広島大学学位請求論文

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

（非正規性の下での高次元正準相関分析における変数選択
のための対数尤度関数に基づく情報量規準の一致性）

２０１５年

広島大学大学院理学研究科
数学専攻

福井　敬祐

目　次

主論文

Hiroshima Math. J.
00 (0000), 1–27

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

Keisuke Fukui

(Received Xxx 00, 0000)

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.

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 $\omega = \{1, \ldots, q\}$ containing q_j elements,

2010 *Mathematics Subject Classification.* Primary 62H12; Secondary 62H20.
Key words and phrases. AIC, assumption of normality, bias-corrected AIC, BIC, consistent AIC, high-dimensional asymptotic framework, HQC, nonnormality, selection of redundancy model, selection probability.

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 $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ be n independent random vectors from $\boldsymbol{z}$, and let $\bar{\boldsymbol{z}}$ be the sample mean of $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ 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 $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n \sim i.i.d.\ N_{p+q}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. 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$_\mathrm{c}$) 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

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

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:

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

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 $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$, $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ and $\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n$ 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

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

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

(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 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 AIC$_\mathrm{c}$ 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, AIC$_\mathrm{c}$, BIC, CAIC, and HQC.

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, AIC$_\mathrm{c}$, BIC, CAIC, and HQC were evaluated by Monte Carlo simulations with 10,000 iterations each.

Table 1. Conditions for consistency

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

Note) c_a and c_b are given in Corollary 1.

Let $\boldsymbol{\nu}_1 = (\nu_{1,1}, \ldots, \nu_{1,p})' \sim N_p(\mathbf{0}_p, \boldsymbol{I}_p)$, $\boldsymbol{\nu}_2 = (\nu_{2,1}, \ldots, \nu_{2,q})' \sim N_q(\mathbf{0}_q, \boldsymbol{I}_q)$, $\delta_1, \delta_2 \sim \chi^2_6$, $\omega_{1,1}, \ldots, \omega_{2,p} \sim i.i.d.\chi^2_5$ and $\omega_{2,1}, \ldots, \omega_{2,q} \sim i.i.d.\chi^2_5$ be mutually independent random vectors and variables. Then, $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_p)'$ and $\boldsymbol{x} = (x_1, \ldots, x_q)'$ 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 AIC$_\mathrm{c}$, 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$_\mathrm{c}$ 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$_\text{c}$	BIC	CAIC	HQC	AIC	AIC$_\text{c}$	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$_\text{c}$	BIC	CAIC	HQC	AIC	AIC$_\text{c}$	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$_\text{c}$	BIC	CAIC	HQC	AIC	AIC$_\text{c}$	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$_\text{c}$	BIC	CAIC	HQC	AIC	AIC$_\text{c}$	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$_\text{c}$	BIC	CAIC	HQC	AIC	AIC$_\text{c}$	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$_\text{c}$	BIC	CAIC	HQC	AIC	AIC$_\text{c}$	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$_\text{c}$	BIC	CAIC	HQC	AIC	AIC$_\text{c}$	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$_\text{c}$	BIC	CAIC	HQC	AIC	AIC$_\text{c}$	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$_\mathrm{c}$	BIC	CAIC	HQC	AIC	AIC$_\mathrm{c}$	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$_\mathrm{c}$	BIC	CAIC	HQC	AIC	AIC$_\mathrm{c}$	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$_\mathrm{c}$	BIC	CAIC	HQC	AIC	AIC$_\mathrm{c}$	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$_\mathrm{c}$	BIC	CAIC	HQC	AIC	AIC$_\mathrm{c}$	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$_\mathrm{c}$	BIC	CAIC	HQC	AIC	AIC$_\mathrm{c}$	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$_\mathrm{c}$	BIC	CAIC	HQC	AIC	AIC$_\mathrm{c}$	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$_\mathrm{c}$	BIC	CAIC	HQC	AIC	AIC$_\mathrm{c}$	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$_\mathrm{c}$	BIC	CAIC	HQC	AIC	AIC$_\mathrm{c}$	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	AIC$_c$	BIC	CAIC	HQC	AIC	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	AIC$_c$	BIC	CAIC	HQC	AIC	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	AIC$_c$	BIC	CAIC	HQC	AIC	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	AIC$_c$	BIC	CAIC	HQC	AIC	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	AIC$_c$	BIC	CAIC	HQC	AIC	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	AIC$_c$	BIC	CAIC	HQC	AIC	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	AIC$_c$	BIC	CAIC	HQC	AIC	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	AIC$_c$	BIC	CAIC	HQC	AIC	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$_\mathrm{c}$	BIC	CAIC	HQC	AIC	AIC$_\mathrm{c}$	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$_\mathrm{c}$	BIC	CAIC	HQC	AIC	AIC$_\mathrm{c}$	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$_\mathrm{c}$	BIC	CAIC	HQC	AIC	AIC$_\mathrm{c}$	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$_\mathrm{c}$	BIC	CAIC	HQC	AIC	AIC$_\mathrm{c}$	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$_\mathrm{c}$	BIC	CAIC	HQC	AIC	AIC$_\mathrm{c}$	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$_\mathrm{c}$	BIC	CAIC	HQC	AIC	AIC$_\mathrm{c}$	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$_\mathrm{c}$	BIC	CAIC	HQC	AIC	AIC$_\mathrm{c}$	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$_\mathrm{c}$	BIC	CAIC	HQC	AIC	AIC$_\mathrm{c}$	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, 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 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 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 $j \in \mathscr{J}_+ \setminus \{j_*\}$ under the given assumptions:

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

Hence, by using Lemma 3 and $(n-1)^{-1}\boldsymbol{X}_\ell'\boldsymbol{X}_\ell \xrightarrow{p} \boldsymbol{\Sigma}_{\ell\ell}$ for all $\ell \in \mathscr{J}$, 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 $j \in \mathscr{J}_-$. It follows that for all $j \in \mathscr{J}_-$,

$$
\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|}{\left|\boldsymbol{\mathcal{E}}'(\boldsymbol{I}_n - \boldsymbol{P}_{j_*})\boldsymbol{\mathcal{E}}\right|} \\
&= \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|}{\left|\boldsymbol{\mathcal{E}}'(\boldsymbol{I}_n - \boldsymbol{P}_{j_*})\boldsymbol{\mathcal{E}}\right|} \\
&\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|}{\left|\boldsymbol{\mathcal{E}}'(\boldsymbol{I}_n - \boldsymbol{P}_{j_*})\boldsymbol{\mathcal{E}}\right|} \\
&= \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|}{\left|\boldsymbol{\mathcal{E}}'(\boldsymbol{I}_n - \boldsymbol{P}_{j_*})\boldsymbol{\mathcal{E}}\right|} \\
&\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|}{\left|\boldsymbol{\mathcal{E}}'(\boldsymbol{I}_n - \boldsymbol{P}_{j_*})\boldsymbol{\mathcal{E}}\right|} - \log \lambda_{\max}(\boldsymbol{V}_j) \\
&= D_1(j) + D_2(j) + D_3(j),
\end{aligned} \tag{A2}
$$

where

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

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,

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

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

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

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

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

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 $-\log x \le -x + 1$ for any $x \ge 0$, we have

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

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,

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

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 AIC$_\mathrm{c}$. According to an expansion of $m(j) - m(j_*)$ in the AIC$_\mathrm{c}$, the differences between the penalty terms of the AIC$_\mathrm{c}$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

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

Hence, the convergence of the differences between the penalty terms of the AICs and those of the AIC$_\mathrm{c}$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 AIC$_\mathrm{c}$. Furthermore, it follows from equation (A10) that

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

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 AIC$_\mathrm{c}$ 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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

A6. Proof of Lemma 5

It follows from elementary linear algebra that

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

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

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

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.

Acknowledgement

First of all, I am deeply grateful to Prof. H. Yanagihara, who offered continuing support and constant encouragement. If I had not met him, I would not be here today. I am very happy to be one of his students. I also owe a very important debt to Prof. H. Wakaki and Prof. Y. Fujikoshi, who gave me invaluable comments and warm encouragement. I'm grateful to Dr. S. Imori of Osaka University, Dr. K. Matsubara, and Mr. H. Miyazaki of Hiroshima University for their encouragement. I would also like to express my deepest gratitude to my wife, Ayumi, for her moral support and warm encouragement. She is always my anchor. I also grateful to the associate editor and the referee for the careful reading and helpful suggestions which led to an improvement of my original manuscript.

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.

Keisuke Fukui
Department of Mathematics
Graduate School of Science
Hiroshima University
1-3-1 *Kagamiyama, Higashi-Hiroshima* 739-8526*, JAPAN*
E-mail: d126313@hiroshima-u.ac.jp

参考論文

(1) Choosing the number of repetitions in the multiple plug-in optimization method for the ridge parameters in multivariate generalized ridge regression.
Isamu Nagai, Keisuke Fukui and Hirokazu Yanagihara,
Bulletin of Informatics and Cybernetics, Vol.45, 2013, 25-35.

(2) Selecting a shrinkage parameter in structural equation modeling
with a near singular covariance matrix by the GIC minimization method.
Ami Kamada, Hirokazu Yanagihara, Hirofumi Wakaki and Keisuke Fukui,
Hiroshima Mathematical Journal, Vol.44 (3), 2014, 315-326.

(3) Comparison with RSS-based model selection criteria for selecting growth functions.
Keisuke Fukui, Mariko Yamamura and Hirokazu Yanagihara,
FORMATH, Vol.14 に掲載予定.

Bulletin of Informatics and Cybernetics, Vol. 45, 2013

CHOOSING THE NUMBER OF REPETITIONS IN THE MULTIPLE PLUG-IN OPTIMIZATION METHOD FOR THE RIDGE PARAMETERS IN MULTIVARIATE GENERALIZED RIDGE REGRESSION

By

Isamu Nagai[*], **Keisuke Fukui**[†] and **Hirokazu Yanagihara**[‡]

Abstract

Multivariate generalized ridge (MGR) regression was proposed by Yanagihara, Nagai, and Satoh (2009) in order to avoid the multicollinearity problem in multivariate linear regression models. The MGR estimator is defined by using multiple nonnegative ridge parameters in an ordinary least-squares (LS) estimator. In order to optimize these ridge parameters, Yanagihara, Nagai, and Satoh (2009) and Nagai, Yanagihara, and Satoh (2012) proposed several optimization methods. We focus on the plug-in optimization method, which is an estimation method for the principal optimal ridge parameters that minimizes the predicted mean squared error. The plug-in optimization method is a repeating method that uses the current ridge parameters estimate as input in order to obtain an improved estimate. In the present paper, we propose two criteria for choosing the number of repetitions. We conducted several numerical studies using the proposed information criteria to compare the LS and MGR estimators with the optimized ridge parameters based on some ordinary plug-in optimization methods, and those obtained by using the optimized multiple plug-in optimization method.

Key Words and Phrases: Generalized ridge regression, Multivariate linear regression model, Plug-in optimization method, Shrinkage estimator.

1. Introduction

In the present paper, we consider a multivariate linear regression model with n observations of a p-dimensional vector of response variables and a k-dimensional vector of regressors (for more detailed information, see for example, Srivastava, 2002, Chapter 9; Timm, 2002, Chapter 4). Let $\boldsymbol{Y}$, $\boldsymbol{X}$, and $\boldsymbol{\mathcal{E}}$ be the $n \times p$ matrix of response variables, the $n \times k$ matrix of nonstochastic centered explanatory variables (i.e., $\boldsymbol{X}'\mathbf{1}_n = \mathbf{0}_k$) of $\mathrm{rank}(\boldsymbol{X}) = k$, and the $n \times p$ matrix of error variables, respectively, where n is the sample size, $\mathbf{1}_n$ is an n-dimensional vector of ones, and $\mathbf{0}_k$ is a k-dimensional vector of zeros.

[*] School of Science and Technology, Kwansei Gakuin University, 2-1 Gakuen, Sanda, Hyogo 669-1337, Japan. inagai@kwansei.ac.jp

[†] Department of Mathematics, Graduate School of Science, Hiroshima University, 1-3-1 Kagamiyama, Higashi-Hiroshima, Hiroshima 739-8526, Japan.

[‡] Department of Mathematics, Graduate School of Science, Hiroshima University, 1-3-1 Kagamiyama, Higashi-Hiroshima, Hiroshima 739-8526, Japan.

Suppose that $n-k-p-2>0$ and $\boldsymbol{\mathcal{E}} \sim N_{n\times p}(\boldsymbol{O}_{n\times p}, \boldsymbol{\Sigma}\otimes \boldsymbol{I}_n)$, where $\boldsymbol{\Sigma}$ is a $p\times p$ unknown covariance matrix, $\mathrm{rank}(\boldsymbol{\Sigma})=p$, and $\boldsymbol{O}_{n\times p}$ is an $n\times p$ matrix of zeros. Then the matrix form of the multivariate linear regression model is expressed as

$$\boldsymbol{Y} = \boldsymbol{1}_n\boldsymbol{\mu}' + \boldsymbol{X}\boldsymbol{\Xi} + \boldsymbol{\mathcal{E}}, \tag{1}$$

where $\boldsymbol{\mu}$ is a p-dimensional unknown location vector and $\boldsymbol{\Xi}$ is a $k\times p$ unknown regression coefficient matrix. We can also express the model (1) as $\boldsymbol{Y} \sim N_{n\times p}(\boldsymbol{1}_n\boldsymbol{\mu}'+\boldsymbol{X}\boldsymbol{\Xi}, \boldsymbol{\Sigma}\otimes\boldsymbol{I}_n)$.

The maximum-likelihood or least-squares (LS) estimators of $\boldsymbol{\mu}$ and $\boldsymbol{\Xi}$ are given by $\hat{\boldsymbol{\mu}} = \boldsymbol{Y}'\boldsymbol{1}_n/n$ and $\hat{\boldsymbol{\Xi}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$, respectively, because $\boldsymbol{X}'\boldsymbol{1}_n = \boldsymbol{0}_k$. Since $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Xi}}$ are simple forms and are unbiased estimators of $\boldsymbol{\mu}$ and $\boldsymbol{\Xi}$, they are widely used in actual data analysis, see, e.g., Dien *et al.* (2006), Sârbu *et al.* (2008), Saxén and Sundell (2006), Skagerberg, Macgregor, and Kiparissides (1992), and Yoshimoto, Yanagihara, and Ninomiya (2005). However, when multicollinearity occurs in $\boldsymbol{X}$, the estimator of $\boldsymbol{\Xi}$ becomes unstable. In order to avoid this problem, multivariate generalized ridge (MGR) regression for the model in (1) was proposed by Yanagihara, Nagai, and Satoh (2009) (the original generalized ridge regression when $p=1$ in the model (1) was proposed by Hoerl and Kennard (1970)). The MGR estimator is defined by using multiple ridge parameters $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_k)'$, $(\theta_i \geq 0,\ i=1,\ldots,k)$. Nagai, Yanagihara, and Satoh (2012) showed that the principal optimal $\boldsymbol{\theta}$ that minimizes the predicted mean squared error (PMSE) is obtained in closed form with the unknown regression coefficient vector and covariance matrix.

In order to estimate the principal optimal $\boldsymbol{\theta}$, Nagai, Yanagihara, and Satoh (2012) proposed the plug-in optimization method. By replacing the LS estimator and unbiased estimator for $\boldsymbol{\Sigma}$ with unknown values, respectively, the single plug-in optimized ridge parameters were derived. However, when multicollinearity occurs, the optimized ridge parameters tend to be too small since the LS estimator tends to have a large variance. Thus, to avoid under evaluation, Nagai, Yanagihara, and Satoh (2012) considered using the MGR estimator based on the single plug-in optimized ridge parameters instead of using the LS estimator. The double plug-in optimized ridge parameters were also derived. Repeating this renewal method, we obtained the multiple plug-in optimized ridge parameters derived from the MGR estimator based on the initial optimized ridge parameters. Let s $(s=1,2,3,\ldots)$ be the number of repetitions in the multiple plug-in optimization method. Note that the single plug-in optimized ridge parameters are obtained when $s=1$ and the double ones are obtained when $s=2$. In the present paper, we wish to find the value of s that minimizes the PMSE. In order to choose s, we propose two information criteria.

The remainder of the present paper is organized as follows: In Section 2, we illustrate the MGR estimator and a target PMSE. We also introduce the multiple plug-in optimization method. In Section 3, we propose some criteria for choosing the number of repetitions in the multiple plug-in optimization method. In Section 4, we compare the optimization methods by conducting numerical studies.

2. Preliminaries

In this section, we introduce the MGR estimator and the principal optimal $\boldsymbol{\theta}$ that minimizes the PMSE. Yanagihara, Nagai, and Satoh (2009) proposed the MGR estimator, which is defined as follows:

$$\hat{\boldsymbol{\Xi}}_{\boldsymbol{\theta}} = (\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{Q}\boldsymbol{\Theta}\boldsymbol{Q}')^{-1}\boldsymbol{X}'\boldsymbol{Y}, \tag{2}$$

where $\boldsymbol{\Theta} = \mathrm{diag}(\boldsymbol{\theta})$ is a $k \times k$ diagonal matrix and $\boldsymbol{Q}$ is a $k \times k$ orthogonal matrix that diagonalizes $\boldsymbol{X}'\boldsymbol{X}$, i.e., $\boldsymbol{Q}$ satisfies $\boldsymbol{Q}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{Q} = \boldsymbol{D}$, where $\boldsymbol{D} = \mathrm{diag}(d_1, \ldots, d_k)$ and $d_1, \ldots, d_k$ are the eigenvalues of $\boldsymbol{X}'\boldsymbol{X}$. This estimator coincides with the LS estimator when $\boldsymbol{\theta} = \boldsymbol{0}_k$, and it coincides with the ridge estimator for the model (1) proposed by Yanagihara and Satoh (2010) when $\boldsymbol{\theta} = \lambda \boldsymbol{1}_k$ for $\lambda \geq 0$.

