Jackknife Bias Correction of the AIC for Selecting Variables in Canonical Correlation Analysis under Model Misspecification

(Last Modified: May 22, 2012)

Yusuke Hashiyama, Hirokazu Yanagihara[1] and Yasunori Fujikoshi[2]

Department of Mathematics, Graduate School of Science, Hiroshima University

1-3-1 Kagamiyama, Higashi-Hiroshima 739-8626, Japan

Abstract

In this paper, we deal with a bias correction of the Akaike information criterion (AIC) for selecting variables in the canonical correlation analysis when a goodness of fit of the model is assessed by the risk function consisting of the expected Kullback-Leibler loss function with a normal assumption. Although the bias of the AIC to the risk function is $O(n^{-1})$ when the model is correctly specified, its order turns into $O(1)$ when the model is misspecified. By using the jackknife method with a constant adjustment, we propose a new criterion that reduces the AIC's bias to $O(n^{-2})$ even when the model is misspecified, and is an exact unbiased estimator of the risk function when data is generated from the normal distribution. By conducting numerical experiments, we verify that our proposed criteria perform better than the existing criteria.

AMS 2010 *subject classifications*: Primary 62F40; Secondary 62H20.

Key words: Bias correction, Canonical correlation analysis, Jackknife method, Model misspecification, Nonnormality, Variable selection.

1. Introduction

Canonical correlation analysis (CCA), which analyzes the correlation of two linearly combined variables, is an important method in multivariate analysis. CCA has been introduced in many textbooks for applied statistical analysis (see e.g., Srivastava, 2002,

[1] Corresponding author, *E-mail address*: yanagi@math.sci.hiroshima-u.ac.jp. The second author's research is partially supported by the Ministry of Education, Science, Sports, and Culture, Grant-in-Aid for Challenging Exploratory Research, #22650058, 2010-2012.

[2] The third author's research is partially supported by the Ministry of Education, Science, Sports, and Culture, Grant-in-Aid for Scientific Research (C), #22500259, 2010–2012.

Chapter 14.7; Timm, 2002, Chapter 8.7), and even now it is widely used in many applied fields (e.g., Doeswijk *et al.*, 2011; Khalil, Ouarda & St-Hilaire, 2011; Vahedia, 2011). Determining the variables to be used is an important problem in CCA as well as selecting the variables in the multivariate linear regression model. Hence, variable selection in CCA has been investigated in many papers, e.g., MacKay (1977), Fujikoshi (1982; 1985), Al-Kandari and Jolliffe (1997), Noble, Smith and Ye (2004), and Ogura (2010).

The choice of variables based on the minimization of an information criterion as typified by the Akaike information criterion (AIC), which was proposed by Akaike (1973), is one of the major variable selection methods. Fujikoshi (1985) identified the problem of variable selection in CCA as one of selecting corresponding covariance structures, and proposed the AIC as a selector of covariance structures. The Kullback-Leibler (KL) discrepancy (Kullback & Leibler, 1951) function, consisting of a density function of the Wishart distribution, is frequently used in covariance structure analysis. To use the KL discrepancy based on the Wishart density naturally means that the normality of the variables is assumed. The AIC is an asymptotic unbiased estimator of the risk function consisting of the expected KL loss function when the model being considered is completely specified. Under this assumption, Fujikoshi (1985) and Fujikoshi and Kurata (2008) proposed a bias-corrected AIC (corrected AIC: CAIC), which is an unbiased estimator of the risk function. However, if the model being considered is not specified, a bias with constant order will appear in the AIC.

One of the information criteria for correcting the bias of the AIC under model misspecification is the Takeuchi information criterion (TIC) proposed by Takeuchi (1976), whose bias-correction term is given by a moment estimator of the first term in an asymptotic expansion of the bias to the risk function. Another criterion correcting the bias of the AIC under model misspecification is the extended information criterion (EIC) proposed by Ishiguro, Sakamoto, and Kitagawa (1997), whose bias-correction term is evaluated from the bootstrap method. The TIC and EIC for selecting covariance structures were studied by Ichikawa and Konishi (1999). Furthermore, Fujikoshi *et al.* (2008) and Ogura (2010) dealt with the EIC for selecting variables in CCA. The orders of biases of both criteria are $O(n^{-1})$ even when the model is misspecified.

A purpose of this paper is to propose a bias-corrected AIC that reduces higher-order bias even when the model being considered is misspecified. Since we use the jackknife method for evaluating the bias-correction term, we call this new criterion the jackknife bias-corrected AIC (JAIC). By using a property of the jackknife estimator, we can reduce the bias of the AIC to $O(n^{-2})$. Besides this, we adjust the JAIC to an exact unbiased estimator of the risk function when the distribution of the true model is the normal distribution, which has been attempted in the multivariate linear regression model by Yanagihara (2006) and Yanagihara, Kamo, and Tonda (2011). This adjustment will remove the negative effects of an increase in dimensions.

This paper is organized in the following way. In Section 2, we describe the bias of the AIC under model misspecification. In Section 3, we propose the JAIC for selecting variables in CCA. In Section 4, we verify by numerical experiments that our criteria perform better than the existing criteria, namely, the AIC, CAIC, TIC, and EIC. Technical details are provided in the Appendix.

2. Bias of the AIC under Model Misspecification

Let $\boldsymbol{z} = (\boldsymbol{x}', \boldsymbol{y}')' = (x_1, \ldots, x_p, y_1, \ldots, y_q)'$ be a $(p+q)$-dimensional vector with

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

Without loss of generality, we divide $\boldsymbol{x}$ and $\boldsymbol{y}$ into two sub-vectors $\boldsymbol{x} = (\boldsymbol{x}_1', \boldsymbol{x}_2')'$ and $\boldsymbol{y} = (\boldsymbol{y}_3', \boldsymbol{y}_4')'$, where $\boldsymbol{x}_1$ and $\boldsymbol{y}_3$ are p_1- and q_1-dimensional random vectors, respectively. Another expression of $\boldsymbol{\Sigma}$ corresponding to the divisions is

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} & \boldsymbol{\Sigma}_{13} & \boldsymbol{\Sigma}_{14} \\ \boldsymbol{\Sigma}_{12}' & \boldsymbol{\Sigma}_{22} & \boldsymbol{\Sigma}_{23} & \boldsymbol{\Sigma}_{24} \\ \boldsymbol{\Sigma}_{13}' & \boldsymbol{\Sigma}_{23}' & \boldsymbol{\Sigma}_{33} & \boldsymbol{\Sigma}_{34} \\ \boldsymbol{\Sigma}_{14}' & \boldsymbol{\Sigma}_{24}' & \boldsymbol{\Sigma}_{34}' & \boldsymbol{\Sigma}_{44} \end{pmatrix}. \tag{2.1}$$

Then, we are interested in whether $\boldsymbol{x}_2$ and $\boldsymbol{y}_4$ have any additional information. Let $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ be n independent random vectors from $\boldsymbol{z}$, and let $\boldsymbol{S}$ be the usual unbiased estimator of $\boldsymbol{\Sigma}$ given by $\boldsymbol{S} = (n-1)^{-1}\sum_{i=1}^{n}(\boldsymbol{z}_i - \bar{\boldsymbol{z}})(\boldsymbol{z}_i - \bar{\boldsymbol{z}})'$, where $\bar{\boldsymbol{z}} = n^{-1}\sum_{i=1}^{n}\boldsymbol{z}_i$. 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 $\boldsymbol{x}_2$ and $\boldsymbol{y}_4$ have no additional information is expressed as

$$M:\ (n-1)\boldsymbol{S} \sim W_{p+q}(n-1, \boldsymbol{\Sigma})\ s.t.\ \mathrm{tr}(\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\Sigma}_{xy}') = \mathrm{tr}(\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{13}\boldsymbol{\Sigma}_{33}^{-1}\boldsymbol{\Sigma}_{13}'). \tag{2.2}$$

An estimator of $\boldsymbol{\Sigma}$ under M in (2.2) is given by

$$\hat{\boldsymbol{\Sigma}} = \arg\min_{\boldsymbol{\Sigma}} \left\{ F(\boldsymbol{S}, \boldsymbol{\Sigma})\ s.t.\ \mathrm{tr}(\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\Sigma}_{xy}') = \mathrm{tr}(\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{13}\boldsymbol{\Sigma}_{33}^{-1}\boldsymbol{\Sigma}_{13}') \right\},$$

where $F(\boldsymbol{S}, \boldsymbol{\Sigma})$ is the KL discrepancy function assessed by the Wishart density, which is given by

