TR-No. 13-08, Hiroshima Statistical Research Group, 1–23

A Study on the Bias-Correction Effect of the AIC for Selecting Variables in Normal Multivariate Linear Regression Models under Model Misspecification

Hirokazu Yanagihara[1*], **Ken-ichi Kamo**[2], **Shinpei Imori**[1]
and Mariko Yamamura[3]

[1]Department of Mathematics, Graduate School of Science, Hiroshima University
1-3-1 Kagamiyama, Higashi-Hiroshima, Hiroshima 739-8626, Japan
[2]Department of Liberal Arts and Sciences, Sapporo Medical University
South 1, West 17, Chuo-ku, Sapporo, Hokkaido 060-8543, Japan
[3]Department of Mathematics Education, Graduate School of Education, Hiroshima University
1-1-1 Kagamiyama, Higashi-Hiroshima, Hiroshima 739-8524, Japan

(Last Modified: May 23, 2013)

Abstract

By numerically comparing a variable-selection method using the crude AIC with those using the bias-corrected AICs, we find out knowledge about what kind of bias correction gives a positive effect to variable selection under model misspecification. Actually, since it can be proved theoretically that all the variable-selection methods considered in this paper asymptotically choose the same model as the best model, we conduct numerical examinations using small and moderate sample sizes. Our results show that bias correction under assumption that the mean structure is misspecified has a better effect on variable selection than that under the assumption that the distribution of the model is misspecified.

Key words: AIC, bias correction, KL information, loss function, model misspecification, nonnormality, normal multivariate linear regression model, risk function, variable selection.

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

1. Introduction

In the analysis of real data, it is important to determine which statistical model best fits the data; there are many candidate models, and they each estimate different results, which may lead to different points of view. In order to improve the accuracy of predictions, the “best” model can be chosen as the one that has the smallest risk function when assessing the goodness of fit of the model, using the Kullback-Leibler (KL) information (Kullback & Leibler, 1951). In practice, an estimator of the risk function is used, because the risk function involves unknown parameters. The Akaike’s information criterion (AIC; proposed by Akaike, 1973; 1974) is the asymptotic unbiased estimator of the risk function under the condition that the candidate model is correctly specified. It is defined by the simple equation $-2 \times$ (the maximum log-likelihood) $+ 2 \times$ (the number of parameters in the model) and is commonly used in actual data analysis.

Since the AIC is the asymptotic unbiased estimator of the risk function, the bias of the AIC to the risk function may become large when the sample size becomes small and the number of parameters used in the candidate model becomes large. In particular, when the candidate models include the true model, the larger the number of parameters in the candidate model, the more the AIC underestimates the risk function. Then the AICs of those candidate models often do not have notable differences. In addition, the variance of the AIC may increase as the number of parameters increases (see e.g., Yanagihara & Ohmoto, 2005). Thus, the model with the most parameters tends to have the smallest AIC, and so the AIC often selects the model with the most parameters as the best model. This fault of the AIC is avoided by using the bias-corrected AIC, which is derived by correcting the bias to the risk function. This has been studied under various different conditions and with various different correction methods (as a general theory correcting the bias of the AIC, see, e.g., Konishi, 1999; Burnham & Anderson, 2002; Konishi & Kitagawa, 2008). Sugiura (1978) and Hurvich & Tsai (1989) proposed a corrected AIC for linear regression models (multiple regression models) by fully removing the bias of the AIC to the risk function under the condition that the candidate model is correctly specified. The bias-corrected AIC then becomes the uniformly minimum-variance unbiased estimator (UMVUE) for the risk function of the candidate model (see Davies *et al*., 2006), and many authors have verified by numerical experiments that a variable-selection method using the corrected AIC performs better in selecting the best model than does one that uses the crude AIC.

The basic concept of bias correction is that we expect that an unbiased estimator of the risk function will allow us to correctly evaluate the risk function, which will further facilitate the selection of the best model. However, there is no theory that promises that the best model chosen by minimizing a bias-corrected AIC has a higher predictive accuracy than that chosen by minimizing the crude AIC. Generally, a bias-corrected estimator has a larger variance than does one without a bias correction. An impairment of the mean square error of the bias-corrected AIC with respect to the risk function, which results from an increase in the variance, may cause a drop in the performance of the model selection when using a bias-corrected AIC.

In this paper, we compare the AIC and eight bias-corrected AICs to study what kind of bias correction gives a positive effect for selecting variables for a multivariate linear regression model (MLRM) with a normal distributed assumption (called the normal MLRM), under a model misspecification. The performances of the model selection methods using the nine criteria are examined by numerical experiments. We do not use large samples, because it has been confirmed theoretically that the variable-selection methods using these nine criteria select the same model as “best” when $n \to \infty$. Our result is that correcting the bias has a greater positive effect on variable selection when the mean structure is misspecified than when the distribution of the model is misspecified.

This paper is organized as follows: In Section 2, the normal MLRM and the risk function based on the KL information are described. In Section 3, the AIC and the bias-corrected AICs for the normal MLRM are summarized. In Section 4, we show that variable-selection methods using the information criteria considered in this paper select the same best model when $n \to \infty$. We then use numerical experiments with small and moderate samples to compare the performance of model selection methods using the AIC and the bias-corrected AICs. Our conclusions and a discussion are presented in Section 5. Technical details are provided in the Appendix.

2. Risk Function Based on the KL Information

The normal MLRM is used when we are interested in predicting not just one response variable but several correlated response variables based on k nonstochastic explanatory variables (for details, see, e.g., Srivastava, 2002, chap. 9; Timm, 2002, chap. 4). Let $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ be p-dimensional independent random vectors of response variables, and let $\boldsymbol{x}_{\omega,1}, \ldots, \boldsymbol{x}_{\omega,n}$ be k_ω-dimensional vectors of the full explanatory variables, where n is the sample size. Furthermore, let $\boldsymbol{x}_i$ be a k-dimensional vector of the candidate explanatory variables, which is a subset of the full explanatory variables $\boldsymbol{x}_{\omega,i}$ $(i = 1, \ldots, n)$. Then, we consider the following normal MLRM as the candidate model:

$$M : \ \boldsymbol{y}_i \sim N_p(\boldsymbol{\Xi}'\boldsymbol{x}_i, \boldsymbol{\Sigma}), \ (i = 1, \ldots, n), \tag{1}$$

where $\boldsymbol{\Xi}$ is a $k \times p$ matrix of the unknown regression coefficients, and $\boldsymbol{\Sigma}$ is a $p \times p$ unknown covariance matrix.

Let $\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)'$ be an $n \times p$ matrix of response variables, and let $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)'$ be an $n \times k$ matrix of candidate explanatory variables. Suppose that an $n \times k_\omega$ matrix of the full explanatory variables, $\boldsymbol{X}_\omega = (\boldsymbol{x}_{\omega,1}, \ldots, \boldsymbol{x}_{\omega,n})'$, is a column full-rank matrix, i.e., $\mathrm{rank}(\boldsymbol{X}_\omega) = k_\omega < n$. Needless to say, $\boldsymbol{X}$ consists of some columns of $\boldsymbol{X}_\omega$ and is also a column full-rank matrix. Moreover, we assume that $\boldsymbol{X}$ and $\boldsymbol{X}_\omega$ each always have $\boldsymbol{1}_n$ as a column vector that corresponds to an intercept, where $\boldsymbol{1}_n$ is an n-dimensional vector of ones. The matrix form of the candidate model (1) is given by

$$M : \ \boldsymbol{Y} \sim N_{n \times p}(\boldsymbol{X}\boldsymbol{\Xi}, \boldsymbol{\Sigma} \otimes \boldsymbol{I}_n). \tag{2}$$

The following normal MLRM using the full explanatory variables is called the full model:

$$M_\omega : \ \boldsymbol{Y} \sim N_{n \times p}(\boldsymbol{X}_\omega\boldsymbol{\Xi}_\omega, \boldsymbol{\Sigma}_\omega \otimes \boldsymbol{I}_n). \tag{3}$$

Although the normal distribution is assumed, we are not able to see whether the assumption is actually correct. A natural assumption for the generating mechanism of $\boldsymbol{Y}$ is

$$\begin{aligned} M_* : \ \boldsymbol{Y} = \boldsymbol{\Gamma}_* + \boldsymbol{\mathcal{E}}\boldsymbol{\Sigma}_*^{1/2}, \quad \boldsymbol{\mathcal{E}} = (\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n)', \\ \boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n \sim i.i.d.\ \boldsymbol{\varepsilon}, \ E[\boldsymbol{\varepsilon}] = \boldsymbol{0}_p, \ Cov[\boldsymbol{\varepsilon}] = \boldsymbol{I}_p, \ E[(\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon})^2] = \kappa_4^{(1)} + p(p+2), \end{aligned} \tag{4}$$

where $\boldsymbol{0}_p$ is a p-dimensional vector of zeros. Here, $\kappa_4^{(1)}$ is called the multivariate kurtosis, which was proposed by Mardia (1970).

In order to clarify assumptions for deriving the information criteria, we separate the candidate models into the following two models:

- Underspecified model: the mean structure does not include that of the true model, i.e., $\boldsymbol{P_X}\boldsymbol{\Gamma}_* \neq \boldsymbol{\Gamma}_*$.
- Overspecified model: the mean structure includes that of the true model, i.e., $\boldsymbol{P_X}\boldsymbol{\Gamma}_* = \boldsymbol{\Gamma}_*$.

Here, $\boldsymbol{P_X}$ is the projection matrix to the subspace spanned by the columns of $\boldsymbol{X}$, i.e., $\boldsymbol{P_X} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$. Furthermore, the candidate model whose mean structure dovetails perfectly with that of model M_* will be called the true model. Although Fujikoshi and Satoh (1997) used the same terminology, they divided the candidate models by whether the candidate model included the true model. This emphasizes that we are separating the candidate models based only on the mean structure. Hence, our separation does not depend on whether a distribution of the true model is the normal distribution. Furthermore, we assume that the full model M_ω is the overspecified model and the true model is included in a set of the candidate models. For an additional characteristic of the candidate model, a $p \times p$ matrix of noncentrality parameters is defined by

$$\boldsymbol{\Omega} = \frac{1}{n}\boldsymbol{\Sigma}_*^{-1/2}\boldsymbol{\Gamma}_*'(\boldsymbol{I}_n - \boldsymbol{P_X})\boldsymbol{\Gamma}_*\boldsymbol{\Sigma}_*^{-1/2}. \tag{5}$$

It should be noted that $\boldsymbol{\Omega}$ is positive semidefinite and $\boldsymbol{\Omega} = \boldsymbol{O}_{p,p}$ (where $\boldsymbol{O}_{p,p}$ is a $p \times p$ matrix of zeros) holds if and only if M is the overspecified model.