Using the same orthogonal transform as was used by Nagai, Yanagihara, and Satoh (2012), the model (1) can be rewritten as follows:

$$\boldsymbol{Z} = \boldsymbol{L}\begin{pmatrix}\boldsymbol{\Gamma} \\ \boldsymbol{\mu}'\end{pmatrix} + \boldsymbol{\mathcal{V}},$$

where $\boldsymbol{\Gamma} = (\gamma_1, \ldots, \gamma_k)' = \boldsymbol{Q}'\boldsymbol{\Xi}$, $\boldsymbol{Z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)' = \boldsymbol{P}_1'\boldsymbol{Y}$, $\boldsymbol{\mathcal{V}} = (\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_n)' = \boldsymbol{P}_1'\boldsymbol{\mathcal{E}}$, $\boldsymbol{L} = (\mathrm{diag}(\sqrt{d_1}, \ldots, \sqrt{d_k}, \sqrt{n}), \boldsymbol{O}_{k+1,n-k-1})'$, and $\boldsymbol{P}_1$ is an $n \times n$ orthogonal matrix that diagonalizes $(\boldsymbol{X}, \boldsymbol{1}_n)(\boldsymbol{X}, \boldsymbol{1}_n)'$, that is, $\boldsymbol{P}_1$ satisfies $\boldsymbol{P}_1'(\boldsymbol{X}, \boldsymbol{1}_n)(\boldsymbol{X}, \boldsymbol{1}_n)'\boldsymbol{P}_1 = \boldsymbol{L}\boldsymbol{L}'$. When $p = 1$, this transformation was used in Goldstein and Smith (1974), and by others. Nagai, Yanagihara, and Satoh (2012) showed that $\boldsymbol{Z} \sim N_{n \times p}(\boldsymbol{\mathcal{M}}, \boldsymbol{\Sigma} \otimes \boldsymbol{I}_p)$, where $\boldsymbol{\mathcal{M}} = (\boldsymbol{m}_1, \ldots, \boldsymbol{m}_n)' = \boldsymbol{L}(\boldsymbol{\Gamma}', \boldsymbol{\mu})'$. The MGR estimator of $\boldsymbol{\Gamma}$ is defined by $\hat{\boldsymbol{\Gamma}}_{\dot{\boldsymbol{\theta}}} = \boldsymbol{Q}'\hat{\boldsymbol{\Xi}}_{\boldsymbol{\theta}}$, thus $\hat{\boldsymbol{\Gamma}}_{\boldsymbol{\theta}} = (\boldsymbol{D} + \boldsymbol{\Theta})^{-1}\boldsymbol{C}'\boldsymbol{Z}$ where $\boldsymbol{C} = (\boldsymbol{D}^{1/2}, \boldsymbol{O}_{k,n-k})'$, which is equivalent to the estimator obtained by substituting $\boldsymbol{D} + \boldsymbol{\Theta}$ into $\boldsymbol{D}$ in the LS estimator of $\boldsymbol{\Gamma}$, i.e., $\hat{\boldsymbol{\Gamma}} = \boldsymbol{D}^{-1}\boldsymbol{C}'\boldsymbol{Z}$.

Then the PMSE of $\hat{\boldsymbol{Z}}_{\boldsymbol{\theta}} = \boldsymbol{L}(\hat{\boldsymbol{\Gamma}}'_{\boldsymbol{\theta}}, \hat{\boldsymbol{\mu}})'$, which is the predictor of $\boldsymbol{Z}$, is defined as follows:

$$\mathrm{PMSE}[\hat{\boldsymbol{Z}}_{\boldsymbol{\theta}}] = E[r(\boldsymbol{V}, \hat{\boldsymbol{Z}}_{\boldsymbol{\theta}})],$$

where $\boldsymbol{V} \sim N_{n \times p}(\boldsymbol{\mathcal{M}}, \boldsymbol{\Sigma} \otimes \boldsymbol{I}_n)$, $\boldsymbol{V} \perp\!\!\!\perp \boldsymbol{Z}$, and the function $r(\cdot, \cdot)$ is defined by the following discrepancy function for measuring the distance between any $n \times p$ matrices $\boldsymbol{A}$ and $\boldsymbol{B}$:

$$r(\boldsymbol{A}, \boldsymbol{B}) = \mathrm{tr}\{(\boldsymbol{A} - \boldsymbol{B})\boldsymbol{\Sigma}^{-1}(\boldsymbol{A} - \boldsymbol{B})'\}.$$

From some simple calculations, we obtain $\mathrm{PMSE}[\hat{\boldsymbol{Z}}_{\boldsymbol{\theta}}] = np + E[r(\hat{\boldsymbol{Z}}_{\boldsymbol{\theta}}, \boldsymbol{\mathcal{M}})]$. Nagai, Yanagihara, and Satoh (2012) showed that $\boldsymbol{\theta}^* = (\theta_1^*, \ldots, \theta_k^*)'$, the principal optimal $\boldsymbol{\theta}$ with minimized $E[r(\hat{\boldsymbol{Z}}_{\boldsymbol{\theta}}, \boldsymbol{\mathcal{M}})]$, is derived as follows:

$$\theta_i^* = \frac{p}{\gamma_i'\boldsymbol{\Sigma}^{-1}\gamma_i}, \quad (i = 1, \ldots, k). \tag{3}$$

However, θ_i^*, $(i = 1, \ldots, k)$ cannot be directly used for estimating $\boldsymbol{\Xi}$ since it includes the unknowns γ_i and $\boldsymbol{\Sigma}$.

Nagai, Yanagihara, and Satoh (2012) proposed the single plug-in optimization method by substituting $\hat{\gamma}_i = \boldsymbol{z}_i/\sqrt{d_i}$, $(i = 1, \ldots, k)$, which is the ith row of $\hat{\boldsymbol{\Gamma}}$, for γ_i, $(i = 1, \ldots, k)$; and $\boldsymbol{S} = \sum_{i=k+2}^{n} \boldsymbol{z}_i\boldsymbol{z}_i'/(n - k - 1)$, which is an unbiased estimator for $\boldsymbol{\Sigma}$, for $\boldsymbol{\Sigma}$ in θ_i^*, $(i = 1, \ldots, k)$ which is in equation (3). Then the estimator for θ_i^*, $(i = 1, \ldots, k)$ from the single plug-in optimization method is derived as follows:

$$\hat{\theta}_i^{[1]} = \frac{p}{\hat{\gamma}_i'\boldsymbol{S}^{-1}\hat{\gamma}_i} = \frac{d_i p}{t_i}, \quad (i = 1, \ldots, k), \tag{4}$$

where $t_i = \boldsymbol{z}_i' \boldsymbol{S}^{-1} \boldsymbol{z}_i$, $(i = 1, \ldots, k)$.

When multicollinearity occurs, we consider using the MGR estimator. However, in that case, $\hat{\theta}_i^{[1]}$, $(i = 1, \ldots, k)$ tends to be too small since it depends heavily on $\hat{\boldsymbol{\gamma}}_i$ and the variance of $\hat{\boldsymbol{\gamma}}_i$ becomes large. In order to avoid this problem, we can use the MGR estimator based on the single plug-in optimization method instead of using the LS estimator because the variance of the MGR estimator is smaller than that of the LS estimator. We then derive the double plug-in optimization method as

$$\hat{\theta}_i^{[2]} = \frac{p}{\hat{\boldsymbol{\gamma}}_i^{[1]\prime} \boldsymbol{S}^{-1} \hat{\boldsymbol{\gamma}}_i^{[1]}} = \left(1 + \frac{\hat{\theta}_i^{[1]}}{d_i}\right)^2 \hat{\theta}_i^{[1]} = \left(1 + \frac{p}{t_i}\right)^2 \frac{d_i p}{t_i}, \quad (i = 1, \ldots, k), \tag{5}$$

where $\hat{\boldsymbol{\gamma}}_i^{[1]} = \sqrt{d_i} \boldsymbol{z}_i / (d_i + \hat{\theta}_i^{[1]})$ is the ith row of $\hat{\boldsymbol{\Gamma}}_{\hat{\boldsymbol{\theta}}^{[1]}}$ obtained by substituting $\hat{\boldsymbol{\theta}}^{[1]} = (\hat{\theta}_1^{[1]}, \ldots, \hat{\theta}_k^{[1]})'$ for $\boldsymbol{\theta}$ in $\hat{\boldsymbol{\Gamma}}_{\boldsymbol{\theta}}$. When we iterate this method, that is, we obtain new plug-in optimized ridge parameters by using the MGR estimator based on the current optimized ridge parameters, the multiple plug-in optimization method is derived. This was also proposed by Nagai, Yanagihara, and Satoh (2012), as follows:

$$\hat{\theta}_i^{[s]} = \frac{p}{\hat{\boldsymbol{\gamma}}_i^{[s-1]\prime} \boldsymbol{S}^{-1} \hat{\boldsymbol{\gamma}}_i^{[s-1]}} = \left(1 + \frac{\hat{\theta}_i^{[s-1]}}{d_i}\right)^2 \hat{\theta}_i^{[1]}, \quad (s = 1, 2, 3, \ldots;\ i = 1, \ldots, k), \tag{6}$$

where $\hat{\theta}_i^{[0]} = 0$ and $\hat{\boldsymbol{\gamma}}_i^{[s-1]}$ is the ith row of $\hat{\boldsymbol{\Gamma}}_{\hat{\boldsymbol{\theta}}^{[s-1]}}$; note that $\hat{\boldsymbol{\gamma}}_i^{[0]} = \hat{\boldsymbol{\gamma}}_i$, $(i = 1, \ldots, k)$. In the case of $p = 1$, $\hat{\theta}_i^{[1]}$ and $\hat{\theta}_i^{[2]}$ correspond with the optimization methods in Hoerl and Kennard (1970), and $\hat{\theta}_i^{[s]}$ coincides with the optimization method in Hemmerle (1975). Numerical studies in previous papers have compared only the single or double optimized ridge parameters. However, $\hat{\boldsymbol{\theta}}^{[s]} = (\hat{\theta}_1^{[s]}, \ldots, \hat{\theta}_k^{[s]})'$ is derived by using $\hat{\boldsymbol{\theta}}^{[s-1]}$ and (6) for any natural number s, and there is no method for choosing s. Hence we consider determining the value of s that reduces the PMSE of $\hat{\boldsymbol{Z}}_{\hat{\boldsymbol{\theta}}^{[s]}} = \boldsymbol{L}(\hat{\boldsymbol{\Gamma}}'_{\hat{\boldsymbol{\theta}}^{[s]}}, \hat{\boldsymbol{\mu}})'$, which is the predictor of $\boldsymbol{Z}$ based on the multiple plug-in optimization method.

3. Method for Choosing s

In this section, we consider a method for choosing s, which was defined in (6) as the number of repetitions in the multiple plug-in optimization method. We regard the s which minimizes the PMSE$[\hat{\boldsymbol{Z}}_{\hat{\boldsymbol{\theta}}^{[s]}}]$ as an optimal number and we will propose information criteria to get it. Note that $\hat{\theta}_i^{[s]}$ depends on t_i. In order to get right this dependence, we express $\hat{\theta}_i^{[s]}$ as $\hat{\theta}_i^{[s]}(t_i)$, $(s = 1, 2, \ldots;\ i = 1, \ldots, k)$. Letting $w^{[s]}(t_i) = d_i/(d_i + \hat{\theta}_i^{[s]}(t_i))$ for $i = 1, \ldots, k$, the ith row of $\hat{\boldsymbol{\Gamma}}_{\hat{\boldsymbol{\theta}}^{[s]}}$ is obtained by $\hat{\boldsymbol{\gamma}}_i^{[s]} = w^{[s]}(t_i) \hat{\boldsymbol{\gamma}}_i$. We now consider how to estimate the PMSE$[\hat{\boldsymbol{Z}}_{\hat{\boldsymbol{\theta}}^{[s]}}]$ for a fixed s. Hence we consider evaluating $\hat{\boldsymbol{Z}}_{\hat{\boldsymbol{\theta}}^{[s]}}$, which is obtained by using $\hat{\boldsymbol{\theta}}^{[s]} = (\hat{\theta}_1^{[s]}(t_1), \ldots, \hat{\theta}_k^{[s]}(t_k))'$, by stating the PMSE as follows:

$$\text{PMSE}[\hat{\boldsymbol{Z}}_{\hat{\boldsymbol{\theta}}^{[s]}}] = E[r(\boldsymbol{V}, \hat{\boldsymbol{Z}}_{\hat{\boldsymbol{\theta}}^{[s]}})].$$

The predicted value $\hat{\boldsymbol{Z}}_{\hat{\boldsymbol{\theta}}^{[s]}}$ can be expressed as $\hat{\boldsymbol{Z}}_{\hat{\boldsymbol{\theta}}^{[s]}} = (\hat{\boldsymbol{z}}_1(\hat{\theta}_1^{[s]}(t_1)), \ldots, \hat{\boldsymbol{z}}_k(\hat{\theta}_k^{[s]}(t_k)), \hat{\boldsymbol{z}}_{k+1}, \ldots, \hat{\boldsymbol{z}}_n)'$ since the ith row of $\hat{\boldsymbol{Z}}_{\hat{\boldsymbol{\theta}}^{[s]}}$ depends on $\hat{\theta}_i^{[s]}(t_i)$ for $i = 1, \ldots, k$, and it does not depend on $\hat{\boldsymbol{\theta}}^{[s]}$ for $i = k+1, \ldots, n$. Additionally, it should be kept in mind that

$\hat{\boldsymbol{z}}_i(\hat{\theta}_i^{[s]}(t_i)) = w^{[s]}(t_i)\boldsymbol{z}_i$, $(i = 1, \ldots, k)$, $\hat{\boldsymbol{z}}_{k+1} = \boldsymbol{z}_{k+1}$, $\hat{\boldsymbol{z}}_i = \boldsymbol{0}_p$, $(i = k+2, \ldots, n)$, and $\boldsymbol{z}_i \overset{\text{i.i.d.}}{\sim} N_p(\boldsymbol{m}_i, \boldsymbol{\Sigma})$ where $\boldsymbol{m}_i$ is the ith row of $\boldsymbol{\mathcal{M}} = \boldsymbol{L}(\boldsymbol{\Gamma}', \boldsymbol{\mu})'$. From a simple calculation, we obtain

$$\text{PMSE}[\hat{\boldsymbol{Z}}_{\hat{\boldsymbol{\theta}}^{[s]}}] = E[r(\boldsymbol{Z}, \hat{\boldsymbol{Z}}_{\hat{\boldsymbol{\theta}}^{[s]}})] + 2E[\text{tr}\{(\hat{\boldsymbol{Z}}_{\hat{\boldsymbol{\theta}}^{[s]}} - \boldsymbol{\mathcal{M}})'(\boldsymbol{Z} - \boldsymbol{\mathcal{M}})\boldsymbol{\Sigma}^{-1}\}]. \tag{7}$$

Then we can see that

$$\begin{aligned}
&E[\text{tr}\{(\hat{\boldsymbol{Z}}_{\hat{\boldsymbol{\theta}}^{[s]}} - \boldsymbol{\mathcal{M}})'(\boldsymbol{Z} - \boldsymbol{\mathcal{M}})\boldsymbol{\Sigma}^{-1}\}] \\
&= \sum_{i=1}^{k} E\left[\left(w^{[s]}(t_i)\boldsymbol{z}_i - \boldsymbol{m}_i\right)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{z}_i - \boldsymbol{m}_i)\right] + E[(\boldsymbol{z}_{k+1} - \boldsymbol{m}_{k+1})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{z}_{k+1} - \boldsymbol{m}_{k+1})] \\
&\quad + \sum_{i=k+2}^{n} E[\boldsymbol{m}_i'\boldsymbol{\Sigma}^{-1}(\boldsymbol{z}_i - \boldsymbol{m}_i)] \\
&= \sum_{i=1}^{k} E[w^{[s]}(t_i)\boldsymbol{z}_i'\boldsymbol{\Sigma}^{-1}(\boldsymbol{z}_i - \boldsymbol{m}_i)] + p,
\end{aligned}$$

because $E[(\boldsymbol{z}_{k+1} - \boldsymbol{m}_{k+1})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{z}_{k+1} - \boldsymbol{m}_{k+1})] = \text{tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}) = p$ and $E[\boldsymbol{m}_i'\boldsymbol{\Sigma}^{-1}(\boldsymbol{z}_i - \boldsymbol{m}_i)] = 0$. If we let $\boldsymbol{u}_i = (u_{i1}, \ldots, u_{ip})' = \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{z}_i - \boldsymbol{m}_i)$, where $\boldsymbol{\Sigma}^{1/2}$ satisfies $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\Sigma}^{1/2}$, then $\boldsymbol{u}_i \overset{\text{i.i.d.}}{\sim} N_p(\boldsymbol{0}_p, \boldsymbol{I}_p)$ and the following result is derived:

$$\begin{aligned}
E[\text{tr}\{(\hat{\boldsymbol{Z}}_{\hat{\boldsymbol{\theta}}^{[s]}} - \boldsymbol{\mathcal{M}})'(\boldsymbol{Z} - \boldsymbol{\mathcal{M}})\boldsymbol{\Sigma}^{-1}\}] &= \sum_{i=1}^{k} E[w^{[s]}(t_i)(\boldsymbol{u}_i + \boldsymbol{\Sigma}^{-1/2}\boldsymbol{m}_i)'\boldsymbol{u}_i] + p \\
&= \sum_{i=1}^{k}\sum_{j=1}^{p} E[w^{[s]}(t_i)(u_{ij} + \boldsymbol{m}_i'\boldsymbol{\Sigma}^{-1/2}\boldsymbol{e}_{p\cdot j})u_{ij}] + p,
\end{aligned}$$

where $\boldsymbol{e}_{p\cdot j}$ is a p-dimensional vector in which only the jth element is equal to one and the other elements are zeros. Using the formula in Stein (1981) (see, e.g., Efron (2004)), we obtain

$$\begin{aligned}
E[w^{[s]}(t_i)(u_{ij} + \boldsymbol{m}_i'\boldsymbol{\Sigma}^{-1/2}\boldsymbol{e}_{p\cdot j})u_{ij}] &= E\left[\frac{\partial\ \{w^{[s]}(t_i)(u_{ij} + \boldsymbol{m}_i'\boldsymbol{\Sigma}^{-1/2}\boldsymbol{e}_{p\cdot j})\}}{\partial u_{ij}}\right] \\
&= E\left[\frac{\partial w^{[s]}(t_i)}{\partial u_{ij}}(u_{ij} + \boldsymbol{m}_i'\boldsymbol{\Sigma}^{-1/2}\boldsymbol{e}_{p\cdot j}) + w^{[s]}(t_i)\right] \\
&= E\left[\boldsymbol{e}_{p\cdot j}'\frac{\partial w^{[s]}(t_i)}{\partial \boldsymbol{u}_i}\boldsymbol{e}_{n\cdot i}'(\boldsymbol{U} + \boldsymbol{\mathcal{M}}\boldsymbol{\Sigma}^{-1/2})\boldsymbol{e}_{p\cdot j} + w^{[s]}(t_i)\right],
\end{aligned}$$

where $\boldsymbol{U} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n)'$. Since $\boldsymbol{U} + \boldsymbol{\mathcal{M}}\boldsymbol{\Sigma}^{-1/2} = \boldsymbol{Z}\boldsymbol{\Sigma}^{-1/2}$ and t_i depends on $\boldsymbol{u}_i$, the following equation is derived:

$$E[w^{[s]}(t_i)(u_{ij} + \boldsymbol{m}_i'\boldsymbol{\Sigma}^{-1/2}\boldsymbol{e}_{p\cdot j})u_{ij}] = E\left[\boldsymbol{e}_{p\cdot j}'\frac{\partial t_i}{\partial \boldsymbol{u}_i}\frac{\partial w^{[s]}(t_i)}{\partial t_i}\boldsymbol{z}_i'\boldsymbol{\Sigma}^{-1/2}\boldsymbol{e}_{p\cdot j} + w^{[s]}(t_i)\right],$$

for $i = 1, \ldots, k$ and $j = 1, \ldots, p$. Note that $t_i = \boldsymbol{z}_i'\boldsymbol{S}^{-1}\boldsymbol{z}_i = (\boldsymbol{\Sigma}^{1/2}\boldsymbol{u}_i + \boldsymbol{m}_i)'\boldsymbol{S}^{-1}(\boldsymbol{\Sigma}^{1/2}\boldsymbol{u}_i + \boldsymbol{m}_i)$ and that $\boldsymbol{S}^{-1}$ does not depend on $\boldsymbol{u}_i$, $(i = 1, \ldots, k)$ since $\boldsymbol{S}$ does not depend on $\boldsymbol{z}_i$,