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

In the analysis of covariance structure, the above discrepancy function is frequently called the maximum likelihood discrepancy function (Jöreskog, 1967). Let $\boldsymbol{S}$ be divided in the same way as $\boldsymbol{\Sigma}$ in (2.1), i.e.,

$$\boldsymbol{S} = \begin{pmatrix} \boldsymbol{S}_{11} & \boldsymbol{S}_{12} & \boldsymbol{S}_{13} & \boldsymbol{S}_{14} \\ \boldsymbol{S}_{12}' & \boldsymbol{S}_{22} & \boldsymbol{S}_{23} & \boldsymbol{S}_{24} \\ \boldsymbol{S}_{13}' & \boldsymbol{S}_{23}' & \boldsymbol{S}_{33} & \boldsymbol{S}_{34} \\ \boldsymbol{S}_{14}' & \boldsymbol{S}_{24}' & \boldsymbol{S}_{34}' & \boldsymbol{S}_{44} \end{pmatrix}.$$

Then, from Fujikoshi and Kurata (2008) or Fujikoshi, Shimizu, and Ulyanov (2010, Chapter 11.5), we can see that an explicit form of $\hat{\boldsymbol{\Sigma}}$ is given by

$$\begin{aligned}\hat{\boldsymbol{\Sigma}} &= \begin{pmatrix} \hat{\boldsymbol{\Sigma}}_{11} & \hat{\boldsymbol{\Sigma}}_{12} & \hat{\boldsymbol{\Sigma}}_{13} & \hat{\boldsymbol{\Sigma}}_{14} \\ \hat{\boldsymbol{\Sigma}}_{12}' & \hat{\boldsymbol{\Sigma}}_{22} & \hat{\boldsymbol{\Sigma}}_{23} & \hat{\boldsymbol{\Sigma}}_{24} \\ \hat{\boldsymbol{\Sigma}}_{13}' & \hat{\boldsymbol{\Sigma}}_{23}' & \hat{\boldsymbol{\Sigma}}_{33} & \hat{\boldsymbol{\Sigma}}_{34} \\ \hat{\boldsymbol{\Sigma}}_{14}' & \hat{\boldsymbol{\Sigma}}_{24}' & \hat{\boldsymbol{\Sigma}}_{34}' & \hat{\boldsymbol{\Sigma}}_{44} \end{pmatrix} \\ &= \begin{pmatrix} \boldsymbol{S}_{11} & \boldsymbol{S}_{12} & \boldsymbol{S}_{13} & \boldsymbol{S}_{13}\boldsymbol{S}_{33}^{-1}\boldsymbol{S}_{34} \\ \boldsymbol{S}_{12}' & \boldsymbol{S}_{22} & \boldsymbol{S}_{12}'\boldsymbol{S}_{11}^{-1}\boldsymbol{S}_{13} & \boldsymbol{S}_{12}'\boldsymbol{S}_{11}^{-1}\boldsymbol{S}_{13}\boldsymbol{S}_{33}^{-1}\boldsymbol{S}_{34} \\ \boldsymbol{S}_{13}' & \boldsymbol{S}_{13}'\boldsymbol{S}_{11}^{-1}\boldsymbol{S}_{12} & \boldsymbol{S}_{33} & \boldsymbol{S}_{34} \\ \boldsymbol{S}_{34}'\boldsymbol{S}_{33}^{-1}\boldsymbol{S}_{13}' & \boldsymbol{S}_{34}'\boldsymbol{S}_{33}^{-1}\boldsymbol{S}_{13}'\boldsymbol{S}_{11}^{-1}\boldsymbol{S}_{12} & \boldsymbol{S}_{34}' & \boldsymbol{S}_{44} \end{pmatrix}. \qquad (2.3)\end{aligned}$$

Actually, this interesting model is likely misspecified, i.e., the constraint of $\boldsymbol{\Sigma}$ in (2.2) might be incorrect, and the distribution of $(n-1)\boldsymbol{S}$ may not necessarily correspond to the Wishart distribution (or, equivalently, it may be that there is no guarantee of the normality of $\boldsymbol{z}$). Hence, we write the true model as

$$M_* : \ Cov[\boldsymbol{S}] = \boldsymbol{\Sigma}. \qquad (2.4)$$

In practice, we may simultaneously consider several candidate models. Among all the candidate models, the model with the fewest number of parameters that fits the data well is regarded as a good one. This idea can be executed by means of the so-called risk function defined by

$$R = E[\mathcal{L}(\hat{\boldsymbol{\Sigma}})], \qquad (2.5)$$

where $\mathcal{L}(\boldsymbol{A})$ is the KL loss function expressed as

$$\mathcal{L}(\boldsymbol{A}) = E[F(\boldsymbol{S}, \boldsymbol{A})] = (n-1)\left\{\mathrm{tr}(\boldsymbol{A}^{-1}\boldsymbol{\Sigma}) - E[\log|\boldsymbol{A}^{-1}\boldsymbol{S}|] - (p+q)\right\}. \qquad (2.6)$$

The candidate model having the smallest risk function is regarded as the best model among all the candidate models.

Although the minimum discrepancy $F(\boldsymbol{S}, \hat{\boldsymbol{\Sigma}})$ is a rough estimator of R, it has a bias with constant order as follows:

$$B = R - E[F(\boldsymbol{S}, \hat{\boldsymbol{\Sigma}})] = (n-1)E[\mathrm{tr}\{\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{\Sigma} - \boldsymbol{S})\}].$$

Let $\boldsymbol{D}$ be a $(p+q) \times d$ matrix expressing the operator that extracts d-elements from $\boldsymbol{z}$ by $\boldsymbol{D}'\boldsymbol{z}$. Here, all the elements of $\boldsymbol{D}$ are 0 or 1 and satisfy $\boldsymbol{D}'\boldsymbol{D} = \boldsymbol{I}_d$. In particular, the following matrices are used in this paper:

$$\boldsymbol{D}_x = \begin{pmatrix} \boldsymbol{I}_p \\ \boldsymbol{O}_{q,p} \end{pmatrix}, \ \boldsymbol{D}_y = \begin{pmatrix} \boldsymbol{O}_{p,q} \\ \boldsymbol{I}_q \end{pmatrix}, \ \boldsymbol{D}_1 = \begin{pmatrix} \boldsymbol{I}_{p_1} \\ \boldsymbol{O}_{p+q-p_1,p_1} \end{pmatrix},$$
$$\boldsymbol{D}_3 = \begin{pmatrix} \boldsymbol{O}_{p,q_1} \\ \boldsymbol{I}_{q_1} \\ \boldsymbol{O}_{p+q-q_1,q_1} \end{pmatrix}, \ \boldsymbol{D}_{(13)} = (\boldsymbol{D}_1, \boldsymbol{D}_3),$$

where $\boldsymbol{O}_{p,q}$ is a $p \times p$ matrix of zeros. Then we have

$$\boldsymbol{x} = \boldsymbol{D}_x'\boldsymbol{z}, \ \boldsymbol{y} = \boldsymbol{D}_y'\boldsymbol{z}, \ \boldsymbol{x}_1 = \boldsymbol{D}_1'\boldsymbol{z}, \ \boldsymbol{y}_3 = \boldsymbol{D}_3'\boldsymbol{z}, \ \begin{pmatrix} \boldsymbol{x}_1 \\ \boldsymbol{y}_3 \end{pmatrix} = \boldsymbol{D}_{(13)}'\boldsymbol{z}.$$

By using the fact that $\mathrm{tr}(\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{S}) = p + q$ and the results in Fujikoshi and Kurata (2008), we can rewrite the bias of $F(\boldsymbol{S}, \hat{\boldsymbol{\Sigma}})$ for R as

$$B = (n-1)\left\{\alpha(\boldsymbol{D}_x) + \alpha(\boldsymbol{D}_y) + \alpha(\boldsymbol{D}_{(13)}) - \alpha(\boldsymbol{D}_1) - \alpha(\boldsymbol{D}_3) - (p+q)\right\}, \tag{2.7}$$

where $\alpha(\boldsymbol{D})$ is defined by

$$\alpha(\boldsymbol{D}) = E[\tau(\boldsymbol{S}|\boldsymbol{D})], \tag{2.8}$$

and

$$\tau(\boldsymbol{S}|\boldsymbol{D}) = \mathrm{tr}\{(\boldsymbol{D}'\boldsymbol{S}\boldsymbol{D})^{-1}\boldsymbol{D}'\boldsymbol{\Sigma}\boldsymbol{D}\}. \tag{2.9}$$

It should be emphasized that the bias expressed as (2.7) is satisfied whether or not the constraint of $\boldsymbol{\Sigma}$ in (2.2) holds.