Let $f(\boldsymbol{y}|\boldsymbol{\eta}, \boldsymbol{\Sigma})$ be the probability density function of $N_p(\boldsymbol{\eta}, \boldsymbol{\Sigma})$. Then, the log-likelihood function of the candidate model M in (2) is derived as

$$\begin{aligned}\ell(\boldsymbol{\Xi}, \boldsymbol{\Sigma}|\boldsymbol{Y}, \boldsymbol{X}) &= \sum_{i=1}^{n} \log f(\boldsymbol{y}_i|\boldsymbol{\Xi}'\boldsymbol{x}_i, \boldsymbol{\Sigma}) \\ &= -\frac{1}{2}\left\{np\log 2\pi + n\log|\boldsymbol{\Sigma}| + \mathrm{tr}(\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\Xi})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\Xi}))\right\}. \end{aligned} \tag{6}$$

By maximizing $\ell(\boldsymbol{\Xi}, \boldsymbol{\Sigma}|\boldsymbol{Y}, \boldsymbol{X})$, or equivalently solving the likelihood equations $\partial\ell(\boldsymbol{\Xi}, \boldsymbol{\Sigma}|\boldsymbol{Y}, \boldsymbol{X})/\partial\boldsymbol{\Xi} = \boldsymbol{O}_{k,p}$ and $\partial\ell(\boldsymbol{\Xi}, \boldsymbol{\Sigma}|\boldsymbol{Y}, \boldsymbol{X})/\partial\boldsymbol{\Sigma} = \boldsymbol{O}_{p,p}$, the maximum likelihood estimators (MLE) of the unknown parameter matrices $\boldsymbol{\Xi}$ and $\boldsymbol{\Sigma}$ in the candidate model M are obtained as

$$\hat{\boldsymbol{\Xi}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}, \quad \hat{\boldsymbol{\Sigma}} = \frac{1}{n}\boldsymbol{Y}'(\boldsymbol{I}_n - \boldsymbol{P_X})\boldsymbol{Y}.$$

Substituting the MLEs into (6) yields the maximum log-likelihood of the candidate model M as

$$\ell(\hat{\boldsymbol{\Xi}}, \hat{\boldsymbol{\Sigma}}|\boldsymbol{Y}, \boldsymbol{X}) = -\frac{n}{2}\left\{p(\log 2\pi + 1) + \log|\hat{\boldsymbol{\Sigma}}|\right\}. \tag{7}$$

Let $\mathcal{L}(\boldsymbol{\Xi}, \boldsymbol{\Sigma})$ be the expected negative twofold log-likelihood function:

$$\mathcal{L}(\boldsymbol{\Xi}, \boldsymbol{\Sigma}) = E_{\boldsymbol{Y}}^*[-2\ell(\boldsymbol{\Xi}, \boldsymbol{\Sigma}|\boldsymbol{Y}, \boldsymbol{X})] = -2\ell(\boldsymbol{\Xi}, \boldsymbol{\Sigma}|\boldsymbol{\Gamma}_*, \boldsymbol{X}) + n\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_*), \tag{8}$$

where $E_{\boldsymbol{Y}}^*$ means the expectation with respect to $\boldsymbol{Y}$ under the true model M_* in (4). We define the loss function measured by the KL information as $\mathcal{L}(\hat{\boldsymbol{\Xi}}, \hat{\boldsymbol{\Sigma}})$. Then, a risk function that uses the KL information to assess the gap between the true model and the candidate model is defined by the expectation of the loss function, i.e.,

$$R_{\mathrm{KL}} = E_{\boldsymbol{Y}}^{*}[\mathcal{L}(\hat{\boldsymbol{\Xi}}, \hat{\boldsymbol{\Sigma}})]. \tag{9}$$

In this paper, the candidate model that makes the risk function the smallest is called the principle best model. The following theorem is satisfied for the principle best model (the proof is given in Appendix A):

Theorem 1 *The principle best model is either the true model or an underspecified model. When $n \to \infty$, the principle best model becomes the true model under the assumption that all the multivariate fourth moments of ε exist and that $\lim_{n\to\infty} \boldsymbol{X}_{\omega}'\boldsymbol{X}_{\omega}/n$ exists and is positive definite.*

3. AIC and Bias-corrected AICs in Normal MLRMs

Although the risk function R_{KL} in (9) assesses the goodness of fit of the model, we cannot use R_{KL} directly because R_{KL} involves unknown parameters. Hence, in practice, an estimator of R_{KL} is needed to select the best model among the candidates. It is easy to see that a naive estimator of R_{KL} is $-2\ell(\hat{\boldsymbol{\Xi}}, \hat{\boldsymbol{\Sigma}}|\boldsymbol{Y}, \boldsymbol{X})$. Unfortunately, when R_{KL} is estimated by $-2\ell(\hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\Sigma}}|\boldsymbol{Y}, \boldsymbol{X})$, the following constant bias appears:

$$B = R_{\mathrm{KL}} - E_{\boldsymbol{Y}}^{*}[-2\ell(\hat{\boldsymbol{\Xi}}, \hat{\boldsymbol{\Sigma}}|\boldsymbol{Y}, \boldsymbol{X})]. \tag{10}$$

Thus, an information criterion for selecting the best model is defined by adding an estimated B to $-2\ell(\hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\Sigma}}|\boldsymbol{Y}, \boldsymbol{X})$ as

$$\mathrm{IC} = -2\ell(\hat{\boldsymbol{\Xi}}, \hat{\boldsymbol{\Sigma}}|\boldsymbol{Y}, \boldsymbol{X}) + \hat{B}, \tag{11}$$

where $\hat{B}$ is an estimator of B. The information criterion is specified by the individual $\hat{B}$, because $\hat{B}$ changes based on the assumptions of the model M and by the estimation method. In this paper, we consider the following two assumptions:

(A1) The candidate model M in (2) is the overspecified model.

(A2) The distribution of the true model M_* in (4), called the true distribution, is the normal distribution, i.e., $\varepsilon \sim N_p(\boldsymbol{0}_p, \boldsymbol{I}_p)$.

Nine information criteria used to estimate R_{KL} are enumerated below. The order of the bias of each information criterion for R_{KL} is summarized in Table 1.

- **AIC**: When assumptions A1 and A2 are satisfied simultaneously, Akaike (1973; 1974) showed that B in (10) is asymptotically equal to twice the number of parameters in the candidate model. Since the number of parameters of M is $pk + p(p+1)/2$, $\hat{B}_{\mathrm{AIC}} = 2pk + p(p+1)$ is used as $\hat{B}$. By using the general formula in (11), the AIC in the model M is given by

$$\mathrm{AIC} = np(\log 2\pi + 1) + n\log|\hat{\boldsymbol{\Sigma}}| + 2pk + p(p+1).$$

Recall that $\hat{B}_{\mathrm{AIC}}$ is derived under the assumption that A1 and A2 are satisfied simultaneously. Hence, the bias of the AIC to $R_{\rm KL}$ becomes $O(n^{-1})$ when assumptions A1 and A2 are satisfied simultaneously. However, the order of the bias changes to $O(1)$, i.e., AIC has constant bias, when either of the assumptions A1 or A2 are violated.

• **Corrected AIC** (**CAIC**): The $\hat{B}_{\rm AIC}$ gives an inaccurate approximation to B when n is not large or k and p are relatively large, because $\hat{B}_{\rm AIC}$ is an asymptotic approximation to B. In order to avoid this problem, when the assumptions A1, A2, and an additional assumption $n > p + k + 1$ are satisfied simultaneously, Bedrick and Tsai (1994) calculated the exact bias as $B = n(n+k)p/(n-k-p-1)-np$, and proposed the CAIC[1] by replacing $\hat{B}_{\rm AIC}$ in (11) with $\hat{B}_{\rm CAIC} = n(n + k)p/(n - k - p - 1) - np$. When $n > p + k + 1$, the CAIC in the model M is given by

$$\mathrm{CAIC} = np \log 2\pi + n \log |\hat{\boldsymbol{\Sigma}}| + \frac{n(n+k)p}{n-p-k-1} = \mathrm{AIC} + \frac{(p+k+1)(p+2k+1)p}{n-p-k-1}. \tag{12}$$

The CAIC in (12) is an unbiased estimator of $R_{\rm KL}$ under the assumptions A1 and A2, and it is congruent with the bias-corrected AIC proposed by Sugiura (1978) and Hurvich and Tsai (1989) when $p = 1$. From the equation (12) and the unbiasedness of the CAIC under the assumptions A1 and A2, we can see that the AIC in the overspecified model underestimates $R_{\rm KL}$, and the strength of the underestimation becomes large as k increases. The number of explanatory variables of the best model selected by the CAIC will be less than or equal to the number selected by the AIC (the proof is given in Appendix B). Additionally, extending the result of Davies *et al*. (2006) to the multivariate case provides that the CAIC is a UMVUE of the risk function $R_{\rm KL}$ when the assumptions A1 and A2 are satisfied simultaneously (for a short proof, see Yanagihara *et al*., 2012). However, as in the case of the AIC, the order of the bias of the CAIC to $R_{\rm KL}$ becomes $O(1)$, i.e., the CAIC has a constant bias, when either of the assumptions A1 or A2 are violated.

• **Modified AIC** (**MAIC**): When the assumption A2 holds but the assumption A1 does not hold, the AIC and CAIC have constant biases to $R_{\rm KL}$. Fujikoshi and Satoh (1997) reduced these biases by using an additional moment estimator for an asymptotic value of B in the underspecified model. Let $\hat{B}_{\rm MAIC} = \hat{B}_{\rm CAIC} + 2k\mathrm{tr}(\boldsymbol{L}) - \mathrm{tr}(\boldsymbol{L})^2 - \mathrm{tr}(\boldsymbol{L}^2)$, where $\boldsymbol{L}$ is a $p \times p$ matrix defined by $\boldsymbol{L} = (n - k)\hat{\boldsymbol{\Sigma}}_\omega \hat{\boldsymbol{\Sigma}}^{-1}/(n - k_\omega) - \boldsymbol{I}_p$. Here $\hat{\boldsymbol{\Sigma}}_\omega$ is the MLE of $\boldsymbol{\Sigma}_\omega$ in the full model M_ω in (3). We note that $\mathrm{tr}(\boldsymbol{L})$, $\mathrm{tr}(\boldsymbol{L})^2$, and $\mathrm{tr}(\boldsymbol{L}^2)$ are consistent estimators of $\mathrm{tr}(\boldsymbol{\Omega})$, $\mathrm{tr}(\boldsymbol{\Omega})^2$, and $\mathrm{tr}(\boldsymbol{\Omega}^2)$, respectively, when $\boldsymbol{\Omega}$ is given by (5). When $n > p + k + 1$, by replacing $\hat{B}_{\rm AIC}$ in (11) with $\hat{B}_{\rm MAIC}$, MAIC in the model M is given by

$$\mathrm{MAIC} = \mathrm{CAIC} + 2k\mathrm{tr}(\boldsymbol{L}) - \mathrm{tr}(\boldsymbol{L})^2 - \mathrm{tr}(\boldsymbol{L}^2).$$

The bias of the MAIC to $R_{\rm KL}$ becomes $O(n^{-2})$ when assumptions A1 and A2 are satisfied simultaneously, and it becomes $O(n^{-1})$ when assumption A2 holds but assumption A1 does not. This implies that the MAIC reduces the constant biases of the AIC and the CAIC in the underspecified model to $O(n^{-1})$ when assumption A2 holds. However, the bias changes to $O(1)$, i.e., the MAIC also has

[1] Although Bedrick and Tsai (1994) used $\mathrm{AIC_C}$ as the abbreviated symbol, we use CAIC following the notation of Fujikoshi and Satoh (1997).

constant bias, when assumption A2 is violated.

• **Takeuchi's Information Criterion** (**TIC**): The CAIC and MAIC correct the bias of the AIC to R_{KL} when the assumption A2 is satisfied. However, it is unknown if the true distribution is normal. The TIC (proposed by Takeuchi, 1976) corrects the bias of the AIC even if the true distribution is not normal. Let a squared standardized residual of the ith individual be denoted by

$$\hat{r}_i^2 = (\boldsymbol{y}_i - \hat{\boldsymbol{\Xi}}'\boldsymbol{x}_i)'\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{y}_i - \hat{\boldsymbol{\Xi}}'\boldsymbol{x}_i), \tag{13}$$

and let an estimator of the multivariate kurtosis $\kappa_4^{(1)}$ in (4) be denoted by

$$\hat{\kappa}_4^{(1)} = \frac{1}{n}\sum_{i=1}^{n} \hat{r}_i^4 - p(p+1). \tag{14}$$

Furthermore, let

$$h_i = 1 - \boldsymbol{x}_i'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i. \tag{15}$$

Then, by defining B in (11) as $\hat{B}_{\mathrm{TIC}} = \hat{B}_{\mathrm{AIC}} + \hat{\kappa}_4^{(1)} + 2\sum_{i=1}^{n}(1-h_i)(\hat{r}_i^2 - p)$, the TIC in the model M is given as follows (for details of the derivation, see Fujikoshi *et al.*, 2005):

$$\mathrm{TIC} = \mathrm{AIC} + \hat{\kappa}_4^{(1)} + 2\sum_{i=1}^{n}(1-h_i)(\hat{r}_i^2 - p). \tag{16}$$

When $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ are independently and identically distributed, the bias of the TIC to the risk function is $O(n^{-1})$ under any model misspecification. However, in the case of multivariate linear regression, the $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ are independent but not identically distributed. This leads to the less well-known fact that the TIC also has constant bias (as do the AIC and CAIC) when assumption A1 is violated. Although the TIC theoretically reduces the bias caused by violating normality, the TIC cannot reduce the bias successfully unless the sample size is huge. Many authors have verified this with numerical experiments (see, e.g., Fujikoshi *et al.*, 2005; Yanagihara, 2006). This occurs because the TIC has an estimator for the multivariate kurtosis $\hat{\kappa}_4^{(1)}$. Yanagihara (2007) presented numerical results that showed that $\hat{\kappa}_4^{(1)}$ has a huge bias to $\kappa_4^{(1)}$ if n is not huge. Hence, the TIC also has a huge bias to R_{KL} if n is not huge.