$(i = 1, \ldots, k+1)$ and $\boldsymbol{u}_i$ is obtained from $\boldsymbol{z}_i$. Thus we obtain the following differential result:

$$\frac{\partial t_i}{\partial \boldsymbol{u}_i} = 2\boldsymbol{\Sigma}^{1/2}\boldsymbol{S}^{-1}\boldsymbol{z}_i.$$

Hence we calculate (7) as follows:

$$\begin{aligned}\mathrm{PMSE}[\hat{\boldsymbol{Z}}_{\hat{\boldsymbol{\theta}}^{[s]}}] &= E[r(\boldsymbol{Z}, \hat{\boldsymbol{Z}}_{\hat{\boldsymbol{\theta}}^{[s]}})] + 2p + 2\sum_{i=1}^{k} E\left[\boldsymbol{z}_i'\boldsymbol{\Sigma}^{-1/2}\frac{\partial t_i}{\partial \boldsymbol{u}_i}\frac{\partial w^{[s]}(t_i)}{\partial t_i} + pw^{[s]}(t_i)\right] \\ &= E[r(\boldsymbol{Z}, \hat{\boldsymbol{Z}}_{\hat{\boldsymbol{\theta}}^{[s]}})] + 2p + 2\sum_{i=1}^{k} E\left[2t_i\frac{\partial w^{[s]}(t_i)}{\partial t_i} + pw^{[s]}(t_i)\right].\end{aligned}$$

From this result, the unbiased estimator of (7) can be defined as follows:

$$C_p^* = r(\boldsymbol{Z}, \hat{\boldsymbol{Z}}_{\hat{\boldsymbol{\theta}}^{[s]}}) + 2p + 2\sum_{i=1}^{k}\left(2t_i\frac{\partial w^{[s]}(t_i)}{\partial t_i} + pw^{[s]}(t_i)\right). \tag{8}$$

However, we cannot use this criterion since $r(\boldsymbol{Z}, \hat{\boldsymbol{Z}}_{\hat{\boldsymbol{\theta}}^{[s]}})$ requires the unknown covariance matrix $\boldsymbol{\Sigma}$.

In (8), we will consider estimating C_p^* by using the idea for the C_p and MC_p criteria that was put forth in Yanagihara, Nagai, and Satoh (2009). We thus estimate the C_p^* criterion as follows:

DEFINITION 3.1. The criteria for choosing s are defined by

$$C_p^{\#} = \hat{r}(\boldsymbol{Z}, \hat{\boldsymbol{Z}}_{\hat{\boldsymbol{\theta}}^{[s]}}) + 2p + 2\sum_{i=1}^{k}\left(2t_i\frac{\partial w^{[s]}(t_i)}{\partial t_i} + pw^{[s]}(t_i)\right), \tag{9}$$

$$MC_p^{\#} = c_{\mathrm{M}}\hat{r}(\boldsymbol{Z}, \hat{\boldsymbol{Z}}_{\hat{\boldsymbol{\theta}}^{[s]}}) + 2p + 2\sum_{i=1}^{k}\left(2t_i\frac{\partial w^{[s]}(t_i)}{\partial t_i} + pw^{[s]}(t_i)\right) + p(p+1), \tag{10}$$

where $c_{\mathrm{M}} = 1 - (p+1)/(n-k-1)$, and $\hat{r}(\boldsymbol{Z}, \hat{\boldsymbol{Z}}_{\hat{\boldsymbol{\theta}}^{[s]}})$ is obtained by substituting $\boldsymbol{S}$ for $\boldsymbol{\Sigma}$ in $r(\boldsymbol{Z}, \hat{\boldsymbol{Z}}_{\hat{\boldsymbol{\theta}}^{[s]}})$.

Minimizing each criterion, we can obtain several estimators for the optimal s. Let s_{C}^* and s_{M}^* be obtained by minimizing the $C_p^{\#}$ and $MC_p^{\#}$ criteria, respectively. From (6), we can obtain $\hat{\boldsymbol{\theta}}^{[s_{\mathrm{C}}^*]} = (\hat{\theta}_1^{[s_{\mathrm{C}}^*]}, \ldots, \hat{\theta}_k^{[s_{\mathrm{C}}^*]})'$ and $\hat{\boldsymbol{\theta}}^{[s_{\mathrm{M}}^*]} = (\hat{\theta}_1^{[s_{\mathrm{M}}^*]}, \ldots, \hat{\theta}_k^{[s_{\mathrm{M}}^*]})'$ in closed forms, respectively. Then, we can obtain $\hat{\boldsymbol{\Xi}}_{\hat{\boldsymbol{\theta}}^{[s_{\mathrm{C}}^*]}}$ and $\hat{\boldsymbol{\Xi}}_{\hat{\boldsymbol{\theta}}^{[s_{\mathrm{M}}^*]}}$ by substituting $\hat{\boldsymbol{\theta}}^{[s_{\mathrm{C}}^*]}$ and $\hat{\boldsymbol{\theta}}^{[s_{\mathrm{M}}^*]}$ into (2), respectively.

4. Numerical Studies

By conducting numerical studies, we compare the PMSEs of the predictors of $\boldsymbol{Y}$ consisting of the ridge regression estimators with the optimized ridge parameters by using each method. Let $\boldsymbol{R}_q = \mathrm{diag}(\sqrt{1}, \ldots, \sqrt{q})$, which is a $q \times q$ diagonal matrix, and let $\boldsymbol{\Delta}_q(\rho)$ be a $q \times q$ matrix whose (i,j)th element is $\rho^{|i-j|}$. Then the explanatory matrix is $\boldsymbol{X} = \boldsymbol{W}\boldsymbol{\Psi}^{1/2}$, where $\boldsymbol{\Psi} = \boldsymbol{R}_k\boldsymbol{\Delta}_k(\rho_x)\boldsymbol{R}_k$, and $\boldsymbol{W}$ is an $n \times k$ matrix whose elements were generated independently from the uniform distribution on $(-1, 1)$. The $k \times p$ unknown regression coefficient matrix $\boldsymbol{\Xi}$ is defined by $\boldsymbol{\Xi} = \delta\boldsymbol{F}\boldsymbol{\Pi}$, where δ is a constant term, $\boldsymbol{F}$ is defined as $\boldsymbol{F} = \mathrm{diag}(\boldsymbol{1}_\kappa, \boldsymbol{0}_{k-\kappa})$, which is a $k \times k$ diagonal matrix, and $\boldsymbol{\Pi}$ is defined by $\boldsymbol{1}'_{p/3} \otimes \boldsymbol{\Pi}_1$ when $k = 10$ and by $\boldsymbol{1}'_{p/3} \otimes \boldsymbol{\Pi}_2$ when $k = 15$. Here, $\boldsymbol{\Pi}_1$ and $\boldsymbol{\Pi}_2$ are given by

$$\boldsymbol{\Pi}_1 = \begin{pmatrix} 0.8501 & 0.6571 & 0.2159 \\ -0.2753 & -0.2432 & -0.1187 \\ -0.3193 & -0.2926 & -0.1671 \\ 0.2754 & 0.2608 & 0.1766 \\ 0.2693 & 0.2164 & 0.2066 \\ -0.0676 & -0.0663 & -0.0561 \\ 0.2239 & 0.2197 & 0.1880 \\ -0.0352 & -0.0346 & -0.0305 \\ 0.3240 & 0.3199 & 0.2868 \\ -0.3747 & -0.3727 & -0.3554 \end{pmatrix}, \quad \boldsymbol{\Pi}_2 = \begin{pmatrix} 1.3794 & 0.0645 & 0.0330 \\ -0.0766 & -0.0241 & -0.0143 \\ -0.2618 & -0.1396 & -0.0951 \\ -0.4619 & -0.2589 & -0.1798 \\ 0.2381 & 0.1488 & 0.1082 \\ 0.2140 & 0.1463 & 0.1112 \\ 0.3002 & 0.2364 & 0.1950 \\ 0.1155 & 0.0953 & 0.0812 \\ -0.2774 & -0.2395 & -0.2091 \\ 0.3392 & 0.3072 & 0.2807 \\ 0.0016 & 0.0107 & 0.0100 \\ 0.0438 & 0.0408 & 0.0381 \\ -0.3187 & -0.3039 & -0.2904 \\ 0.0529 & 0.0510 & 0.0493 \\ 0.2505 & 0.2451 & 0.2399 \end{pmatrix}.$$

Here, δ controls the scale of the regression coefficient matrix, and $\boldsymbol{F}$ controls the number of nonzero regression coefficients via κ. The values of the elements of $\boldsymbol{\Pi}_1$ and $\boldsymbol{\Pi}_2$, which are an essential regression coefficient matrix, are the same as in Lawless (1981). The simulated data $\boldsymbol{Y}$ were generated iteratively from $N_{n\times p}(\boldsymbol{X}\boldsymbol{\Xi}, \boldsymbol{\Sigma} \otimes \boldsymbol{I}_n)$ under several selections of n, k, κ, δ, ρ_y, and ρ_x, where $\boldsymbol{\Sigma} = \boldsymbol{R}_p\boldsymbol{\Delta}_p(\rho_y)\boldsymbol{R}_p$, and the number of iterations was 10,000. At each iteration, we evaluated $r(\boldsymbol{X}\boldsymbol{\Xi}, \hat{\boldsymbol{Y}}_{\hat{\boldsymbol{\theta}}})$ where $\hat{\boldsymbol{Y}}_{\hat{\boldsymbol{\theta}}} = \boldsymbol{1}_n\hat{\boldsymbol{\mu}}' + \boldsymbol{X}\hat{\boldsymbol{\Xi}}_{\hat{\boldsymbol{\theta}}}$, which is the predicted value of $\boldsymbol{Y}$ obtained from each method. The average of $np + r(\boldsymbol{X}\boldsymbol{\Xi}, \hat{\boldsymbol{Y}}_{\hat{\boldsymbol{\theta}}})$ across 10,000 iterations was regarded as the PMSE of $\hat{\boldsymbol{Y}}_{\hat{\boldsymbol{\theta}}}$. In the simulation, a standardized $\boldsymbol{X}$ was used to estimate the regression coefficients.

We obtained the optimized ridge parameter $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_k)'$ from the following two methods:

Method 1 : $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}^{[s^*_{\mathrm{C}}]}$ where $s^*_{\mathrm{C}} = \arg\min_{s\in\mathcal{S}} C_p^{\#}$ and $C_p^{\#}$ is defined in (9).

Method 2 : $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}^{[s^*_{\mathrm{M}}]}$ where $s^*_{\mathrm{M}} = \arg\min_{s\in\mathcal{S}} MC_p^{\#}$ and $MC_p^{\#}$ is defined in (10).

In this paper, we set $\mathcal{S} = \{1, 2, 3, 4, 5, 10, 15, 20, 50\}$ and let $\chi(i) = i$ for $i = 1, \ldots, 5$, $\chi(i) = 5(i - 4)$ for $i = 6, 7, 8$, and $\chi(9) = 50$, and we let $\sharp(\mathcal{S})$ be the number of elements in the set $\mathcal{S}$. To reduce the number of computations, we applied the selection method proposed by Ruppert (2002) to select $s \in \mathcal{S}$, as follows:

1. Set i to 1.
2. Calculate several information criteria (IC) for $\chi(i)$ and $\chi(i + 1)$, and denote these IC values as IC(i) and IC$(i + 1)$, respectively.
3. If IC$(i + 1) > 0.98 \times$ IC(i), stop iterating and go to Step 5.

4. If $i+1 \leq \sharp(\mathcal{S})$, update i as $i+1$ and go to Step 2. Otherwise, if $i+1 > \sharp(\mathcal{S})$, do not update i and go to Step 5.

5. When $\mathrm{IC}(i+1) < \mathrm{IC}(i)$, let $i^* = i+1$; otherwise, let $i^* = i$.

6. Obtain the optimized s as $\chi(i^*)$.

By using this selection method, we can reduce the number of computations for selecting s since it stops when there is little improvement in the information criteria when i is large, i.e., when s becomes large. When we use the $C_p^{\#}$ criterion to obtain s, we calculate (9) to obtain $\mathrm{IC}(i)$. As was done for $C_p^{\#}$, $MC_p^{\#}$ in (10) is calculated for each $s \in \mathcal{S}$. For the purpose of comparison with the proposed methods, we prepared the two conventional optimization methods, as follows:

Method 3 : $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}^{[1]} = (\hat{\theta}_1^{[1]}, \ldots, \hat{\theta}_k^{[1]})'$, which is the single plug-in optimization method in (4).

Method 4 : $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}^{[2]} = (\hat{\theta}_1^{[2]}, \ldots, \hat{\theta}_k^{[2]})'$, which is the double plug-in optimization method in (5).

Table 1 shows the simulation results for $\mathrm{PMSE}[\hat{\boldsymbol{Y}}_{\hat{\boldsymbol{\theta}}}]/\{p(n+k+1)\} \times 100$ for the cases in which $(k,n) = (10,30)$ and $(10,50)$, and Table 2 shows the results for the cases in which $(15,30)$ and $(15,50)$. In both tables, $p=6$, where $p(n+k+1)$ is the theoretical value of the PMSE for the predictor of $\boldsymbol{Y}$ based on the LS estimators.

We can see that all of the methods improved the PMSEs of the LS estimators in all cases since none of the values in the tables exceed 100. When $k=10$, Method 2 is almost always the best method for small δ and n. Methods 1 and 2 resulted in a greater improvement than did Method 3 in almost all cases when $k=10$ and $n=30$. Methods 1 and 2 resulted in a greater improvement than did Method 3 in almost all cases when $k=10$, $n=50$, and δ was small. When $k=15$, Method 2 is always the best method for small δ and large ρ_x. Methods 1 and 2 resulted in a greater improvement than did Method 3 in almost cases when $k=15$. Methods 1 and 2 also resulted in a greater improvement than did Method 4 in all cases when δ was small and ρ_x was large. When δ was small and ρ_x was large, Method 2 also resulted in a greater improvement than did Method 1 in almost all cases. Methods 1 and 2 resulted in the greatest improvement when k became large and ρ_y was small. In almost all cases, there was greater improvement when κ was smaller . When n or δ became small, each method was improved. On average, Method 2 was the best method, and Method 1 was the second best. Hence, we recommend using the $MC_p^{\#}$ criterion in (10) to choose the number of repetitions in the multiple plug-in optimization method.

Acknowledgement

The authors thank the editor and the referee for reading the paper very carefully and their valuable suggestions.