Let $\kappa_4(\boldsymbol{D})$ be a multivariate kurtosis of $\boldsymbol{D}'\boldsymbol{z}$ defined as

$$\kappa_4(\boldsymbol{D}) = E[\{(\boldsymbol{z} - \boldsymbol{\mu})'\boldsymbol{D}(\boldsymbol{D}'\boldsymbol{\Sigma}\boldsymbol{D})^{-1}\boldsymbol{D}'(\boldsymbol{z} - \boldsymbol{\mu})\}^2] - d(d+2). \tag{2.10}$$

The bias B in (2.7) is evaluated as in the following theorem (the proof is given in Appendix A.1):

Theorem 1. *Suppose that* $E[\mathrm{tr}(\boldsymbol{S}^{-2})] < \infty$ *and all the sixteenth multivariate moments of* $\boldsymbol{z}$ *exist. Whether the constraint of* $\boldsymbol{\Sigma}$ *in* (2.2) *holds or not, the bias* B *is expanded as*

$$B = \kappa_4(\boldsymbol{D}_x) + \kappa_4(\boldsymbol{D}_y) + \kappa_4(\boldsymbol{D}_{(13)}) - \kappa_4(\boldsymbol{D}_1) - \kappa_4(\boldsymbol{D}_3) + f(p_1, q_1) + O(n^{-1}),$$

where $f(p_1, q_1) = p^2 + q^2 + p + q + 2p_1q_1$. *Especially, when* $\boldsymbol{z} \sim N_{p+q}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, *the bias* B *is explicitly expressed as*

$$B = m(p_1, q_1) = (n-1)\left\{c_n(p) + c_n(q) + c_n(p_1 + q_1) - c_n(p_1) - c_n(q_1) - (p+q)\right\},$$

where

$$c_n(k) = \frac{(n-1)k}{n-k-2}. \tag{2.11}$$

The AIC and CAIC for selecting variables in CCA, which were proposed by Fujikoshi (1985) and Fujikoshi and Kurata (2008), are given by

$$\mathrm{AIC} = F(\boldsymbol{S}, \hat{\boldsymbol{\Sigma}}) + f(p_1, q_1), \tag{2.12}$$

$$\mathrm{CAIC} = F(\boldsymbol{S}, \hat{\boldsymbol{\Sigma}}) + m(p_1, q_1). \tag{2.13}$$

Notice that $m(p_1, q_1) = f(p_1, q_1) + O(n^{-1})$. Furthermore, it follows from the same method as in Theorem 1 that $B = O(1)$ if $E[\mathrm{tr}(\boldsymbol{S}^{-2})] < \infty$ and all the eighth multivariate moments of $\boldsymbol{z}$ exist. Thus, from the above results and Theorem 1, we obtain the following corollary.

Corollary 1. *Suppose that* $E[\mathrm{tr}(\boldsymbol{S}^{-2})] < \infty$ *and all the eighth multivariate moments of* $\boldsymbol{z}$ *exist. Whether the constraint of* $\boldsymbol{\Sigma}$ *in* (2.2) *holds or not, the orders of biases of the AIC and CAIC become*

$$R - E[\mathrm{AIC}] = \begin{cases} O(n^{-1}) & (\boldsymbol{z} \sim N_{p+q}(\boldsymbol{\mu}, \boldsymbol{\Sigma})) \\ O(1) & (\textit{otherwise}) \end{cases}, \quad R - E[\mathrm{CAIC}] = \begin{cases} 0 & (\boldsymbol{z} \sim N_{p+q}(\boldsymbol{\mu}, \boldsymbol{\Sigma})) \\ O(1) & (\textit{otherwise}) \end{cases}.$$

Generally, the AIC will have a bias with constant order when the structure of the model is misspecified. Furthermore, the CAIC in (2.13) was proposed under the assumption that the covariance structure of the candidate model is true. However, Corollary 1 indicates that if the normality of $\boldsymbol{z}$ holds, the AIC is an asymptotic unbiased estimator and the CAIC is an unbiased estimator of R even when the covariance structure of the model is misspecified. Additionally, by using the same idea as in Davies, Neath, and Cavanaugh

(2006), i.e., the complete efficiency of $\boldsymbol{S}$ (see e.g., Siotani, Hayakawa & Fujikoshi, 1985, pp. 18–20) and the Lehman-Scheffé theorem, we can prove that the CAIC is a uniformly minimum-variance unbiased estimator (UMVUE) of R when $\boldsymbol{z} \sim N_{p+q}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. On another front, Corollary 1 also points out that the AIC and CAIC have biases with constant order if the normality of $\boldsymbol{z}$ is not satisfied.

3. Bias-corrected AICs under Model Misspecification

3.1. Existing Criteria

Based on the evaluation of the bias of the AIC in the previous section, in this section we consider a bias correction of the AIC under model misspecification. A simple bias-correction method under model misspecification is to estimate a bias-correction term by the moment method. A bias-corrected AIC by the moment method is called the TIC, which was proposed by Takeuchi (1976). In the selection of variables in CCA, estimating the bias of the AIC by the moment method corresponds to estimating $\alpha(\boldsymbol{D})$ by $\hat{\kappa}_4(\boldsymbol{D}) + d(d+1)$, where $\hat{\kappa}_4(\boldsymbol{D})$ is an estimator of $\kappa_4(\boldsymbol{D})$ given by

$$\hat{\kappa}_4(\boldsymbol{D}) = \frac{1}{n}\sum_{i=1}^{n} w_{ii}(\boldsymbol{D})^2 - d(d+2),$$

where $w_{ij}(\boldsymbol{D})$ is defined by

$$w_{ij}(\boldsymbol{D}) = (\boldsymbol{z}_i - \bar{\boldsymbol{z}})'\boldsymbol{D}(\boldsymbol{D}'\boldsymbol{S}\boldsymbol{D})^{-1}\boldsymbol{D}'(\boldsymbol{z}_i - \bar{\boldsymbol{z}}). \tag{3.1}$$

By using this estimator, the TIC for selecting variables in CCA is defined by

$$\mathrm{TIC} = \mathrm{AIC} + \hat{\kappa}_4(\boldsymbol{D}_x) + \hat{\kappa}_4(\boldsymbol{D}_y) + \hat{\kappa}_4(\boldsymbol{D}_{(13)}) - \hat{\kappa}_4(\boldsymbol{D}_1) - \hat{\kappa}_4(\boldsymbol{D}_3). \tag{3.2}$$

The criterion in (3.2) can be regarded as the special case of selecting general covariance structures, which was proposed by Ichikawa and Konishi (1999), although they did not treat covariance structures in this paper. Theoretically, the TIC reduces the bias of the AIC to $O(n^{-1})$ under model misspecification. However, there is a possibility that the TIC overly underestimates the risk function under the small sample because the bias-correction term depends on estimators of fourth cumulants of the true distribution. These estimators frequently give poor estimates under the small sample (see e.g., Yanagihara, 2007). Consequently, numerically, the TIC sometimes does not work well under the small sample.

The bias-correction by bootstrap is an effective bias-correction method. The AIC bias-corrected by the bootstrap method is called the EIC, which was proposed by Ishiguro, Sakamoto, and Kitagawa (1997). Let $\hat{\boldsymbol{\Sigma}}_b$ be the bth bootstrap estimator of $\boldsymbol{\Sigma}$ evaluated from the bth bootstrap resample. From Fujikoshi *et al.* (2008), the EIC for selecting variables in CCA is given by

$$\mathrm{EIC} = F(\boldsymbol{S}, \hat{\boldsymbol{\Sigma}}) + (n-1)\left\{\frac{1}{m}\left(1-\frac{1}{n}\right)\sum_{b=1}^{m}\mathrm{tr}(\hat{\boldsymbol{\Sigma}}_b^{-1}\boldsymbol{S}) - (p+q)\right\}, \tag{3.3}$$

where m is the number of bootstrap iterations. The criterion in (3.3) is also a special case of that for selecting general covariance structures, which was proposed by Ichikawa and Konishi (1999). Theoretically, the EIC reduces the bias of the AIC to $O(n^{-1})$ under model misspecification as well as the TIC.

