi_j \Sigma_{jj,ss}^{-1} \right).
\]
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

\[
\frac{1}{np} \alpha_{j,1} = \frac{1}{np} \lambda_{\text{max}}(L_j) \geq \frac{1}{\gamma_j np} \text{tr}(L_j)
\]

\[
= \frac{1}{\gamma_j np} \text{tr}\left\{ X_n'(I_n - P_j) X_n \Sigma_{jj}^{-1} \Sigma_{yy,j} \Sigma_{yy,j}^{-1} \Sigma_{y,j} \Sigma_{y,j}^{-1} \right\}
\]

\[
\overset{p!}{\rightarrow} \text{tr}\left\{ \Sigma_{jj}^{-1} \Sigma_{jj}^{-1} \right\}.
\]

Hence, with assumption A4, this implies that

\[
\liminf_{c_n \rightarrow c_0} \frac{1}{np} \alpha_{j,1} > 0.
\]

Consequently, Lemma 5 is proved.

References