References

Table 1: The values of $\mathrm{PMSE}[\hat{\boldsymbol{Y}}_{\hat{\boldsymbol{\theta}}}]/\{p(n+k+1)\} \times 100$ for each method when $k = 10$

	$n = 30$				$n = 50$			
	Method				Method			
$(\kappa,\delta,\rho_y,\rho_x)$	1	2	3	4	1	2	3	4
(0,0,0.2,0.2)	*79.71*	**78.25**	85.46	80.92	*85.28*	**84.82**	89.30	86.10
(0,0,0.2,0.95)	*79.75*	**78.29**	85.50	80.96	*85.31*	**84.85**	89.32	86.12
(0,0,0.95,0.2)	*79.76*	**78.31**	85.50	80.97	*85.33*	**84.86**	89.34	86.13
(0,0,0.95,0.95)	*79.73*	**78.28**	85.49	80.94	*85.35*	**84.88**	89.36	86.16
(5,1,0.2,0.2)	86.47	**86.32**	88.98	*86.46*	*91.45*	91.54	92.73	**91.39**
(5,1,0.2,0.95)	*82.28*	**81.16**	86.88	83.02	*87.50*	**87.09**	90.58	87.93
(5,1,0.95,0.2)	*90.29*	90.63	91.62	**90.21**	*94.20*	94.29	94.81	**94.19**
(5,1,0.95,0.95)	*83.35*	**82.24**	87.58	83.97	*88.72*	**88.39**	91.27	88.94
(5,3,0.2,0.2)	*93.63*	93.90	94.31	**93.59**	97.33	97.24	**97.13**	*97.23*
(5,3,0.2,0.95)	*85.44*	**84.72**	88.69	85.68	90.62	**90.52**	92.36	*90.60*
(5,3,0.95,0.2)	*96.20*	*96.20*	96.36	**96.08**	98.23	98.24	**98.09**	*98.12*
(5,3,0.95,0.95)	*89.00*	89.04	90.88	**88.95**	*92.93*	93.00	94.00	**92.90**
(10,1,0.2,0.2)	*87.12*	**86.67**	89.75	87.21	*94.19*	94.29	94.67	**94.17**
(10,1,0.2,0.95)	*83.75*	**82.77**	87.69	84.22	*88.83*	**88.51**	91.29	89.01
(10,1,0.95,0.2)	*91.09*	91.43	92.25	**91.03**	96.47	96.45	**96.30**	*96.44*
(10,1,0.95,0.95)	*84.65*	**83.76**	88.27	85.02	*89.68*	**89.46**	91.79	89.75
(10,3,0.2,0.2)	*93.53*	93.95	94.14	**93.49**	*98.76*	98.87	**98.72**	98.96
(10,3,0.2,0.95)	*89.76*	89.91	91.37	**89.69**	*93.26*	93.35	94.21	**93.23**
(10,3,0.95,0.2)	98.12	98.09	**97.74**	*98.05*	**99.51**	**99.51**	**99.51**	*99.67*
(10,3,0.95,0.95)	*91.94*	92.36	92.88	**91.89**	*94.78*	94.84	95.29	**94.77**
Average	*87.28*	**86.81**	90.07	87.62	*91.88*	**91.75**	93.50	92.09

Table 2: The values of PMSE[$\hat{\mathbf{Y}}_{\hat{\boldsymbol{\theta}}}$]$/\{p(n+k+1)\} \times 100$ for each method when $k = 15$

	$n = 30$				$n = 50$			
	Method				Method			
$(\kappa,\delta,\rho_y,\rho_x)$	1	2	3	4	1	2	3	4
(0,0,0.2,0.2)	*74.89*	**72.16**	82.13	76.43	*79.46*	**78.83**	85.39	80.96
(0,0,0.2,0.95)	*74.90*	**72.18**	82.10	76.42	*79.42*	**78.81**	85.36	80.94
(0,0,0.95,0.2)	*74.87*	**72.15**	82.09	76.40	*79.47*	**78.84**	85.42	80.98
(0,0,0.95,0.95)	*74.91*	**72.20**	82.13	76.43	*79.43*	**78.82**	85.37	80.95
(5,1,0.2,0.2)	*81.80*	**81.09**	85.64	81.99	*86.56*	**86.40**	88.98	86.57
(5,1,0.2,0.95)	*77.29*	**75.13**	83.28	78.26	*82.00*	**81.54**	86.56	82.85
(5,1,0.95,0.2)	*90.51*	90.89	91.80	**90.47**	*94.27*	94.36	94.61	**94.27**
(5,1,0.95,0.95)	*84.04*	**83.56**	87.26	84.14	*87.74*	**87.55**	89.98	*87.74*
(5,3,0.2,0.2)	*90.43*	90.79	91.75	**90.40**	**93.84**	*93.95*	94.35	**93.84**
(5,3,0.2,0.95)	*83.64*	**83.21**	86.90	83.72	*87.39*	**87.19**	89.68	87.40
(5,3,0.95,0.2)	*96.06*	*96.06*	96.32	**95.91**	98.10	98.05	*98.03*	**97.90**
(5,3,0.95,0.95)	90.43	**90.39**	92.03	*90.40*	*92.12*	92.17	93.36	**92.08**
(10,1,0.2,0.2)	*84.73*	84.93	87.37	**84.68**	*90.53*	90.91	91.49	**90.45**
(10,1,0.2,0.95)	*78.56*	**76.48**	84.17	79.44	*83.20*	**82.71**	87.43	83.95
(10,1,0.95,0.2)	*93.55*	94.03	93.97	**93.52**	96.50	*96.49*	**96.19**	*96.49*
(10,1,0.95,0.95)	*83.57*	**82.79**	87.06	83.74	*87.64*	**87.40**	89.98	87.69
(10,3,0.2,0.2)	*94.13*	94.48	94.57	**94.10**	98.00	97.99	**97.65**	*97.96*
(10,3,0.2,0.95)	*85.13*	**84.92**	87.94	85.14	*89.86*	90.01	91.29	**89.77**
(10,3,0.95,0.2)	*98.60*	98.73	**98.48**	98.68	*99.34*	*99.34*	**99.33**	99.46
(10,3,0.95,0.95)	*91.32*	91.67	92.51	**91.29**	*94.01*	*94.12*	94.58	**94.01**
(15,1,0.2,0.2)	*88.25*	89.06	89.86	**88.19**	*92.51*	92.76	93.08	**92.49**
(15,1,0.2,0.95)	*79.46*	**77.67**	84.59	80.16	*84.13*	**83.72**	87.88	84.66
(15,1,0.95,0.2)	*93.14*	93.53	93.81	**93.12**	96.90	*96.86*	**96.70**	*96.86*
(15,1,0.95,0.95)	*84.09*	**83.47**	87.38	84.22	*88.20*	**88.01**	90.36	88.21
(15,3,0.2,0.2)	96.43	96.46	*96.39*	**96.31**	*98.47*	98.68	**98.29**	98.66
(15,3,0.2,0.95)	86.45	**86.28**	88.91	*86.44*	*91.05*	91.28	92.16	**90.99**
(15,3,0.95,0.2)	*98.61*	98.76	**98.49**	98.71	**99.36**	**99.36**	**99.36**	*99.48*
(15,3,0.95,0.95)	*90.91*	91.03	92.34	**90.88**	*94.13*	94.22	94.70	**94.12**
Average	*86.45*	**85.86**	89.33	86.77	*90.13*	**90.01**	92.06	90.42

Dien, S. J. V., Iwatani, S. Usuda, Y. and Matsui, K. (2006). *Theoretical analysis of amino acid-producing* Eschenrichia coli *using a stoixhiometrix model and multivariate linear regression.* J. Biosci. Bioeng., **102**, 34–40.

Efron, B. (2004). *The estimation of prediction error: covariance penalties and cross-validation.* J. Amer. Statist. Assoc., **99**, 619–632.

Goldstein, M. and Smith, A. F. M. (1974). *Ridge-type estimators for regression analysis.* J. Roy. Statist. Soc. Ser. B, **36**, 284–291.

Hemmerle, W. J. (1975). *An explicit solution for generalized ridge regression.* Technometrics, **17**, 309–314.

Hoerl, A. E. and Kennard, R. W. (1970). *Ridge regression: biased estimation for nonorthogonal problems.* Technometrics, **12**, 55–67.

Lawless, J. F. (1981). *Mean squared error properties of generalized ridge estimators.* J. Amer. Statist. Assoc., **76**, 462–466.

Nagai, I., Yanagihara, H. and Satoh, K. (2012). *Optimization of ridge parameters in multivariate generalized ridge regression by plug-in methods.* Hiroshima Math. J., **42**, 301–324.

Ruppert, D. (2002). *Selection the number of knots for penalized splines.* J. Comput. Graph. Statist., **11**, 735–757.

Sârbu, C., Onişor, C., Posa, M., Kevresan, S. and Kuhajda, K. (2008). *Modeling and prediction (correction) of partition coefficients of bile acids and their derivatives by multivariate regression methods.* Talanta, **75**, 651–657.

Saxén, R. and Sundell, J. (2006). *^{137}Cs in freshwater fish in Finland since 1986– a statistical analysis with multivariate linear regression models.* J. Environ. Radioactiv., **87**, 62–76.

Skagerberg, B. , MacGregor, J. and Kiparissides, C. (1992). *Multivariate data analysis applied to low-density polyethylene reactors.* Chemometr. Intell. Lab. Syst., **14**, 341–356.

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

Stein, C. M. (1981). *Estimation of the mean of a multivariate normal distribution.* Ann. Statist., **9**, 1135–1151.

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

Yanagihara, H., Nagai, I. and Satoh, K (2009). *A bias-corected C_p criterion for optimizing ridge parameters in multivariate generalized ridge regression.* Japanese J. Appl. Statist., **38**, 151-172 (in Japanese).

Yanagihara, H. and Satoh, K. (2010). *An unbiased C_p criterion for multivariate ridge regression.* J. Multivariate Anal., **101**, 1226–1238.

Yoshimoto, A., Yanagihara, H. and Ninomiya, Y. (2005). *Finding factors affecting a forest stand growth through multivariate linear modeling.* J. Jpn. For. Res., **87**, 504–512 (in Japanese).

Received October 11, 2012
Revised August 6, 2013

Hiroshima Math. J.
44 (2014), 315–326

Selecting a shrinkage parameter in structural equation modeling with a near singular covariance matrix by the GIC minimization method

Ami Kamada, Hirokazu Yanagihara, Hirofumi Wakaki and Keisuke Fukui

(Received February 13, 2012)
(Revised May 7, 2014)

Abstract. In the structural equation modeling, unknown parameters of a covariance matrix are derived by minimizing the discrepancy between a sample covariance matrix and a covariance matrix having a specified structure. When a sample covariance matrix is a near singular matrix, Yuan and Chan (2008) proposed the estimation method to use an adjusted sample covariance matrix instead of the sample covariance matrix in the discrepancy function. The adjusted sample covariance matrix is defined by adding a scalar matrix with a shrinkage parameter to the existing sample covariance matrix. They used a constant value as the shrinkage parameter, which was chosen based solely on the sample size and the number of dimensions of the observation, and not on the data itself. However, selecting the shrinkage parameter from the data may lead to a greater improvement in prediction compared to the use of a constant shrinkage parameter. Hence, we propose an information criterion for selecting the shrinkage parameter, and attempt to select the shrinkage parameter by an information criterion minimization method. The proposed information criterion is based on the discrepancy function measured by the normal theory maximum likelihood. Using the Monte Carlo method, we demonstrate that the proposed criterion works well in the sense that the prediction accuracy of an estimated covariance matrix is improved.

1. Introduction

Structural equation modeling (SEM) has been widely used in many fields, especially in social and behavioral sciences (see e.g., Bollen (1989), and Yuan and Bentler (2007)). In SEM, unknown parameters of a covariance matrix are

The Second author is supported by the Ministry of Education, Science, Sports, and Culture, a Grant-in-Aid for Challenging Exploratory Research, #25540012, 2013–2015.

The Third author is supported by the Ministry of Education, Science, Sports, and Culture, Grant-in-Aid for Scientific Research (C), #24500343, 2012–2014.

2010 *Mathematics Subject Classification*. 62H12, 62F07.

Key words and phrases. bias correction, GIC, model selection, near singular covariance matrix, SEM, shrinkage parameter.

derived by minimizing the discrepancy between a sample covariance matrix and a covariance matrix having a specified structure.

Let $\boldsymbol{x}_1, \dots, \boldsymbol{x}_N$ be independent random samples from $\boldsymbol{x}$ distributed according to a p-variate normal distribution $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where N is the sample size. We are interested in modeling the population covariance matrix $\boldsymbol{\Sigma}$. Denote the model of interest as $\boldsymbol{\Sigma}(\boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\theta_1, \dots, \theta_q)'$. For simplicity, we write $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ as $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$. Let $\boldsymbol{S}$ be an unbiased estimator of $\boldsymbol{\Sigma}$, i.e.,

$$\boldsymbol{S} = \frac{1}{N-1}\sum_{i=1}^{N}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})',$$

where $\bar{\boldsymbol{x}}$ is the sample mean of $\boldsymbol{x}_1, \dots, \boldsymbol{x}_N$ defined by $\bar{\boldsymbol{x}} = N^{-1}\sum_{i=1}^{N}\boldsymbol{x}_i$. Then, the candidate model is represented by

$$M : n\boldsymbol{S} \sim W_p(n, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}), \tag{1}$$

where $n = N - 1$. Suppose that $\boldsymbol{\Sigma}_0$ is the true covariance matrix, i.e., $Cov[\boldsymbol{x}] = \boldsymbol{\Sigma}_0$. The true model is represented by

$$M_0 : n\boldsymbol{S} \sim W_p(n, \boldsymbol{\Sigma}_0). \tag{2}$$

If the covariance structure can be correctly specified, then there exists $\boldsymbol{\theta}_0$ such that $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_{\boldsymbol{\theta}_0}$. The classical approach to SEM fits the sample covariance matrix $\boldsymbol{S}$ by $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ through minimizing the normal theory maximum likelihood (ML) discrepancy function as

$$F(\boldsymbol{S}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}) = \mathrm{tr}(\boldsymbol{S}\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}) - \log|\boldsymbol{S}\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}| - p. \tag{3}$$

Then, the ML estimator of $\boldsymbol{\theta}$, which is represented by $\hat{\boldsymbol{\theta}}$, is defined by

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} F(\boldsymbol{S}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}).$$

In general, $\hat{\boldsymbol{\theta}}$ is obtained using a modification of Newton's algorithm (see e.g., Lee and Jennrich (1979)), which requires an iteration process to solve the estimating equation. When $\boldsymbol{S}$ is near singular (not full rank), the iteration process for obtaining $\hat{\boldsymbol{\theta}}$ will be very unstable and may require hundreds of iterations to reach convergence (e.g., Boomsma (1985)). A near singular $\boldsymbol{S}$ often occurs in practical data analysis due to not only small samples but also multicollinearity or missing data even when sample size is quite large (Wothke (1993)). When $\boldsymbol{S}$ is literally singular, it is very likely that the iteration will never converge.

In order to avoid such a problem, Yuan and Chan (2008) proposed a new method in which $\boldsymbol{\theta}$ is estimated by minimizing $F(\boldsymbol{S}_a, \boldsymbol{\Sigma}_{\boldsymbol{\theta}})$, where $\boldsymbol{S}_a = \boldsymbol{S} + a\boldsymbol{I}_p$, a is a small positive value and $\boldsymbol{I}_p$ is a p-dimensional identity matrix. Here, a is commonly referred to as the shrinkage parameter. Hence, a new estimator of $\boldsymbol{\theta}$ is defined by

$$\hat{\boldsymbol{\theta}}_a = \arg \min_{\boldsymbol{\theta}} F(\boldsymbol{S}_a, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}).$$

Although $\hat{\boldsymbol{\theta}}_a$ has a constant bias, under LISREL models (see Jöreskog and Sörbom (1996), pp. 1–3), Yuan and Chan (2008) reported that $\hat{\boldsymbol{\theta}}_a$ can be adjusted to a consistent estimator through a simple procedure when the covariance structure is the correct model. The adjusted estimator is defined as

$$\tilde{\boldsymbol{\theta}}_a = \hat{\boldsymbol{\theta}}_a - a\boldsymbol{j},$$

where $\boldsymbol{j}$ is a q-dimensional vector whose elements are ones corresponding to the parameters on the diagonals of the covariance matrix, and otherwise are zero. They also studied for the case that $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ is not correctly specified. There exists a unique vector $\boldsymbol{\theta}_*$ such that

$$\boldsymbol{\Sigma}_{\boldsymbol{\theta}_{a*}} = \boldsymbol{\Sigma}_{\boldsymbol{\theta}_*} + a\boldsymbol{I}_p, \tag{4}$$

where $\boldsymbol{\theta}_{a*}$ is a population parameter minimizing $F(\boldsymbol{\Sigma}_0 + a\boldsymbol{I}_p, \boldsymbol{\Sigma}_{\boldsymbol{\theta}})$, i.e.,

$$\boldsymbol{\theta}_{a*} = \arg \min_{\boldsymbol{\theta}} F(\boldsymbol{\Sigma}_0 + a\boldsymbol{I}_p, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}). \tag{5}$$

Then, $\hat{\boldsymbol{\theta}}_a$ and $\tilde{\boldsymbol{\theta}}_a$ are consistent for $\boldsymbol{\theta}_{a*}$ and $\boldsymbol{\theta}_*$, respectively. If $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ is correctly specified, $\boldsymbol{\theta}_{a*} = \boldsymbol{\theta}_0 + a\boldsymbol{j}$ and $\boldsymbol{\theta}_* = \boldsymbol{\theta}_0$.

The selection of the shrinkage parameter is crucial because if the shrinkage parameter is changed, the estimate will be also changed. In Yuan and Chan (2008), the shrinkage parameter was taken to be a constant, determined by only N and p. This means that the shrinkage parameter was not chosen based on the data. However, it is possible that the prediction could be improved by basing the shrinkage parameter on the data itself. Furthermore, it does not always guarantee that the estimator is proper solution by fixed a. Therefore, we attempt to select the shrinkage parameter based on the predictive Kullback-Leibler (KL) discrepancy (Kullback and Leibler (1951)). The basic idea is to measure the goodness of fit of the model by the risk function assessed by the predictive KL discrepancy. In the present paper, our objective is to select the appropriate value of a by minimizing the risk function. However, we cannot directly use the risk function to select a because the risk function includes unknown parameters. Hence, instead of the risk function itself, we use its estimator.

Akaike's information criterion (AIC) (Akaike (1973)) is an estimator of the risk function assessed by the predictive KL information (for the AIC for SEM, see, e.g., Cudeck and Brown (1983), Akaike (1987), Ichikawa and Konishi (1999), Yanagihara (2005)). The objective of the present study may be achieved by minimizing the AIC rather than the risk function. In general, the AIC is defined by adding the bias to the risk function, i.e., the number of independent parameters divided by n, to the KL discrepancy function with an estimated parameter, which is referred to as a sample discrepancy function. However, the bias term of the AIC is obtained under the situation that the discrepancy function for estimating $\boldsymbol{\theta}$ is the same as that for evaluating the model fit. In the present paper, the discrepancy function for estimating $\boldsymbol{\theta}$ is

$$F(\boldsymbol{S}_a, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}) = F(\boldsymbol{S}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}) + a \operatorname{tr}(\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}) - \log|\boldsymbol{S}_a| + \log|\boldsymbol{S}|,$$

and that for evaluating the model is $F(\boldsymbol{S}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}})$. Since the two functions are different, we cannot use the bias term of the ordinary AIC. Therefore, we must revaluate the bias using the same approach as the generalized information criterion (GIC) proposed by Konishi and Kitagawa (1996). Hence, we denote the proposed information criterion as GIC(a). We define GIC(a) by adding an estimator of the revaluated bias to the sample discrepancy function $F(\boldsymbol{S}, \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\theta}}_a})$. Then, the best a is chosen by minimizing GIC(a).

The remainder of the present paper is organized as follows: In Section 2, we obtain GIC(a) from a stochastic expansion of $\hat{\boldsymbol{\theta}}_a$. In Section 3, we verify the performance of our criteria using the Monte Carlo method. In Section 4, we present conclusions and discussions. The proof of the theorem presented herein is provided in the Appendix.