3.2. Proposed Criterion

In this subsection, we propose a bias-corrected AIC which reduces the bias of the AIC to $O(n^{-2})$ by using the jackknife method with a constant adjustment. Let $\boldsymbol{S}_{[-i,-j]}$ be the (i,j)th jackknife unbiased estimator of $\boldsymbol{\Sigma}$, which is given by

$$\boldsymbol{S}_{[-i,-j]} = \frac{1}{n-3}\sum_{k\neq i,j}^{n}(\boldsymbol{z}_k - \bar{\boldsymbol{z}}_{[-i,-j]})(\boldsymbol{z}_k - \bar{\boldsymbol{z}}_{[-i,-j]})', \tag{3.4}$$

where $\bar{\boldsymbol{z}}_{[-i,j]}$ is the (i,j)th jackknife sample mean, which is given by $\bar{\boldsymbol{z}}_{[-i,-j]} = (n-2)^{-1}\sum_{k\neq i,j}^{n}\boldsymbol{z}_k$. Then, we define the following jackknife residuals sum of squares:

$$r(\boldsymbol{D}) = \frac{1}{n(n-1)}\sum_{k\neq i,j}^{n}(\boldsymbol{z}_i - \boldsymbol{z}_j)'\boldsymbol{D}(\boldsymbol{D}'\boldsymbol{S}_{[-i,-j]}\boldsymbol{D})^{-1}\boldsymbol{D}'(\boldsymbol{z}_i - \boldsymbol{z}_j). \tag{3.5}$$

Since $\boldsymbol{z}_i$ and $\boldsymbol{z}_j$ are independent of $\boldsymbol{S}_{[-i,-j]}$, and $E[(\boldsymbol{z}_i - \boldsymbol{z}_j)(\boldsymbol{z}_i - \boldsymbol{z}_j)'] = 2\boldsymbol{\Sigma}$, we obtain

$$E[r(\boldsymbol{D})] = E[\mathrm{tr}\{(\boldsymbol{D}'\boldsymbol{S}_{[-i,-j]}\boldsymbol{D})^{-1}\boldsymbol{D}'\boldsymbol{\Sigma}\boldsymbol{D}\}] = E[\tau(\boldsymbol{S}_{[-i,-j]}|\boldsymbol{D})], \tag{3.6}$$

where $\tau(\boldsymbol{A}|\boldsymbol{D})$ is given in (2.9). Since $\boldsymbol{S}_{[-i,-j]} = \boldsymbol{S} + O_p(n^{-1})$ holds, $r(\boldsymbol{D})$ becomes an asymptotic unbiased estimator of $\alpha(\boldsymbol{D})$. By slightly modifying $r(\boldsymbol{D})$, we define an estimator of $\alpha(\boldsymbol{D})$ as

$$\hat{\alpha}(\boldsymbol{D}) = \frac{(n-1)(n-d-4)}{(n-d-2)(n^2-3n-2d-2)}\{2d + (n-2)r(\boldsymbol{D})\}. \tag{3.7}$$

Then, we propose a new criterion called the jackknife bias-corrected AIC (JAIC) as

$$\begin{aligned}\mathrm{JAIC} &= F(\boldsymbol{S}, \hat{\boldsymbol{\Sigma}}) \\ &\quad + (n-1)\left\{\hat{\alpha}(\boldsymbol{D}_x) + \hat{\alpha}(\boldsymbol{D}_y) + \hat{\alpha}(\boldsymbol{D}_{(13)}) - \hat{\alpha}(\boldsymbol{D}_1) - \hat{\alpha}(\boldsymbol{D}_3) - (p+q)\right\}.\end{aligned} \tag{3.8}$$

The order of the bias of the JAIC to R is stated by the following theorem (the proof is given in Appendix A.2):

Theorem 2. *Suppose that $E[\mathrm{tr}(\boldsymbol{S}^{-2})] < \infty$ and all the twenty-fourth multivariate moments of $\boldsymbol{z}$ exist. Whether the constraint of $\boldsymbol{\Sigma}$ in* (2.2) *holds or not, the order of bias of the JAIC becomes*

$$R - E[\mathrm{JAIC}] = \begin{cases} 0 & (\boldsymbol{z} \sim N_{p+q}(\boldsymbol{\mu}, \boldsymbol{\Sigma})) \\ O(n^{-2}) & (\textit{otherwise}) \end{cases}.$$

We summarize the results on the orders of the biases of the information criteria handled in this paper in Table 1. It may be noted that the JAIC has the highest performance on the bias correction among all the criteria.

Table 1. The order of the bias of each criterion

	AIC	CAIC	TIC	EIC	JAIC
Under normality	$O(n^{-1})$	0	$O(n^{-1})$	$O(n^{-1})$	0
Under nonnormality	$O(1)^*$	$O(1)^*$	$O(n^{-1})^{**}$	$O(n^{-1})^{**}$	$O(n^{-2})^{***}$

Note) $E[\mathrm{tr}(\boldsymbol{S}^{-2})] < \infty$, and existences of all the eighth, sixteenth and twenty-fourth multivariate moments of $\boldsymbol{z}$ are required for obtaining the orders superscripted by *, **, and ***, respectively.

Cauchy-Schwarz's Although the formula of $r(\boldsymbol{D})$ is simple, its computation time is long when n or d is large. This is caused by the necessity for $n(n-1)/2$ calculations of inverse matrices. Such a problem can be avoided by using another expression of $r(\boldsymbol{D})$ as in the following theorem (the proof of this theorem is given in Appendix A.3):

Theorem 3. *Let $\boldsymbol{H}(\boldsymbol{D})$ be the $n \times n$ symmetric matrix whose (i,j)th element is given as*

$$\begin{aligned} &h_{ij}(\boldsymbol{D}) \\ &= \frac{n^2(w_{ii}(\boldsymbol{D}) + w_{jj}(\boldsymbol{D}) - 2w_{ij}(\boldsymbol{D})) - 2nb_1b_2(w_{ii}(\boldsymbol{D})w_{jj}(\boldsymbol{D}) - w_{ij}(\boldsymbol{D})^2)}{n^2 - nb_2(w_{ii}(\boldsymbol{D}) + w_{jj}(\boldsymbol{D})) - b_1b_2\{2w_{ij}(\boldsymbol{D}) - b_1(w_{ii}(\boldsymbol{D})w_{jj}(\boldsymbol{D}) - w_{ij}(\boldsymbol{D})^2)\}}, \end{aligned} \tag{3.9}$$

where $w_{ij}(\boldsymbol{D})$ is given by (3.1) *and b_j is defined by*

$$b_j = \frac{n}{n-j}. \tag{3.10}$$

Then, $r(\boldsymbol{D})$ in (3.5) is rewritten as

$$r(\boldsymbol{D}) = \frac{b_1^2}{2b_3n^2}\mathbf{1}_n'\boldsymbol{H}(\boldsymbol{D})\mathbf{1}_n = \frac{(n-3)}{n(n-1)^2}\sum_{i>j}^{n} h_{ij}(\boldsymbol{D}), \tag{3.11}$$

where $\mathbf{1}_n$ is an n-dimensional vector of ones.

Since only one calculation of an inverse matrix is required for the computation of (3.11), the computation time of $r(\boldsymbol{D})$ using (3.11) is faster than that using (3.5). Table 2 shows the ratio: {average computation times of $r(\boldsymbol{D})$ using (3.11)}/{average computation times of $r(\boldsymbol{D})$ using (3.5)} at 100 repetitions when we generate $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ from the normal distribution and $\boldsymbol{D} = \boldsymbol{I}_{p+q}$. From the table, we can see that the formula in (3.11) dramatically reduces the computation time of $r(\boldsymbol{D})$, especially when n or $p+q$ are large.