• **Extended Information Criterion** (**EIC**): The serious problem with the TIC comes from the moment estimator of $\kappa_4^{(1)}$. This problem can be avoided by using the bootstrap method for an estimation of the bias; this is the EIC, proposed by Ishiguro *et al*. (1997). In order to express the bth bootstrap resample of $\boldsymbol{Y}$, the following $n \times n$ random matrix is prepared:

$$\boldsymbol{D}_b = (\boldsymbol{d}_{b,1}, \ldots, \boldsymbol{d}_{b,n})', \quad \boldsymbol{d}_{b,1}, \ldots, \boldsymbol{d}_{b,n} \sim i.i.d.\ MN_n(1; n^{-1}\boldsymbol{1}_n), \tag{17}$$

where $MN_n(1; n^{-1}\boldsymbol{1}_n)$ denotes the n-variate one-trial multinomial distribution with the same cell probabilities $1/n$. Following Freedman (1981), the bth bootstrap resample of $\boldsymbol{Y}$ is $\tilde{\boldsymbol{Y}}_b = \boldsymbol{X}\hat{\boldsymbol{\Xi}} + \boldsymbol{D}_b(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{Y}$. Let $\tilde{\boldsymbol{\Sigma}}_b$ be the MLE of $\boldsymbol{\Sigma}$ evaluated from $(\tilde{\boldsymbol{Y}}_b, \boldsymbol{X})$. From the formula in Konishi (1999), an estimator of the bias obtained from the bootstrap method with m repetitions is given by

$\hat{B}_{\rm EIC} = m^{-1}\sum_{b=1}^{m}\mathrm{tr}\left\{\tilde{\boldsymbol{\Sigma}}_b^{-1}(\boldsymbol{Y}-\boldsymbol{P_X}\tilde{\boldsymbol{Y}}_b)'(\boldsymbol{Y}-\boldsymbol{P_X}\tilde{\boldsymbol{Y}}_b)\right\}-np$. Then, by using (11), the EIC in the model M is given as follows (see Fujikoshi *et al.*, 2005):

$$\mathrm{EIC} = np\log 2\pi + n\log|\hat{\boldsymbol{\Sigma}}| + \frac{1}{m}\sum_{b=1}^{m}\mathrm{tr}\left\{\tilde{\boldsymbol{\Sigma}}_b^{-1}(\boldsymbol{Y}-\boldsymbol{P_X}\tilde{\boldsymbol{Y}}_b)'(\boldsymbol{Y}-\boldsymbol{P_X}\tilde{\boldsymbol{Y}}_b)\right\}. \tag{18}$$

When $\boldsymbol{y}_1,\ldots,\boldsymbol{y}_n$ are independently and identically distributed, the bias of the EIC to the risk function is $O(n^{-1})$ under any model misspecification like the TIC. However, in the case of multivariate linear regression, $\boldsymbol{y}_1,\ldots,\boldsymbol{y}_n$ are independent but not identically distributed. Hence, the bias of EIC is $O(n^{-1})$ under the assumption A1, but that changes to $O(1)$, i.e., the EIC has constant bias (as does the TIC), when assumption A1 is violated (see Yanagihara, 2006). In particular, EIC = TIC + $O_p(n^{-1/2})$ holds when both $m\to\infty$ and assumption A1 holds (the proof is given in Appendix C). Although the theoretical bias of the EIC has the same order as that of the TIC, the bias of the EIC tends to become smaller than that of the TIC (see Yanagihara, 2006) because the EIC does not directly use $\hat{\kappa}_4^{(1)}$ for estimating the bias.

• **Adjusted EIC (EIC$_{\rm A}$)**: By using a full-model-based resampling, the bias of the EIC in (18) can be reduced to $O(n^{-1})$ even if assumption A1 does not hold (see, e.g., Fujikoshi *et al.*, 2005). We call this the adjusted EIC (EIC$_{\rm A}$). Let $\bar{\boldsymbol{Y}}_b$ be the bth bootstrap resample of $\boldsymbol{Y}$ based on the full model M_ω given by $\bar{\boldsymbol{Y}}_b = \boldsymbol{X}_\omega\hat{\boldsymbol{\Xi}}_\omega + \boldsymbol{D}_b(\boldsymbol{I}_n-\boldsymbol{P}_{\boldsymbol{X}_\omega})\boldsymbol{Y}$, and let $\bar{\boldsymbol{\Sigma}}_b$ be the MLE of $\boldsymbol{\Sigma}$ evaluated from $(\bar{\boldsymbol{Y}}_b,\boldsymbol{X})$, where $\hat{\boldsymbol{\Xi}}_\omega$ is the MLE of $\boldsymbol{\Xi}_\omega$ in the full model M_ω, and $\boldsymbol{D}_b$ is given by (17). We define an estimator of the bias $\hat{B}_{{\rm EIC}_{\rm A}}$ by replacing $\tilde{\boldsymbol{Y}}_b$ and $\tilde{\boldsymbol{\Sigma}}_b$ in $\hat{B}_{\rm EIC}$ with $\bar{\boldsymbol{Y}}_b$ and $\bar{\boldsymbol{\Sigma}}_b$. Then, by using (11), the EIC$_{\rm A}$ in the model M is as follows:

$$\mathrm{EIC_A} = np\log 2\pi + n\log|\hat{\boldsymbol{\Sigma}}| + \frac{1}{m}\sum_{b=1}^{m}\mathrm{tr}\left\{\bar{\boldsymbol{\Sigma}}_b^{-1}(\boldsymbol{Y}-\boldsymbol{P_X}\bar{\boldsymbol{Y}}_b)'(\boldsymbol{Y}-\boldsymbol{P_X}\bar{\boldsymbol{Y}}_b)\right\}.$$

The bias of the EIC$_{\rm A}$ to the risk function is always $O(n^{-1})$. In particular, EIC$_{\rm A}$ = TIC + $O_p(n^{-1/2})$ holds when $m\to\infty$ and assumption A1 holds (the proof is given in Appendix C).

• **Cross-Validation (CV) Criterion**: The information criteria introduced above are based on (11), but the CV criterion proposed by Stone (1974) is not based on (11) but instead estimated directly estimates the risk function. Let $\boldsymbol{Y}_{(-i)}$ be a $(n-1)\times p$ matrix obtained from $\boldsymbol{Y}$ by deleting $\boldsymbol{y}_i$, let $\boldsymbol{X}_{(-i)}$ be a $(n-1)\times k$ matrix obtained from $\boldsymbol{X}$ by deleting $\boldsymbol{x}_i$, and let $\hat{\boldsymbol{\Xi}}_{[-i]}$ and $\hat{\boldsymbol{\Sigma}}_{[-i]}$ be the MLEs of $\boldsymbol{\Xi}$ and $\boldsymbol{\Sigma}$, respectively, evaluated from $(\boldsymbol{Y}_{(-i)},\boldsymbol{X}_{(-i)})$. Then, the CV criterion for the model M is

$$\begin{aligned}\mathrm{CV} &= -2\sum_{i=1}^{n}\log f(\boldsymbol{y}_i|\hat{\boldsymbol{\Xi}}_{[-i]}'\boldsymbol{x}_i,\hat{\boldsymbol{\Sigma}}_{[-i]})\\ &= np\log 2\pi + \sum_{i=1}^{n}\left\{\log|\hat{\boldsymbol{\Sigma}}_{[-i]}| + (\boldsymbol{y}_i-\hat{\boldsymbol{\Xi}}_{[-i]}'\boldsymbol{x}_i)'\hat{\boldsymbol{\Sigma}}_{[-i]}^{-1}(\boldsymbol{y}_i-\hat{\boldsymbol{\Xi}}_{[-i]}'\boldsymbol{x}_i)\right\}.\end{aligned} \tag{19}$$

From Stone (1977), CV = TIC + $O_p(n^{-1})$ always holds if $\boldsymbol{y}_1,\ldots,\boldsymbol{y}_n$ are independently and identically distributed. In the case of multivariate linear regression, although $\boldsymbol{y}_1,\ldots,\boldsymbol{y}_n$ are not identically distributed, we can prove that CV $=$ TIC $+ O_p(n^{-1})$ always holds (the proof is given in Appendix D). From this result, the bias of the CV criterion is $O(n^{-1})$ under assumption A1, but like the TIC, it has a constant bias when assumption A1 is violated. On the other hand, Yoshimoto *et al*. (2005) showed that the CV criterion in (19) can be rewritten as

$$\mathrm{CV} = np \log\left(\frac{2\pi n}{n-1}\right) + n\log|\hat{\mathbf{\Sigma}}| + \sum_{i=1}^{n}\left\{\log\left(1 - \frac{\hat{r}_i^2}{nh_i}\right) + \frac{(n-1)\hat{r}_i^2}{h_i(nh_i - \hat{r}_i^2)}\right\}, \tag{20}$$

where $\hat{r}_i^2$ and h_i are given by (13) and (15), respectively. Equation (19) indicates that n repetitions of the calculations for $\hat{\mathbf{\Xi}}_{[-i]}$ and $\hat{\mathbf{\Sigma}}_{[-i]}$ are needed to derive the CV criterion. However, by using (20), we only need to calculate $\hat{\mathbf{\Xi}}$ and $\hat{\mathbf{\Sigma}}$ to calculate the CV criterion.

• **Jackknifed AIC (AIC$_\mathrm{J}$)**: The CV method which is equivalent to the jackknife method can estimate not only the risk function R_{KL} but also the bias B. We can then derive an estimator of B that is unbiased when assumptions A1 and A2 are satisfied simultaneously, and that is asymptotically unbiased when assumption A1 is satisfied but assumption A2 is not. Yanagihara (2006) proposed such a bias-corrected AIC, called the jackknifed AIC (AIC$_\mathrm{J}$). Although it is also necessary to calculate $\hat{\mathbf{\Xi}}_{[-i]}$ and $\hat{\mathbf{\Sigma}}_{[-i]}$ for a jackknife estimation, by using the same calculation as for the CV criterion, the AIC$_\mathrm{J}$ can be obtained from only $\hat{\mathbf{\Xi}}$ and $\hat{\mathbf{\Sigma}}$. Let $\hat{B}_{\mathrm{AIC_J}} = c\sum_{i=1}^{n} Q(\hat{r}_i^2/h_i; 1)/h_i - np$, where $\hat{r}_i^2$ and h_i are given by (13) and (15), respectively, then $Q(x;\lambda)$ is a function with respect to x and c is a positive constant, as follows:

$$Q(x;\lambda) = x\left(1 - \frac{x}{n}\right)^{-\lambda}, \quad c = \frac{(n+k)(n-k-p-2)}{(n-k-p-1)\sum_{i=1}^{n} h_i^{-1}}. \tag{21}$$

Then, by using (11), the AIC$_\mathrm{J}$ for the model M is (see Yanagihara, 2006):

$$\mathrm{AIC_J} = np\log 2\pi + n\log|\hat{\mathbf{\Sigma}}| + c\sum_{i=1}^{n}\frac{Q(\hat{r}_i^2/h_i;1)}{h_i}.$$

From Yanagihara (2006), AIC$_\mathrm{J}$ $=$ TIC $+ O_p(n^{-1})$ always holds. Hence, like the TIC, the bias of the AIC$_\mathrm{J}$ is $O(n^{-1})$ under the assumption A1, but it has a constant bias when assumption A1 is violated. On the other hand, when assumptions A1 and A2 are satisfied simultaneously, the AIC$_\mathrm{J}$ is an unbiased estimator of R_{KL}. Although the order of the bias of the AIC$_\mathrm{J}$ is the same as that of the bias of the TIC and EIC, it has been verified numerically that the bias of the AIC$_\mathrm{J}$ in the overspecified model becomes very small. Yanagihara (2006) showed a theoretical result that the absolute value of the bias of the AIC$_\mathrm{J}$ becomes smaller than those of either the TIC or EIC under assumption A1 when we neglect the terms of B that are $O(n^{-2})$.