2. GIC for selecting the shrinkage parameter

In order to select the best a, we consider the risk function between the true model and the candidate model. Let $\mathscr{L}(\boldsymbol{\Sigma})$ be an expected ML discrepancy function defined by

$$\begin{aligned}\mathscr{L}(\boldsymbol{\Sigma}) &= E[F(\boldsymbol{S}, \boldsymbol{\Sigma})] \\ &= \operatorname{tr}(\boldsymbol{\Sigma}_0 \boldsymbol{\Sigma}^{-1}) - E[\log|\boldsymbol{S}|] + \log|\boldsymbol{\Sigma}| - p.\end{aligned}$$

In this paper, E denotes the expectation under the true model M_0 in (2) with respect to $\boldsymbol{S}$. We measure the discrepancy between the candidate model M in (1) and the true model M_0 in (2) by the predictive KL discrepancy function. Then, we define the risk function assessed by the predictive ML discrepancy in (3) as

$$R = E[\mathscr{L}(\boldsymbol{\Sigma}_{\tilde{\boldsymbol{\theta}}_a})].$$

We regard the shrinkage parameter a having the smallest R as the principle best model. Obtaining an unbiased estimator of R will allow us to correctly evaluate the discrepancy between the data and the model, which will further facilitate the selection of the best shrinkage parameter. A rough estimator of R is the sample ML discrepancy function $F(\boldsymbol{S}, \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\theta}}_a})$. However, since $F(\boldsymbol{S}, \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\theta}}_a})$ has a bias, the information criterion can be defined as $F(\boldsymbol{S}, \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\theta}}_a}) + \hat{B}$, where $\hat{B}$ is an estimator of the bias given as

$$B = R - E[F(\boldsymbol{S}, \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\theta}}_a})]. \tag{6}$$

Henceforth, in order to derive $\hat{B}$, we calculate a limiting value of B.

Let

$$\boldsymbol{\Delta}_{\boldsymbol{\theta}} = \frac{\partial}{\partial \boldsymbol{\theta}'} \operatorname{vec}(\boldsymbol{\Sigma}_{\boldsymbol{\theta}}), \tag{7}$$

and

$$\boldsymbol{G}_{\boldsymbol{\theta}_{a*}} = \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} F(\boldsymbol{\Sigma}_0 + a\boldsymbol{I}_p, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}) \bigg|_{\boldsymbol{\theta} = \boldsymbol{\theta}_{a*}}, \tag{8}$$

where

$$\begin{aligned} &\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} F(\boldsymbol{\Sigma}_0 + a\boldsymbol{I}_p, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}) \\ &\quad = 2\boldsymbol{\Delta}'_{\boldsymbol{\theta}}(\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}(\boldsymbol{\Sigma}_0 + a\boldsymbol{I}_p)\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} \otimes \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1})\boldsymbol{\Delta}_{\boldsymbol{\theta}} - \boldsymbol{\Delta}'_{\boldsymbol{\theta}}(\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} \otimes \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1})\boldsymbol{\Delta}_{\boldsymbol{\theta}} \\ &\qquad - \sum_{i,j}^{q} \operatorname{tr}\{\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}(\boldsymbol{\Sigma}_0 + a\boldsymbol{I}_p - \boldsymbol{\Sigma}_{\boldsymbol{\theta}})\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\ddot{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}ij}\}\boldsymbol{e}_i\boldsymbol{e}_j'. \end{aligned}$$

Here, $\boldsymbol{e}_i$ is a q-dimensional vector, the ith element of which is 1, with all others being 0, and $\ddot{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}ij} = \partial^2\boldsymbol{\Sigma}_{\boldsymbol{\theta}}/\partial\theta_i\partial\theta_j$. Since $\boldsymbol{\theta}_{a*}$ is the minimizer of $F(\boldsymbol{\Sigma}_0 + a\boldsymbol{I}_p, \boldsymbol{\Sigma}_{\boldsymbol{\theta}})$, $\boldsymbol{G}_{\boldsymbol{\theta}_{a*}}$ is a nonsingular matrix. Using the above notation, we have the following theorem for the bias.

Theorem 1. *Suppose that a set of standard regularity conditions, as given in Browne* (1984) *or Yuan and Bentler* (1997), *is satisfied. Then, the bias of* $E[F(\boldsymbol{S}, \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\theta}}_a})]$ *is expanded as*

$$B = \frac{2}{n} \operatorname{tr}\{\boldsymbol{\Delta}_{\boldsymbol{\theta}_*}\boldsymbol{G}_{\boldsymbol{\theta}_{a*}}^{-1}\boldsymbol{\Delta}'_{\boldsymbol{\theta}_{a*}}(\boldsymbol{\Sigma}_{\boldsymbol{\theta}_{a*}}^{-1}\boldsymbol{\Sigma}_0\boldsymbol{\Sigma}_{\boldsymbol{\theta}_*}^{-1} \otimes \boldsymbol{\Sigma}_{\boldsymbol{\theta}_{a*}}^{-1}\boldsymbol{\Sigma}_0\boldsymbol{\Sigma}_{\boldsymbol{\theta}_*}^{-1})\} + O(n^{-2}). \tag{9}$$

The proof of this theorem, which is derived by modifying the results presented in Yanagihara, Himeno, and Yuan (2010), is given in the Appendix.

By replacing $\boldsymbol{\theta}_{a*}$, $\boldsymbol{\theta}_*$, and $\boldsymbol{\Sigma}_0$ by neglecting $O(n^{-2})$ in (9) with $\hat{\boldsymbol{\theta}}_a$, $\tilde{\boldsymbol{\theta}}_a$, and $\boldsymbol{S}$, respectively, an estimator of B is given by

$$\hat{B} = \frac{2}{n} \operatorname{tr}\{\boldsymbol{\Delta}_{\tilde{\boldsymbol{\theta}}_a}\boldsymbol{G}_{\hat{\boldsymbol{\theta}}_a}^{-1}\boldsymbol{\Delta}'_{\hat{\boldsymbol{\theta}}_a}(\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}_a}^{-1}\boldsymbol{S}\boldsymbol{\Sigma}_{\tilde{\boldsymbol{\theta}}_a}^{-1} \otimes \boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}_a}^{-1}\boldsymbol{S}\boldsymbol{\Sigma}_{\tilde{\boldsymbol{\theta}}_a}^{-1})\}.$$

Thus, the information criterion for selecting a (GIC(a)) is defined by

$$\mathrm{GIC}(a) = F(\boldsymbol{S}, \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\theta}}_a}) + \hat{B}.$$

Let A be a set $A = \{a \mid a \geq 0$ and $\tilde{\boldsymbol{\theta}}_a$ gives a proper solution$\}$. Then, the best a is chosen by minimizing GIC(a), i.e.,

$$\hat{a} = \arg \min_{a \in A} \mathrm{GIC}(a).$$

When the candidate model is correctly specified, $\boldsymbol{\Sigma}_{\boldsymbol{\theta}_{a*}} = \boldsymbol{\Sigma}_a$. Then, the bias becomes simple, as in the following corollary.

Corollary 1. *If the candidate model is correctly specified, the bias of* $E[F(\boldsymbol{S}, \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\theta}}_a})]$ *is expanded as*

$$B = \frac{2}{n} q + O(n^{-2}).$$

This corollary indicates that the bias does not depend on a by neglecting the $O(n^{-2})$ term when the candidate model is correctly specified. Hence, the best a is the value that minimizes $F(\boldsymbol{S}, \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\theta}}_a})$ in A.

3. Monte Carlo results

In this section, we compare the risk functions of estimated $\boldsymbol{\Sigma}$ obtained from the following methods.

- Method 1 (new method): We estimate $\boldsymbol{\Sigma}$ by $\boldsymbol{\Sigma}_{\tilde{\boldsymbol{\theta}}_{\hat{a}}}$, where $\hat{a}$ is selected by minimizing GIC(a).
- Method 2 (Yuan and Chan's (YC) method): We estimate $\boldsymbol{\Sigma}$ by $\boldsymbol{\Sigma}_{\tilde{\boldsymbol{\theta}}_{p/N}}$.
- Method 3 (ordinary ML method): We estimate $\boldsymbol{\Sigma}$ by $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}}$.

Actually, since $-E[\log|\boldsymbol{S}|] - p$ in the expected ML discrepancy does not depend on the result of a selection of a, we evaluated the following expectations:

$$R_{\mathrm{new}} = E[\mathscr{L}(\boldsymbol{\Sigma}_{\tilde{\boldsymbol{\theta}}_{\hat{a}}})] + \alpha, \qquad R_{\mathrm{YC}} = E[\mathscr{L}(\boldsymbol{\Sigma}_{\tilde{\boldsymbol{\theta}}_{p/N}})] + \alpha, \qquad R_{\mathrm{ML}} = E[\mathscr{L}(\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}})] + \alpha,$$

where $\alpha = E[\log|\boldsymbol{S}|] + p$. In the simulation, we used the confirmatory factor model, which is included in the LISREL model, as the true model M_0, i.e., the true covariance matrix is $\boldsymbol{\Sigma}_0 = \boldsymbol{\Lambda}_0 \boldsymbol{\Phi}_0 \boldsymbol{\Lambda}_0' + \boldsymbol{\Psi}_0$, where $\boldsymbol{\Lambda}_0$ is the true factor loading matrix, $\boldsymbol{\Phi}_0$ is the true correlation matrix, and $\boldsymbol{\Psi}_0$ is the true covariance matrix of the measurement errors. In this simulation, we defined $\boldsymbol{\Psi}_0 = \boldsymbol{I}_p - \mathrm{diag}(\boldsymbol{\Lambda}_0 \boldsymbol{\Phi}_0 \boldsymbol{\Lambda}_0')$. As the true model, we used the two models specified by the following parameters:

$$\text{Case 1:}\quad \boldsymbol{\Lambda}_0 = \begin{pmatrix} \boldsymbol{b} & \mathbf{0}_5 \\ \mathbf{0}_5 & \boldsymbol{b} \\ \mathbf{0}_5 & \boldsymbol{b} \end{pmatrix}, \qquad \boldsymbol{\Phi}_0 = \begin{pmatrix} 1.0 & .30 \\ .30 & 1.0 \end{pmatrix},$$

$$\text{Case 2:}\quad \boldsymbol{\Lambda}_0 = \begin{pmatrix} \boldsymbol{b} & \mathbf{0}_5 & \mathbf{0}_5 \\ \mathbf{0}_5 & \boldsymbol{b} & \mathbf{0}_5 \\ \mathbf{0}_5 & \mathbf{0}_5 & \boldsymbol{b} \end{pmatrix}, \qquad \boldsymbol{\Phi}_0 = \begin{pmatrix} 1.0 & .30 & .40 \\ .30 & 1.0 & .30 \\ .40 & .30 & 1.0 \end{pmatrix},$$

where $\boldsymbol{b} = (.70, .70, .75, .80, .80)'$ and $\mathbf{0}_q$ is a q-dimensional vector of zeros. The candidate model used in the simulation was also the confirmatory factor model, i.e., the covariance matrix $\boldsymbol{\Sigma}_\theta = \boldsymbol{\Lambda\Phi\Lambda}' + \boldsymbol{\Psi}$, where $\boldsymbol{\Psi} = \mathrm{diag}(\psi_1, \dots, \psi_p)$. In the case 1, we used the confirmatory three-factor model as the candidate model. On the other hand, the confirmatory two-factor model was used as the candidate model in the case 2. Hence, $\boldsymbol{\lambda}$ and $\boldsymbol{\Phi}$ in the candidate models were

$$\text{Case 1:}\quad \boldsymbol{\Lambda} = \begin{pmatrix} \boldsymbol{\lambda}_1 & \mathbf{0}_5 & \mathbf{0}_5 \\ \mathbf{0}_5 & \boldsymbol{\lambda}_2 & \mathbf{0}_5 \\ \mathbf{0}_5 & \mathbf{0}_5 & \boldsymbol{\lambda}_3 \end{pmatrix}, \qquad \boldsymbol{\Phi} = \begin{pmatrix} 1.0 & \phi_{12} & \phi_{13} \\ \phi_{12} & 1.0 & \phi_{23} \\ \phi_{13} & \phi_{23} & 1.0 \end{pmatrix},$$

$$\text{Case 2:}\quad \boldsymbol{\Lambda} = \begin{pmatrix} \boldsymbol{\lambda}_1 & \mathbf{0}_5 \\ \mathbf{0}_5 & \boldsymbol{\lambda}_2 \\ \mathbf{0}_5 & \boldsymbol{\lambda}_3 \end{pmatrix}, \qquad \boldsymbol{\Phi} = \begin{pmatrix} 1.0 & \phi_{12} \\ \phi_{12} & 1.0 \end{pmatrix}.$$

It is easy to see that the candidate model in the case 1 is overspecified, and that in the case 2 is underspecified. In order to obtain smaller sample sizes, we chose $N = 30$, 50, and 100. The number of replications is 1000.

In order to calculate R_{new}, R_{YC}, and R_{ML}, we first obtained an estimator of $\boldsymbol{\theta}$ for each method using R ver. 2.12.1. We then counted the frequencies when the estimate of $\boldsymbol{\theta}$ is the proper solution (i.e., an estimator of $\boldsymbol{\Sigma}$ is positive define). Next, we recorded the value of $\mathscr{L}(\hat{\boldsymbol{\Sigma}})$ for each method, where $\hat{\boldsymbol{\Sigma}}$ is an estimated $\boldsymbol{\Sigma}$ for each method. After the replication was finished, we obtained the arithmetic mean of $\mathscr{L}(\hat{\boldsymbol{\Sigma}})$ for each method. If all of the estimators are proper solutions, then the arithmetic mean is regarded as a target risk function.

From Table 1, when $N = 30$ in the case 1, the R_{new} was obtained, but R_{YC} and R_{ML} were not obtained because there were several improper solutions for $a = p/N$ and 0. When $N = 50$ and 100 in the case 1, since there were no improper solutions, we could obtain all risk functions. Then, R_{new} was the smallest. On the other hand, in the case 2, R_{new} and R_{YC} were obtained, but R_{ML} was not obtained. Then, R_{new} was smaller than R_{YC}. Hence, the

Table 1. Frequencies of the proper solutions and the risk functions for each method

Case	N	Frequency			Risk		
		New	YC	ML	New	YC	ML
1	30	1000	996	987	16.8295	—	—
	50	1000	1000	1000	15.9808	15.9858	16.0088
	100	1000	1000	1000	15.5024	15.5044	15.5067
2	30	1000	1000	972	19.2521	19.3887	—
	50	1000	1000	987	16.1748	16.2869	—
	100	1000	1000	990	14.1732	14.2618	—

proposed information criterion works well in the sense that the prediction accuracy of an estimated covariance matrix is improved.

4. Conclusion and discussion

In the present paper, we proposed a GIC for selecting the shrinkage parameter, which is used to obtain the estimator for SEM with a near singular covariance matrix. In order to derive the GIC, we revaluated the bias of the risk function. Then, GIC(a) was obtained by adding the estimator of the revaluated bias to the sample discrepancy function. We have observed that when the candidate model is correctly specified, the bias does not depend on a when the $O(n^{-2})$ term is neglected, i.e., the bias term is equivalent to that of the AIC. This means that the best a is the value that minimizes $F(\boldsymbol{S}, \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\theta}}_a})$ under the condition that $\tilde{\boldsymbol{\theta}}_a$ gives a proper solution. In the Monte Carlo results, an estimate of $\tilde{\boldsymbol{\theta}}_{\hat{a}}$ was always a proper solution, and the risk function of the estimated covariance matrix based on $\tilde{\boldsymbol{\theta}}_{\hat{a}}$ with the selected a was the smallest.

In this paper, we assumed that data has normality. If we do not assume normality to data, a kurtosis will appear in the bias to the risk function. Hence, an estimator of kurtosis will be required to estimate the bias. Unfortunately, Yanagihara (2007) reported that such an estimator gives a poor value unless the sample is huge. When the sample size is large enough, a sample covariance matrix will not become a near singular matrix in most cases. A near singular sample covariance matrix occurs frequently under the small or moderate sample sizes. This is almost the same as a well-known fact that a multicollinearity frequently occur under the small or moderate sample. In practice, we confirmed such results through many simulation experiments. Hence, it is suitable to assume not the large sample case but the small or moderate sample case under a near singular sample covariance matrix. Therefore, at present, we judge that it is necessary to deal with the case of nonnormal when a sample covariance matrix is a near singular matrix.

Appendix

The derivation of the risk function and the proof of Theorem 1 are presented in this appendix. First, we derive the risk function. In this paper, we measure the discrepancy between the candidate model M in (1) and the true model M_0 in (2) by the following discrepancy function:

$$\int \log \frac{f(\boldsymbol{W}|n, \boldsymbol{\Sigma}_0)}{f(\boldsymbol{W}|n, \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\theta}}_a})} f(\boldsymbol{W}|n, \boldsymbol{\Sigma}_0) d\boldsymbol{W} = \frac{n}{2}\{\mathscr{L}(\boldsymbol{\Sigma}_{\tilde{\boldsymbol{\theta}}_a}) - \mathscr{L}(\boldsymbol{\Sigma}_0)\}.$$

By omitting the terms that do not depend on a, we have

$$\int F(\boldsymbol{W}, \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\theta}}_a}) f(\boldsymbol{W}|n, \boldsymbol{\Sigma}_0) d\boldsymbol{W} = \mathscr{L}(\boldsymbol{\Sigma}_{\tilde{\boldsymbol{\theta}}_a}).$$

Hence, we define the risk function as R in Section 2.