Table 2. Ratio of the average computation time of (3.11) to that of (3.5)

n $p+q$	50	100	250	500
4	0.21	0.18	0.17	0.15
8	0.22	0.18	0.15	0.14
16	0.18	0.15	0.12	0.10
32	0.12	0.10	0.08	0.05
64	—-	0.04	0.03	0.02

4. Numerical Study

In this section, we conduct numerical studies to show that the JAIC in (3.8) works better than the AIC in (2.12), the CAIC in (2.13), the TIC in (3.2), or the EIC in (3.3). Simulation data was generated from $\boldsymbol{z} = (z_1, \ldots, z_8)' = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon}$ is an 8-dimensional error vector distributed according to the distribution with $E[\boldsymbol{\varepsilon}] = \mathbf{0}_8$ and $Cov[\boldsymbol{\varepsilon}] = \boldsymbol{I}_8$, and $\boldsymbol{\Sigma}$ is a 8×8 matrix defined by

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1.000 & 0.000 & 1.000 & 1.300 & 0.100 & 0.200 & 0.470 & 0.530 \\ 0.000 & 1.000 & 1.200 & 1.400 & 0.200 & 0.100 & 0.460 & 0.520 \\ 1.000 & 1.200 & 5.440 & 3.580 & 0.340 & 0.320 & 1.022 & 1.154 \\ 1.300 & 1.400 & 3.580 & 8.050 & 0.410 & 0.400 & 1.255 & 1.417 \\ 0.100 & 0.200 & 0.340 & 0.410 & 1.000 & 0.000 & 1.500 & 1.700 \\ 0.200 & 0.100 & 0.320 & 0.400 & 0.000 & 1.000 & 1.600 & 1.800 \\ 0.470 & 0.460 & 1.022 & 1.255 & 1.500 & 1.600 & 6.810 & 5.630 \\ 0.530 & 0.520 & 1.154 & 1.417 & 1.700 & 1.800 & 5.630 & 11.730 \end{pmatrix}.$$

Here, $\mathbf{0}_m$ is an m-dimensional vector of zeros. The sixteenth candidate models $M_{i,j}$ $(i = 1, \ldots, 4; j = 1, \ldots, 4)$, with $n = 100$ and 250, were prepared for simulations. We divided $\boldsymbol{z}$ into $\boldsymbol{x} = (x_1, \ldots, x_4)' = (z_1, \ldots, z_4)'$ and $\boldsymbol{y} = (y_1, \ldots, y_4)' = (z_5, \ldots, z_8)'$. The $\boldsymbol{x}_1$ and $\boldsymbol{y}_3$ in the model $M_{i,j}$ are $\boldsymbol{x}_1 = (x_1, \ldots, x_i)'$ and $\boldsymbol{y}_3 = (y_1, \ldots, y_j)'$. In this case, the true model is $M_{2,2}$.

Let $\boldsymbol{\nu} \sim N_8(\mathbf{0}_8, \boldsymbol{I}_8)$ and $\delta \sim \chi_6^2$ be a mutually independent random vector and variable, and let $\boldsymbol{\Psi}$ be an 8×8 matrix defined by $\boldsymbol{\Psi} = \boldsymbol{I}_8 + \mathbf{1}_8\mathbf{1}_8'/3$. Then, $\boldsymbol{\varepsilon}$ was generated from the following three distributions:

- Distribution 1 (multivariate normal distribution): $\boldsymbol{\varepsilon} = \boldsymbol{\nu}$,
- Distribution 2 (scale mixture of multivariate normal distribution): $\boldsymbol{\varepsilon} = \sqrt{\delta/6}\boldsymbol{\nu}$,
- Distribution 3 (scale and location mixtures of multivariate normal distribution): $\boldsymbol{\varepsilon} = \boldsymbol{\Psi}^{-1/2}\{(\delta/6 - 1)\mathbf{1}_8 + \sqrt{\delta/6}\boldsymbol{\nu}\}$.

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

First, we studied the bias of each information criterion. Figure 1 shows R in (2.5) and E[AIC], E[CAIC], E[TIC], E[EIC], and E[JAIC]. These values were obtained for 10,000 iterations. The horizontal axis of each figure expresses the candidate model (or the subindex i, j of $M_{i,j}$). From these figures, we can see that the biases of the CAIC were very small when $\boldsymbol{\varepsilon} \sim N_8(\mathbf{0}_8, \boldsymbol{I}_8)$. However, when the distribution of $\boldsymbol{\varepsilon}$ was not normal, the biases of the CAIC became large. On the other hand, the biases of the JAIC were very small even when the distribution of $\boldsymbol{\varepsilon}$ was not normal. The biases of the EIC were not so large, but they were larger than those of the JAIC. When the sample size was not large, the biases of the TIC were large. This is because the estimation of $\alpha(\boldsymbol{D})$ did not work well.

Next, we compared the performances of variable selections using the AIC, CAIC, TIC, EIC, and JAIC by the following two properties:

(i) the selection probability: the frequency of the model chosen by minimizing the information criterion.

(ii) the prediction error of the best model (PE$_\text{B}$): the risk function of the best model chosen by the information criterion, which is defined by

$$\text{PE}_\text{B} = E[\mathcal{L}(\hat{\boldsymbol{\Sigma}}_\text{B})],$$

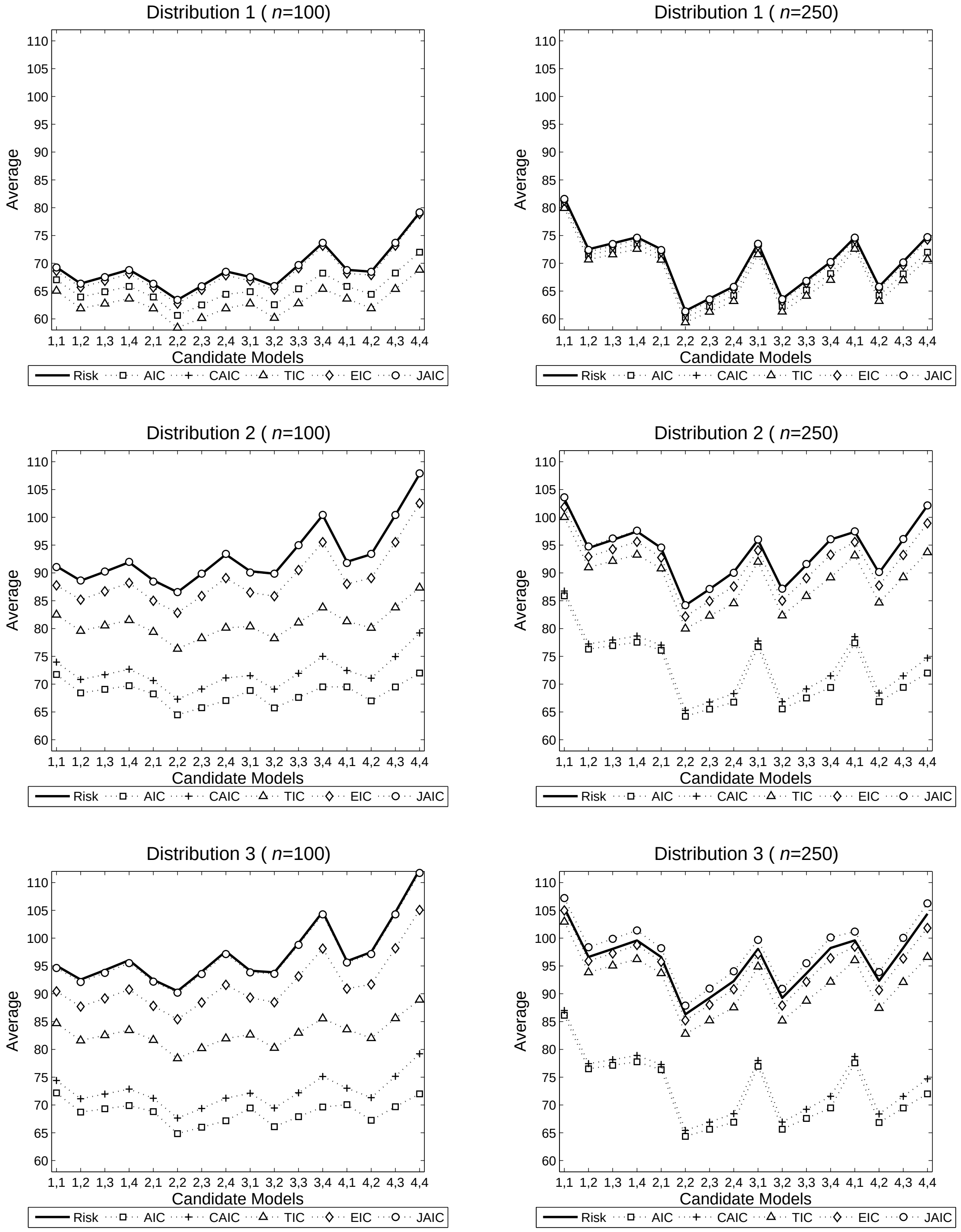

Figure 1. Risk function and average of each criterion

where $\mathcal{L}(\boldsymbol{A})$ is the loss function given by (2.6) and $\hat{\boldsymbol{\Sigma}}_{\mathrm{B}}$ is the estimator of $\boldsymbol{\Sigma}$ in (2.3) under the best model.

