REVSTAT – Statistical Journal
Volume 15, Number 3, July 2017, 299–332

A STUDY ON THE BIAS-CORRECTION EFFECT OF THE AIC FOR SELECTING VARIABLES IN NORMAL MULTIVARIATE LINEAR REGRESSION MODELS UNDER MODEL MISSPECIFICATION

Authors: Hirokazu Yanagihara
– Department of Mathematics, Graduate School of Science,
Hiroshima University, Hiroshima, Japan
yanagi@math.sci.hiroshima-u.ac.jp

Ken-ichi Kamo
– Department of Liberal Arts and Sciences,
Sapporo Medical University, Hokkaido, Japan
kamo@sapmed.ac.jp

Shinpei Imori
– Division of Mathematical Science, Graduate School of Engineering Science,
Osaka University, Osaka, Japan
imori@sigmath.es.osaka-u.ac.jp

Mariko Yamamura
– Department of Mathematics Education, Graduate School of Education,
Hiroshima University, Hiroshima, Japan
yamamura@hiroshima-u.ac.jp

Received: October 2013 Revised: March 2015 Accepted: July 2015

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 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 gives a better effect to a variable-selection method than that under the assumption that the distribution of the model is misspecified.

Key-Words:

- *AIC; bias-corrected AIC; KL information; loss function; nonnormality; risk function; variable selection.*

AMS Subject Classification:

- 62J05, 62F07.

1. INTRODUCTION

In the analysis of real data, it is important to determine which statistical model best fits the data, because there are many candidate models, and they each estimate different results, which may lead to different conclusions. One of the aims of model selection is to choose a statistical model having a high predictive accuracy. In order to achieve the aim, it is common that the risk function based on the Kullback–Leibler (KL) information [18] is used for assessing a goodness of fit of a statistical model. Then, the model making the risk function the smallest is regarded as the "best" model. Hence, in order to seek a statistical model having a high predictive accuracy, we have to compare with risk functions of each of candidate models. In practice, an estimate of the risk function is used, because the risk function involves unknown parameters. The most famous asymptotic unbiased estimator of the risk function is Akaike's information criterion (AIC; proposed by [1, 2]), which is derived 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 only asymptotically unbiased, the bias of the AIC to the risk function may be considerable when the sample size is not large enough and the number of parameters is large. Then, the AIC of a candidate model which is overspecified and has a large number of parameters tends to underestimate the risk function overly. This tendency causes that 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., [31]). Thus, the model with the most parameters tends to make AIC the smallest, and so the AIC often selects the model with the most parameters as the best model. Since this fault of AIC is due to the bias, it is frequently avoided 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., [4, 14, 16, 20]). Sugiura [24] and Hurvich and Tsai [12] proposed a bias-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, [5]), and many authors have verified by numerical experiments that a variable-selection method using the bias-corrected AIC performs better than that using the crude AIC.

A basic concept of bias correction is that we expect that an unbiased estimate 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 a crude estimator before 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 performances of 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. Performances of variable-selection methods using the nine criteria are examined by numerical experiments. We do not conduct numerical experiments under the large sample, because it has been confirmed theoretically that the variable-selection methods using the nine criteria select the same model as "best" when the sample size goes to ∞. Our result is that correcting the bias gives a greater positive effect to 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 use numerical experiments with small and moderate samples to compare performances of variable-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., [6], [21, chap. 9], [26, chap. 4]). Let $\boldsymbol{y}_1, ..., \boldsymbol{y}_n$ be p-dimensional independent random vectors of response variables, and let $\boldsymbol{x}_{\omega,1}, ..., \boldsymbol{x}_{\omega,n}$ 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 candidate explanatory variables, which is a subset of the full explanatory variables $\boldsymbol{x}_{\omega,i}$ $(i = 1, ..., n)$. Then, we consider the following normal MLRM as the candidate model:

$$M\colon\ \boldsymbol{y}_i \sim N_p(\boldsymbol{\Xi}'\boldsymbol{x}_i, \boldsymbol{\Sigma})\ , \qquad (i = 1, ..., n)\ , \tag{2.1}$$

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

Let $\boldsymbol{Y} = (\boldsymbol{y}_1, ..., \boldsymbol{y}_n)'$ be an $n \times p$ matrix of response variables, and let $\boldsymbol{X} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_n)'$ be an $n \times k$ matrix of candidate explanatory variables. Suppose that an $n \times k_\omega$ matrix of the full explanatory variables, $\boldsymbol{X}_\omega = (\boldsymbol{x}_{\omega,1}, ..., \boldsymbol{x}_{\omega,n})'$, is a column full-rank matrix, i.e., $\text{rank}(\boldsymbol{X}_\omega) = k_\omega < n$. 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, and $\lim_{n\to\infty} \boldsymbol{X}_\omega' \boldsymbol{X}_\omega / n$ exists and is positive definite. The matrix form of the candidate model (2.1) is given by

$$M\colon\ \boldsymbol{Y} \sim N_{n\times p}(\boldsymbol{X}\boldsymbol{\Xi}, \boldsymbol{\Sigma}\otimes\boldsymbol{I}_n)\ , \tag{2.2}$$

where $\boldsymbol{I}_n$ is an identity matrix of size n. Here, $\boldsymbol{A}\otimes\boldsymbol{B}$ denotes an the Kronecker product of an $m\times n$ matrix $\boldsymbol{A}$ and a $p\times q$ matrix $\boldsymbol{B}$, which is an $mp\times nq$ matrix defined by

$$\boldsymbol{A}\otimes\boldsymbol{B} = \begin{pmatrix} a_{11}\boldsymbol{B} & \cdots & a_{1n}\boldsymbol{B} \\ \vdots & \ddots & \vdots \\ a_{m1}\boldsymbol{B} & \cdots & a_{mn}\boldsymbol{B} \end{pmatrix},$$

where a_{ij} is the (i,j)-th element of $\boldsymbol{A}$ (see, e.g., [10, chap. 16.1]). Additionally, $\boldsymbol{Z} \sim N_{n\times p}(\boldsymbol{A}, \boldsymbol{B}\otimes\boldsymbol{C})$ denotes that an $n\times p$ random matrix $\boldsymbol{Z}$ is distributed according to the $n\times p$ matrix normal distribution with a mean matrix $E[\boldsymbol{Z}] = \boldsymbol{A}$ and a covariance matrix $Cov[(\boldsymbol{Z})] = \boldsymbol{B}\otimes\boldsymbol{C}$ (see, e.g., [26, p. 91, def. 3.3.1]), i.e., $\text{vec}(\boldsymbol{Z}) \sim N_{np}(\text{vec}(\boldsymbol{A}), \boldsymbol{B}\otimes\boldsymbol{C})$, where $\text{vec}(\boldsymbol{Z})$ is an operator that transforms a matrix to a vector by stacking the first to the last columns of $\boldsymbol{Z}$, i.e., $\text{vec}(\boldsymbol{Z}) = (\boldsymbol{z}_1', ..., \boldsymbol{z}_p')'$ when $\boldsymbol{Z} = (\boldsymbol{z}_1, ..., \boldsymbol{z}_p)$ (see, e.g., [10, chap. 16.2]). The following normal MLRM using the full explanatory variables is called the full model:

$$M_\omega\colon\ \boldsymbol{Y} \sim N_{n\times p}(\boldsymbol{X}_\omega\boldsymbol{\Xi}_\omega, \boldsymbol{\Sigma}_\omega\otimes\boldsymbol{I}_n)\ , \tag{2.3}$$

where $\boldsymbol{\Xi}_\omega$ and $\boldsymbol{\Sigma}_\omega$ denote a matrix of the unknown regression coefficients and a covariance matrix of the full model, respectively. 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

$$\begin{gathered} M_*\colon\ \boldsymbol{Y} = \boldsymbol{\Gamma}_* + \boldsymbol{\mathcal{E}}\boldsymbol{\Sigma}_*^{1/2}\ , \qquad \boldsymbol{\mathcal{E}} = (\boldsymbol{\varepsilon}_1, ..., \boldsymbol{\varepsilon}_n)'\ , \qquad \boldsymbol{\varepsilon}_1, ..., \boldsymbol{\varepsilon}_n \sim i.i.d.\ \boldsymbol{\varepsilon}\ , \\ E[\boldsymbol{\varepsilon}] = \boldsymbol{0}_p\ , \qquad Cov[\boldsymbol{\varepsilon}] = \boldsymbol{I}_p\ , \qquad E[\|\boldsymbol{\varepsilon}\|^4] = \kappa_4^{(1)} + p\,(p+2)\ , \end{gathered} \tag{2.4}$$

where $\boldsymbol{\Gamma}_*$ and $\boldsymbol{\Sigma}_*$ are the true mean and covariance matrices, respectively, $\boldsymbol{0}_p$ is a p-dimensional vector of zeros, and $\|\boldsymbol{a}\|$ is the Euclidean norm of the vector $\boldsymbol{a} = (a_1, ..., a_m)'$, i.e, $\|\boldsymbol{a}\| = (a_1^2 + \cdots + a_m^2)^{1/2}$. Here, $\kappa_4^{(1)}$ is called the multivariate kurtosis, which was proposed by [19].