Next, we prove Theorem 1. The bias of $F(\boldsymbol{S}, \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\theta}}_a})$, defined in (6), can be written as

$$B = E[\mathscr{L}(\boldsymbol{\Sigma}_{\tilde{\boldsymbol{\theta}}_a}) - F(\boldsymbol{S}, \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\theta}}_a})] = E[\mathrm{tr}\{\boldsymbol{\Sigma}_{\tilde{\boldsymbol{\theta}}_a}^{-1}(\boldsymbol{\Sigma}_0 - \boldsymbol{S})\}]. \tag{A1}$$

Since $\boldsymbol{\Sigma}_0 - \boldsymbol{S} = O_p(n^{-1/2})$ and $E[\boldsymbol{S}] = \boldsymbol{\Sigma}_0$, by applying the Taylor expansion to $\mathrm{tr}\{\boldsymbol{\Sigma}_{\tilde{\boldsymbol{\theta}}_a}^{-1}(\boldsymbol{\Sigma}_0 - \boldsymbol{S})\}$ at $\tilde{\boldsymbol{\theta}}_a = \boldsymbol{\theta}_*$, we derive

$$E[\mathrm{tr}\{\boldsymbol{\Sigma}_{\tilde{\boldsymbol{\theta}}_a}^{-1}(\boldsymbol{\Sigma}_0 - \boldsymbol{S})\}] = E[\boldsymbol{d}_{\boldsymbol{\theta}_*}(\tilde{\boldsymbol{\theta}}_a - \boldsymbol{\theta}_*)] + O(n^{-2}),$$

where $\boldsymbol{\theta}_*$ is given by (4), and

$$\boldsymbol{d}_{\boldsymbol{\theta}_*} = \frac{\partial}{\partial \boldsymbol{\theta}'} \mathrm{tr}\{\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}(\boldsymbol{\Sigma}_0 - \boldsymbol{S})\}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_*}.$$

The remainder term of the above expectation is $O(n^{-2})$ because $\tilde{\boldsymbol{\theta}}_a$ can be expressed as a function of $\boldsymbol{V} = n^{1/2}(\boldsymbol{S} - \boldsymbol{\Sigma}_0)$ which has an asymptotic normality and general cumulants of elements of $\boldsymbol{V}$ may be expanded as a power series in n^{-1} (see e.g., Hall, 1992, p. 46). Indeed, an $n^{-3/2}$ term of the stochastic expansion of $\mathrm{tr}\{\boldsymbol{\Sigma}_{\tilde{\boldsymbol{\theta}}_a}^{-1}(\boldsymbol{\Sigma}_0 - \boldsymbol{S})\}$ can be expressed as the third-order polynomial of elements of $\boldsymbol{V}$. Since $\boldsymbol{V}$ has an asymptotic normality, an expectation of the odd-order polynomial of element $\boldsymbol{V}$ becomes $O(n^{-1/2})$. Consequently, the expectation of the $n^{-3/2}$ term of the stochastic expansion becomes not $O(n^{-3/2})$ but $O(n^{-2})$. Let $\boldsymbol{\Gamma}_{\boldsymbol{\theta}} = (\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} \otimes \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1})$. From this expression, we obtain

$$\boldsymbol{d}_{\boldsymbol{\theta}_*} = \mathrm{vec}'\{\boldsymbol{\Sigma}_{\boldsymbol{\theta}_*}^{-1}(\boldsymbol{S} - \boldsymbol{\Sigma}_0)\boldsymbol{\Sigma}_{\boldsymbol{\theta}_*}^{-1}\}\boldsymbol{\Delta}_{\boldsymbol{\theta}_*} = \frac{1}{\sqrt{n}} \mathrm{vec}'(\boldsymbol{V})\boldsymbol{\Gamma}_{\boldsymbol{\theta}_*}\boldsymbol{\Delta}_{\boldsymbol{\theta}_*}. \tag{A2}$$

Since $\hat{\boldsymbol{\theta}}_a$ is the minimizer of $F(\boldsymbol{S}_a, \boldsymbol{\Sigma}_{\boldsymbol{\theta}})$, $\partial F(\boldsymbol{S}_a, \boldsymbol{\Sigma}_{\boldsymbol{\theta}})/\partial \boldsymbol{\theta}|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_a} = \boldsymbol{0}_q$ is satisfied. Then, under a set of standard regularity conditions, the following equation is derived.

$$\boldsymbol{0}_q = \boldsymbol{\Delta}'_{\hat{\boldsymbol{\theta}}_a} \operatorname{vec}\{\boldsymbol{\Sigma}^{-1}_{\hat{\boldsymbol{\theta}}_a}(\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}_a} - \boldsymbol{\Sigma}_0 - a\boldsymbol{I}_p)\boldsymbol{\Sigma}^{-1}_{\hat{\boldsymbol{\theta}}_a}\} - \frac{1}{\sqrt{n}}\boldsymbol{\Delta}'_{\hat{\boldsymbol{\theta}}_a}\boldsymbol{\Gamma}_{\hat{\boldsymbol{\theta}}_a} \operatorname{vec}(\boldsymbol{V}).$$

Hence, we obtain

$$\boldsymbol{\Delta}'_{\hat{\boldsymbol{\theta}}_a} \operatorname{vec}\{\boldsymbol{\Sigma}^{-1}_{\hat{\boldsymbol{\theta}}_a}(\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}_a} - \boldsymbol{\Sigma}_0 - a\boldsymbol{I}_p)\boldsymbol{\Sigma}^{-1}_{\hat{\boldsymbol{\theta}}_a}\} = \frac{1}{\sqrt{n}}\boldsymbol{\Delta}'_{\hat{\boldsymbol{\theta}}_a}\boldsymbol{\Gamma}_{\hat{\boldsymbol{\theta}}_a} \operatorname{vec}(\boldsymbol{V}). \tag{A3}$$

Note that $n^{1/2}(\hat{\boldsymbol{\theta}}_a - \boldsymbol{\theta}_{a*}) = O_p(1)$ and that both sides of (A3) are functions of $\hat{\boldsymbol{\theta}}_a$, where $\boldsymbol{\theta}_{a*}$ is given by (5). Applying the Taylor expansion to (A3) at $\hat{\boldsymbol{\theta}}_a = \boldsymbol{\theta}_{a*}$ and comparing the $O_p(n^{-1})$ term on both sides of the resulting equation, we obtain

$$\hat{\boldsymbol{\theta}}_a - \boldsymbol{\theta}_{a*} = \frac{1}{\sqrt{n}}\boldsymbol{G}^{-1}_{\boldsymbol{\theta}_{a*}}\boldsymbol{\Delta}'_{\boldsymbol{\theta}_{a*}}\boldsymbol{\Gamma}_{\boldsymbol{\theta}_{a*}} \operatorname{vec}(\boldsymbol{V}) + O_p(n^{-1}),$$

where $\boldsymbol{\Delta}_{\boldsymbol{\theta}}$ and $\boldsymbol{G}_{\boldsymbol{\theta}}$ are given by (7) and (8), respectively. Note that

$$\begin{aligned} E[\operatorname{vec}(\boldsymbol{V}) \operatorname{vec}'(\boldsymbol{V})] &= nE[\operatorname{vec}(\boldsymbol{S} - \boldsymbol{\Sigma}_0) \operatorname{vec}'(\boldsymbol{S} - \boldsymbol{\Sigma}_0)] \\ &= nCov[\operatorname{vec}(\boldsymbol{S})] \\ &= (\boldsymbol{I}_{p^2} + \boldsymbol{K}_p)(\boldsymbol{\Sigma}_0 \otimes \boldsymbol{\Sigma}_0), \end{aligned}$$

where $\boldsymbol{K}_p$ is the commutation matrix (see Magnus and Neudecker (1999), p. 48). Therefore,

$$\begin{aligned} B &= E[\boldsymbol{d}_{\boldsymbol{\theta}_*}(\hat{\boldsymbol{\theta}}_a - \boldsymbol{\theta}_{a*})] + O(n^{-2}) \\ &= \frac{1}{n} \operatorname{tr}\{\boldsymbol{\Gamma}_{\boldsymbol{\theta}_*}\boldsymbol{\Delta}_{\boldsymbol{\theta}_*}\boldsymbol{G}^{-1}_{\boldsymbol{\theta}_{a*}}\boldsymbol{\Delta}'_{\boldsymbol{\theta}_{a*}}\boldsymbol{\Gamma}_{\boldsymbol{\theta}_{a*}}(\boldsymbol{I}_{p^2} + \boldsymbol{K}_p)(\boldsymbol{\Sigma}_0 \otimes \boldsymbol{\Sigma}_0)\} + O(n^{-2}). \end{aligned} \tag{A4}$$

Consequently, by using the equations $\boldsymbol{K}_p(\boldsymbol{A} \otimes \boldsymbol{C}) = (\boldsymbol{C} \otimes \boldsymbol{A})\boldsymbol{K}_p$ and $\boldsymbol{K}_p \operatorname{vec}(\boldsymbol{C}) = \operatorname{vec}(\boldsymbol{C}')$ (see Magnus and Neudecker (1999), p. 47), the equation (9) in Theorem 1 is derived.

Acknowledgement

We wish to express our deepest gratitude to Prof. Y. Kano of Osaka University for his valuable advices. We would also like to thank Dr. I. Nagai, Chukyo University, and Mr. S. Imori and Mr. Y. Hashiyama, Hiroshima

University, for their encouragements. Furthermore, the authors wish to thank the referee for helpful suggestions.

References

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

[2] Akaike, H. (1987). Factor analysis and AIC. Psychometrika, **52**, 317–332.

[3] Bollen, K. A. (1989). Structural Equations with Latent Variables. John Wiley & Sons, Inc., New York.

[4] Boomsma, A. (1985). Nonconvergence, improper solutions, and starting values in LISREL maximum likelihood estimation. Psychometrika, **50**, 229–242.

[5] Browne, M. W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. British J. Math. Statist. Psych., **37**, 62–83.

[6] Cudeck, R. and Brown, M. W. (1983). Cross-validation of covariance structures. Multivariate Behav. Res., **18**, 147–167.

[7] Hall, P. (1992). The Bootstrap and Edgeworth Expansion. Springer, New York.

[8] Ichikawa, M. and Konishi, S. (1999). Model evaluation and Information criteria in covariance structure analysis. British J. Math. Statist. Psych., **52**, 285–302.

[9] Jöreskog, K. G. and Sörbom, D. (1996). LISREL 8 User's Reference Guide. Scientific Software International, Chicago.

[10] Konishi, S. and Kitagawa, G. (1996). Generalised information criteria in model selection. Biometrika, **83**, 875–890.

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

[12] Lee, S. Y. and Jennrich, R. I. (1979). A study of algorithms for covariance structure analysis with specific comparisons using factor analysis. Psychometrika, **44**, 99–113.

[13] Magnus, J. R. and Neudecker, H. (1999). Matrix Differential Calculus with Applications in Statistics and Econometrics (revised ed.). John Wiley & Sons, Inc., New York.

[14] Yanagihara, H. (2005). Selection of covariance structure models in nonnormal data by using information criterion: an application to data from the survey of the Japanese notional character. Proc. Inst. Statist. Math., **53**, 133–157 (in Japanese).

[15] Yanagihara, H. (2007). A family of estimators for multivariate kurtosis in a nonnormal linear regression model. J. Multivariate Anal., **98**, 1–29.

[16] Yanagihara, H., Himeno, T. and Yuan, K.-H. (2010). GLS discrepancy based information criteria for selecting covariance structure models. Behaviormetrika, **37**, 71–86.

[17] Yuan, K.-H. and Bentler, P. M. (1997). Mean and covariance structure analysis: theoretical and practical improvements. J. Amer. Statist. Assoc., **92**, 767–774.

[18] Yuan, K.-H. and Bentler, P. M. (2007). Structural equation modeling. In Handbook of Statistics **27**: Psychometrics (C. R. Rao & S. Sinharay eds.), Elsevier/North-Holland, Amsterdam, 297–358.

[19] Yuan, K.-H. & Chan, W. (2008). Structural equation modeling with near singular covariance matrices. Comput. Statist. Data Anal., **52**, 4842–4858.

[20] Wothke, W. (1993). Nonpositive definite matrices in structural modeling. In Testing Structural Equation Models (K. A. Bollen & J. S. Long eds.), Sage, Newbury park, CA, 256–293.

Ami Kamada
Biostatistics, Clinical Data Science Department
Takeda Development Center Japan
Pharmaceutical Development Division
Takeda Pharmaceutical Company Limited
1-1 *Doshomachi* 4-*chome, Chuo-ku, Osaka* 540-8645*, Japan*
E-mail: ami.kamada@gmail.com

Hirokazu Yanagihara
Department of Mathematics
Hiroshima University
1-3-1 *Kagamiyama, Higashi-Hiroshima, Hiroshima* 739-8626*, Japan*
E-mail: yanagi@math.sci.hiroshima-u.ac.jp

Hirofumi Wakaki
Department of Mathematics
Hiroshima University
1-3-1 *Kagamiyama, Higashi-Hiroshima, Hiroshima* 739-8626*, Japan*
E-mail: wakaki@math.sci.hiroshima-u.ac.jp

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

Advance Publication J-STAGE: Jan. 23, 2015
©2015 FORMATH Research Group

FORMATH Vol. 14 (2015): xx–xx, DOI:10.15684/formath.14.004

Comparison with Residual-Sum-of-Squares-Based Model Selection Criteria for Selecting Growth Functions

Keisuke Fukui[1], Mariko Yamamura[1], Hirokazu Yanagihara[1*]

Abstract: A growth curve model used for analyzing growth is characterized by a mathematical function with respect to time, called a growth function. As the results of analysis from a growth curve model strongly depend on the growth function used for the analysis, the selection of growth functions is important. A choice of growth function based on the minimization of a model selection criterion is one of the major selection methods. In this paper, we compare the performances of growth-function selection methods using these criteria (e.g., Mallows' Cp criterion) through Monte Carlo simulations. As a result, we recommend the use of a method employing the Bayesian information criterion for the selection of growth functions.

Keywords: growth curve model, growth-function-selection, model selection criterion, residual sum of squares

1. Introduction

A growth curve model used for analyzing growth is specified by a mathematical function, called the growth function. A number of growth functions may be used for analysis; therefore, growth-function-selection (GF-selection) is important because the results of analysis from a growth curve model vary according to the growth function used. Naturally, a growth function with high prediction performance is regarded as a better growth function. Hence, during GF-selection, the best model should be chosen to improve prediction accuracy.

Choosing growth functions based on the minimization of a model selection criterion (MSC) is one of the major selection methods. An MSC consists of two terms; a goodness-of-fit term and a penalty term based on the complexity of the model. Particularly, an MSC whose goodness-of-fit term is the residual sum of squares (RSS) is called an RSS-based MSC in this paper. An RSS-based MSC is often used to select the best model in many fields. Because several RSS-based MSC approaches can be used to estimate the risk function assessing the standardized mean square error (MSE) of the prediction, we can expect that the accuracy of a growth prediction will be improved in the sense of making the MSE small by minimizing an RSS-based MSC. However, numerous RSS-based MSC approaches, e.g., Mallows' Cp criterion (Mallows, 1973), are available, and the chosen growth function will depend upon the MSC employed for GF-selection. Hence, the purpose of this study is to compare the performances of GF-selection methods using RSS-based MSC through Monte Carlo simulations.

The remainder of this paper is organized as follows. In Section 2, we introduce the growth curve model and the growth functions used. In Section 3, we describe the RSS-based MSC approaches considered for GF-selection. In Section 4, we compare the GF-selection methods considered through numerical experiments and discuss the results.

2. Growth Curve Model

2.1 True and Candidate Models

Let $y(t_i)$ be the extent of growth at a time t_i $(i = 1, \ldots, n)$, where n is the sample size. Suppose that $y(t_i)$ is generated from the following true model:

$$[1] \qquad y(t_i) = \mu_*(t_i) + \varepsilon_*(t_i),$$

where $\mu_*(t_i)$ is the true expected value of $y(t_i)$, and $\varepsilon_*(t_1), \ldots, \varepsilon_*(t_n)$ are mutually independent true error variables derived from the same distribution with a mean 0 and variance σ_*^2. As $\mu_*(t)$

Received Aug.15, 2014; Accepted Dec.22, 2014

[1] Hiroshima University, Japan
[*] Corresponding Author:yanagi@math.sci.hiroshima-u.ac.jp

expresses the average value of the true growth, $\mu_*(t)$ is denoted by the growth function. However, the true model is unknown. Hence, the following candidate model is assumed for $y(t_i)$:

[2] $$y(t_i) = \mu(t_i) + \varepsilon(t_i),$$

where $\mu(t_i)$ is the expected value of $y(t_i)$ under the candidate model, and $\varepsilon(t_1), \ldots, \varepsilon(t_n)$ are mutually independent error variables derived from the same distribution with a mean 0 and variance σ^2. Here, $\mu(t_i)$ is denoted as the candidate growth function. In practice, we must prepare a specific function with respect to t, whose shape is determined by unknown parameters, as the candidate growth function.

Let $\mu(t; \boldsymbol{\theta}_\mu)$ denote the candidate growth function, where $\boldsymbol{\theta}_\mu$ represents a $q(\mu)$-dimensional vector. Note that $q(\mu)$ denotes the number of unknown parameters of a candidate growth function μ. To use the growth curve model, $\boldsymbol{\theta}_\mu$ must be estimated from growth data. In this paper, $\boldsymbol{\theta}_\mu$ is obtained by least squares (LS) estimation. Let the RSS be denoted by

[3] $$\mathrm{RSS}(\boldsymbol{\theta}_\mu; \mu) = \sum_{i=1}^{n} \{y(t_i) - \mu(t_i; \boldsymbol{\theta}_\mu)\}^2 .$$

Then, the LS estimator of $\boldsymbol{\theta}_\mu$ is derived by minimizing $\mathrm{RSS}(\boldsymbol{\theta}_\mu; \mu)$ as

[4] $$\hat{\boldsymbol{\theta}}_\mu = \arg\min_{\boldsymbol{\theta}_\mu} \mathrm{RSS}(\boldsymbol{\theta}_\mu; \mu).$$

Using $\hat{\boldsymbol{\theta}}_\mu$, a growth curve can be estimated by $\mu(t; \hat{\boldsymbol{\theta}}_\mu)$.

2.2 Selection of Growth Functions

Numerous growth functions have been proposed in the literature. In this paper, we consider the following twelve candidate growth functions that were described in Zeide (1993).

(1) Bertalanffy: $\mu_1(t; \boldsymbol{\theta}) = \alpha(1 - e^{-\beta t})^3$ $(\boldsymbol{\theta} = (\alpha, \beta)')$.

(2) Chapman-Richards: $\mu_2(t; \boldsymbol{\theta}) = \alpha(1 - e^{-\beta t})^\gamma$ $(\boldsymbol{\theta} = (\alpha, \beta, \gamma)')$.

(3) Gompertz: $\mu_3(t; \boldsymbol{\theta}) = \alpha \exp(-\beta e^{-\gamma t})$ $(\boldsymbol{\theta} = (\alpha, \beta, \gamma)')$.

(4) Hossfeld-4: $\mu_4(t; \boldsymbol{\theta}) = \alpha(1 + \beta t^{-\gamma})^{-1}$ $(\boldsymbol{\theta} = (\alpha, \beta, \gamma)')$.

(5) Korf: $\mu_5(t; \boldsymbol{\theta}) = \alpha \exp(-\beta t^{-\gamma})$ $(\boldsymbol{\theta} = (\alpha, \beta, \gamma)')$.

(6) Levakovic-3: $\mu_6(t; \boldsymbol{\theta}) = \alpha(1 + \beta t^{-2})^{-\gamma}$ $(\boldsymbol{\theta} = (\alpha, \beta, \gamma)')$.