The information criterion with the highest selection probability of the true model and

Table 3. Selection probability of the model and the prediction error of the best model ($\mathrm{PE_B}$) in the case of the multivariate normal distribution

	$n = 100$						$n = 250$					
		Selection Probability (%)						Selection Probability (%)				
Model	Risk	AIC	CAIC	TIC	EIC	JAIC	Risk	AIC	CAIC	TIC	EIC	JAIC
1,1	69.3	3.7	5.1	3.3	4.7	5.1	81.7	0.0	0.0	0.0	0.2	0.0
1,2	66.4	13.4	16.0	12.3	12.9	15.8	72.5	2.2	2.3	2.1	2.7	2.4
1,3	67.6	3.2	3.4	3.2	6.5	3.7	73.6	0.6	0.7	0.6	1.7	0.7
1,4	68.9	2.7	2.5	2.8	3.6	2.5	74.7	0.4	0.4	0.5	1.1	0.5
2,1	66.3	12.8	15.4	11.9	12.7	15.3	72.5	2.1	2.3	2.0	3.0	2.4
2,2	63.4	33.3	34.6	32.0	26.8	33.4	61.5	61.4	64.6	60.1	37.0	64.0
2,3	65.9	6.9	5.5	7.6	7.6	5.9	63.7	9.7	9.4	10.2	15.7	9.4
2,4	68.5	3.6	2.2	4.2	3.3	2.3	65.8	4.8	4.0	5.0	6.9	4.1
3,1	67.6	3.5	3.7	3.5	5.8	3.8	73.6	0.5	0.5	0.5	1.4	0.5
3,2	65.9	6.8	5.5	7.3	8.0	5.6	63.6	9.1	8.4	9.3	15.4	8.6
3,3	69.7	1.6	1.0	1.9	1.2	1.0	66.9	2.0	1.6	2.0	3.7	1.5
3,4	73.7	0.8	0.4	1.0	0.5	0.4	70.3	0.8	0.6	1.1	1.3	0.6
4,1	68.8	2.8	2.3	2.8	3.1	2.5	74.6	0.5	0.4	0.5	1.1	0.5
4,2	68.5	3.7	2.1	4.5	3.0	2.4	65.8	4.8	4.1	5.1	7.3	4.2
4,3	73.7	0.8	0.3	1.0	0.4	0.3	70.3	0.8	0.5	0.9	1.3	0.5
4,4	79.1	0.5	0.1	0.7	0.1	0.1	74.8	0.3	0.2	0.4	0.3	0.2
$\mathrm{PE_B}$		68.6	68.0	68.9	68.2	68.1		65.2	64.9	65.3	65.9	64.9

Note) The bold face indicates the model having the true covariance structure.

the smallest prediction error of the best model is regarded as a high-performance model selector. In the basic concept of the AIC, a good model-selection method is one that chooses the best model so that the prediction is improved. Hence, $\mathrm{PE_B}$ is a more important property than the selection probability.

Tables 3, 4, and 5 show the selection probability and prediction error when the distributions of $\boldsymbol{\varepsilon}$ are 1, 2, and 3, respectively. From the tables, we can see that the selection probabilities of the CAIC were highest among those of all criteria when $n = 100$ and the distribution of $\boldsymbol{\varepsilon}$ was normal. However, the differences between those of the CAIC and the JAIC were not so large. When $n = 250$ and the distribution of $\boldsymbol{\varepsilon}$ was not normal, the selection probabilities of the CAIC were far below those of the JAIC. On the other hand, the prediction errors of the JAIC when the distribution of $\boldsymbol{\varepsilon}$ was not normal were the smallest among all criteria. Although the prediction errors of the CAIC were the smallest when the distribution of $\boldsymbol{\varepsilon}$ was normal, the differences between those of the CAIC and the JAIC were very small. On the other hand, the prediction errors of the EIC were not so large, but the selection probabilities of the EIC were the lowest among all the criteria.

Table 4. Selection probability of the model and the prediction error of the best model (PE$_{\mathrm{B}}$) in the case of the scale mixture of multivariate normal distribution

	$n = 100$						$n = 250$					
		Selection Probability (%)						Selection Probability (%)				
Model	Risk	AIC	CAIC	TIC	EIC	JAIC	Risk	AIC	CAIC	TIC	EIC	JAIC
1,1	91.2	3.6	4.8	6.0	7.8	10.1	103.3	0.2	0.2	0.5	0.8	0.6
1,2	88.6	9.4	12.1	12.4	12.7	17.3	94.5	2.0	2.3	4.4	5.5	5.6
1,3	90.2	3.7	4.3	4.3	7.3	4.8	95.9	0.7	0.8	1.1	2.8	1.3
1,4	91.9	3.3	3.5	3.6	4.8	3.0	97.5	0.9	1.0	1.1	2.1	1.2
2,1	88.7	9.8	12.5	12.7	14.0	17.9	94.5	2.1	2.5	4.2	4.9	5.4
2,2	86.5	23.3	26.1	23.3	18.8	24.0	84.2	41.1	44.5	51.1	30.0	56.1
2,3	89.8	7.7	6.9	6.5	6.6	4.4	87.1	10.9	10.7	9.5	13.6	8.5
2,4	93.3	6.0	4.2	4.5	3.1	2.1	90.1	8.9	8.2	5.8	7.0	4.1
3,1	90.3	3.7	4.2	4.4	7.5	4.7	95.9	0.7	0.8	1.3	3.0	1.4
3,2	89.9	8.2	7.6	6.6	6.7	4.9	87.0	11.2	11.1	9.2	13.0	7.9
3,3	95.0	3.3	2.6	2.5	1.8	1.1	91.4	4.0	3.5	2.4	4.3	1.7
3,4	100.4	2.8	1.4	1.7	0.6	0.4	96.0	3.1	2.4	1.2	1.8	0.5
4,1	92.0	3.9	3.7	3.9	4.7	3.1	97.4	0.9	1.0	1.0	2.0	1.0
4,2	93.4	6.2	4.0	4.4	3.1	2.0	90.0	7.9	7.1	5.2	7.0	3.7
4,3	100.4	2.9	1.5	1.8	0.5	0.2	96.0	3.3	2.5	1.1	1.7	0.6
4,4	107.8	2.3	0.7	1.3	0.2	0.1	102.1	2.3	1.6	0.8	0.6	0.3
PE$_{\mathrm{B}}$		95.5	93.9	94.3	92.4	92.3		91.8	91.3	90.3	90.9	89.6

Note) The bold face indicates the model having the true covariance structure.

Furthermore, although the performance of the TIC when the sample size was large was not so bad, the performance when the sample size was small was bad. We simulated several other models and obtained similar results. Hence, we recommend using the JAIC for selecting variables in CCA.

Appendix

A.1. The Proof of Theorem 1

In order to prove Theorem 1, it is sufficient to evaluate $\alpha(\boldsymbol{D})$ given by (2.8). Let $\boldsymbol{\varepsilon}_i = \boldsymbol{z}_i - \boldsymbol{\mu}$, and

$$\boldsymbol{V} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i' - \boldsymbol{\Sigma}), \quad \boldsymbol{u} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{\varepsilon}_i.$$

When all the fourth multivariate moments of $\boldsymbol{z}$ exist, $\boldsymbol{V}$ and $\boldsymbol{u}$ have asymptotic normality, thus $\boldsymbol{V} = O_p(1)$ and $\boldsymbol{u} = O_p(1)$ as $n \to \infty$. Hence, a stochastic expansion of $\boldsymbol{D}'\boldsymbol{S}\boldsymbol{D}$ is

Table 5. Selection probability of the model and the prediction error of the best model ($\mathrm{PE_B}$) in the case of the scale and location mixture of multivariate normal distribution