In order to clarify assumptions for deriving 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_* is here called the true model. Although Fujikoshi and Satoh [8] used the same terminology, they divided the candidate models by whether a candidate model includes the true model. It 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

$$\boldsymbol{\Omega} = \frac{1}{n}\boldsymbol{\Sigma}_*^{-1/2}\boldsymbol{\Gamma}_*'(\boldsymbol{I}_n - \boldsymbol{P_X})\boldsymbol{\Gamma}_*\boldsymbol{\Sigma}_*^{-1/2}. \tag{2.5}$$

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

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 (2.2) is derived as

$$\begin{aligned}\ell(\boldsymbol{\Xi}, \boldsymbol{\Sigma}\,|\,\boldsymbol{Y}, \boldsymbol{X}) &= \sum_{i=1}^{n} \log f(\boldsymbol{y}_i\,|\,\boldsymbol{\Xi}'\boldsymbol{x}_i, \boldsymbol{\Sigma}) \\ &= -\frac{1}{2}\Big[np\log 2\pi + n\log|\boldsymbol{\Sigma}| \\ &\qquad + \operatorname{tr}\big\{\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\Xi})'(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\Xi})\big\}\Big].\end{aligned} \tag{2.6}$$

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 (ML) estimators of the unknown parameter matrices $\boldsymbol{\Xi}$ and $\boldsymbol{\Sigma}$ in the candidate model M are obtained as

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

In particular, $\boldsymbol{S}$ denotes a standardized $\hat{\boldsymbol{\Sigma}}$ defined by $\boldsymbol{S} = \boldsymbol{\Sigma}_*^{-1/2}\hat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}_*^{-1/2}$. Furthermore, $(\hat{\boldsymbol{\Xi}}_\omega, \hat{\boldsymbol{\Sigma}}_\omega, \boldsymbol{S}_\omega)$ denotes $(\hat{\boldsymbol{\Xi}}, \hat{\boldsymbol{\Sigma}}, \boldsymbol{S})$ in the full model M_ω in (2.3). By substituting $(\hat{\boldsymbol{\Xi}}, \hat{\boldsymbol{\Sigma}})$ into (2.6), the maximum log-likelihood of the candidate model M is derived as

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

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

$$\begin{aligned}\mathcal{L}(\boldsymbol{\Xi},\boldsymbol{\Sigma}\,|\,\boldsymbol{X}) &= E_*\big[-2\ell(\boldsymbol{\Xi},\boldsymbol{\Sigma}\,|\,\boldsymbol{Y},\boldsymbol{X})\big]\\ &= np\log 2\pi + n\log|\boldsymbol{\Sigma}|\\ &\quad+\ \mathrm{tr}\Big[\big\{n\boldsymbol{\Sigma}_* + (\boldsymbol{\Gamma}_*-\boldsymbol{X}\boldsymbol{\Xi})'\,(\boldsymbol{\Gamma}_*-\boldsymbol{X}\boldsymbol{\Xi})\big\}\,\boldsymbol{\Sigma}^{-1}\Big]\,,\end{aligned} \tag{2.7}$$

where E_* means the expectation under the true model M_* in (2.4). We define the loss function of the model M measured by the KL information as $\mathcal{L}(\hat{\boldsymbol{\Xi}},\hat{\boldsymbol{\Sigma}}\,|\,\boldsymbol{X})$. 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.,

$$R_{\mathrm{KL}} = E_*\big[\mathcal{L}(\hat{\boldsymbol{\Xi}},\hat{\boldsymbol{\Sigma}}\,|\,\boldsymbol{X})\big]\,. \tag{2.8}$$

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.1):

Theorem 2.1. *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 $E[\mathrm{tr}(\boldsymbol{S}_\omega^{-2})] = O(1)$.*

3. AIC AND BIAS-CORRECTED AICS IN NORMAL MLRMS

Although the risk function R_{KL} in (2.8) assesses the goodness of fit of the model, we cannot use R_{KL} directly because R_{KL} involves unknown parameters. Hence, in practice, an estimate of R_{KL} is needed to select the best model among the candidates. Although a naive estimator of R_{KL} is $-2\ell(\hat{\boldsymbol{\Xi}},\hat{\boldsymbol{\Sigma}}\,|\,\boldsymbol{Y},\boldsymbol{X})$, it has the following constant bias:

$$B\,=\,R_{\mathrm{KL}} - E_*\big[-2\ell(\hat{\boldsymbol{\Xi}},\hat{\boldsymbol{\Sigma}}\,|\,\boldsymbol{Y},\boldsymbol{X})\big]\,. \tag{3.1}$$

Thus, an information criterion for selecting the best model is defined by adding $\hat{B}$, an estimator of B, to $-2\ell(\hat{\boldsymbol{\Xi}},\hat{\boldsymbol{\Sigma}}\,|\,\boldsymbol{Y},\boldsymbol{X})$, i.e.,

$$\mathrm{IC}\,=\,-2\ell(\hat{\boldsymbol{\Xi}},\hat{\boldsymbol{\Sigma}}\,|\,\boldsymbol{Y},\boldsymbol{X}) + \hat{B}\,. \tag{3.2}$$

The information criterion is specified by the individual $\hat{B}$, because $\hat{B}$ changes based on assumptions of the model M and by an estimation method. As for such assumptions, the following two assumptions are considered:

(A1) The candidate model M in (2.2) is an overspecified model;

(A2) The distribution of the true model M_* in (2.4), called the true distribution, is the normal distribution, i.e., $\boldsymbol{\varepsilon} \sim N_p(\mathbf{0}_p, \boldsymbol{I}_p)$.

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. As for information criteria in the NMLR model other than the nine information criteria used in this paper, see [20, chap. 4].

Table 1: The order of the bias of each criterion.

	Criterion	Bias-Correction Method	Normality		Nonnormality	
			Under-specified	Over-specified	Under-specified	Over-specified
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 asymptotically 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$_{\mathrm{J}}$.

3.1. AIC

Under the assumption that the candidate model is completely specified, Akaike [1, 2] proposed AIC by estimating a bias of a negative twofold maximum log-likelihood to a risk function as twice the number of parameters. According to the general formula of AIC, $\hat{B}$ in (3.2) is $\hat{B}_{\mathrm{AIC}} = 2pk + p(p+1)$. Thus, the AIC in the model M is expressed as

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

From the assumption to derive an bias of AIC, the bias of the AIC in the model M to R_{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 assumptions A1 or A2 are violated (for details, see, e.g., [8, 9, 27]).

3.2. Corrected AIC

When assumptions A1, A2, and an additional assumption $n > p + k + 1$ are satisfied, Bedrick and Tsai [3] calculated the exact form of B as $\hat{B}_{\mathrm{CAIC}} = n(n+k)p/(n-k-p-1) - np$ and proposed the corrected AIC (CAIC)[1] by replacing $\hat{B}$ in (3.2) with $\hat{B}_{\mathrm{CAIC}}$ as

$$\begin{aligned} \mathrm{CAIC} &= np\log 2\pi + n\log|\hat{\boldsymbol{\Sigma}}| + \frac{n(n+k)p}{n-p-k-1} \\ &= \mathrm{AIC} + \frac{(p+k+1)\,(p+2k+1)p}{n-p-k-1}\,. \end{aligned}$$

The CAIC is an unbiased estimator of R_{KL} under assumptions A1 and A2, and is congruent with the bias-corrected AIC proposed by [12, 24] when $p = 1$. Additionally, extending the result of [5] to the multivariate case provides that the CAIC is a UMVUE of the risk function R_{KL} when assumptions A1 and A2 are satisfied simultaneously (for a short proof, see [34]). From the definition of the CAIC and its unbiasedness under assumptions A1 and A2, we can see that the AIC in an overspecified model underestimates R_{KL}, and the amount of the underestimation becomes large as k increases. This will cause the undesirable property of the AIC that the AIC has a tendency to overestimate the best model when the sample size is not large enough and the number of candidate models is large. The problem of the AIC can be avoided by using CAIC instead of the AIC, because 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 A.2). Because of CAIC $=$ AIC $+\, O(n^{-1})$, as in the case of the AIC, the order of the bias of the CAIC to R_{KL} becomes $O(1)$, i.e., the CAIC has a constant bias, when either of assumptions A1 or A2 are violated.

3.3. Modified AIC

When assumption A2 holds but assumption A1 does not hold, and $n > p + k + 1$, Fujikoshi and Satoh [8] estimated B by $\hat{B}_{\mathrm{MAIC}} = \hat{B}_{\mathrm{CAIC}} + 2k\,\mathrm{tr}(\boldsymbol{L}) - \mathrm{tr}(\boldsymbol{L})^2 - \mathrm{tr}(\boldsymbol{L}^2)$, where $\boldsymbol{L}$ is a $p\times p$ matrix defined by $\boldsymbol{L} = (n-k)\hat{\boldsymbol{\Sigma}}_\omega\hat{\boldsymbol{\Sigma}}^{-1}/$

[1] Although Bedrick and Tsai [3] used $\mathrm{AIC_c}$ as the abbreviated symbol, we use CAIC following the notation of [8].

$(n - k_\omega) - \boldsymbol{I}_p$, and proposed the modified AIC (MAIC) by replacing $\hat{B}$ in (3.2) with $\hat{B}_{\mathrm{MAIC}}$ as

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

The bias of the MAIC to R_{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 (see, [8]). However, the bias changes to $O(1)$, i.e., the MAIC also has constant bias, when assumption A2 is violated, because B_{AIC} depends on a nonnormality of the true model when assumption A2 is violated (see, e.g., [9, 27]).