• **Corrected AIC$_\mathrm{J}$ (CAIC$_\mathrm{J}$)**: Although the bias of the AIC$_\mathrm{J}$ becomes very small, in theory, it does not disappear. Thus, Yanagihara *et al*. (2011) proposed a bias-corrected AIC$_\mathrm{J}$ (CAIC$_\mathrm{J}$) that corrects the bias while maintaining the desirable characteristic of keeping the bias very small. Let $\hat{B}_{\mathrm{CAIC_J}} = c^+\sum_{i=1}^{n}\{1 + a_1(1-h_i)\}Q(\tilde{r}_i^2/h_i; a_0) - np$, where $\hat{r}_i^2$, h_i, and $Q(x;\lambda)$ are given by (13), (15), and (21), respectively, and c^+ and a_j are positive constants such that

Table 1. The order of the bias of each criterion

		Bias Correction	Normality		Nonnormality	
	Criterion	Method	Underspecified	Overspecified	Underspecified	Overspecified
Proposed under Normality	AIC[*1]	—	$O(1)$	$O(n^{-1})$	$O(1)$	$O(1)$
	CAIC[*1,*2]	Exact	$O(1)$	0	$O(1)$	$O(1)$
	MAIC	Moment, Exact	$O(n^{-1})$	$O(n^{-2})$	$O(1)$	$O(1)$
Proposed without Normality	TIC[*3,*4,*5]	Moment	$O(1)$	$O(n^{-1})$	$O(1)$	$O(n^{-1})$
	EIC[*3,*5,*6]	Bootstrap	$O(1)$	$O(n^{-1})$	$O(1)$	$O(n^{-1})$
	EIC$_{\mathrm{A}}$[*3,*6]	Bootstrap	$O(n^{-1})$	$O(n^{-1})$	$O(n^{-1})$	$O(n^{-1})$
	CV[*4]	Cross-validation	$O(1)$	$O(n^{-1})$	$O(1)$	$O(n^{-1})$
	AIC$_{\mathrm{J}}$[*4,*5,*7]	Jackknife, Exact	$O(1)$	0	$O(1)$	$O(n^{-1})$
	CAIC$_{\mathrm{J}}$[*4,*7]	Jackknife, Exact	$O(1)$	0	$O(1)$	$O(n^{-2})$

[*1] The number of explanatory variables in the best model selected by the CAIC is less than or equal to that in the best model selected by the AIC.

[*2] This is the UMVUE of the risk function when assumptions A1 and A2 hold.

[*3] These are asymptotic equivalent when assumption A1 holds. The differences are $O_p(n^{-1/2})$.

[*4] These are asymptotically equivalent. The differences are $O_p(n^{-1})$.

[*5] When $O(n^{-2})$ term is neglected and assumption A1 holds, the absolute value of the bias of the AIC$_{\mathrm{J}}$ is smaller than those of the TIC and EIC.

[*6] The only difference between these two criteria is the resampling method.

[*7] When the $O(n^{-2})$ term is neglected and assumption A1 holds, the variance of the CAIC$_{\mathrm{J}}$ is smaller than that of the AIC$_{\mathrm{J}}$.

$$c^{+} = \frac{(n+k)(n-k-p-2a_0)\Gamma\left(\frac{n-k}{2}+\frac{1}{n}\right)\Gamma\left(\frac{n-k-p}{2}\right)}{(n+a_1k)(n-k-p-1)\Gamma\left(\frac{n-k}{2}\right)\Gamma\left(\frac{n-k-p}{2}+\frac{1}{n}\right)}, \quad a_j = \frac{n+j-1}{n+j}.$$

Here, $\Gamma(x)$ is the gamma function. Then, by using (11), the CAIC$_{\mathrm{J}}$ for the model M is (see Yanagihara *et al.*, 2011)

$$\mathrm{CAIC_J} = np\log 2\pi + n\log|\hat{\boldsymbol{\Sigma}}| + c^{+}\sum_{i=1}^{n}\{1+a_1(1-h_i)\}Q(\tilde{r}_i^2/h_i;a_0).$$

When assumptions A1 and A2 are satisfied simultaneously, like the AIC$_{\mathrm{J}}$, the CAIC$_{\mathrm{J}}$ is an unbiased estimator of R_{KL}. Although, like the AIC$_{\mathrm{J}}$, the CAIC$_{\mathrm{J}}$ has constant bias when assumption A1 is violated, the CAIC$_{\mathrm{J}}$ reduces the bias of the AIC$_{\mathrm{J}}$ to $O(n^{-2})$ when assumption A1 holds. Moreover, Yanagihara *et al.* (2011) showed a theoretical result under assumption A1 that CAIC$_{\mathrm{J}}$ reduces not only the bias of AIC$_{\mathrm{J}}$ but also the variance of AIC$_{\mathrm{J}}$ when we neglect the $O(n^{-2})$ terms.

4. Numerical Comparison

The best models chosen by nine information criteria described in the previous section have the following characteristic when $n \to \infty$ (the proof is given in Appendix E):

Theorem 2 *Suppose that all the multivariate fourth moments of* ε *exist, and* $\lim_{n\to\infty}\boldsymbol{X}_\omega'\boldsymbol{X}_\omega/n$ *exists and is positive definite. Then the best models selected by the nine information criteria con-*

sidered in this paper are asymptotically equivalent. In particular, an underspecified model is never selected as the best model when $n \to \infty$.

Yanagihara *et al.* (2012) derived asymptotic probabilities of selecting the candidate models by AIC and CAIC under assumption A2. In addition to Theorem 2, we can prove that the asymptotic probabilities of selecting the candidate models by nine information criteria become the same as those derived by Yanagihara *et al.* (2012) even if assumption A2 is violated. Theorem 2 indicates that numerical comparisons with variable-selection methods using the nine information criteria are meaningless when the sample size is large. Hence, we conduct numerical comparisons using smaller sample sizes. The model in Yanagihara *et al.* (2011) was used as the basic simulation model for generating data. The expectations and probabilities in the numerical studies were evaluated by a Monte Carlo simulation with 10,000 repetitions. The $\hat{B}_{\mathrm{EIC}}$ and $\hat{B}_{\mathrm{EIC_A}}$ were obtained by resampling $1,000$ times, i.e., $m = 1000$.

We prepared the $k_\omega - 1$ candidate models M_j $(j = 1, \ldots, k_\omega - 1)$ with $p = 4$ and $n = 30$ or 100. First, we generated $z_1, \ldots, z_n \sim i.i.d.\ U(-1, 1)$. Using these $z_1, \ldots, z_n$, we constructed the $n \times k_\omega$ matrix of explanatory variables $\boldsymbol{X}_\omega$, whose (i, j)th element is defined by $\{(z_i - \bar{z})/s_z\}^{j-1}$ $(i = 1, \ldots, n;\ j = 1, \ldots, k_\omega)$, and where $\bar{z}$ and s_z are the sample mean and standard deviation, respectively, of $z_1, \ldots, z_n$. The true model was determined by $\boldsymbol{\Gamma}_* = \boldsymbol{X}_\omega \boldsymbol{\mu}_* \boldsymbol{1}_4'$ and $\boldsymbol{\Sigma}_*$, whose (i, j)th element is defined by $(0.8)^{|i-j|}$ $(i = 1, \ldots, 4;\ j = 1, \ldots, 4)$, where $\boldsymbol{1}_p$ is a p-dimensional vector of ones. In this simulation study, we prepared the six $\boldsymbol{\mu}_*$ as

$$\begin{array}{lll}
\text{Case 1:} & \boldsymbol{\mu}_* = (0,1,2,4,0,0,0,0)', & (k_\omega = 8), \\
\text{Case 2:} & \boldsymbol{\mu}_* = (0,1,2,4,0.5,0.5,0,0)', & (k_\omega = 8), \\
\text{Case 3:} & \boldsymbol{\mu}_* = (0,1,2,4,0,0,0,0,0,0,0,0,0,0,0,0)', & (k_\omega = 16), \\
\text{Case 4:} & \boldsymbol{\mu}_* = (0,1,2,4,0.5,0.5,0,0,0,0,0,0,0,0,0,0)', & (k_\omega = 16), \\
\text{Case 5:} & \boldsymbol{\mu}_* = (0,1,1,1,-1,-1,2,2,4,0,0,0,0,0,0,0)', & (k_\omega = 16), \\
\text{Case 6:} & \boldsymbol{\mu}_* = (0,1,1,1,-1,-1,2,2,4,0.5,0.5,0,0,0,0,0)', & (k_\omega = 16).
\end{array}$$

The matrix of explanatory variables in M_j $(j = 1, \ldots, k_\omega - 1)$ consists of the first $(j + 1)$ columns of $\boldsymbol{X}_\omega$. Thus, the true models M_* in the cases 1, 2, 3, 4, 5, and 6 are M_3, M_5, M_3, M_5, M_8, and M_{10}, respectively. In a sense, the subindex j expresses the degree of the polynomial regression in M_j.

Next, in order to generate multivariate nonnormal data, we prepared the data model introduced by Yuan and Bentler (1997).

Data Model: Let $w_1, \ldots, w_q$ $(q \geq p)$ be independent random variables with $E[w_j] = 0$, $E[w_j^2] = 1$ and $E[w_j^4] - 3 = \psi$, and $\boldsymbol{w} = (w_1, \ldots, w_q)'$. Further, let r be a random variable that is independent of $\boldsymbol{w}$, $E[r^2] = 1$ and $E[r^4] = \beta$. Then, an error vector is generated by $\boldsymbol{\varepsilon} = r\boldsymbol{C}'\boldsymbol{w}$, where $\boldsymbol{C}$ is a $q \times p$ matrix defined by $\boldsymbol{C} = (\boldsymbol{c}_1, \ldots, \boldsymbol{c}_q)'$ with full rank p, and $\boldsymbol{C}'\boldsymbol{C} = \boldsymbol{I}_p$. Then, the multivariate kurtosis of this model becomes $\kappa_4^{(1)} = \beta\psi \sum_{j=1}^q (\boldsymbol{c}_j'\boldsymbol{c}_j)^2 + (\beta - 1)p(p + 2)$.

Let χ_f be a random variable from the chi-square distribution with f degrees of freedom, and let $\boldsymbol{C}_0$ be a $(p + 1) \times p$ matrix defined by $\boldsymbol{C}_0 = (\boldsymbol{I}_p, \boldsymbol{1}_p)'(\boldsymbol{I}_p + \boldsymbol{1}_p\boldsymbol{1}_p')^{-1/2}$. By using the data model, we generate error vectors with the following three distributions:

1. *Normal Distribution*: $w_j \sim N(0,1)$, $r = 1$ and $\boldsymbol{C} = \boldsymbol{I}_p$ ($\kappa_4^{(1)} = 0$).
2. *Laplace Distribution*: w_j is generated from a Laplace distribution with mean 0 and standard deviation 1, $r = \sqrt{6/\chi_8^2}$ and $\boldsymbol{C} = \boldsymbol{C}_0$ ($\kappa_4^{(1)} = 4.5 \times p^2(p+1)^{-1} + p(p+2)/2$).
3. *Skew Laplace Distribution*: w_j is generated from a skew Laplace distribution with location parameter 0, dispersion parameter 1, and skew parameter 1, standardized by mean 3/4 and standard deviation $\sqrt{23}/4$, $r = \sqrt{6/\chi_8^2}$ and $\boldsymbol{C} = \boldsymbol{C}_0$ ($\kappa_4^{(1)} \approx 4.88 \times p^2(p+1)^{-1} + p(p+2)/2$).

For details of the skew Laplace distribution, see, e.g., Kotz *et al*. (2001). It is easy to see that data models 1 and 2 are symmetric distributions, and data model 3 is a skewed distribution. Moreover, the size of the kurtosis $\kappa_4^{(1)}$ in each model satisfies the following inequality: model 1 < model 2 < model 3.