	$n = 100$						$n = 250$					
		Selection Probability (%)						Selection Probability (%)				
Model	Risk	AIC	CAIC	TIC	EIC	JAIC	Risk	AIC	CAIC	TIC	EIC	JAIC
1,1	95.1	3.8	5.0	5.8	7.3	10.2	105.4	0.1	0.1	0.2	0.9	0.3
1,2	92.6	9.3	11.7	12.3	13.7	17.7	96.6	2.4	2.8	4.9	5.0	6.4
1,3	94.2	4.1	4.7	4.6	7.6	5.1	98.1	1.0	1.0	1.4	3.2	1.6
1,4	96.0	3.4	3.5	3.7	4.6	3.1	99.6	1.0	1.0	1.1	2.3	1.1
2,1	92.5	9.8	12.3	12.7	13.2	17.7	96.6	2.1	2.4	4.5	4.9	6.2
2,2	90.5	21.6	24.3	21.6	18.3	22.5	86.3	39.9	43.0	50.4	28.6	54.6
2,3	94.0	7.6	7.4	6.7	7.1	4.8	89.3	10.7	10.7	8.9	13.1	8.0
2,4	97.6	7.1	5.2	5.5	3.3	2.4	92.3	8.7	7.8	5.6	6.6	3.9
3,1	94.1	3.9	4.3	4.2	6.9	4.5	98.1	0.9	1.0	1.3	3.4	1.5
3,2	93.9	7.4	7.1	6.5	6.7	4.6	89.3	10.6	10.7	9.2	13.9	8.0
3,3	99.1	3.6	2.5	2.7	1.8	1.2	93.7	4.0	3.5	2.2	4.3	1.5
3,4	104.7	3.1	1.5	2.0	0.7	0.4	98.2	3.3	2.7	1.5	1.7	0.8
4,1	95.9	3.4	3.4	3.5	4.5	3.0	99.6	0.9	0.9	1.2	2.6	1.1
4,2	97.5	6.8	5.0	5.2	3.6	2.4	92.3	8.8	8.1	5.8	7.2	4.2
4,3	104.7	3.2	1.6	2.0	0.6	0.3	98.2	3.2	2.6	1.3	1.6	0.6
4,4	112.3	2.1	0.6	1.2	0.2	0.1	104.4	2.4	1.7	0.6	0.6	0.2
$\mathrm{PE_B}$		99.8	98.2	98.7	96.5	96.4		94.3	93.7	92.6	93.2	91.9

Note) The bold face indicates the model having the true covariance structure.

derived as

$$\boldsymbol{D}'\boldsymbol{S}\boldsymbol{D} = \boldsymbol{D}'\boldsymbol{\Sigma}\boldsymbol{D} + \frac{1}{\sqrt{n}}\boldsymbol{D}'\boldsymbol{V}\boldsymbol{D} - \frac{1}{n}\boldsymbol{D}'(\boldsymbol{u}\boldsymbol{u}' - \boldsymbol{\Sigma})\boldsymbol{D} + O_p(n^{-3/2}).$$

This expansion implies a stochastic expansion of $\tau(\boldsymbol{S}|\boldsymbol{D})$ given in (2.9) as

$$\begin{aligned}\tau(\boldsymbol{S}|\boldsymbol{D}) = d - \frac{1}{\sqrt{n}}\tau(\boldsymbol{V}|\boldsymbol{D}) + \frac{1}{n}&\left[\mathrm{tr}\left(\{\boldsymbol{D}'\boldsymbol{V}\boldsymbol{D}(\boldsymbol{D}'\boldsymbol{\Sigma}\boldsymbol{D})^{-1}\}^2\right)\right.\\ &\left.+\mathrm{tr}\{\boldsymbol{D}'(\boldsymbol{u}\boldsymbol{u}' - \boldsymbol{\Sigma})\boldsymbol{D}(\boldsymbol{D}'\boldsymbol{\Sigma}\boldsymbol{D})^{-1}\}\right] + O_p(n^{-3/2}).\end{aligned} \tag{A.1}$$

Notice that $E[\boldsymbol{V}] = \boldsymbol{O}_{p+q,p+q}$ and $E[\boldsymbol{u}\boldsymbol{u}'] = \boldsymbol{\Sigma}$, and

$$E\left[\mathrm{tr}\left(\{\boldsymbol{D}'\boldsymbol{V}\boldsymbol{D}(\boldsymbol{D}'\boldsymbol{\Sigma}\boldsymbol{D})^{-1}\}^2\right)\right] = \kappa_4(\boldsymbol{D}) + d(d+1),$$

where $\kappa_4(\boldsymbol{D})$ is given by (2.10). From the fact that the top term in the remainder term in (A.1) can be expressed as an odd polynomial function of $\boldsymbol{V}$ and $\boldsymbol{u}$ and the Cauchy-Schwarz inequality, the order of the expectation of the $O_p(n^{-3/2})$ term in (A.1) is $O(n^{-2})$ when $E[\mathrm{tr}(\boldsymbol{S}^{-2})] < \infty$ and all the sixteenth multivariate moments of $\boldsymbol{z}$ exist. Therefore, when $E[\mathrm{tr}(\boldsymbol{S}^{-2})] < \infty$ and all the sixteenth multivariate moments of $\boldsymbol{z}$ exist, $\alpha(\boldsymbol{D})$ can be expanded as

$$\alpha(\boldsymbol{D}) = d + \frac{1}{n}\{\kappa_4(\boldsymbol{D}) + d(d+1)\} + O(n^{-2}). \tag{A.2}$$

Moreover, if $\boldsymbol{z} \sim N_{p+q}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $(n-1)\boldsymbol{D}'\boldsymbol{S}\boldsymbol{D} \sim W_d(n-1, \boldsymbol{D}'\boldsymbol{\Sigma}\boldsymbol{D})$. Hence, from the property of the Wishart distribution (see e.g., Siotani, Hayakawa & Fujikoshi, 1985, p. 74, Theorem 2.4.6), we derive

$$\alpha(\boldsymbol{D}) = c_n(d), \tag{A.3}$$

where $c_n(d)$ is given by (2.11). Substituting (A.2) or (A.3) into (2.7) yields the expression for B in Theorem 1.

A.2. The Proof of Theorem 2

In order to prove Theorem 2, it is sufficient to show that

$$E[\hat{\alpha}(\boldsymbol{D})] = \left\{ \begin{array}{ll} c_n(d) & (\boldsymbol{z} \sim N_{p+q}(\boldsymbol{\mu}, \boldsymbol{\Sigma})) \\ \alpha(\boldsymbol{D}) + O(n^{-3}) & (\text{otherwise}) \end{array} \right. . \tag{A.4}$$

By using the same idea as in the proof of Theorem 1, when $E[\mathrm{tr}(\boldsymbol{S}^{-2})] < \infty$ and all the twenty-fourth multivariate moments of $\boldsymbol{z}$ exist, $\alpha(\boldsymbol{D})$ can be expanded as

$$\alpha(\boldsymbol{D}) = d + \frac{1}{n}\beta_1(\boldsymbol{D}) + \frac{1}{n^2}\beta_2(\boldsymbol{D}) + O(n^{-3}). \tag{A.5}$$

An explicit form of $\beta_1(\boldsymbol{D})$ is given by (A.2). By comparing with (2.8) and (3.6), we can see that the difference between $\alpha(\boldsymbol{D})$ and $E[r(\boldsymbol{D})]$ is only the sample size of an unbiased estimator of $\boldsymbol{\Sigma}$. Hence, when $E[\mathrm{tr}(\boldsymbol{S}^{-2})] < \infty$ and all the twenty-fourth multivariate moments of $\boldsymbol{z}$ exist, $E[r(\boldsymbol{D})]$ can be also expanded as

$$E[r(\boldsymbol{D})] = d + \frac{1}{n-2}\beta_1(\boldsymbol{D}) + \frac{1}{(n-2)^2}\beta_2(\boldsymbol{D}) + O(n^{-3}). \tag{A.6}$$

Notice that $\{2d + (n-2)r(\boldsymbol{D})\}/n = d + (n-2)\{r(\boldsymbol{D}) - d\}/n$. Hence, by using this equation and the equations (A.5) and (A.6), we have

$$\begin{aligned} E\left[\frac{1}{n}\{2d + (n-2)r(\boldsymbol{D})\}\right] &= d + \frac{1}{n}\beta_1(\boldsymbol{D}) + \frac{1}{n(n-2)}\beta_2(\boldsymbol{D}) + O(n^{-3}) \\ &= \alpha(\boldsymbol{D}) + O(n^{-3}). \end{aligned} \tag{A.7}$$

Meanwhile, it is clear that

$$\frac{n(n-1)(n-d-4)}{(n-d-2)(n^2-3n-2d-2)} = 1 + O(n^{-3}). \tag{A.8}$$

Hence, combining (A.7) and (A.8) yields the lower result in (A.4). Furthermore, by using the same idea as in the derivation of the equation (A.3), if $\boldsymbol{z} \sim N_{p+q}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, we have

$$E[r(\boldsymbol{D})] = E[\tau(\boldsymbol{S}_{[-i,-j]}|\boldsymbol{D})] = c_{n-2}(d),$$

where $c_n(m)$ is given by (2.11). Therefore, when $\boldsymbol{z} \sim N_{p+q}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, the upper result in (A.4) is also proved as

$$E[\hat{\alpha}(\boldsymbol{D})] = \frac{(n-1)(n-d-4)}{(n-d-2)(n^2-3n-2d-2)} \left\{ 2d + \frac{(n-2)(n-3)d}{n-d-4} \right\} = c_n(d).$$