3.4. Takeuchi's Information Criterion

Takeuchi [25] revaluated an asymptotic bias under model misspecification and proposed Takeuchi's information criterion (TIC) by estimating such an asymptotic bias with a moment-estimation method. According to the general formula of the TIC, a bias-correction term of the TIC in the model M can be calculated as $\hat{B}_{\mathrm{TIC}} = \hat{B}_{\mathrm{AIC}} + \hat{\kappa}_4^{(1)} + 2\sum_{i=1}^n (1-h_i)\,(\hat{r}_i^2 - p)$ (for details of the derivation, see [9]), where $\hat{r}_i$ is a squared standardized residual of the i-th individual, $\hat{\kappa}_4^{(1)}$ is an estimator of the multivariate kurtosis $\kappa_4^{(1)}$ in (2.4), and h_i is a constant, which are given by

$$\begin{aligned} \hat{r}_i^2 \;&=\; (\boldsymbol{y}_i - \hat{\boldsymbol{\Xi}}'\boldsymbol{x}_i)'\,\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{y}_i - \hat{\boldsymbol{\Xi}}'\boldsymbol{x}_i) \;, \\ \hat{\kappa}_4^{(1)} \;&=\; \frac{1}{n}\sum_{i=1}^n \hat{r}_i^4 - p(p+2) \;, \qquad h_i = 1 - \boldsymbol{x}_i'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i \;. \end{aligned} \tag{3.3}$$

Hence, the TIC in the model M is expressed as

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

When $\boldsymbol{y}_1, ..., \boldsymbol{y}_n$ 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, $\boldsymbol{y}_1, ..., \boldsymbol{y}_n$ are independent but not identically distributed. This leads to the less well-known fact that the bias of the TIC in the model M to R_{KL} is $O(n^{-1})$ when assumption A1 holds but assumption A2 does not, and becomes $O(1)$ when assumption A1 is violated (see, [9]). By conducting numerical experiments, many authors have verified a fact that although the TIC theoretically reduces the bias caused by violating normality, the TIC cannot reduce the bias successfully unless the sample size is huge (see, e.g., [9, 27]). This occurs because the TIC consists of an estimator for the multivariate kurtosis $\hat{\kappa}_4^{(1)}$.

Yanagihara [28] 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.

When $\boldsymbol{y}_1, ..., \boldsymbol{y}_n$ are independently and identically distributed, the bias of TIC can be reduced to $O(n^{-2})$ by using a formula in [35], which is a special case of those in [15] and [30]. However, as stated already, $\boldsymbol{y}_1, ..., \boldsymbol{y}_n$ are independent but not identically distributed in the case of the multivariate linear regression. Regrettably, we cannot correct the bias of TIC by using their formula.

3.5. Extended Information Criterion

The serious problem with TIC comes from the moment estimation of a bias. Ishiguro *et al.* [13] cleared this problem by using the bootstrap method for an estimation of the bias, and proposed the extended information criterion (EIC). Let $\boldsymbol{D}_b$ be an $n \times n$ matrix to express the b-th bootstrap resample of $\boldsymbol{Y}$ as

$$(3.4) \qquad \boldsymbol{D}_b = (\boldsymbol{d}_{b,1}, ..., \boldsymbol{d}_{b,n})' , \qquad \boldsymbol{d}_{b,1}, ..., \boldsymbol{d}_{b,n} \sim i.i.d.\ MN_n(1; n^{-1}\boldsymbol{1}_n) ,$$

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 [7], the b-th 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 ML estimator of $\boldsymbol{\Sigma}$ evaluated from $(\tilde{\boldsymbol{Y}}_b, \boldsymbol{X})$. From the general formula of EIC in [14], an estimator of the bias obtained from the bootstrap method with m repetitions is given by $\hat{B}_{\mathrm{EIC}} = m^{-1}\sum_{b=1}^m \mathrm{tr}\{\tilde{\boldsymbol{\Sigma}}_b^{-1}(\boldsymbol{Y} - \boldsymbol{P}_{\boldsymbol{X}}\tilde{\boldsymbol{Y}}_b)'(\boldsymbol{Y} - \boldsymbol{P}_{\boldsymbol{X}}\tilde{\boldsymbol{Y}}_b)\} - np$. Then, by using (3.2), the EIC in the model M is expressed as follows (see, [27]):

$$\mathrm{EIC} \;=\; np \log 2\pi + n \log|\hat{\boldsymbol{\Sigma}}| + \frac{1}{m}\sum_{b=1}^{m} \mathrm{tr}\Big\{\tilde{\boldsymbol{\Sigma}}_b^{-1}(\boldsymbol{Y} - \boldsymbol{P}_{\boldsymbol{X}}\tilde{\boldsymbol{Y}}_b)'(\boldsymbol{Y} - \boldsymbol{P}_{\boldsymbol{X}}\tilde{\boldsymbol{Y}}_b)\Big\} .$$

When $\boldsymbol{y}_1, ..., \boldsymbol{y}_n$ 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, $\boldsymbol{y}_1, ..., \boldsymbol{y}_n$ are independent but not identically distributed. Hence, the bias of EIC is $O(n^{-1})$ under 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, [27]). 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 A.3). Although the theoretical bias of the EIC has the same order as that of the TIC, the bias of the EIC tends to be smaller than that of the TIC (see, [27]) because the EIC does not directly use $\hat{\kappa}_4^{(1)}$ for estimating the bias.

3.6. Adjusted EIC

Fujikoshi *et al.* [9] proposed an adjusted version of the EIC in the model M by using a full-model-based resampling instead of a candidate-model-based resampling. We call this the adjusted EIC ($\mathrm{EIC_A}$). Let $\bar{\boldsymbol{Y}}_b$ be the b-th bootstrap resample of $\boldsymbol{Y}$ based on the full model M_ω given by $\bar{\boldsymbol{Y}}_b = \boldsymbol{X}_\omega \hat{\boldsymbol{\Xi}}_\omega + \boldsymbol{D}_b(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}_\omega})\boldsymbol{Y}$, where $\boldsymbol{D}_b$ is given by (3.4), and let $\bar{\boldsymbol{\Sigma}}_b$ be the ML estimator of $\boldsymbol{\Sigma}$ evaluated from $(\bar{\boldsymbol{Y}}_b, \boldsymbol{X})$. Then, $\hat{B}_{\mathrm{EIC_A}}$, which is an estimator of the bias obtained from a full-model-based bootstrap method with m repetitions, is given by replacing $\tilde{\boldsymbol{Y}}_b$ and $\tilde{\boldsymbol{\Sigma}}_b$ in $\hat{B}_{\mathrm{EIC}}$ with $\bar{\boldsymbol{Y}}_b$ and $\bar{\boldsymbol{\Sigma}}_b$. By using (3.2), the $\mathrm{EIC_A}$ in the model M is expressed as follows (see, [9]):

$$\mathrm{EIC_A} \;=\; np\log 2\pi + n\log|\hat{\boldsymbol{\Sigma}}| + \frac{1}{m}\sum_{b=1}^{m}\mathrm{tr}\Big\{\bar{\boldsymbol{\Sigma}}_b^{-1}(\boldsymbol{Y}-\boldsymbol{P}_{\boldsymbol{X}}\bar{\boldsymbol{Y}}_b)'(\boldsymbol{Y}-\boldsymbol{P}_{\boldsymbol{X}}\bar{\boldsymbol{Y}}_b)\Big\}\,.$$

The bias of the $\mathrm{EIC_A}$ to the risk function is always $O(n^{-1})$ (see, [9]). In particular, $\mathrm{EIC_A} = \mathrm{TIC} + O_p(n^{-1/2})$ holds when $m \to \infty$ and assumption A1 holds (the proof is given in Appendix A.3).

3.7. Cross-Validation Criterion

The cross-validation (CV) criterion proposed by [22] estimates a risk function directly, and it can be defined without an estimator of a bias of a negative twofold maximum log-likelihood to a risk function. We know that n repetitions of the calculations for the ML estimator of $(\boldsymbol{\Xi}, \boldsymbol{\Sigma})$ are needed for the CV criterion in the model M. However, Yoshimoto *et al.* [36] gave the formula to derive the CV criterion in the model M without the n repetitions as

$$(3.5)\qquad \mathrm{CV} \;=\; np\log\left(\frac{2\pi n}{n-1}\right) + n\log|\hat{\boldsymbol{\Sigma}}| + \sum_{i=1}^{n}\left\{\log\left(1-\frac{\hat{r}_i^2}{nh_i}\right) + \frac{(n-1)\,\hat{r}_i^2}{h_i(nh_i-\hat{r}_i^2)}\right\},$$

where $\hat{r}_i^2$ and h_i are given by (3.3). From [23], $\mathrm{CV} = \mathrm{TIC} + O_p(n^{-1})$ always holds if $\boldsymbol{y}_1, ..., \boldsymbol{y}_n$ are independently and identically distributed. In the case of multivariate linear regression, we can prove that $\mathrm{CV} = \mathrm{TIC} + O_p(n^{-1})$ always holds (the proof is given in Appendix A.4). 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.