First, we examine the average biases of the criteria. Figure 1 shows R_{KL} and the average of each criterion in case 1. Since the shape of the figure was almost the same, we omit the results for cases 2 to 6 to save space. The horizontal axis of the figures expresses the number of candidate models, i.e., the subindex j of M_j. We see that the biases of the AIC$_{\mathrm{J}}$ and CAIC$_{\mathrm{J}}$ were very small under any distribution. As for the size of the bias, the AIC most underestimated the risk function, and the CV criterion overestimated the risk function in the most cases. The size of the bias of the TIC was almost the same as that of the AIC. This is because the estimate of the multivariate kurtosis $\hat{\kappa}_4^{(1)}$ for the TIC was close to 0 when the sample size was not large enough. Moreover, as the number of variables in the model increases, the biases of the AIC and TIC increase.

Tables 2 and 3 show, for case 1 and for each information criterion, the standard deviation ($\{Var[\mathrm{IC}]\}^{1/2}$) and the root-mean-square error (RMSE) ($\{Var[\mathrm{IC}] + (E[\mathrm{IC}] - R_{\mathrm{KL}})^2\}^{1/2}$). Since the tendency was almost the same, to save space, we omit the results for M_2, M_3, M_4, M_5, and M_6, and in cases 2 to 6. We can see in the tables that the standard deviations of the AIC and CAIC were the smallest and those of the MAIC and TIC were the second smallest. The standard deviation of the EIC and EIC$_{\mathrm{A}}$ were larger than that of the AIC, but smaller than those of the CV, AIC$_{\mathrm{J}}$, and CAIC$_{\mathrm{J}}$. The standard deviation of the CV criterion was the largest among all the information criteria considered. On the other hand, the RMSEs of the AIC and TIC became large when the sample size was small because their biases became large. The RMSEs of the CV criterion, the AIC$_{\mathrm{J}}$, and CAIC$_{\mathrm{J}}$ were also large because their standard deviations became large. In all cases, there was a tendency for the standard deviation and RMSE to become large when κ_4 was large.

To compare the nine information criteria for their performance as a model selector, the following two properties were considered:

- the probability of selecting the principle best model: the frequency with which the principle best model is selected as the best model.
- the prediction error (PE) of the best model: the risk function of the best model which is chosen by the information criterion; PE is defined as follows:

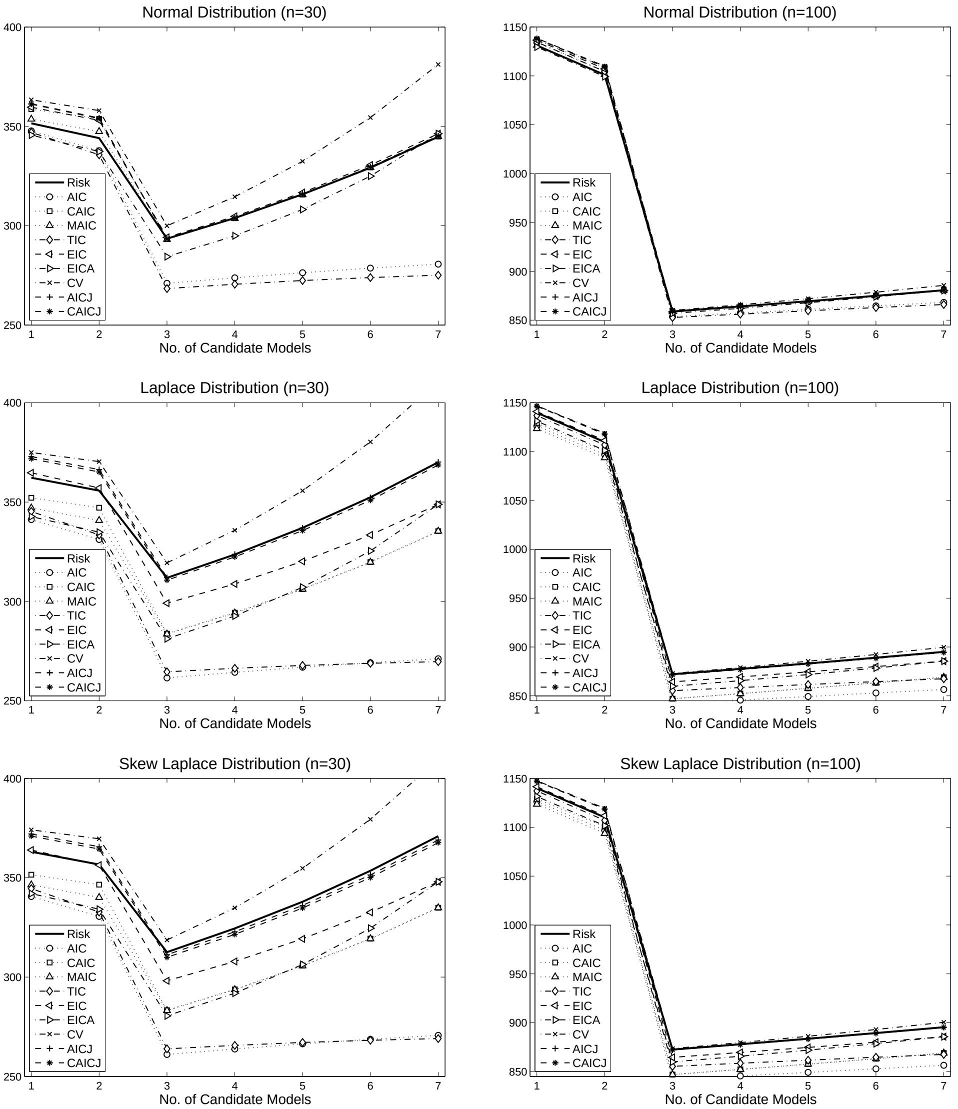

Figure 1. Risk function and the average of each criterion (Case 1)

$$\mathrm{PE} = E^{*}_{\boldsymbol{Y}}[\mathcal{L}(\hat{\boldsymbol{\Xi}}_{\mathrm{best}}, \hat{\boldsymbol{\Sigma}}_{\mathrm{best}})],$$

where $\mathcal{L}(\boldsymbol{\Xi}, \boldsymbol{\Sigma})$ is given by (8) and $\hat{\boldsymbol{\Xi}}_{\mathrm{best}}$ and $\hat{\boldsymbol{\Sigma}}_{\mathrm{best}}$ are the MLEs of $\boldsymbol{\Xi}$ and $\boldsymbol{\Sigma}$, respectively, under the best model.

A high-performance model selector is considered to be an information criterion with a high probability of selecting the principle best model and a small prediction error. According to the basic concept of the model selection based on the risk function minimization, a good model selection method is one that can choose the best model for improving the predictive accuracy. Hence, the PE is a more important property than is the probability of selecting the principle best model.

Table 2. Standard deviation of each criterion (Case 1)

n	Dist.	Model	AIC	CAIC	MAIC	TIC	EIC	EIC$_\text{A}$	CV	AIC$_\text{J}$	CAIC$_\text{J}$
30	1	1	15.010	15.010	15.033	15.106	15.342	15.179	16.007	15.998	15.899
		3	17.416	17.416	17.476	17.567	17.842	17.813	19.465	19.358	19.010
		7	19.007	19.007	19.007	19.228	19.680	19.680	30.358	28.239	24.748
	2	1	24.300	24.300	24.359	25.931	30.636	25.933	39.426	39.264	38.073
		3	29.050	29.050	29.123	30.758	35.666	31.824	51.824	50.891	48.977
		7	30.194	30.194	30.194	31.440	35.972	35.972	70.243	64.135	59.042
	3	1	24.539	24.539	24.626	26.264	31.183	26.330	39.878	39.717	38.532
		3	29.102	29.102	29.199	30.828	35.906	31.930	53.943	52.920	50.881
		7	30.317	30.317	30.317	31.546	36.130	36.130	72.282	65.915	61.491
100	1	1	25.465	25.465	25.460	25.490	25.519	25.501	25.519	25.519	25.518
		3	29.346	29.346	29.343	29.401	29.410	29.403	29.457	29.457	29.449
		7	29.896	29.896	29.896	29.995	29.968	29.968	30.268	30.263	30.171
	2	1	45.873	45.873	45.892	48.881	50.177	48.966	54.003	54.025	53.871
		3	54.960	54.960	54.964	58.601	60.232	59.079	65.510	65.512	65.312
		7	55.323	55.323	55.323	58.706	60.240	60.240	66.751	66.645	66.355
	3	1	46.667	46.667	46.682	50.057	51.413	50.127	55.152	55.176	55.033
		3	55.358	55.358	55.358	59.470	61.296	60.043	66.796	66.801	66.601
		7	55.669	55.669	55.669	59.438	61.244	61.244	67.987	67.877	67.623

Table 3. RMSE of each criterion (Case 1)

n	Dist.	Model	AIC	CAIC	MAIC	TIC	EIC	EIC$_\text{A}$	CV	AIC$_\text{J}$	CAIC$_\text{J}$
30	1	1	15.486	16.625	15.181	15.772	17.357	16.290	19.905	18.803	18.599
		3	28.397	17.416	17.478	30.642	17.855	19.981	20.531	19.358	19.010
		7	66.895	19.007	19.007	72.312	19.740	19.740	47.359	28.242	24.749
	2	1	32.159	26.318	28.698	30.994	30.735	32.465	41.417	40.677	39.253
		3	58.144	40.404	40.567	56.425	37.878	44.191	52.376	50.891	48.990
		7	103.424	45.985	45.985	105.162	41.763	41.763	81.197	64.135	59.059
	3	1	33.300	27.123	29.715	32.153	31.195	33.695	41.371	40.715	39.331
		3	59.137	41.222	41.410	57.603	38.675	45.242	54.292	52.935	50.948
		7	104.755	47.094	47.094	106.657	42.810	42.810	81.953	65.943	61.577
100	1	1	25.637	26.089	25.462	25.719	26.102	25.552	26.554	26.460	26.449
		3	29.818	29.346	29.344	30.044	29.413	29.471	29.475	29.458	29.450
		7	32.396	29.896	29.896	33.371	29.969	29.969	30.669	30.263	30.171
	2	1	47.841	47.144	48.692	48.967	50.191	49.714	54.467	54.451	54.270
		3	62.714	60.405	60.411	60.963	60.729	60.356	65.514	65.514	65.316
		7	67.442	61.137	61.137	64.859	60.990	60.990	66.914	66.646	66.358
	3	1	48.672	47.973	49.517	50.139	51.431	50.850	55.661	55.646	55.473
		3	63.288	60.962	60.964	61.888	61.811	61.352	66.804	66.801	66.602
		7	67.982	61.641	61.641	65.645	62.010	62.010	68.174	67.877	67.624

Tables 4 and 5 show the selection probability and PE, respectively. When $n = 30$, the principle best models were different from the true models in the cases 2, 4, 5, and 6, in which the principle best models were M_3, M_3, M_6, and M_7, respectively. On the other hand, when $n = 100$, the principle best model was different from the true model only in case 6, in which the principle best model