A.3. The Proof of Theorem 3

For simplicity, in this subsection, we write $w_{ij}(\boldsymbol{D})$ given in (3.1) and $h_{ij}(\boldsymbol{D})$ given in (3.9) as w_{ij} and h_{ij}, respectively. Let $\boldsymbol{a}_i$ be a $(p+q)$-dimensional vector defined by $\boldsymbol{a}_i = \boldsymbol{z}_i - \bar{\boldsymbol{z}}$, and $\boldsymbol{A}_{ij}$ and $\boldsymbol{C}$ be the $(p+q) \times 2$ and the 2×2 matrices defined by $\boldsymbol{A}_{ij} = (\boldsymbol{a}_i, \boldsymbol{a}_j)$ and $\boldsymbol{C} = \boldsymbol{I}_2 + (n-2)^{-1}\boldsymbol{1}_2\boldsymbol{1}_2'$, respectively. Notice that $\boldsymbol{S}_{[-i,-j]}$ in (3.4) can be rewritten as

$$\boldsymbol{S}_{[-i,-j]} = \frac{n-1}{n-3}\left(\boldsymbol{S} - \frac{1}{n-1}\boldsymbol{A}_{ij}\boldsymbol{C}\boldsymbol{A}_{ij}'\right).$$

By applying the general formula of the inversion of a matrix (see e.g., Siotani, Hayakawa & Fujikoshi, 1985, p. 591, Theorem A.2.1) to the above equality, we have

$$\begin{aligned}&(\boldsymbol{D}'\boldsymbol{S}_{[-i,-j]}\boldsymbol{D})^{-1} \\ &= \frac{n-3}{n-1}\left\{(\boldsymbol{D}'\boldsymbol{S}\boldsymbol{D})^{-1} - \frac{1}{n-1}(\boldsymbol{D}'\boldsymbol{S}\boldsymbol{D})^{-1}\boldsymbol{D}'\boldsymbol{A}_{ij}\boldsymbol{G}_{ij}^{-1}\boldsymbol{A}_{ij}'\boldsymbol{D}(\boldsymbol{D}'\boldsymbol{S}\boldsymbol{D})^{-1}\right\},\end{aligned} \tag{A.9}$$

where $\boldsymbol{G}_{ij} = \boldsymbol{C}^{-1} - (n-1)^{-1}\boldsymbol{A}_{ij}'\boldsymbol{D}(\boldsymbol{D}'\boldsymbol{S}\boldsymbol{D})^{-1}\boldsymbol{D}'\boldsymbol{A}_{ij}$. It is easy to obtain

$$\boldsymbol{G}_{ij}^{-1} = \frac{1}{n\delta_{ij}}\begin{pmatrix} n-1-b_1w_{jj} & 1+b_1w_{ij} \\ 1+b_1w_{ij} & n-1-b_1w_{ii}\end{pmatrix},$$

where b_j is given by (3.10) and δ_{ij} is the determinant of $\boldsymbol{G}_{ij}$ given by

$$\delta_{ij} = \frac{n-2}{n}\left(1 - \frac{1}{n}b_2(w_{ii}+w_{jj}) - \frac{1}{n^2}b_1b_2\{2w_{ij} - b_1(w_{ii}w_{jj} - w_{ij}^2)\}\right).$$

Since $\boldsymbol{z}_i - \boldsymbol{z}_j = \boldsymbol{a}_i - \boldsymbol{a}_j$, we obtain

$$r(\boldsymbol{D}) = \frac{1}{2n(n-1)}\sum_{i,j}^{n}(\boldsymbol{a}_i - \boldsymbol{a}_j)'\boldsymbol{D}(\boldsymbol{D}'\boldsymbol{S}_{[-i,-j]}\boldsymbol{D})^{-1}\boldsymbol{D}'(\boldsymbol{a}_i - \boldsymbol{a}_j).$$

Substituting (A.9) into the above equality yields

$$r(\boldsymbol{D}) = \frac{(n-3)}{2n(n-1)^2} r_1(\boldsymbol{D}) + \frac{(n-3)}{2n(n-1)^3} r_2(\boldsymbol{D}), \tag{A.10}$$

where

$$\begin{aligned} r_1(\boldsymbol{D}) &= \sum_{i,j}^{n} (\boldsymbol{a}_i - \boldsymbol{a}_j)' \boldsymbol{D}(\boldsymbol{D}'\boldsymbol{S}\boldsymbol{D})^{-1}\boldsymbol{D}'(\boldsymbol{a}_i - \boldsymbol{a}_j), \\ r_2(\boldsymbol{D}) &= \sum_{i,j}^{n} (\boldsymbol{a}_i - \boldsymbol{a}_j)' \boldsymbol{D}(\boldsymbol{D}'\boldsymbol{S}\boldsymbol{D})^{-1}\boldsymbol{D}'\boldsymbol{A}_{ij}\boldsymbol{G}_{ij}^{-1}\boldsymbol{A}_{ij}'\boldsymbol{D}(\boldsymbol{D}'\boldsymbol{S}\boldsymbol{D})^{-1}(\boldsymbol{a}_i - \boldsymbol{a}_j). \end{aligned}$$

It follows from the equalities $\sum_{i=1}^{n} \boldsymbol{a}_i = \boldsymbol{0}_{p+q}$ and $\sum_{i=1}^{n} \boldsymbol{a}_i\boldsymbol{a}_i' = (n-1)\boldsymbol{S}$ that

$$r_1(\boldsymbol{D}) = 2n(n-1)p. \tag{A.11}$$

Moreover, we obtain

$$r_2(\boldsymbol{D}) = (n-1)^2 \sum_{i,j}^{n} \boldsymbol{\ell}'(\boldsymbol{G}_{ij}^{-1} - 2\boldsymbol{C}^{-1} + \boldsymbol{G}_{ij})\boldsymbol{\ell},$$

where $\boldsymbol{\ell} = (1, -1)'$. Notice that

$$\begin{aligned} \boldsymbol{\ell}'\boldsymbol{G}_{ij}^{-1}\boldsymbol{\ell} &= \frac{1}{n\delta_{ij}} \left\{2(n-2) - b_1(w_{ii} + w_{jj} + 2w_{ij})\right\}, \\ \boldsymbol{\ell}'\boldsymbol{C}^{-1}\boldsymbol{\ell} &= 2, \quad \boldsymbol{\ell}'\boldsymbol{G}_{ij}\boldsymbol{\ell} = 2 - \frac{1}{n} b_1(w_{ii} + w_{jj} - 2w_{ij}). \end{aligned}$$

These equalities imply that

$$\begin{aligned} r_2(\boldsymbol{D}) &= -2n(n-1)^2(n+p) + (n-1)^2 \sum_{i,j}^{n} \frac{1}{n\delta_{ij}} \{2(n-2) - b_1(w_{ii} + w_{jj} + 2w_{ij})\} \\ &= -2n(n-1)^2 p + (n-1) \sum_{i,j}^{n} h_{ij}. \end{aligned} \tag{A.12}$$

Substituting (A.11) and (A.12) into (A.10) yields the result (3.11).

Acknowledgments

The authors thank Prof. Hirofumi Wakaki, Hiroshima University, for helpful comments on the moment condition of the expansion.

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 and F. Csáki), 267–281, Akadémiai Kiadó, Budapest.

[2] Al-Kandari, N. M. & Jolliffe, I. T. (1997). Variable selection and interpretation in canonical correlation analysis. *Comm. Statist. Simulation Comput.*, **26**, 873–900.

[3] Davies, S. J., Neath, A. A. & Cavanaugh, J. E. (2006). Estimation optimality of corrected AIC and modified C_p in linear regression model. *International Statist. Review*, **74**, 161–168.

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

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

[9] 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).

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

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

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

[13] 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.

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

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

[16] Noble, R., Smith, E. P. & Ye, K. (2004). Model selection in canonical correlation analysis (CCA) using Bayesian model averaging. *Environmetrics*, **15**, 291–311.

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

[18] Siotani, M., Hayakawa, T. & Fujikoshi, Y. (1985). *Modern Multivariate Statistical Analysis*: *A Graduate Course and Handbook*. American Sciences Press, Columbus, Ohio.

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

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

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

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

[23] Yanagihara, H. (2006). Corrected version of AIC for selecting multivariate normal linear regression models in a general nonnormal case. *J. Multivariate Anal.*, **97**, 1070–1089.

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

[25] Yanagihara, H., Kamo, K. & Tonda, T. (2011). Second-order bias-corrected AIC in multivariate normal linear models under nonnormality. *Canad. J. Statist.*, **39**, 126–146.