Yanagihara and Fujisawa [30], and Yanagihara *et al.* [33, 35] proposed bias-corrected CV criteria, which are criteria correcting the bias of CV to the risk function, under general statistical models. It should be noted that their results cannot be applied to the case of multivariate linear regression because they proposed the bias-corrected CV under the assumption that $\boldsymbol{y}_1, ..., \boldsymbol{y}_n$ are independently and identically distributed.

3.8. Jackknife AIC

Yanagihara [32] proposed a bias-corrected AIC by using a jackknife method for estimating B and by adjusting such an estimator of B to become an unbiased estimator when assumptions A1 and A2 are satisfied simultaneously. We call this a jackknife AIC ($\mathrm{AIC_J}$). 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 (3.3), $Q(x;\lambda)$ is a function with respect to x and c is a positive constant, as follows:

$$Q(x;\lambda) = x\left(1-\frac{x}{n}\right)^{-\lambda}, \qquad c = \frac{(n+k)\,(n-k-p-2)}{(n-k-p-1)\sum_{i=1}^{n} h_i^{-1}}\,. \tag{3.6}$$

Then, by using (3.2), the $\mathrm{AIC_J}$ for the model M is (see [27]):

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

From [27], $\mathrm{AIC_J} = \mathrm{TIC} + O_p(n^{-1})$ always holds. Hence, like the TIC, the bias of the $\mathrm{AIC_J}$ is $O(n^{-1})$ under assumption A1, but it has a constant bias when assumption A1 is violated (see, [27]). On the other hand, when assumptions A1 and A2 are satisfied simultaneously, the $\mathrm{AIC_J}$ is an unbiased estimator of R_{KL}. Although the order of the bias of the $\mathrm{AIC_J}$ is the same as that of the bias of the TIC and EIC, it has been verified numerically that the bias of the $\mathrm{AIC_J}$ in an overspecified model becomes very small (see, [27]). Moreover, Yanagihara [27] showed a theoretical result that the absolute value of the bias of the $\mathrm{AIC_J}$ is smaller than those of either the TIC or EIC under assumption A1 when the $O(n^{-2})$ term of B is neglected.

3.9. Corrected Jackknife AIC

Since the bias of the $\mathrm{AIC_J}$ does not disappear in theory, Yanagihara *et al.* [32] proposed a corrected $\mathrm{AIC_J}$ ($\mathrm{CAIC_J}$) that corrects the bias while maintaining the desirable characteristic of keeping the bias very small numerically. 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 $\hat{r}_i^2$ and h_i are given by (3.3) and $Q(x;\lambda)$ is given by (3.6), c^{+} and a_j $(j=0,1)$ being positive constants given by

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

Here, $\Gamma(x)$ is the gamma function. Then, by using (3.2), the $\mathrm{CAIC_J}$ for the model M is (see [32])

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

When assumptions A1 and A2 are satisfied simultaneously, like the $\mathrm{AIC_J}$, the $\mathrm{CAIC_J}$ is an unbiased estimator of R_{KL}. Although, like the $\mathrm{AIC_J}$, the $\mathrm{CAIC_J}$ has constant bias when assumption A1 is violated, the $\mathrm{CAIC_J}$ reduces the bias of the $\mathrm{AIC_J}$ to $O(n^{-2})$ when assumption A1 holds (see, [32]). Moreover, Yanagihara *et al.* [32] showed a theoretical result under assumption A1 that $\mathrm{CAIC_J}$ reduces not only the bias of $\mathrm{AIC_J}$ but also the variance of $\mathrm{AIC_J}$ when we neglect the $O(n^{-2})$ terms.

4. NUMERICAL COMPARISON

In this section, we numerically compare performances of variable-selection methods using the nine information criteria described in the previous section. The best models selected by the nine information criteria are asymptotically equivalent, and in particular, an underspecified model is never selected as the best model when $n \to \infty$ (the proof is given in Appendix A.5). This 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 experiments using small and moderate sample sizes. We study performances of the nine information criteria by applying variable-selection methods to simulation data first, and by applying variable-selection methods to real data later.

4.1. A Simulation Study

4.1.1. Target Characteristics

In the simulation study, performances as an estimator of the risk function are studied at first, and performances as a model selector are studied later. In a numerical experiment to check performances as an estimator of the risk function, we compare the nine information criteria by the following three characteristics of an estimator:

(C-1) The mean of the information criterion $E[\mathrm{IC}]$;

(C-2) The standard deviation of the information criterion $\sqrt{Var[\mathrm{IC}]}$;

(C-3) The root-mean-square error (RMSE) of the information criterion $\sqrt{Var[\mathrm{IC}] + \left(E[\mathrm{IC}] - R_{\mathrm{KL}}\right)^2}$.

On the other hand, in a numerical experiment to check performances as a model selector, we compare the nine information criteria by the following two characteristics of a model selector:

(C-4) The probability of selecting the principle best model: the frequency with which the principle best model is selected as the best model;

(C-5) The prediction error (PE) of the best model: the expected loss function of the best model which is chosen by the information criterion; PE is defined as follows:

$$\mathrm{PE} = \frac{1}{n} E_* \left[\mathcal{L}(\hat{\boldsymbol{\Xi}}_{\mathrm{best}}, \hat{\boldsymbol{\Sigma}}_{\mathrm{best}} \,|\, \boldsymbol{X}_{\mathrm{best}})\right] ,$$

where $\mathcal{L}(\boldsymbol{\Xi}, \boldsymbol{\Sigma} \,|\, \boldsymbol{X})$ is the expected negative twofold log-likelihood function given by (2.7), and $(\hat{\boldsymbol{\Xi}}_{\mathrm{best}}, \hat{\boldsymbol{\Sigma}}_{\mathrm{best}}, \boldsymbol{X}_{\mathrm{best}})$ is $(\hat{\boldsymbol{\Xi}}, \hat{\boldsymbol{\Sigma}}, \boldsymbol{X})$ in 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 variable-selection method is one that can choose the best model for improving the predictive accuracy. Hence, the PE is a more important property than the probability of selecting the principle best model.

The expectations and probabilities in the simulation 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., $m = 1,000$.

4.1.2. Simulation Model

The model in [32] was used as the basic simulation model for generating data. We prepared the $k_\omega - 1$ candidate models M_j $(j = 1, ..., k_\omega - 1)$ with $p = 4$ and $n = 30$ or 100. First, we generated $z_1, ..., z_n \sim i.i.d.\ U(-1, 1)$. Using these $z_1, ..., z_n$, we constructed the $n \times k_\omega$ matrix of explanatory variables $\boldsymbol{X}_\omega$, whose (i, j)-th element is given by $\{(z_i - \bar{z})/s_z\}^{j-1}$ $(i = 1, ..., n;\ j = 1, ..., k_\omega)$, where $\bar{z}$ and s_z are the sample mean and standard deviation, respectively, of $z_1, ..., z_n$. 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 $(0.8)^{|i-j|}$ $(i = 1, ..., 4;\ j = 1, ..., 4)$. In this simulation study, we arranged 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 $(j = 1, ..., k_\omega - 1)$ 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.

For generating multivariate nonnormal data, the following data model introduced by [37] was used:

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

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, $r = (6/\chi_8^2)^{1/2}$ 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 $(23)^{1/2}/4$, $r = (6/\chi_8^2)^{1/2}$ 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., [17]. 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 inequality: model 1 < model 2 < model 3.

4.1.3. Results of Simulation Study

Figure 1 shows R_{KL} and the mean of each criterion in case 1. Since the shapes of the figures were almost the same, we omit the results for cases 2 to 6 to save space. The horizontal axis of the figures expresses numbers of candidate models, i.e., the subindex j of M_j. We see that the biases of the $\mathrm{AIC_J}$ and $\mathrm{CAIC_J}$ 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