Table 4. Probabilities of selecting the principle best model

Case	n	Dist.	AIC	CAIC	MAIC	TIC	EIC	EIC_A	CV	AIC_J	CAIC_J
1	30	1	69.07	98.44	*99.41*	60.20	97.92	**99.62**	98.66	95.12	96.07
		2	70.19	98.46	*99.55*	54.35	94.19	**99.64**	95.02	91.59	92.65
		3	69.68	98.35	*99.41*	53.84	94.42	**99.74**	95.18	91.73	92.84
	100	1	85.11	92.59	*93.82*	82.51	92.51	**94.28**	93.63	91.75	91.87
		2	85.50	92.94	*94.18*	79.39	90.22	**96.22**	93.01	91.13	91.21
		3	85.04	92.20	*93.70*	79.09	89.87	**96.22**	92.79	90.78	90.96
2	30	1	34.70	87.34	*93.48*	26.83	86.92	**95.33**	90.71	79.01	80.98
		2	30.82	84.57	*91.54*	21.99	80.84	**95.27**	88.84	77.52	79.96
		3	30.27	84.07	91.04	22.15	80.26	95.07	88.92	77.08	79.19
	100	1	**56.85**	50.78	47.78	*56.66*	50.40	46.13	47.42	51.00	51.03
		2	**58.45**	52.19	49.07	*54.17*	46.90	39.82	41.18	44.48	44.73
		3	**58.55**	52.08	49.58	*54.50*	47.60	40.46	41.86	45.09	45.01
3	30	1	50.70	98.20	**99.04**	15.16	97.56	89.42	*98.40*	94.24	96.10
		2	48.98	*98.26*	**99.46**	12.22	94.18	89.08	95.22	90.12	92.86
		3	49.86	*98.40*	**99.28**	12.54	94.58	89.78	95.08	90.08	92.54
	100	1	84.64	92.40	*93.59*	81.22	92.21	91.36	**93.62**	91.45	91.57
		2	84.39	92.22	**93.25**	76.86	89.33	*92.68*	92.57	90.40	90.57
		3	84.63	92.54	**93.82**	76.68	89.64	92.97	*93.14*	91.01	91.20
4	30	1	23.10	86.92	**92.48**	6.04	86.08	63.20	*89.32*	76.76	80.28
		2	20.14	83.68	**89.82**	3.64	78.44	60.52	*87.80*	73.84	78.14
		3	20.60	83.80	**90.28**	4.80	80.30	59.94	*88.42*	75.48	78.72
	100	1	**55.03**	49.49	46.27	*52.55*	49.38	50.64	46.02	49.64	50.02
		2	**57.20**	*52.13*	48.85	50.83	47.24	48.80	41.48	44.49	44.66
		3	**57.01**	*52.57*	49.51	50.34	47.63	49.27	41.95	45.03	45.32
5	30	1	0.00	13.14	*32.36*	0.00	16.97	9.35	**52.99**	14.86	16.92
		2	0.01	12.04	*27.49*	0.00	19.93	11.14	**59.57**	24.44	27.32
		3	0.03	11.98	*27.77*	0.01	18.17	10.45	**58.67**	23.61	26.23
	100	1	81.26	93.78	*96.24*	69.55	93.84	**96.94**	94.02	85.14	90.15
		2	80.96	93.57	*96.04*	65.05	91.92	**97.77**	93.62	83.40	89.14
		3	80.31	93.72	*96.19*	65.35	92.00	**97.70**	93.28	83.50	89.07
6	30	1	0.00	12.43	**43.81**	0.00	17.70	*35.74*	29.85	9.50	11.53
		2	0.02	12.39	**36.86**	0.00	24.16	*34.66*	32.36	16.43	22.50
		3	0.01	12.24	**38.17**	0.01	24.08	*35.10*	33.40	17.43	23.29
	100	1	58.23	80.14	*85.79*	45.61	80.25	**87.91**	80.46	65.66	72.51
		2	57.59	79.48	*85.09*	42.72	78.61	**90.24**	81.20	65.54	72.78
		3	58.58	79.45	*85.18*	43.79	78.75	**89.62**	81.20	66.21	73.27

was M_7. In the tables, bold and italic fonts indicate the highest and second highest probabilities of selecting the principle best model and the smallest and second smallest prediction errors of the best models. We see from the tables that , except for the TIC, the bias-corrected AICs resulted in improved performance for variable selection, compared to the uncorrected AIC. This indicates that correcting the bias of the AIC is effective for improving the performance of the AIC as a model selector when the sample size is not large. Although in theory, the TIC reduces the bias of the AIC,

Table 5. Prediction errors of the best model

Case	n	Dist.	AIC	CAIC	MAIC	TIC	EIC	$\mathrm{EIC_A}$	CV	$\mathrm{AIC_J}$	$\mathrm{CAIC_J}$
1	30	1	310.13	294.30	*293.84*	314.82	294.49	**293.70**	294.08	295.59	295.21
		2	330.42	312.82	*312.16*	339.12	314.07	**312.07**	314.42	315.81	315.36
		3	331.31	313.57	*312.97*	340.15	314.77	**312.73**	315.08	316.52	316.03
	100	1	861.86	860.29	*860.05*	862.42	860.29	**859.94**	860.07	860.45	860.41
		2	875.17	873.53	*873.28*	876.41	874.00	**872.90**	873.51	873.87	873.86
		3	875.59	874.04	*873.74*	876.81	874.45	**873.24**	873.91	874.30	874.22
2	30	1	319.84	300.60	*299.14*	323.80	300.51	**298.56**	299.59	302.26	301.78
		2	342.00	318.78	*316.75*	348.57	319.27	**315.47**	317.39	320.58	319.98
		3	342.88	319.13	*316.92*	349.43	319.86	**315.52**	317.55	321.06	320.31
	100	1	873.04	872.53	*872.50*	873.38	872.55	**872.38**	872.55	872.65	872.62
		2	887.13	*886.45*	**886.38**	887.96	887.05	886.71	887.02	887.12	887.08
		3	887.16	*886.57*	**886.50**	887.98	887.13	886.75	887.10	887.14	887.14
3	30	1	439.72	294.46	**294.06**	578.07	294.65	297.77	*294.27*	300.32	295.15
		2	498.77	*312.61*	**312.02**	651.61	313.86	316.68	314.05	322.85	315.00
		3	498.99	*313.24*	**312.77**	654.32	314.39	317.14	314.67	323.77	315.98
	100	1	862.26	860.23	*860.02*	863.88	860.27	860.38	**859.99**	860.47	860.40
		2	876.39	873.94	**873.70**	879.92	874.49	873.77	*873.74*	874.42	874.28
		3	876.94	874.22	**873.91**	880.92	874.86	874.07	*874.01*	874.77	874.69
4	30	1	468.43	300.60	**299.30**	594.48	300.52	305.25	*299.69*	310.28	301.73
		2	523.01	318.88	**317.08**	662.42	319.95	324.96	*317.77*	330.44	320.72
		3	535.53	319.01	**317.05**	668.18	319.76	325.20	*317.64*	329.55	320.37
	100	1	874.65	872.81	*872.71*	877.19	872.79	**872.52**	872.79	873.26	872.96
		2	888.63	886.30	*886.26*	893.32	886.97	**886.12**	886.78	887.56	887.21
		3	889.36	886.98	*886.80*	893.87	887.63	**886.72**	887.39	888.16	887.85
5	30	1	534.94	358.40	**354.18**	606.00	357.82	354.57	*354.41*	393.43	362.78
		2	599.70	384.29	*377.31*	676.70	382.62	378.95	**374.86**	420.67	386.40
		3	598.80	382.88	*376.24*	676.24	381.25	377.87	**374.05**	420.32	385.74
	100	1	891.83	888.07	*887.50*	896.26	888.03	**887.34**	887.92	891.38	888.88
		2	907.82	903.71	*903.06*	914.29	904.06	**902.63**	903.73	908.18	905.09
		3	907.98	903.62	*902.98*	914.20	904.07	**902.64**	903.87	908.19	905.21
6	30	1	543.46	**364.39**	364.69	607.88	365.17	*364.54*	366.57	402.97	369.09
		2	615.90	*392.18*	**391.43**	686.35	394.43	392.48	392.97	437.17	398.68
		3	618.30	393.04	**392.34**	688.63	393.99	*392.46*	393.87	438.28	399.14
	100	1	896.97	892.20	*891.39*	901.53	892.17	**891.04**	892.05	896.66	893.43
		2	912.35	907.04	*906.18*	918.32	907.23	**905.37**	906.76	912.26	908.63
		3	912.71	907.56	*906.57*	918.60	907.70	**905.89**	907.27	912.64	909.08

its performance as a model selector was inferior. This is because the TIC only minimally corrects the bias of the AIC. As stated earlier, the $\mathrm{AIC_J}$ and $\mathrm{CAIC_J}$ have the smallest biases. Nevertheless, their performance for variable selection was not the best. This leads us to the conclusion that it is not necessary to bring the bias close to 0 as much as possible, although bias correction is effective. The best performance in the sense of high selection probability and small PE was by the MAIC and $\mathrm{EIC_A}$. This is because the candidate model that minimizes the loss function is either the true model or an underspecified model, as described in the proof of Theorem 1. Hence, this result indicates that the bias correction in the underspecified model is important for improving the model-selecting performance of an information criterion. The performance of the $\mathrm{EIC_A}$ was slightly better than that of the MAIC; this is because the $\mathrm{EIC_A}$ reduces the influence of nonnormality more effectively than does the MAIC. However, when the sample size was small and the number of explanatory variables was large, i.e., cases 4 to 6, the performance of the $\mathrm{EIC_A}$ as a model selector was reduced. One reason for this is that the $\mathrm{EIC_A}$ is constructed by resampling the full model. When the sample size is small and the number of explanatory variables is large, we anticipate that the accuracy of resampling will be decreased due to a decrease in the asymptotic approximation. The performance of the CV criterion as a model selector was not bad even though it has a large bias. This is because the variable-selection method using the CV criterion is conscious of improving for a prediction of a validation sample. Although the performance was not bad, it was not as good as either the MAIC or $\mathrm{EIC_A}$.

In this section, we listed simulation results of the variable selections using nested models. We also conducted simulations using nonnested models. However, we omit the results because they were very similar to those for the nested models.

5. Conclusions and Discussion

In this paper, we considered variable-selection methods for normal MLRM models, using the original AIC and eight bias-corrected AICs. From a theoretical aspect, we derived asymptotic results ($n \to \infty$) under the assumption that the true distribution is not always normal, while from a numerical aspect, we performed a comparative study for small- to medium-sized samples. Our results are summarized as follows:

- When $n \to \infty$, the best models selected by all the criteria become the same, even though the biases of the criteria were corrected under the assumption of nonnormality. Moreover, in this case, an underspecified model will never be selected as the best model.

- Except for the TIC, the performances of the variable-selection methods using the bias-corrected AIC were better than that using the uncorrected AIC. This suggests that exact correction, bootstrapping, or cross-validation work better than the moment method for correcting the bias. It will be that correcting only the top term in an asymptotic expansion of the bias, as do AIC and TIC, is insufficient in the overspecified models.

- Theoretically, the bias of the $\mathrm{CAIC_J}$ becomes the smallest among all the criterion mentioned in this paper, but by numerical examination, we verified that the $\mathrm{CAIC_J}$ is not the best model selector. This indicates that the performance of a criterion is not necessarily improved even if the bias of the risk function for the overspecified model is reduced to as small as possible.

- The CAIC and MAIC perform well as model selectors, even though they have constant bias when the true distribution is not normal. The reason for this is that the correction for the bias caused by nonnormality cannot be estimated accurately when the sample size is small. Thus, if we try to estimate this bias when the sample size is small, it will reduce the accuracy of the estimation.

- Variable-selection methods using the MAIC or EIC_A, which are obtained by correcting the constant bias of the AIC, always perform well. This result leads us to the conclusion that correcting the bias for the underspecified model has a positive effect on the selection of variables. One reason for this is that the model that minimizes the loss function is either the true model or the underspecified model. The EIC_A has the best performance as the model selector except for when the sample size is small and there are a large number of explanatory variables in the full model.

In conclusion, we recommend using the MAIC for a small number of samples and the EIC_A for a moderate number of samples. We note that when the number of samples is sufficiently large, it does not matter which criterion is used.

Acknowledgment Hirokazu Yanagihara's research was partially supported by the Ministry of Education, Science, Sports, and Culture, and a Grant-in-Aid for Challenging Exploratory Research, #25540012, 2013-2015; Shinpei Imori's research was supported by Research Fellowships of the Japan Society for the Promotion of Science for Young Scientists; Mariko Yamamura's research was supported by the project to support female researchers at Hiroshima University through system reforms.

References

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

Akaike, H. (1974): A new look at the statistical model identification. *Institute of Electrical and Electronics Engineers. Transactions on Automatic Control* **AC-19**, 716–723.

Bedrick, E. J. and Tsai, C.-L. (1994): Model selection for multivariate regression in small samples. *Biometrics* **50**, 226–231.

Burnham, K. P. and Anderson, D. R. (2002): *Model Selection and Multimodel Inference*. *A Practical Information-Theoretic Approach* (2nd. ed.). Springer-Verlag, New York.