(7) Logistic: $\mu_7(t; \boldsymbol{\theta}) = \alpha(1 + \beta e^{-\gamma t})^{-1}$ $(\boldsymbol{\theta} = (\alpha, \beta, \gamma)')$.

(8) Monomolecular: $\mu_8(t; \boldsymbol{\theta}) = \alpha(1 - \beta e^{-\gamma t})$ $(\boldsymbol{\theta} = (\alpha, \beta, \gamma)')$.

(9) Weibull: $\mu_9(t; \boldsymbol{\theta}) = \alpha(1 - e^{-\beta t^\gamma})$ $(\boldsymbol{\theta} = (\alpha, \beta, \gamma)')$.

(10) Levakovic-1: $\mu_{10}(t; \boldsymbol{\theta}) = \alpha(1 + \beta t^{-\gamma})^{-\delta}$ $(\boldsymbol{\theta} = (\alpha, \beta, \gamma, \delta)')$.

(11) Sloboda : $\mu_{11}(t; \boldsymbol{\theta}) = \alpha \exp(-\beta e^{-\gamma t^\delta})$ $(\boldsymbol{\theta} = (\alpha, \beta, \gamma, \delta)')$.

(12) Yoshida-1 : $\mu_{12}(t; \boldsymbol{\theta}) = \alpha(1 + \beta t^{-\gamma})^{-1} + \delta$ $(\boldsymbol{\theta} = (\alpha, \beta, \gamma, \delta)')$.

In the above list, t denotes the time, and all parameters are restricted to positive values. The candidate growth functions have been listed in the order of increasing number of unknown parameters, i.e., the function μ_1 includes two parameters, the functions μ_2 to μ_9 include three and the functions μ_{10} to μ_{12} include four.

Although an estimate of a growth curve can be obtained by the LS estimation, the choice of growth function most suited to the obtain growth data is important. In this paper, we select the best growth function by the RSS-based MSC minimization method. Let $\mathrm{MSC}_{\mathrm{RSS}}(\mu)$ denote a general form of a RSS-based MSC. The best growth function is then determined according to

[5] $$\hat{\mu} = \arg\min_{\mu \in \{\mu_1, \ldots, \mu_{12}\}} \mathrm{MSC}_{\mathrm{RSS}}(\mu).$$

2.3. Underspecified and Overspecified Models

An evaluation of the growth function equations given above indicate that several growth functions are equivalent under certain conditions (e.g., Chapman-Richards with $\gamma = 3$ corresponds perfectly to Bertalanffy). In model selection, these relationships sometimes play key roles because several MSC approaches are derived under the assumption that a candidate model includes the true model. We define the following two specific candidate models.

- An overspecified model: a growth function of a candidate model includes that of the true model, i.e., the true growth function can be expressed as a special case of the growth function of the overspecified model. In general, the true model is the overspecified model. However, in this paper, we rule out the true model from the definition of an overspecified model.
- An underspecified model: the model is neither the overspecified model nor the true model.

In practice, there is no overspecified model in most cases. An overspecified model does not exist except under the following three cases:

(i) When the true growth function is Bertalanffy, the candidate model whose growth function is Chapman-Richards is the overspecified model.

(ii) When the true growth function is Gompertz, the candidate model whose growth function is Sloboda is the overspecified model.

(iii) When the true growth function are Hossfeld-4 or Levakovic-3, the candidate model whose growth function is Levakovic-1 is the overspecified model.

3. RSS-based Model Selection Criteria

In this section, we describe explicit forms of the RSS-based MSC approaches used in this work for GF-selection.

When the penalty for the complexity of a model is imposed additively, an estimator of σ^2 is required for the use an RSS-based MSC. In the general regression model, an estimator of σ^2 in the full model is typically employed. A full model is the model that includes all candidate models. For example, if we consider growth functions (1)-(12) as candidate models, the full model includes all growth functions (1)-(12). However, constructing the full model in the growth curve model is difficult because there is no candidate model that includes all candidate models. Hence, we use the following estimator of σ^2 derived from a local linear fitting, which was proposed by Gasser, Sroka and Jennen-Steinmetz (1986),

$$[6] \qquad \hat{\sigma}^2_{\mathrm{L}} = \frac{1}{n-2} \sum_{i=2}^{n-1} \frac{(a_i y_{i-1} + b_i y_{i+1} - y_i)^2}{a_i^2 + b_i^2 - 1},$$

where coefficients a_i and b_i are given by

$$[7] \qquad a_i = \frac{t_{i+1} - t_i}{t_{i+1} - t_{i-1}}, \quad b_i = \frac{t_i - t_{i-1}}{t_{i+1} - t_{i-1}}.$$

The representation $\hat{\sigma}^2_{\mathrm{L}}$ has a desirable property as an estimator of σ^2, e.g., $\hat{\sigma}^2_{\mathrm{L}}$ converges to σ^2 as $n \to \infty$ in probability if $\mu_*(t)$ is twice continuously differentiable, $\limsup_{n\to\infty} \max_{i=2,\ldots,n-1} |t_i - t_{i-1}| < \infty$ and $E[\varepsilon_*(t_i)^4] < \infty$.

3.1. Mallows' C_p Criterion

Using $2q(\mu)$ as the penalty term, Mallows' C_p criterion is defined as

$$[8] \qquad C_p(\mu) = \frac{\mathrm{RSS}(\hat{\boldsymbol{\theta}}_\mu; \mu)}{\hat{\sigma}^2_{\mathrm{L}}} + 2q(\mu).$$

The $2q(\mu)$ was derived as the bias of $\mathrm{RSS}(\hat{\boldsymbol{\theta}}_\mu;\mu)/\hat{\sigma}_{\mathrm{L}}^2$ to the risk function assessing the standardized MSE of prediction under the assumption that the candidate model considered is not an underspecified model. Hence, there is a possibility that the C_p may not correctly evaluate the complexity of an underspecified model.

3.2. Modified C_p Criterion

The weakness of the C_p criterion may be overcome using the generalized degree of freedom (GDF), proposed by Ye (1998) instead of $q(\mu)$. The GDF of the growth curve model was calculated by Kamo and Yoshimoto (2013) as

$$[9] \qquad df(\mu) = q(\mu) + \mathrm{tr}\left\{ \left(\boldsymbol{I}_\mu(\hat{\boldsymbol{\theta}}_\mu) - \boldsymbol{J}_\mu(\hat{\boldsymbol{\theta}}_\mu) \right)^{-1} \boldsymbol{I}_\mu(\hat{\boldsymbol{\theta}}_\mu) \right\},$$

where $\boldsymbol{I}_\mu(\hat{\boldsymbol{\theta}}_\mu)$ and $\boldsymbol{J}_\mu(\hat{\boldsymbol{\theta}}_\mu)$ are matrices given by

$$[10] \qquad \boldsymbol{I}_\mu(\hat{\boldsymbol{\theta}}_\mu) = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \mu(t_i;\boldsymbol{\theta}_\mu)}{\partial \boldsymbol{\theta}_\mu} \frac{\partial \mu(t_i;\boldsymbol{\theta}_\mu)}{\partial \boldsymbol{\theta}_\mu'} \Bigg|_{\boldsymbol{\theta}_\mu=\hat{\boldsymbol{\theta}}_\mu},$$

$$[11] \qquad \boldsymbol{J}_\mu(\hat{\boldsymbol{\theta}}_\mu) = \frac{1}{n}\sum_{i=1}^{n} \{y(t_i) - \mu(t_i;\boldsymbol{\theta}_\mu)\} \frac{\partial^2 \mu(t_i;\boldsymbol{\theta}_\mu)}{\partial \boldsymbol{\theta}_\mu \partial \boldsymbol{\theta}_\mu'} \Bigg|_{\boldsymbol{\theta}_\mu=\hat{\boldsymbol{\theta}}_\mu}.$$

In this paper, "$\boldsymbol{a}'$" denotes the transpose of a vector $\boldsymbol{a}$. Kamo and Yoshimoto (2013) proposed the following modified C_p (MC_p) expressed by replacing $q(\mu)$ with $df(\mu)$ in [8] as

$$[12] \qquad MC_p(\mu) = \frac{\mathrm{RSS}(\hat{\boldsymbol{\theta}}_\mu;\mu)}{\hat{\sigma}_{\mathrm{L}}^2} + 2df(\mu).$$

The description of the expression as "modified" indicates that the bias of $\mathrm{RSS}(\hat{\boldsymbol{\theta}}_\mu;\mu)/\hat{\sigma}_{\mathrm{L}}^2$ to the risk function is corrected even under an underspecified model. A modified C_p criterion was originally proposed by Fujikoshi and Satoh (1997) in the multivariate linear regression model. As the MC_p was derived under the assumption that the candidate model may be an underspecified model, the MC_p may correctly evaluate the complexity of an underspecified model. If the candidate model considered is an overspecified model, then $df(\mu)$ converges to $q(\mu)$ as $n \to \infty$ in probability.

3.3. Bayesian Information Criterion(BIC)-type C_p Criterion

The Bayesian information criterion (BIC) proposed by Schwarz (1978) is very well known MSC. In the BIC, the penalty term is given as "(the number of parameters)$\times \log n$". Using $q(\mu)\log n$ instead of $2q(\mu)$ in [8], the BIC-type C_p (BC_p) can be proposed as

$$[13] \qquad BC_p(\mu) = \frac{\mathrm{RSS}(\hat{\boldsymbol{\theta}}_\mu;\mu)}{\hat{\sigma}_{\mathrm{L}}^2} + q(\mu)\log n.$$

Recall that the purpose of GF-selection employed here is to choose a growth function that improves the growth-prediction of the selected model. However, a consistency property wherein the selection probability of the true model by the MSC approaches 1 asymptotically is also an important property of the model selection. Because BIC has a consistency property, we can expect that BC_p has one too.

3.4. Generalized Cross-Validation Criterion

The generalized cross-validation (GCV) criterion proposed by Craven and Wahba (1979) is one of the RSS-based MSC approaches. In the GCV criterion, the penalty attributed to the complexity of a model is imposed not additively but multiplicatively. The GCV based the GDF was proposed by Ye (1998). The GCV for GF-selection is defined by

$$[14] \qquad \mathrm{GCV}(\mu) = \frac{\mathrm{RSS}(\hat{\boldsymbol{\theta}}_\mu;\mu)}{\{1 - df(\mu)/n\}^2}.$$

If $\hat{\sigma}^2_{\mathrm{L}}$ does not work well, there are possibilities that C_p, MC_p and BC_p will possibly become unstable. However, even if $\hat{\sigma}^2_{\mathrm{L}}$ does not work well, the GCV does not become unstable because the GCV in [14] is defined without an estimator of σ^2.

4. Numerical Study

4.1. Setting

In this section, we compare the performance of each criterion by conducting numerical experiments with several sample sizes, variances and true growth functions. At first, we prepared the twelve true growth functions listed as cases 1-12 below.

Case 1: $\mu_*(t)$ is Bertalanffy as $\mu_*(t) = 100(1 - e^{-0.5t})^3$.

Case 2: $\mu_*(t)$ is Chapman-Richards as $\mu_*(t) = 100(1 - e^{-0.4t})^{3.8}$.

Case 3: $\mu_*(t)$ is Gompertz as $\mu_*(t) = 100\exp(-3e^{-0.3t})$.

Case 4: $\mu_*(t)$ is Hossfeld-4 as $\mu_*(t) = 100(1 + 5t^{-1.5})^{-1}$.

Case 5: $\mu_*(t)$ is Korf as $\mu_*(t) = 100\exp(-3t^{-1})$.

Case 6: $\mu_*(t)$ is Levakovic-3 as $\mu_*(t) = 100(1 + 5t^{-2})^{-1.5}$.

Case 7: $\mu_*(t)$ is Logistic as $\mu_*(t) = 100(1 + 5e^{-0.4t})^{-1}$.

Case 8: $\mu_*(t)$ is Monomolecular as $\mu_*(t) = 100(1 - 1.35e^{-0.25t})$.

Case 9: $\mu_*(t)$ is Weibull as $\mu_*(t) = 100(1 - e^{-0.6t^{0.7}})$.

Case 10: $\mu_*(t)$ is Levakovic-1 as $\mu_*(t) = 100(1 + 3t^{-2.3})^{-2}$.

Case 11: $\mu_*(t)$ is Sloboda as $\mu_*(t) = 100\exp(-4e^{-0.5t^{0.8}})$.

Case 12: $\mu_*(t)$ is Yoshida-1 as $\mu_*(t) = 80(1 + 5t^{-1.4t})^{-1} + 20$.

We used $t_i = 2 + 18i/(n-1)$ $(i = 1, \ldots, n)$ as the time series with $n = 30, 50, 100, 300$ and 500, and generated error variables of the true model from $N(0, \sigma^2_*)$ with $\sigma^2_* = 1$ and 2. The shapes of the true growth curves are shown in Figures 1 and 2. In this paper, we assessed the performances of the GF-selection methods according to the following two properties derived from $1,000$ repetitions.

- The prediction error (PE) of the best growth function chosen by minimizing the MSC.
- The selection probability (SP) of the true growth function chosen by minimizing the MSC.

Here, the PE is defined by

$$\text{[15]} \qquad \mathrm{PE} = \frac{1}{n}\sum_{j=n+1}^{n+3n/10}\left\{\mu_*(t_j) - \hat{\mu}(t_j; \hat{\boldsymbol{\theta}}_{\hat{\mu}})\right\}^2,$$

where $t_j = 2 + 18j/(n-1)$. Note that the PE is a more important property because the aim of our study is to select a growth function that improves the growth prediction of the selection model.

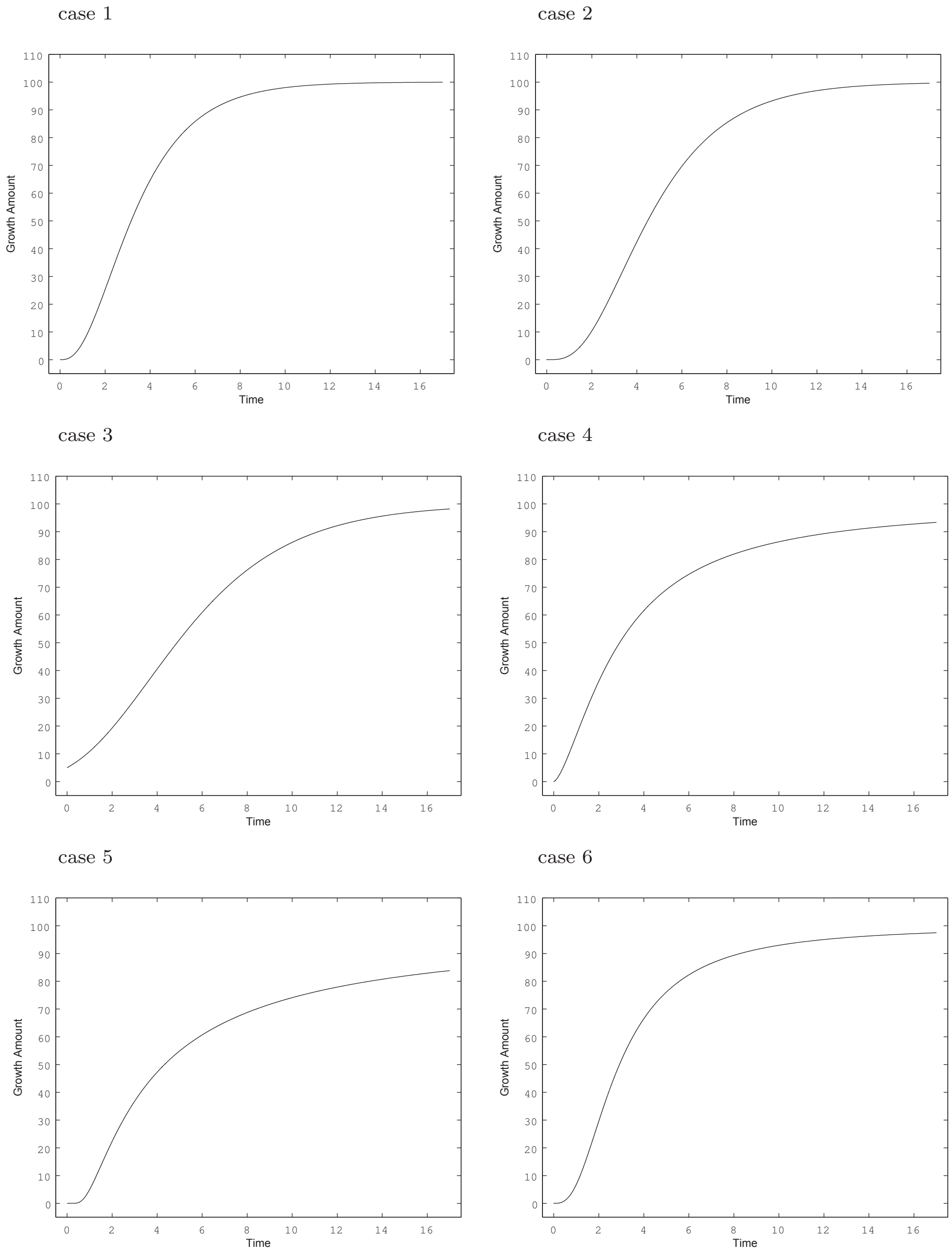
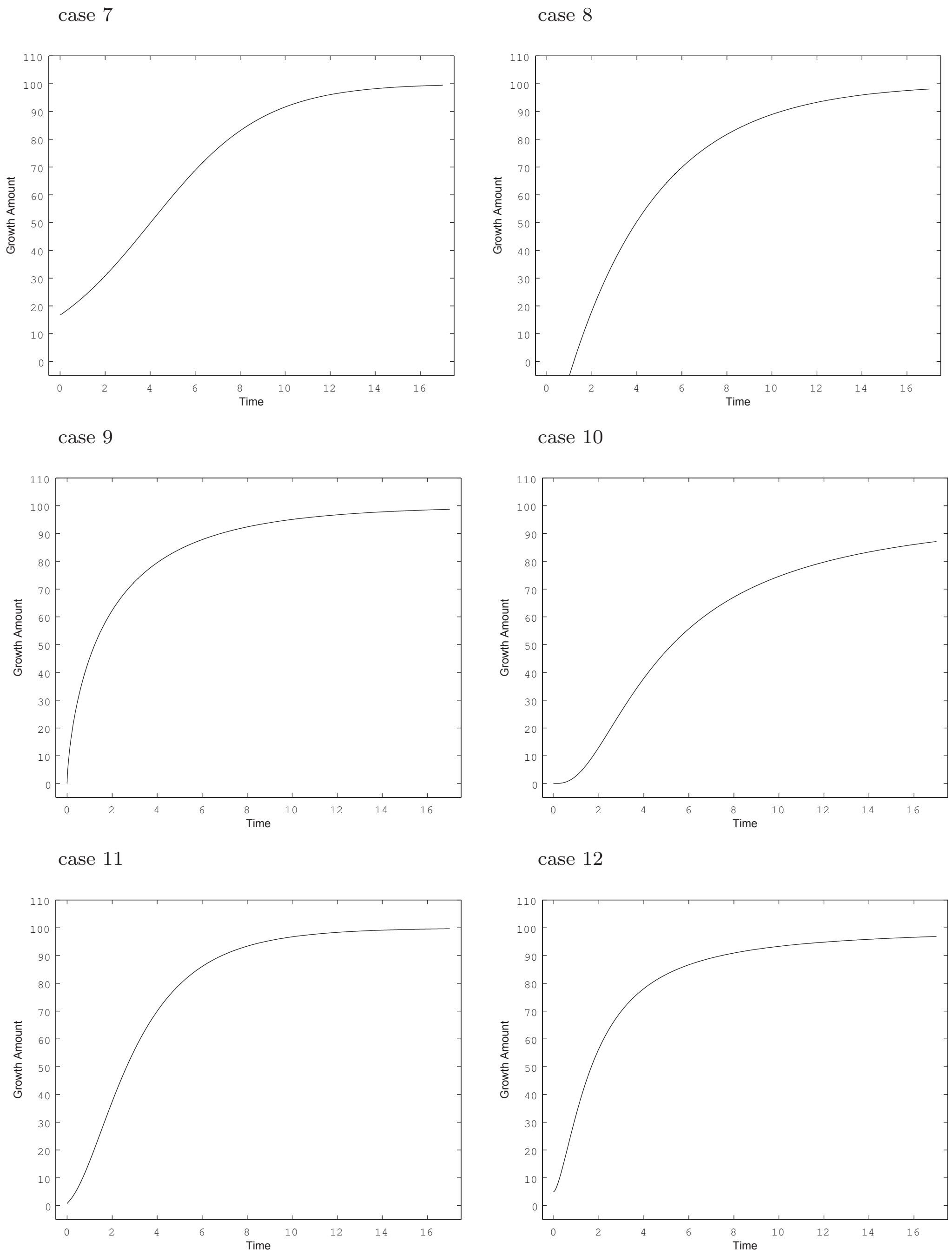


Figure 1. The shapes of the true growth curves (case 1 to case 6).


case 7

case 8

case 9

case 10

case 11

case 12

Figure 2. The shapes of the true growth curves (case 7 to case 12).