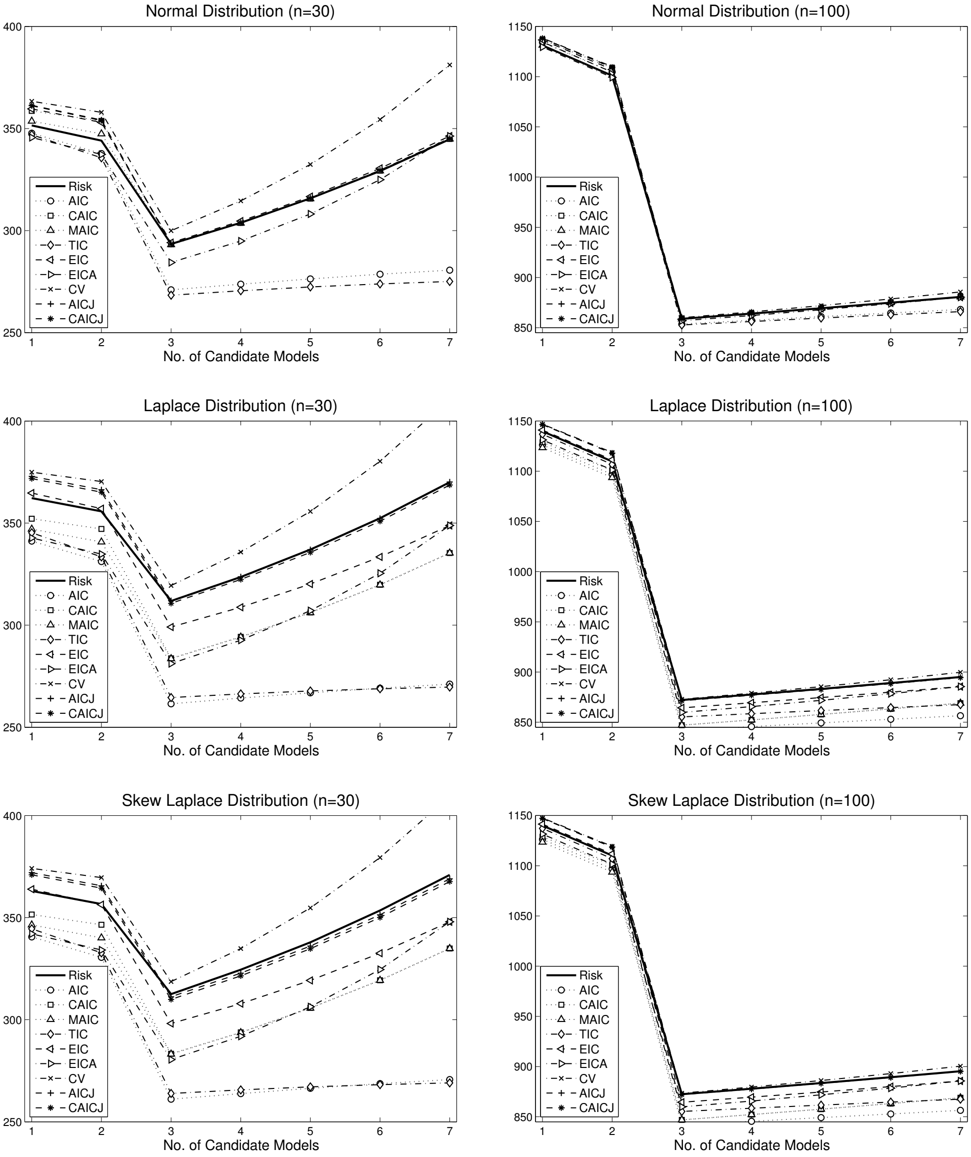

Figure 1: Risk function and the average of each criterion (Case 1).

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 increased, the biases of the AIC and TIC increased.

Tables 2 and 3 show, for case 1 and for each information criterion, the standard deviation and the RMSE. Since the tendencies were almost the same, to save space, we omit the results for M_2, M_4, M_5, and M_6, and in cases 2 to 6.

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

n	Dist.	Model	AIC	CAIC	MAIC	TIC	EIC	EIC$_\mathrm{A}$	CV	AIC$_\mathrm{J}$	CAIC$_\mathrm{J}$
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$_\mathrm{A}$	CV	AIC$_\mathrm{J}$	CAIC$_\mathrm{J}$
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

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 deviations 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 $\kappa_4^{(1)}$ was large.

Table 4: Probabilities of selecting the principle best model.

Case	n	Dist.	AIC	CAIC	MAIC	TIC	EIC	EIC$_\mathrm{A}$	CV	AIC$_\mathrm{J}$	CAIC$_\mathrm{J}$
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

Note: Bold and italic fonts indicate the highest and second highest probabilities of selecting the principle best model.

Table 5: Prediction errors of the best model.

Case	n	Dist.	AIC	CAIC	MAIC	TIC	EIC	EIC$_{\mathrm{A}}$	CV	AIC$_{\mathrm{J}}$	CAIC$_{\mathrm{J}}$
1	30	1	10.338	9.810	*9.795*	10.494	9.816	**9.790**	9.803	9.853	9.840
		2	11.014	10.427	*10.405*	11.304	10.469	**10.402**	10.481	10.527	10.512
		3	11.044	10.452	*10.432*	11.338	10.492	**10.424**	10.503	10.551	10.534
	100	1	8.619	8.603	*8.601*	8.624	8.603	**8.599**	*8.601*	8.605	8.604
		2	8.752	8.735	*8.733*	8.764	8.740	**8.729**	8.735	8.739	8.739
		3	8.756	8.740	*8.737*	8.768	8.745	**8.732**	8.739	8.743	8.742
2	30	1	10.661	10.020	*9.971*	10.793	10.017	**9.952**	9.986	10.075	10.059
		2	11.400	10.626	*10.558*	11.619	10.642	**10.516**	10.580	10.686	10.666
		3	11.429	10.638	*10.564*	11.648	10.662	**10.517**	10.585	10.702	10.677
	100	1	8.730	8.725	8.725	8.734	8.726	**8.724**	8.726	8.727	8.726
		2	8.871	8.865	**8.864**	8.880	8.871	*8.867*	8.870	8.871	8.871
		3	8.872	8.866	**8.865**	8.880	8.871	*8.868*	8.871	8.871	8.871
3	30	1	14.657	9.815	**9.802**	19.269	9.822	9.926	*9.809*	10.011	9.838
		2	16.626	*10.420*	**10.401**	21.720	10.462	10.556	10.468	10.762	10.500
		3	16.633	*10.441*	**10.426**	21.811	10.480	10.571	10.489	10.792	10.533
	100	1	8.623	*8.602*	**8.600**	8.639	8.603	8.604	**8.600**	8.605	8.604
		2	8.764	8.739	**8.737**	8.799	8.745	*8.738*	**8.737**	8.744	8.743
		3	8.769	8.742	**8.739**	8.809	8.749	8.741	*8.740*	8.748	8.747
4	30	1	15.614	10.020	**9.977**	19.816	10.017	10.175	*9.990*	10.343	10.058
		2	17.434	10.629	**10.569**	22.081	10.665	10.832	*10.592*	11.015	10.691
		3	17.851	10.634	**10.568**	22.273	10.659	10.840	*10.588*	10.985	10.679
	100	1	8.747	8.728	*8.727*	8.772	8.728	**8.725**	8.728	8.733	8.730
		2	8.886	*8.863*	*8.863*	8.933	8.870	**8.861**	8.868	8.876	8.872
		3	8.894	8.870	*8.868*	8.939	8.876	**8.867**	8.874	8.882	8.879
5	30	1	17.831	11.947	**11.806**	20.200	11.927	11.819	*11.814*	13.114	12.093
		2	19.990	12.810	*12.577*	22.557	12.754	12.632	**12.495**	14.022	12.880
		3	19.960	12.763	*12.541*	22.541	12.708	12.596	**12.468**	14.011	12.858
	100	1	8.918	8.881	*8.875*	8.963	8.880	**8.873**	8.879	8.914	8.889
		2	9.078	9.037	*9.031*	9.143	9.041	**9.026**	9.037	9.082	9.051
		3	9.080	9.036	*9.030*	9.142	9.041	**9.026**	9.039	9.082	9.052
6	30	1	18.115	**12.146**	12.156	20.263	12.172	*12.151*	12.219	13.432	12.303
		2	20.530	*13.073*	**13.048**	22.878	13.148	13.083	13.099	14.572	13.289
		3	20.610	13.101	**13.078**	22.954	13.133	*13.082*	13.129	14.609	13.305
	100	1	8.970	8.922	*8.914*	9.015	8.922	**8.910**	8.921	8.967	8.934
		2	9.124	9.070	*9.062*	9.183	9.072	**9.054**	9.068	9.123	9.086
		3	9.127	9.076	*9.066*	9.186	9.077	**9.059**	9.073	9.126	9.091

Note: Bold and italic fonts indicate the smallest and second smallest prediction errors of the best models.

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 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, 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 2.1. Hence, this result indicates that the bias correction in an 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 3 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 an increase of variances of ML estimators in the full model. 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 subsection, 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.

4.2. An Example Study

4.2.1. Target Characteristic

In the example study, we study performances of the variable-selection methods using nine information criteria by an estimator of the PE, which is derived as follows: We divide data to two samples, a calibration sample $(\boldsymbol{Y}_c, \boldsymbol{X}_c)$ with n_c and a validation sample $(\boldsymbol{Y}_v, \boldsymbol{X}_v)$ with n_v, randomly, and repeated such division

$N_r = 10,000$ times. In each repetition, we select the best model by minimizing each information criterion from a calibration sample $(\boldsymbol{Y}_c, \boldsymbol{X}_c)$, and record the selected best model. Let $\boldsymbol{X}_{c,\text{best}}$ and $\boldsymbol{X}_{v,\text{best}}$ be matrices of the selected best explanatory variables in $\boldsymbol{X}_c$ and $\boldsymbol{X}_v$, respectively. In order to assess an accuracy of prediction, we calculate as

$$\begin{aligned}\hat{\Delta} &= p \log 2\pi + \log |\hat{\boldsymbol{\Sigma}}_{c,\text{best}}| \\ &\quad + \frac{1}{n_v} \text{tr}\left\{ \left(\boldsymbol{Y}_v - \boldsymbol{X}_{v,\text{best}} \hat{\boldsymbol{\Xi}}_{c,\text{best}} \right)' \left(\boldsymbol{Y}_v - \boldsymbol{X}_{v,\text{best}} \hat{\boldsymbol{\Xi}}_{c,\text{best}} \right) \hat{\boldsymbol{\Sigma}}_{c,\text{best}}^{-1} \right\} .\end{aligned}$$

The average of $\hat{\Delta}$ across the N_r replications, $\widehat{\text{PE}}$, is regarded as an estimate of the prediction error of the best model.