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

Freedman, D. A. (1981): Bootstrapping regression models. *The Annals of Statistics* **9**, 1218–1228.

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

Fujikoshi, Y., Yanagihara, H. and Wakaki, H. (2005): Bias corrections of some criteria for selection multivariate linear regression models in a general case. *American Journal of Mathematical and Management Sciences* **25**, 221–258.

Hurvich, C. M. and Tsai, C.-L. (1989): Regression and times series model selection in small samples. *Biometrika* **50**, 226–231.

Ishiguro, M., Sakamoto, Y. and Kitagawa, G. (1997): Bootstrapping log likelihood and EIC, an extension of AIC. *Annals of the Institute of Statistical Mathematics* **49**, 411–434.

Konishi, S. (1999): Statistical model evaluation and information criteria. In *Multivariate Analysis, Design of Experiments, and Survey Sampling*, Ed. S. Ghosh, pp. 369–399. Marcel Dekker, New York.

Konishi, S. and Kitagawa, G. (2008): *Information Criteria and Statistical Modeling*. Springer Science+Business Media, LLC, New York.

Kotz, S., Kozubowski, T. J. and Podgórski, K. (2001): *The Laplace Distribution and Generalizations*. Birkhäuser Boston, Inc., Boston.

Kullback, S. and Leibler, R. A. (1951): On information and sufficiency. *The Annals of Mathematical Statistics* **22**, 79–86.

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

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

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

Stone, M. (1974): Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society, Series* **B 36**, 111–147.

Stone, M. (1977): An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society, Series* **B 39**, 44–47.

Sugiura, N. (1978): Further analysis of the data by Akaike's information criterion and the finite corrections. *Communications in Statistics, Theory and Methods* **A7**, 13–26.

Takeuchi, K. (1976): Distribution of information statistics and criteria for adequacy of models. *Mathematical Science* **153**, 12–18 (in Japanese).

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

Yanagihara, H. (2006): Corrected version of *AIC* for selecting multivariate normal linear regression models in a general nonnormal case. *Journal of Multivariate Analysis* **97**, 1070–1089.

Yanagihara, H. (2007): A family of estimators for multivariate kurtosis in a nonnormal linear regression model. *Journal of Multivariate Analysis* **98**, 1–29.

Yanagihara, H. and Ohmoto, C. (2005): On distribution of AIC in linear regression models. *Journal of Statistical Planning and Inference* **133**, 417–433.

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

Yanagihara, H., Kamo, K. and Tonda, T. (2011): Second-order bias-corrected AIC in multivariate normal linear models under nonnormality. *The Canadian Journal of Statistics* **39**, 126–146.

Yanagihara, H., Wakaki, H. and Fujikoshi, Y. (2012): A consistency property of the AIC for multivariate linear models when the dimension and the sample size are large. *TR 12-08, Statistical Research Group*, Hiroshima University, Hiroshima.

Yoshimoto, A., Yanagihara, H. and Ninomiya, Y. (2005): Finding factors affecting a forest stand growth through multivariate linear modeling. *Journal of Japanese Forestry Society* **87**, 504–512 (in Japanese).

Yuan, K.-H. and Bentler, P. M. (1997): Generating multivariate distributions with specified marginal skewness and kurtosis. In *SoftStat' 97 – Advances in Statistical Software 6 –* , Eds. W. Bandilla & F. Faulbaum, pp. 385–391. Lucius & Lucius, Stuttgart, Germany.

Appendix

A. Proof of Theorem 1

First, we show that the candidate model minimizing the loss function is either the true model or the underspecified model; we do this in order to prove that the principle best model is either the true model or the underspecified model. Let $\boldsymbol{X}_1 = (\boldsymbol{X}, \boldsymbol{a})$ be a $n \times (k+1)$ matrix of explanatory variables in the model M_1: $\boldsymbol{Y} \sim N_{n\times p}(\boldsymbol{X}_1\boldsymbol{\Xi}_1, \boldsymbol{\Sigma}_1 \otimes \boldsymbol{I}_n)$, where $\boldsymbol{a}$ is an n-dimensional vector that is linearly independent from any combination of the columns of $\boldsymbol{X}$. Let $\hat{\boldsymbol{\Xi}}_1$ and $\hat{\boldsymbol{\Sigma}}_1$ denote the MLEs of $\boldsymbol{\Xi}_1$ and $\boldsymbol{\Sigma}_1$, respectively. From the formula for the inverse matrix (see, e.g., Siotani *et al.*, 1985, p. 592, Theorem A.2.3), we have

$$\boldsymbol{P}_{\boldsymbol{X}_1} = \boldsymbol{P}_{\boldsymbol{X}} + \frac{1}{\boldsymbol{a}'(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{a}}(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{a}\boldsymbol{a}'(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}}) = \boldsymbol{P}_{\boldsymbol{X}} + \boldsymbol{a}_{\mathrm{s}}\boldsymbol{a}_{\mathrm{s}}',$$

where $\boldsymbol{a}_{\mathrm{s}} = (\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{a}/\sqrt{\boldsymbol{a}'(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{a}}$. From the formulas for the determinant and the inverse matrix (see, e.g., Siotani *et al.*, 1985, p. 591, Theorem A.1.3, and p. 592, Theorem A.2.2), $|\hat{\boldsymbol{\Sigma}}_1|$ and $\hat{\boldsymbol{\Sigma}}_1^{-1}$ are rewritten as

$$|\hat{\boldsymbol{\Sigma}}_1| = |\hat{\boldsymbol{\Sigma}}|\,(1 - \boldsymbol{a}_{\mathrm{s}}'\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{a}_{\mathrm{s}})\,, \tag{A.1}$$

$$\hat{\boldsymbol{\Sigma}}_1^{-1} = \hat{\boldsymbol{\Sigma}}^{-1} + \frac{n}{1 - \boldsymbol{a}_{\mathrm{s}}'\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{a}_{\mathrm{s}}}\boldsymbol{\Sigma}_*^{-1/2}(\boldsymbol{W}'\boldsymbol{W})^{-1}\boldsymbol{W}'\boldsymbol{a}_{\mathrm{s}}\boldsymbol{a}_{\mathrm{s}}'\boldsymbol{W}(\boldsymbol{W}'\boldsymbol{W})^{-1}\boldsymbol{\Sigma}_*^{-1/2}, \tag{A.2}$$

where $\boldsymbol{W} = (\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{Y}\boldsymbol{\Sigma}_*^{-1/2}$. Recall that $\hat{\boldsymbol{\Sigma}}_1$ is positive definite and $\boldsymbol{a}$ is linearly independent from any combinations of the columns of $\boldsymbol{X}$. From the proof in Appendix A.5 of Yanagihara and Satoh (2010), we can see $0 < \boldsymbol{a}_{\mathrm{s}}'\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{a}_{\mathrm{s}} < 1$. Suppose that the model M is overspecified. Then, $\boldsymbol{W} = (\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{\mathcal{E}}$ holds, where $\boldsymbol{\mathcal{E}}$ is given by (4). Moreover, since $n\hat{\boldsymbol{\Sigma}}_1 = \boldsymbol{\Sigma}_*^{1/2}\boldsymbol{\mathcal{E}}'(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}_1})\boldsymbol{\mathcal{E}}\boldsymbol{\Sigma}_*^{1/2}$ holds when M is the overspecified model, the loss function under M_1 can be simplified as

$$\mathcal{L}(\hat{\boldsymbol{\Xi}}_1, \hat{\boldsymbol{\Sigma}}_1) = np\log 2\pi + n\log|\hat{\boldsymbol{\Sigma}}_1| + \mathrm{tr}\left\{\hat{\boldsymbol{\Sigma}}_1^{-1}\boldsymbol{\Sigma}_*^{1/2}(n\boldsymbol{I}_p + \boldsymbol{\mathcal{E}}'\boldsymbol{\mathcal{E}})\boldsymbol{\Sigma}_*^{1/2}\right\} - np. \tag{A.3}$$

Substituting (A.1) and (A.2) into (A.3) yields

$$\begin{aligned}\mathcal{L}(\hat{\boldsymbol{\Xi}}_1, \hat{\boldsymbol{\Sigma}}_1) &= np\log 2\pi + n\log|\hat{\boldsymbol{\Sigma}}| + n\log(1 - \boldsymbol{a}_{\mathrm{s}}'\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{a}_{\mathrm{s}}) + \mathrm{tr}\left\{\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Sigma}_*^{1/2}(n\boldsymbol{I}_p + \boldsymbol{\mathcal{E}}'\boldsymbol{\mathcal{E}})\boldsymbol{\Sigma}_*^{1/2}\right\} - np \\ &\quad + \frac{n}{1 - \boldsymbol{a}_{\mathrm{s}}'\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{a}_{\mathrm{s}}}\boldsymbol{a}_{\mathrm{s}}'\boldsymbol{W}(\boldsymbol{W}'\boldsymbol{W})^{-1}\left(\boldsymbol{W}'\boldsymbol{W} + \boldsymbol{\mathcal{E}}'\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{\mathcal{E}} + n\boldsymbol{I}_p\right)(\boldsymbol{W}'\boldsymbol{W})^{-1}\boldsymbol{W}'\boldsymbol{a}_{\mathrm{s}} \\ &\geq \mathcal{L}(\hat{\boldsymbol{\Xi}}, \hat{\boldsymbol{\Sigma}}) + n\left\{\log(1 - \boldsymbol{a}_{\mathrm{s}}'\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{a}_{\mathrm{s}}) + \frac{\boldsymbol{a}_{\mathrm{s}}'\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{a}_{\mathrm{s}}}{1 - \boldsymbol{a}_{\mathrm{s}}'\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{a}_{\mathrm{s}}}\right\}.\end{aligned}$$

Notice that $\log(1 - x) + x/(1 - x) > 0$ when $x \in (0, 1)$. Hence, the inequality $\mathcal{L}(\hat{\boldsymbol{\Xi}}_1, \hat{\boldsymbol{\Sigma}}_1) > \mathcal{L}(\hat{\boldsymbol{\Xi}}, \hat{\boldsymbol{\Sigma}})$ holds. This means that the loss function becomes small when a new explanatory variable is added to the overspecified model. Since the overspecified model that has the smallest number of explanatory variables is the true model, the candidate model that minimizes the loss function is either the true model or the underspecified model.

Next, we show that the candidate model that minimizes the loss function is the true model when $n \to \infty$; we do this in order to prove that the principle best model is asymptotically equivalent to the true model. When the assumptions in Theorem 1 hold, we obtain $\hat{\boldsymbol{\Sigma}} \xrightarrow{p} \boldsymbol{\Sigma}_* + \boldsymbol{\Sigma}_*^{1/2}\boldsymbol{\Omega}\boldsymbol{\Sigma}_*^{1/2}$ and $(\boldsymbol{\Gamma}_* - \boldsymbol{X}\hat{\boldsymbol{\Xi}})'(\boldsymbol{\Gamma}_* - \boldsymbol{X}\hat{\boldsymbol{\Xi}})/n \xrightarrow{p} \boldsymbol{\Sigma}_*^{1/2}\boldsymbol{\Omega}\boldsymbol{\Sigma}_*^{1/2}$ as $n \to \infty$, where $\boldsymbol{\Omega}$ is given by (5). The above results imply that

$$\begin{aligned}\frac{1}{n}\mathcal{L}(\hat{\boldsymbol{\Xi}}, \hat{\boldsymbol{\Sigma}}) &\xrightarrow{p} p\log 2\pi + \log|\boldsymbol{\Sigma}_*| + \log|\boldsymbol{I}_p + \boldsymbol{\Omega}| + \mathrm{tr}\{(\boldsymbol{I}_p + \boldsymbol{\Omega})^{-1}\} + \mathrm{tr}\{(\boldsymbol{I}_p + \boldsymbol{\Omega})^{-1}\boldsymbol{\Omega}\} \\ &= p\log 2\pi + \log|\boldsymbol{\Sigma}_*| + \log|\boldsymbol{I}_p + \boldsymbol{\Omega}| + p \geq p\log 2\pi + \log|\boldsymbol{\Sigma}_*| + p,\end{aligned} \tag{A.4}$$

with equality if and only if M is the overspecified model. Recall that the candidate model making the loss function the smallest is either the true model or the underspecified model. This fact and equation (A.4) indicate that the loss function in the true model is the smallest among the all candidate models when $n \to \infty$. Consequently, Theorem 1 is proved.