4.2. Results

Tables 1 and 2 list the PEs of the best growth functions when $\sigma_*^2 = 1$ and 2, respectively. Additionally Tables 3 and 4 list the SPs of the true growth functions when $\sigma_*^2 = 1$ and 2, respectively. The number in the first column labeled "case" indicates that growth function used as the true growth function. For example, a number 1 in the first column indicates that simulation data were generated from the true growth function of case 1, i.e., Bertalanffy. Furthermore, the addition of an asterisk * denotes that the case is the overspecified model. In the tables, bold fonts indicate the smallest PEs of the best growth functions, and the highest SPs of the true growth functions (although the PEs are rounded at the second decimal place, the smallest value is based on the original values).

From the tables, we obtained the following results:

- When the number of parameters of the true growth function was not large, i.e., cases 1 to 9, BC_p was the high-performance MSC in most cases. Particularly, when the sample size was not small, the SPs of the true growth function by BC_p were always the highest among all MSC approaches. The differences between the SPs were large in cases where an overspecified model existed, i.e., cases 1, 3, 4 and 6. This is because BC_p has a consistency property and C_p, MC_p and GCV do not, i.e., the SPs of BC_p asymptotically converge to 1 although those of C_p, MC_p and GCV do not for cases 1, 3, 4 and 6.

- When the number of parameters of the true growth function was large, i.e., cases 10 to 12, BC_p was not the high-performance MSC. This is because the penalty term of BC_p was too large in cases 10 to 12. In general, BC_p tends to choose a model having a smaller number of known parameters than the true model. Conversely, C_p, MC_p and GCV tend to choose a model having a larger number of known parameters than the true model. In cases 10 to 12, none of the models had a larger number of known parameters than the true model. Hence, the SPs of C_p, MC_p and GCV tended to be higher than those of BC_p. Although the PEs of the best models chosen by C_p, MC_p and GCV tended to be smaller than those chosen by BC_p, the differences were not large.

Based upon the simulation results, using a selection method employing BC_p is recommended for selecting growth functions.

Table 1. The prediction error under each case when $\sigma_*^2 = 1$.

case	n	C_p	MC_p	BC_p	GCV	case	n	C_p	MC_p	BC_p	GCV
1*	30	1.13	1.14	**1.11**	1.14	7	30	1.32	1.34	**1.26**	1.33
	50	1.09	1.09	**1.06**	1.09		50	1.21	1.21	**1.15**	1.21
	100	1.04	1.04	**1.02**	1.04		100	1.10	1.10	**1.06**	1.10
	300	1.01	1.01	**1.01**	1.01		300	1.02	1.02	**1.02**	1.02
	500	1.01	1.01	**1.00**	1.01		500	1.01	f 1.01	**1.01**	1.02
2	30	1.42	1.43	**1.42**	1.43	8	30	1.53	1.56	**1.52**	1.55
	50	1.23	1.23	1.23	**1.23**		50	1.34	1.34	**1.30**	1.33
	100	1.09	1.09	**1.08**	1.09		100	1.15	1.15	**1.12**	1.15
	300	1.02	1.02	**1.02**	1.02		300	1.04	1.04	**1.03**	1.04
	500	1.01	1.03	**1.01**	1.02		500	1.02	1.03	**1.01**	1.03
3*	30	1.53	1.53	**1.41**	1.54	9	30	**1.40**	1.45	1.40	1.45
	50	1.33	1.33	**1.22**	1.34		50	**1.29**	1.31	1.29	1.31
	100	1.18	1.18	**1.11**	1.17		100	1.15	1.17	**1.15**	1.16
	300	1.06	1.06	**1.03**	1.05		300	1.05	1.05	**1.05**	1.05
	500	1.02	1.02	**1.01**	1.03		500	1.03	1.03	**1.02**	1.03
4*	30	1.49	1.49	1.49	**1.48**	10	30	**1.22**	1.25	1.23	1.25
	50	1.29	**1.27**	1.29	1.28		50	**1.15**	1.17	1.16	1.17
	100	1.13	1.13	1.13	**1.12**		100	**1.07**	1.08	1.09	1.08
	300	1.03	1.03	**1.02**	1.03		300	**1.02**	1.02	1.03	1.02
	500	1.02	1.02	**1.01**	1.02		500	1.14	1.14	1.16	**1.13**
5	30	1.36	1.36	**1.36**	1.36	11	30	**1.94**	2.01	1.94	2.03
	50	**1.21**	1.21	1.22	**1.21**		50	**1.68**	1.71	1.70	1.71
	100	**1.09**	1.09	1.09	1.10		100	**1.42**	1.45	1.52	1.45
	300	1.03	1.03	**1.02**	1.03		300	**1.26**	1.30	1.35	1.30
	500	1.01	1.02	**1.01**	1.02		500	**1.04**	1.04	1.05	1.05
6*	30	**1.31**	1.31	1.31	1.31	12	30	1.60	**1.58**	1.60	1.58
	50	1.17	1.16	**1.16**	1.16		50	1.43	**1.43**	1.44	1.43
	100	**1.06**	1.06	1.06	1.06		100	1.27	**1.26**	1.30	1.26
	300	1.02	1.02	**1.02**	1.02		300	**1.12**	1.12	1.24	1.12
	500	1.01	1.01	**1.01**	1.01		500	**1.02**	1.02	1.02	1.02

Table 2. The prediction error under each case when $\sigma_*^2 = 2$.

case	n	C_p	MC_p	BC_p	GCV	case	n	C_p	MC_p	BC_p	GCV
1^*	30	2.53	2.57	**2.43**	2.57	7	30	3.20	3.24	**3.09**	3.25
	50	2.39	2.40	**2.27**	2.40		50	2.85	2.87	**2.67**	2.87
	100	2.16	2.17	**2.11**	2.17		100	2.48	2.49	**2.31**	2.49
	300	2.06	2.06	**2.03**	2.06		300	2.13	2.13	**2.08**	2.13
	500	2.03	2.06	**2.02**	2.06		500	2.06	2.06	**2.04**	2.07
2	30	**3.47**	3.55	3.51	3.57	8	30	4.12	4.26	**4.03**	4.30
	50	**2.91**	2.95	2.95	2.95		50	**3.49**	3.61	3.49	3.59
	100	2.50	2.52	**2.50**	2.52		100	2.81	2.83	**2.76**	2.84
	300	2.14	2.14	**2.12**	2.14		300	2.22	2.22	**2.16**	2.21
	500	2.06	2.12	**2.04**	2.11		500	2.11	2.12	**2.08**	2.12
3^*	30	3.83	3.92	**3.79**	3.95	9	30	**3.18**	3.22	3.18	3.24
	50	3.00	3.10	**2.88**	3.09		50	**2.80**	2.84	2.80	2.84
	100	2.52	2.53	**2.34**	2.54		100	**2.43**	2.48	2.43	2.48
	300	2.23	2.23	**2.12**	2.23		300	**2.18**	2.21	2.18	2.21
	500	2.11	2.14	**2.06**	2.15		500	**2.12**	2.14	2.13	2.14
4^*	30	3.53	**3.49**	3.55	3.50	10	30	**2.76**	2.78	2.92	2.79
	50	2.99	2.96	2.98	**2.96**		50	**2.46**	2.47	2.51	2.47
	100	2.56	2.55	2.55	**2.55**		100	**2.25**	2.27	2.26	2.27
	300	2.18	2.18	2.18	**2.18**		300	**2.09**	2.11	2.11	2.11
	500	2.10	2.10	2.11	**2.10**		500	2.38	2.38	2.72	**2.36**
5	30	3.58	**3.52**	3.59	3.53	11	30	**4.36**	4.56	4.44	4.60
	50	2.95	**2.92**	2.95	2.92		50	**3.57**	3.69	3.62	3.71
	100	2.51	**2.51**	2.51	2.51		100	**2.95**	3.03	3.02	3.03
	300	2.13	2.13	2.13	**2.13**		300	**2.53**	2.56	2.57	2.56
	500	2.08	2.08	**2.07**	2.08		500	2.10	2.13	**2.10**	2.12
6^*	30	3.12	3.05	3.14	**3.04**	12	30	3.59	**3.51**	3.59	3.53
	50	2.69	2.68	2.70	**2.67**		50	3.06	**3.04**	3.07	3.04
	100	2.34	2.34	**2.33**	2.34		100	2.66	**2.65**	2.67	2.66
	300	2.10	**2.10**	2.11	2.10		300	2.33	**2.31**	2.34	2.31
	500	2.05	2.05	**2.05**	2.05		500	2.08	**2.08**	2.09	2.08

Table 3. The selection probability under each case when $\sigma_*^2 = 1$.

case	n	C_p	MC_p	BC_p	GCV	case	n	C_p	MC_p	BC_p	GCV
1*	30	71.9	71.8	**85.5**	71.1	7	30	80.8	79.7	**89.0**	79.4
	50	72.7	72.1	**90.9**	73.3		50	85.5	85.0	**93.8**	85.2
	100	74.8	74.7	**93.8**	74.9		100	89.6	89.3	**97.6**	89.5
	300	81.0	80.7	**97.9**	80.6		300	93.1	93.0	**99.5**	92.7
	500	82.8	83.0	**98.9**	64.7		500	89.3	89.2	**99.3**	77.6
2	30	74.1	74.1	**78.8**	74.2	8	30	76.1	75.3	**77.9**	75.7
	50	79.2	79.3	**85.3**	80.3		50	80.8	80.3	**83.6**	80.7
	100	88.7	89.3	**95.1**	89.2		100	88.2	88.2	**90.1**	88.2
	300	93.8	94.0	**98.8**	93.8		300	97.1	96.8	**97.9**	96.8
	500	97.2	96.5	**98.9**	95.1		500	98.8	98.3	**99.3**	98.2
3*	30	63.1	63.2	**74.2**	63.8	9	30	25.9	18.0	**26.0**	17.9
	50	67.0	67.0	**80.9**	66.8		50	28.2	21.3	**28.5**	21.1
	100	73.8	73.6	**89.6**	73.6		100	38.2	32.0	**38.8**	32.0
	300	77.0	77.0	**96.1**	77.5		300	52.8	50.2	**58.6**	49.9
	500	88.6	88.3	**98.5**	81.2		500	60.8	60.3	**67.0**	55.9
4*	30	57.3	55.2	**57.7**	55.5	10	30	2.3	4.0	0.2	**5.6**
	50	70.3	67.3	**70.9**	67.7		50	**12.7**	12.1	1.6	11.7
	100	80.0	75.7	**83.6**	76.0		100	**38.6**	36.2	7.2	36.4
	300	81.2	75.6	**98.6**	75.5		300	**77.5**	77.3	55.7	77.1
	500	87.2	77.2	**98.9**	68.4		500	49.3	50.1	46.7	**50.4**
5	30	85.3	55.7	**87.5**	56.1	11	30	**1.2**	**1.2**	0.5	**1.2**
	50	87.9	56.9	**90.5**	57.2		50	4.6	**4.8**	1.0	4.5
	100	89.1	55.4	**95.7**	54.9		100	12.9	13.2	2.9	**13.4**
	300	87.4	52.8	**98.4**	52.4		300	24.9	25.3	15.2	**25.6**
	500	95.8	79.4	**99.6**	77.1		500	61.1	61.6	38.0	**64.4**
6*	30	54.8	54.1	**55.4**	54.5	12	30	1.5	**4.9**	0.0	3.3
	50	63.7	63.1	**65.6**	63.0		50	3.6	**6.5**	0.1	5.3
	100	71.4	70.3	**77.4**	70.8		100	12.8	**17.5**	0.3	17.1
	300	83.5	81.9	**90.4**	82.1		300	**53.8**	52.3	11.8	52.0
	500	88.3	88.1	**95.2**	82.3		500	8.7	10.8	2.8	**16.7**

Table 4. The selection probability under each case when $\sigma_*^2 = 2$.

case	n	C_p	MC_p	BC_p	GCV	case	n	C_p	MC_p	BC_p	GCV
1*	30	62.9	61.6	**79.1**	62.1	7	30	65.8	64.2	**75.0**	64.7
	50	65.2	64.0	**85.5**	65.1		50	72.4	71.1	**83.0**	70.5
	100	69.1	68.1	**91.1**	68.2		100	76.2	75.3	**91.6**	76.4
	300	75.7	75.1	**95.8**	75.3		300	87.6	87.0	**98.0**	87.3
	500	79.2	78.6	**98.4**	73.2		500	86.5	86.4	**98.8**	82.3
2	30	**46.2**	45.5	44.9	**46.2**	8	30	48.5	46.0	**48.7**	45.7
	50	56.7	56.1	**56.9**	56.2		50	54.3	52.9	**54.6**	53.0
	100	69.9	69.5	**73.6**	69.9		100	66.0	64.8	**67.1**	65.3
	300	84.6	84.6	**93.8**	84.9		300	83.3	83.3	**86.6**	83.4
	500	90.6	88.3	**98.5**	86.8		500	89.7	89.1	**93.5**	88.4
3*	30	50.9	50.7	**57.8**	51.1	9	30	**10.9**	7.6	**10.9**	7.6
	50	54.2	53.7	**66.1**	53.5		50	14.6	8.2	**14.8**	8.0
	100	61.2	61.2	**76.4**	61.3		100	19.9	12.9	**20.0**	12.9
	300	72.2	72.3	**90.0**	72.1		300	35.2	28.8	**35.5**	28.8
	500	84.4	83.0	**95.2**	79.1		500	42.5	39.9	**42.8**	39.5
4*	30	29.8	27.7	**30.4**	27.9	10	30	1.0	**7.7**	0.5	7.4
	50	37.5	35.0	**38.1**	35.0		50	1.8	7.3	0.6	**7.4**
	100	47.5	45.1	**48.4**	45.1		100	5.2	**10.1**	0.9	10.0
	300	74.2	71.0	**75.0**	71.5		300	**22.5**	20.0	1.1	19.6
	500	83.6	72.7	**86.3**	70.0		500	35.0	37.0	13.0	**38.1**
5	30	69.6	40.3	**71.1**	40.2	11	30	0.1	0.2	0.1	**0.3**
	50	71.7	45.6	**73.2**	45.6		50	0.0	**0.1**	0.0	**0.1**
	100	78.2	47.9	**80.9**	47.9		100	0.1	**0.3**	0.0	**0.3**
	300	86.6	54.1	**93.7**	54.3		300	**5.8**	**5.8**	0.3	5.6
	500	91.5	75.1	**97.7**	74.5		500	17.9	18.3	0.6	**24.1**
6*	30	28.3	28.4	28.4	**28.5**	12	30	0.5	**2.3**	0.0	2.1
	50	37.5	37.3	**37.6**	**37.6**		50	0.7	**3.1**	0.0	3.0
	100	47.0	45.9	**47.1**	45.7		100	1.1	4.3	0.0	4.0
	300	65.9	64.3	**68.7**	64.4		300	5.8	**10.3**	0.0	9.4
	500	74.1	73.3	**78.7**	70.8		500	2.6	6.7	0.3	**7.7**

References

Craven, P. and Wahba, G. (1979) Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation, *Numer. Math.* 31: 377–403.

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

Gasser, T., Sroka, L. and Jennen-Steinmetz, C. (1986) Residual variance and residual pattern in nonlinear regression model, *Biometrika* 73: 625–633.

Kamo, K. and Yoshimoto, A. (2013) Comparative analysis of GFs based on Mallows' C_p type criterion, *FORMATH* 12: 133–147.

Mallows, C. L. (1973) Some Comments on C_p, *Technometrics* 15: 611–675.

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

Ye, J. (1998) On measuring and correcting the effects of data mining and model selection, *J. Amer. Statist. Assoc.* 93: 120–131.

Zeide, B. (1993) Analysis of growth equations, *Forest Sci.* 39: 594–616.