4.2.2. Used Real Data

We used data of 37 kindergarten students ($n = 37$) in a low-socioeconomic-status area, which was provided by Dr. William D. Rohwer of the University of California at Berkeley to examine how well performance on a set of paired-associate (PA) tasks can predict performance on some measures of aptitude and achievement (see, [26, p. 217]). The data gives eight variables; score on the Peabody Picture Vocabulary Test (PPVT); score on the Raven Progressive Matrices Test (RPMT); score on a Student Achievement Test (SAT); performance on a 'named' PA task (N); performance on a 'still' PA task (S); performance on a 'named still' PA task (NS); performance on a 'named action' PA task (NA); performance on a 'sentence still' PA task (SS). We used PPVT, RPMT and SAT as the response variables ($p = 3$) and N, S, NS, NA and SS as explanatory variables. The number of explanatory variables in the full model is $k_\omega = 6$, because we always add a constant term to a regression. We compared with all 32 ($= 2^5$) candidate models by values of nine criteria. When all the samples were used for variable selection, the model having NS, NA, SS was selected as the best model by TIC, and the model having NA was selected as the best model by eight criteria other than TIC. Since we have conducted the numerical examination in the case of $n = 30$, we divided data into 30 and 7, i.e., $n_c = 30$ and $n_v = 7$.

4.2.3. Results of Example Study

Table 6 shows the probability of selecting the model and $\widehat{\text{PE}}$. In the table, "variables" shows used explanatory variables in the candidate model. The set of variables which is not listed in the table indicates that it was not chosen as the best model in every criterion. Superscript symbols * and ** denote the best models selected by each of criteria when the full data was used for variable selection.

Table 6: Results of real data.

Variables	Selection Probability (%)								
	AIC	CAIC	MAIC	TIC	EIC	EIC_A	CV	AIC_J	CAIC_J
N	0.21	0.50	0.51	0.11	0.42	0.39	0.46	0.43	0.48
NS	0.59	3.06	3.37	0.22	4.12	2.22	2.09	1.46	1.60
NA*	51.17	88.03	90.52	33.58	80.86	91.18	88.51	80.86	82.50
SS	0.75	2.44	2.65	0.15	1.70	2.94	2.03	1.43	1.51
N, NS	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01
N, NA	0.77	0.04	0.01	3.04	0.08	0.05	0.04	1.03	0.20
N, SS	0.12	0.02	0.00	0.19	0.03	0.00	0.04	0.18	0.16
S, NA	4.05	0.64	0.30	2.11	1.24	0.73	0.84	1.69	1.55
NS, NA	14.04	1.36	0.57	7.36	4.94	0.66	0.57	1.32	1.60
NS, SS	13.48	3.87	2.07	12.83	6.45	1.81	5.28	10.24	9.78
NA, SS	0.16	0.01	0.00	0.06	0.02	0.01	0.02	0.02	0.02
N, S, NA	0.85	0.00	0.00	2.97	0.00	0.00	0.00	0.18	0.03
N, NS, NA	0.00	0.00	0.00	0.07	0.00	0.00	0.00	0.00	0.00
N, NS, SS	0.23	0.00	0.00	0.42	0.00	0.00	0.00	0.07	0.02
N, NA, SS	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00
S, NS, NA	5.66	0.03	0.00	10.58	0.13	0.01	0.12	0.85	0.42
S, NA, SS	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
NS, NA, SS**	5.68	0.00	0.00	15.72	0.01	0.00	0.00	0.22	0.12
N, S, NS, NA	0.03	0.00	0.00	0.80	0.00	0.00	0.00	0.00	0.00
N, S, NS, SS	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00
N, S, NA, SS	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
N, NS, NA, SS	0.04	0.00	0.00	0.52	0.00	0.00	0.00	0.00	0.00
S, NS, NA, SS	2.05	0.00	0.00	8.37	0.00	0.00	0.00	0.01	0.00
N, S, NS, NA, SS	0.05	0.00	0.00	0.87	0.00	0.00	0.00	0.00	0.00
$\widehat{\text{PE}}$	22.396	22.043	*22.006*	22.582	22.108	**21.991**	22.036	22.135	22.113

Note: The set of variables which is not listed in the table indicates that it was not chosen as the best model in every criterion.
** denotes the best model selected by TIC, and * denotes the best model selected by criteria other than TIC, which were calculated from the full data.
Bold and italic fonts indicate the smallest and second smallest estimates of prediction errors of the best model.

Bold and italic fonts indicate the smallest and second smallest estimates of prediction errors of the best model. From the table, we can find the same tendency as the simulation study, i.e., EIC_A and MAIC were high performance criteria in the sense of improving the prediction. Moreover, we also find that results of variable selections using AIC and TIC tended to have larger variances than those of the other criteria.

5. CONCLUSIONS AND DISCUSSION

In this paper, we studied a bias-correction effect in the AIC to variable-selection methods for normal MLRMs, which are based on a minimization of an information criterion, by numerical examinations. Since all the variable-selection methods considered in this paper asymptotically choose the same model as the best model, we conducted numerical examinations using small and moderate sample sizes. Our results are summarized as follows:

- Except for the TIC, the performances of the variable-selection methods using the bias-corrected AIC were better than that using the original 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 an overspecified models.
- Theoretically, the bias of the $\mathrm{CAIC_J}$ becomes the smallest among all the criteria 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 biases of the risk functions for overspecified models are 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 $\mathrm{EIC_A}$, 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 an 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 an underspecified model. The $\mathrm{EIC_A}$ 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 $\mathrm{EIC_A}$ 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.

APPENDIX

A.1. Proof of Theorem 2.1

First, we show that the candidate model minimizing the risk function is either the true model or an underspecified model. Let $\boldsymbol{X}_1 = (\boldsymbol{X}, \boldsymbol{a})$ be an $n \times (k+1)$ matrix of explanatory variables in the model $M_1 \colon \boldsymbol{Y} \sim N_{n\times p}(\boldsymbol{X}_1\boldsymbol{\Xi}_1, \boldsymbol{\Sigma}_1 \otimes \boldsymbol{I}_n)$, 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 ML estimators of $\boldsymbol{\Xi}_1$ and $\boldsymbol{\Sigma}_1$, respectively. From the formula for the inverse matrix (see, e.g., [10, p. 424, cor. 18.2.10]), 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 $\boldsymbol{a}_{\mathrm{s}} = (\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{a}/\|(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{a}\|$. From the formulas for the determinant and the inverse matrix (see, e.g., [10, p. 416, cor. 18.1.3, and p. 424, thm. 18.2.8]), $|\hat{\boldsymbol{\Sigma}}_1|$ and $\hat{\boldsymbol{\Sigma}}_1^{-1}$ are rewritten as

$$\text{(A.1)} \qquad |\hat{\boldsymbol{\Sigma}}_1| = |\hat{\boldsymbol{\Sigma}}|\left(1 - \boldsymbol{a}_{\mathrm{s}}'\boldsymbol{P}_{\boldsymbol{U}}\boldsymbol{a}_{\mathrm{s}}\right),$$

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

where $\boldsymbol{U} = (\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{Y}\boldsymbol{\Sigma}_*^{-1/2}$. Since $\hat{\boldsymbol{\Sigma}}_1$ is positive definite and $\boldsymbol{P}_{\boldsymbol{U}}$ is positive semidefinite, we can see that $0 \le \boldsymbol{a}_{\mathrm{s}}'\boldsymbol{P}_{\boldsymbol{U}}\boldsymbol{a}_{\mathrm{s}} < 1$ with equality if and only if

$$\text{(A.3)} \qquad \boldsymbol{Y}'(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{a} = \boldsymbol{0}_p ,$$

because of

$$\boldsymbol{a}_{\mathrm{s}}'\boldsymbol{P}_{\boldsymbol{U}}\boldsymbol{a}_{\mathrm{s}} = 0 \iff (\boldsymbol{U}'\boldsymbol{U})^{-1/2}\boldsymbol{U}'\boldsymbol{a}_{\mathrm{s}} = \boldsymbol{0}_p \iff \boldsymbol{U}'\boldsymbol{a}_{\mathrm{s}} = \boldsymbol{0}_p .$$

The condition for equality means that a partial correlation between $\boldsymbol{Y}$ and $\boldsymbol{a}$ given $\boldsymbol{X}$ is exactly 0. Suppose that the model M is overspecified. Then, $\boldsymbol{U} = (\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{\mathcal{E}}$ holds, where $\boldsymbol{\mathcal{E}}$ is given by (2.4). It should be kept in mind that the standardized $\hat{\boldsymbol{\Sigma}}$ is expressed as $\boldsymbol{S} = \boldsymbol{U}'\boldsymbol{U}/n$. Notice that when M is an overspecified model,

$$n\hat{\boldsymbol{\Sigma}}_1 = \boldsymbol{\Sigma}_*^{1/2}\boldsymbol{\mathcal{E}}'(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}_1})\boldsymbol{\mathcal{E}}\boldsymbol{\Sigma}_*^{1/2},$$

$$(\boldsymbol{\Gamma}_* - \boldsymbol{X}_1\hat{\boldsymbol{\Xi}}_1)'(\boldsymbol{\Gamma}_* - \boldsymbol{X}_1\hat{\boldsymbol{\Xi}}_1) = \boldsymbol{\Sigma}_*^{1/2}\boldsymbol{\mathcal{E}}'\boldsymbol{P}_{\boldsymbol{X}_1}\boldsymbol{\mathcal{E}}\boldsymbol{\Sigma}_*^{1/2}.$$