B. Relationship between the best models selected by the AIC and CAIC

Let M_j $(j = 1, \ldots, m_M)$ be the jth candidate model with an $n \times k_j$ matrix of explanatory variables $\boldsymbol{X}_j$, and let AIC_j and CAIC_j be the AIC and CAIC of the model M_j, respectively, where m_M is the number of candidate models. Without loss of generality, we assume that M_1 denotes the best model selected by minimizing the AIC. Let $\mathcal{J}$ be the set of indexes, which is defined by $\mathcal{J} = \{j \in \{2, \ldots, m_M\} | \, k_j \geq k_1\}$, and let $q(k)$ be a function given by $q(k) = (p+k+1)\{2pk+p(p+1)\}/(n-p-k-1)$. Since $q(k)$ is a monotonically increasing function with respect to k, $q(k_j) \geq q(k_1)$ holds when $j \in \mathcal{J}$. Moreover, $\mathrm{AIC}_j - \mathrm{AIC}_1 > 0$ holds for all $j \in \{2, \ldots, m_M\}$, because M_1 is the best model selected by the AIC. By using the above two results and the relation between the AIC and CAIC in (12), the following inequality is derived:

$$\mathrm{CAIC}_j - \mathrm{CAIC}_1 = \mathrm{AIC}_j - \mathrm{AIC}_1 + q(k_j) - q(k_1) > 0, \ (j \in \mathcal{J}). \tag{B.1}$$

The result of (B.1) indicates that a model with more than k_1 explanatory variables will never be selected as the best model by the CAIC. Therefore, the number of explanatory variables in the best model selected by the CAIC is less than or equal to k_1.

C. Asymptotic equivalence of the EIC, EIC$_\mathrm{A}$, and TIC for the overspecified model

From Fujikoshi *et al.* (2005)[2] and Yanagihara (2006), when $m \to \infty$, $\hat{B}_{\mathrm{EIC}}$ and $\hat{B}_{\mathrm{EIC_A}}$ can be expanded as

$$\begin{aligned}
\hat{B}_{\mathrm{EIC}} &= 2pk + p(p+1) + \hat{\kappa}_4^{(1)} - np + O_p(n^{-1}), \\
\hat{B}_{\mathrm{EIC_A}} &= 2(k+p+1)\mathrm{tr}(\boldsymbol{G}) - \mathrm{tr}(\boldsymbol{G}^2) - 2\mathrm{tr}(\boldsymbol{G})^2 + \frac{1}{n}\sum_{i=1}^{n} \hat{r}_{\omega,i}^4 - np + O_p(n^{-1}),
\end{aligned}$$

where $\hat{\kappa}_4^{(1)}$ is given by (14), $\boldsymbol{G} = \hat{\boldsymbol{\Sigma}}_\omega \hat{\boldsymbol{\Sigma}}^{-1}$, and $\hat{r}_{\omega,i}^2 = (\boldsymbol{y}_i - \hat{\boldsymbol{\Xi}}_\omega' \boldsymbol{x}_i)' \hat{\boldsymbol{\Sigma}}^{-1} (\boldsymbol{y}_i - \hat{\boldsymbol{\Xi}}_\omega' \boldsymbol{x}_i)$. When the model M is overspecified, $\boldsymbol{G} = \boldsymbol{I}_p + O_p(n^{-1/2})$, $\hat{\kappa}_4^{(1)} = \kappa_4^{(1)} + O_p(n^{-1/2})$, and $n^{-1}\sum_{i=1}^n \hat{r}_{\omega,i}^2 = p(p+2) + \kappa_4^{(1)} + O_p(n^{-1/2})$ hold, where $\kappa_4^{(1)}$ is given in (4). Hence, $\hat{B}_{\mathrm{EIC}}$ and $\hat{B}_{\mathrm{EIC_A}}$ can be rewritten as follows when the model M is overspecified:

$$\hat{B}_{\mathrm{EIC}} = 2pk + p(p+1) + \kappa_4^{(1)} + O_p(n^{-1/2}), \quad \hat{B}_{\mathrm{EIC_A}} = 2pk + p(p+1) + \kappa_4^{(1)} + O_p(n^{-1/2}). \tag{C.1}$$

On the other hand, when the model M is overspecified, $\sum_{i=1}^n (1-h_i)(\hat{r}_i^2 - p) = O_p(n^{-1/2})$ holds because $\hat{r}_i^2 = \varepsilon_i'\varepsilon_i + O_p(n^{-1/2})$ and $1 - h_i = O(n^{-1})$ are satisfied. Then, $\hat{B}_{\mathrm{TIC}}$ can be expanded as

$$\hat{B}_{\mathrm{TIC}} = 2pk + p(p+1) + \kappa_4^{(1)} - np + O_p(n^{-1/2}). \tag{C.2}$$

Comparing (C.1) with (C.2) yields EIC = TIC + $O_p(n^{-1/2})$ and $\mathrm{EIC_A} = \mathrm{TIC} + O_p(n^{-1/2})$, when the model M is overspecified and $m \to \infty$.

D. Asymptotic equivalence of the CV criterion and the TIC

From Yanagihara (2006), the last term in (19) can be expanded as

[2] At the bottom of p. 240 of Fujikoshi *et al.* (2005), $-\mathrm{tr}(\hat{\boldsymbol{\Lambda}}^2)$ is missing in the equation $E[\hat{B}_A | \boldsymbol{Y}]$.

$$\sum_{i=1}^{n}(\boldsymbol{y}_i-\hat{\boldsymbol{\Xi}}'_{[-i]}\boldsymbol{x}_i)'\hat{\boldsymbol{\Sigma}}^{-1}_{[-i]}(\boldsymbol{y}_i-\hat{\boldsymbol{\Xi}}'_{[-i]}\boldsymbol{x}_i)=np+2pk+p(p+1)+\hat{\kappa}_4^{(1)}+2\sum_{i=1}^{n}(1-h_i)\hat{r}_i^2+O_p(n^{-1}), \tag{D.1}$$

where $\hat{r}_i^2$, $\hat{\kappa}_4^{(1)}$, and h_i are given by (13), (14), and (15), respectively. Moreover, by applying the Taylor expansion to the equation (20), we obtain

$$\sum_{i=1}^{n}\log|\hat{\boldsymbol{\Sigma}}_{[-i]}| = n\log|\hat{\boldsymbol{\Sigma}}| + \frac{1}{n}\sum_{i=1}^{n}\log\left(1-\frac{\hat{r}_i^2}{h_i}\right) = n\log|\hat{\boldsymbol{\Sigma}}| - \frac{1}{n}\sum_{i=1}^{n}\frac{\hat{r}_i^2}{h_i} + O_p(n^{-1}). \tag{D.2}$$

It follows from $h_i = 1 + O(n^{-1})$ and $\sum_{i=1}^{n}\hat{r}_i^2 = np$ that $n^{-1}\sum_{i=1}^{n}\hat{r}_i^2/h_i = n^{-1}\sum_{i=1}^{n}\hat{r}_i^2 + O_p(n^{-1}) = p + O_p(n^{-1})$. By combining the above result with (D.2), we obtain

$$\sum_{i=1}^{n}\log|\hat{\boldsymbol{\Sigma}}_{[-i]}| = n\log|\hat{\boldsymbol{\Sigma}}| - p + O_p(n^{-1}). \tag{D.3}$$

On the other hand, $n\log\{2\pi n/(n-1)\} = p + O(n^{-1})$ holds. Consequently, substituting this result and the equations (D.1) and (D.3) into (20), and comparing the obtained equation with the definition of the TIC in (16), yields CV = TIC + $O_p(n^{-1})$.

E. The proof of Theorem 2

Let IC be a general notation to indicate one of the nine information criteria considered in this paper. Notice that all the bias-correction terms in the information criteria, expect for the CV criterion, are $O_p(1)$, and CV = TIC + $O_p(n^{-1})$ holds. Using the same idea as in Appendix A, we have

$$\frac{1}{n}\mathrm{IC} \xrightarrow{p} p\log 2\pi + \log|\boldsymbol{\Sigma}_*| + \log|\boldsymbol{I}_p + \boldsymbol{\Omega}| + p \geq p\log 2\pi + \log|\boldsymbol{\Sigma}_*| + p,\ n\to\infty, \tag{E.1}$$

where $\boldsymbol{\Omega}$ is given by (5), and with equality if and only if M is the overspecified model. The equation (E.1) indicates that the underspecified models are never selected as the best model when $n\to\infty$. Hence, it is sufficient to consider the selection of the best model among the overspecified models.

Let ICA denote an information criterion proposed under normality (i.e., the AIC, CAIC, or MAIC), and let ICT denote an information criterion proposed under nonnormality (i.e., the TIC, EIC, $\mathrm{EIC_A}$, CV criterion, $\mathrm{AIC_J}$, or $\mathrm{CAIC_J}$). Moreover, let $\boldsymbol{V} = n^{-1/2}(\boldsymbol{\mathcal{E}}'\boldsymbol{\mathcal{E}} - n\boldsymbol{I}_p)$ and $\boldsymbol{Z} = (\boldsymbol{X}'\boldsymbol{X})^{-1/2}\boldsymbol{X}'\boldsymbol{\mathcal{E}}$, where $\boldsymbol{\mathcal{E}}$ is given by (4). Notice that $\boldsymbol{V} = O_p(1)$ and $\boldsymbol{Z} = O_p(1)$ hold under the assumptions in Theorem 2. Hence, when M is the overspecified model, we obtain

$$n\log|\hat{\boldsymbol{\Sigma}}| = n\log|\boldsymbol{\Sigma}_*| + \sqrt{n}\mathrm{tr}(\boldsymbol{V}) - \{\mathrm{tr}(\boldsymbol{V}^2)/2 + \mathrm{tr}(\boldsymbol{Z}'\boldsymbol{Z})\} + o_p(1).$$

On the other hand, from Fujikoshi *et al.* (2005), when M is the overspecified model, the bias B in (10) can be expanded as $B = 2pk + p(p+1) + \kappa_4^{(1)} + O(n^{-1})$, where $\kappa_4^{(1)}$ is given in (4). Recall that ICT = TIC + $o_p(1)$ and the bias of TIC is $O(n^{-1})$ when M is the overspecified model. From a simple calculation, we can see that $\hat{B}_{\mathrm{TIC}}$ is a consistent estimator of B when M is the overspecified model. By using the above results, when M is the overspecified model, the ICA and ICT are expressed as follows:

$$\begin{aligned} \mathrm{ICA} &= n\log|\boldsymbol{\Sigma}_*| + \sqrt{n}\mathrm{tr}(\boldsymbol{V}) - \{\mathrm{tr}(\boldsymbol{V}^2)/2 + \mathrm{tr}(\boldsymbol{Z}'\boldsymbol{Z})\} + np + 2pk + p(p+1) + o_p(1), \\ \mathrm{ICT} &= \mathrm{ICA} + \kappa_4^{(1)} + o_p(1). \end{aligned} \tag{E.2}$$

Let M_1 and M_2 be two different overspecified models, and let ICA_j and ICT_j be information criteria for M_j ($j = 1, 2$). From (E.2), we obtain

$$\mathrm{ICA}_1 - \mathrm{ICA}_2 = \mathrm{tr}(\boldsymbol{Z}_2'\boldsymbol{Z}_2 - \boldsymbol{Z}_1'\boldsymbol{Z}_1) + 2p(k_1 - k_2) + o_p(1), \quad \mathrm{ICT}_1 - \mathrm{ICT}_2 = \mathrm{ICA}_1 - \mathrm{ICA}_2 + o_p(1), \tag{E.3}$$

where $\boldsymbol{Z}_j$ is $\boldsymbol{Z}$ in M_j and k_j is the number of explanatory variables in M_j. The equations (E.3) indicate that the differences between two information criteria for the two different overspecified models are asymptotically equivalent. Consequently, all the information criteria choose the same model as the best one when $n \to \infty$.