Therefore, by using the above equations and (2.7), the loss function under M_1 can be simplified as

$$\begin{aligned} \mathcal{L}(\hat{\boldsymbol{\Xi}}_1, \hat{\boldsymbol{\Sigma}}_1 \,|\, \boldsymbol{X}_1) &= \\ \text{(A.4)} \qquad &= np\log 2\pi + n\log|\hat{\boldsymbol{\Sigma}}_1| + \mathrm{tr}\left\{\hat{\boldsymbol{\Sigma}}_1^{-1}\boldsymbol{\Sigma}_*^{1/2}(n\boldsymbol{I}_p + \boldsymbol{\mathcal{E}}'\boldsymbol{P}_{\boldsymbol{X}_1}\boldsymbol{\mathcal{E}})\boldsymbol{\Sigma}_*^{1/2}\right\} \\ &= np(\log 2\pi - 1) + n\log|\hat{\boldsymbol{\Sigma}}_1| + \mathrm{tr}\left\{\hat{\boldsymbol{\Sigma}}_1^{-1}\boldsymbol{\Sigma}_*^{1/2}(n\boldsymbol{I}_p + \boldsymbol{\mathcal{E}}'\boldsymbol{\mathcal{E}})\boldsymbol{\Sigma}_*^{1/2}\right\} . \end{aligned}$$

Substituting (A.1) and (A.2) into (A.4) yields

$$\mathcal{L}(\hat{\boldsymbol{\Xi}}_1, \hat{\boldsymbol{\Sigma}}_1 | \boldsymbol{X}_1) = \mathcal{L}(\hat{\boldsymbol{\Xi}}, \hat{\boldsymbol{\Sigma}} | \boldsymbol{X}) + \mathcal{L}_1 + \mathcal{L}_2 + \mathcal{L}_3 ,$$

where

$$\begin{aligned} \mathcal{L}_1 &= n \left\{ \log(1 - \boldsymbol{a}_{\mathrm{s}}' \boldsymbol{P}_{\boldsymbol{U}} \boldsymbol{a}_{\mathrm{s}}) + \frac{\boldsymbol{a}_{\mathrm{s}}' \boldsymbol{P}_{\boldsymbol{U}} \boldsymbol{a}_{\mathrm{s}}}{1 - \boldsymbol{a}_{\mathrm{s}}' \boldsymbol{P}_{\boldsymbol{U}} \boldsymbol{a}_{\mathrm{s}}} \right\} , \\ \mathcal{L}_2 &= \frac{n}{1 - \boldsymbol{a}_{\mathrm{s}}' \boldsymbol{P}_{\boldsymbol{U}} \boldsymbol{a}_{\mathrm{s}}} \boldsymbol{a}_{\mathrm{s}}' \boldsymbol{U} (\boldsymbol{U}'\boldsymbol{U})^{-1} \boldsymbol{\mathcal{E}}' \boldsymbol{P}_{\boldsymbol{X}} \boldsymbol{\mathcal{E}} (\boldsymbol{U}'\boldsymbol{U})^{-1} \boldsymbol{U}' \boldsymbol{a}_{\mathrm{s}} , \\ \mathcal{L}_3 &= \frac{n^2}{1 - \boldsymbol{a}_{\mathrm{s}}' \boldsymbol{P}_{\boldsymbol{U}} \boldsymbol{a}_{\mathrm{s}}} \boldsymbol{a}_{\mathrm{s}}' \boldsymbol{U} (\boldsymbol{U}'\boldsymbol{U})^{-2} \boldsymbol{U}' \boldsymbol{a}_{\mathrm{s}} . \end{aligned}$$

Notice that $\log(1 - x) + x/(1 - x) \geq 0$ when $x \in [0, 1)$ with equality if and only if $x = 0$. Hence, $\mathcal{L}_1 \geq 0$ holds with equality if and only if (A.3) holds. Moreover, we have $\mathcal{L}_2 \geq 0$ with equality if (A.3) because $\boldsymbol{P}_{\boldsymbol{X}}$ is positive semidefinite. These equations imply that

$$\mathcal{L}(\hat{\boldsymbol{\Xi}}_1, \hat{\boldsymbol{\Sigma}}_1 | \boldsymbol{X}_1) \geq \mathcal{L}(\hat{\boldsymbol{\Xi}}, \hat{\boldsymbol{\Sigma}} | \boldsymbol{X}) + \mathcal{L}_3 , \tag{A.5}$$

with equality if and only if (A.3) holds. A singular value decomposition of $\boldsymbol{U}$ (see, e.g., [10, chap. 21.12]) implies that $\boldsymbol{U}(\boldsymbol{U}'\boldsymbol{U})^{-2}\boldsymbol{U}' = \boldsymbol{H}\boldsymbol{D}^{-1}\boldsymbol{H}'$, where $\boldsymbol{D}$ is a $p \times p$ diagonal matrix whose diagonal elements are eigenvalues of $\boldsymbol{U}'\boldsymbol{U}$, and $\boldsymbol{H}$ is an $n \times p$ matrix satisfying $\boldsymbol{H}'\boldsymbol{H} = \boldsymbol{I}_p$ and $\boldsymbol{H}\boldsymbol{H}' = \boldsymbol{P}_{\boldsymbol{U}}$. Moreover, $\lambda_{\max}(\boldsymbol{A}) \leq \mathrm{tr}(\boldsymbol{A})$ holds for any positive semidefinite matrix $\boldsymbol{A}$, where $\lambda_{\max}(\boldsymbol{A})$ is maximum eigenvalue of $\boldsymbol{A}$. Using these results and the equation $(1 - \boldsymbol{a}_{\mathrm{s}}' \boldsymbol{P}_{\boldsymbol{U}} \boldsymbol{a}_{\mathrm{s}})^{-1} \geq 1$ yields

$$\begin{aligned} \mathcal{L}_3 &\geq n^2 \boldsymbol{a}_{\mathrm{s}}' \boldsymbol{U} (\boldsymbol{U}'\boldsymbol{U})^{-2} \boldsymbol{U}' \boldsymbol{a}_{\mathrm{s}} = n^2 \boldsymbol{a}_{\mathrm{s}}' \boldsymbol{H} \boldsymbol{D}^{-1} \boldsymbol{H}' \boldsymbol{a}_{\mathrm{s}} \\ &\geq \frac{n^2 \boldsymbol{a}_{\mathrm{s}}' \boldsymbol{H}\boldsymbol{H}' \boldsymbol{a}_{\mathrm{s}}}{\lambda_{\max}(\boldsymbol{U}'\boldsymbol{U})} = \frac{n \boldsymbol{a}_{\mathrm{s}}' \boldsymbol{P}_{\boldsymbol{U}} \boldsymbol{a}_{\mathrm{s}}}{\lambda_{\max}(\boldsymbol{S})} \geq \frac{n \boldsymbol{a}_{\mathrm{s}}' \boldsymbol{P}_{\boldsymbol{U}} \boldsymbol{a}_{\mathrm{s}}}{\mathrm{tr}(\boldsymbol{S})} . \end{aligned} \tag{A.6}$$

Let $\boldsymbol{U}_0 = (\boldsymbol{I}_n - \boldsymbol{J}_n)\boldsymbol{\mathcal{E}}$ and $\boldsymbol{S}_0 = \boldsymbol{U}_0'\boldsymbol{U}_0/n$, where $\boldsymbol{J}_n = \boldsymbol{1}_n\boldsymbol{1}_n'/n$. Since we assume that $\boldsymbol{X}$ always has $\boldsymbol{1}_n$ as a column vector, $(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}})(\boldsymbol{I}_n - \boldsymbol{J}_n) = \boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}}$. This implies that $\boldsymbol{a}_{\mathrm{s}}'\boldsymbol{U} = \boldsymbol{a}_{\mathrm{s}}'\boldsymbol{U}_0$. Moreover, since $\boldsymbol{P}_{\boldsymbol{X}} - \boldsymbol{J}_n$ is a symmetric idempotent matrix with $\mathrm{tr}(\boldsymbol{P}_{\boldsymbol{X}} - \boldsymbol{J}_n) = k - 1$, it is rewritten as $\boldsymbol{P}_{\boldsymbol{X}} - \boldsymbol{J}_n = \boldsymbol{Q}\boldsymbol{Q}'$, where $\boldsymbol{Q}$ is an $n \times (k-1)$ matrix satisfying $\boldsymbol{Q}'\boldsymbol{Q} = \boldsymbol{I}_{k-1}$. Hence, from the formula for the inverse matrix (see, e.g., [10, p. 424, thm. 18.2.8]), we have

$$\begin{aligned} (\boldsymbol{U}'\boldsymbol{U})^{-1} &= \left\{ \boldsymbol{\mathcal{E}}'(\boldsymbol{I}_n - \boldsymbol{J}_n)\boldsymbol{\mathcal{E}} - \boldsymbol{\mathcal{E}}'(\boldsymbol{P}_{\boldsymbol{X}} - \boldsymbol{J}_n)\boldsymbol{\mathcal{E}} \right\}^{-1} \\ &= (\boldsymbol{U}_0'\boldsymbol{U}_0)^{-1} \left\{ \boldsymbol{U}_0'\boldsymbol{U}_0 + \boldsymbol{U}_0'\boldsymbol{Q}(\boldsymbol{I}_{k-1} - \boldsymbol{Q}'\boldsymbol{P}_{\boldsymbol{U}_0}\boldsymbol{Q})^{-1}\boldsymbol{Q}'\boldsymbol{U}_0 \right\} (\boldsymbol{U}_0'\boldsymbol{U}_0)^{-1} . \end{aligned}$$

This implies that for any p-dimensional vector $\boldsymbol{b}$

$$\boldsymbol{b}'(\boldsymbol{U}'\boldsymbol{U})^{-1}\boldsymbol{b} \geq \boldsymbol{b}'(\boldsymbol{U}_0'\boldsymbol{U}_0)^{-1}\boldsymbol{b} . \tag{A.7}$$

Moreover, $\mathrm{tr}(\boldsymbol{S}) \leq \mathrm{tr}(\boldsymbol{S}_0)$ holds, because $n\boldsymbol{S}_0 = n\boldsymbol{S} + \boldsymbol{\mathcal{E}}'(\boldsymbol{P}_{\boldsymbol{X}} - \boldsymbol{J}_n)\boldsymbol{\mathcal{E}}$. Using these results and equation (A.6) yields

$$\begin{aligned} \mathcal{L}_3 &\geq \frac{n \boldsymbol{a}_{\mathrm{s}}' \boldsymbol{P}_{\boldsymbol{U}} \boldsymbol{a}_{\mathrm{s}}}{\mathrm{tr}(\boldsymbol{S})} = \frac{n \boldsymbol{a}_{\mathrm{s}}' \boldsymbol{U}_0 (\boldsymbol{U}'\boldsymbol{U})^{-1} \boldsymbol{U}_0' \boldsymbol{a}_{\mathrm{s}}}{\mathrm{tr}(\boldsymbol{S})} \\ &\geq \frac{n \boldsymbol{a}_{\mathrm{s}}' \boldsymbol{U}_0 (\boldsymbol{U}_0'\boldsymbol{U}_0)^{-1} \boldsymbol{U}_0' \boldsymbol{a}_{\mathrm{s}}}{\mathrm{tr}(\boldsymbol{S}_0)} = n \boldsymbol{a}_{\mathrm{s}}' \boldsymbol{W}_0 \boldsymbol{a}_{\mathrm{s}} , \end{aligned} \tag{A.8}$$

where $\boldsymbol{W}_0 = \boldsymbol{P}_{\boldsymbol{U}_0}/\operatorname{tr}(\boldsymbol{S}_0)$. Using (A.5), (A.6) and (A.8) yields

$$\text{(A.9)} \qquad E_*\big[\mathcal{L}(\hat{\boldsymbol{\Xi}}_1, \hat{\boldsymbol{\Sigma}}_1 \,|\, \boldsymbol{X}_1)\big] - E_*\big[\mathcal{L}(\hat{\boldsymbol{\Xi}}, \hat{\boldsymbol{\Sigma}} \,|\, \boldsymbol{X})\big] \;\geq\; n E_*\big[\boldsymbol{a}_{\mathrm{s}}' \boldsymbol{W}_0 \, \boldsymbol{a}_{\mathrm{s}}\big] \,,$$

Hence, in order to evaluate the right side of equation (A.9), we have to evaluate the expectation of $\boldsymbol{W}_0$. This expectation can be calculated in the same way as in the proof of Lemma 7 in [29]. Notice that

$$\text{(A.10)} \qquad \frac{p}{\operatorname{tr}(\boldsymbol{S}_0)} = \operatorname{tr}(\boldsymbol{W}_0) = \sum_{a=1}^{n} w_{aa} \,, \qquad 0 = \boldsymbol{1}_n' \boldsymbol{W}_0 \boldsymbol{1}_n = \sum_{a=1}^{n} w_{aa} + \sum_{a=1}^{n} \sum_{b \neq a}^{n} w_{aa} \,,$$

where w_{ab} is the (a,b)-th element of $\boldsymbol{W}_0$. Since $w_{ab} = (\boldsymbol{\varepsilon}_a - \bar{\boldsymbol{\varepsilon}})'(\boldsymbol{U}_0 \boldsymbol{U}_0)^{-1}(\boldsymbol{\varepsilon}_b - \bar{\boldsymbol{\varepsilon}})/\operatorname{tr}(\boldsymbol{S}_0)$, where $\bar{\boldsymbol{\varepsilon}}$ is the sample mean of $\boldsymbol{\varepsilon}_1, ..., \boldsymbol{\varepsilon}_n$, i.e., $\bar{\boldsymbol{\varepsilon}} = n^{-1} \sum_{i=1}^{n} \boldsymbol{\varepsilon}_i$, we can see that the diagonal elements of $\boldsymbol{W}_0$ are identically distributed and the upper (or lower) off-diagonal elements of $\boldsymbol{W}_0$ are also identically distributed. These results and the equations in (A.10) imply that

$$n E_*[w_{aa}] = p\alpha \,, \qquad n E_*[w_{aa}] + n(n-1) E_*[w_{ab}] = 0 \quad (a \neq b) \,,$$

where $\alpha = E_*[1/\operatorname{tr}(\boldsymbol{S}_0)]$. Thus, $E_*[\boldsymbol{W}_0] = p\alpha(\boldsymbol{I}_n - \boldsymbol{J}_n)/(n-1)$ is derived. From the Jensen's inequality, we have $\alpha \geq 1/E_*[\operatorname{tr}(\boldsymbol{S}_0)] = n\{(n-1)p\}^{-1}$. Consequently, it follows from these results and equation (A.9) that

$$\text{(A.11)} \qquad E_*\big[\mathcal{L}(\hat{\boldsymbol{\Xi}}_1, \hat{\boldsymbol{\Sigma}}_1 \,|\, \boldsymbol{X}_1)\big] - E_*\big[\mathcal{L}(\hat{\boldsymbol{\Xi}}, \hat{\boldsymbol{\Sigma}} \,|\, \boldsymbol{X})\big] \;\geq\; \left(\frac{n}{n-1}\right)^2 \;>\; 0 \,.$$

This means that the risk function becomes large when a new explanatory variable is added to an overspecified model. Since the overspecified model that has the smallest number of explanatory variables is the true model, the candidate model that minimizes the risk function is either the true model or an underspecified model.

Next, we show that the candidate model that minimizes the risk function is the true model when $n \to \infty$. From (A.11), we can see that overspecified models except the true model do not minimize the risk function even when $n \to \infty$, because the right side of (A.11) converges to a positive value. Hence, we only have to show the proof when the candidate model is an underspecified model. Suppose that the model M is underspecified. Let $\boldsymbol{\Pi} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Gamma}_*$ and $\boldsymbol{\Psi} = \boldsymbol{\Sigma}_*^{1/2}(\boldsymbol{I}_p + \boldsymbol{\Omega})\boldsymbol{\Sigma}_*^{1/2}$, where $\boldsymbol{\Omega}$ is a matrix of noncentrality parameter given by (2.5). By minimizing $\mathcal{L}(\boldsymbol{\Xi}, \boldsymbol{\Sigma} \,|\, \boldsymbol{X})$ in (2.7), or equivalently solving the equations $\partial\mathcal{L}(\boldsymbol{\Xi}, \boldsymbol{\Sigma} \,|\, \boldsymbol{X})/\partial\boldsymbol{\Xi} = \boldsymbol{O}_{k,p}$ and $\partial\mathcal{L}(\boldsymbol{\Xi}, \boldsymbol{\Sigma} \,|\, \boldsymbol{X})/\partial\boldsymbol{\Sigma} = \boldsymbol{O}_{p,p}$, we can see that $(\boldsymbol{\Pi}, \boldsymbol{\Psi})$ makes $\mathcal{L}(\boldsymbol{\Xi}, \boldsymbol{\Sigma} \,|\, \boldsymbol{X})$ the smallest. This implies that

$$\text{(A.12)} \quad \mathcal{L}(\hat{\boldsymbol{\Xi}}, \hat{\boldsymbol{\Sigma}} | \boldsymbol{X}) \;\geq\; \mathcal{L}(\boldsymbol{\Pi}, \boldsymbol{\Psi} | \boldsymbol{X}) \;=\; np(\log 2\pi + 1) + n \log|\boldsymbol{\Sigma}_*| + n \log|\boldsymbol{I}_p + \boldsymbol{\Omega}| \,.$$

On the other hand, since the full model is overspecified and $(\hat{\boldsymbol{\Xi}}_\omega, \hat{\boldsymbol{\Sigma}}_\omega)$ makes the negative twofold log-likelihood function of the full model the smallest, it follows from equation (A.4) that

$$
\begin{aligned}
\mathcal{L}(\hat{\boldsymbol{\Xi}}_\omega, \hat{\boldsymbol{\Sigma}}_\omega | \boldsymbol{X}_\omega) &= -2\ell(\hat{\boldsymbol{\Xi}}_\omega, \hat{\boldsymbol{\Sigma}}_\omega | \boldsymbol{Y}, \boldsymbol{X}_\omega) - np + \mathrm{tr}\big\{(n\boldsymbol{I}_p + \boldsymbol{\mathcal{E}}'\boldsymbol{P}_{\boldsymbol{X}_\omega}\boldsymbol{\mathcal{E}})\,\boldsymbol{S}_\omega^{-1}\big\} \\
&\le -2\ell(\boldsymbol{\Pi}_\omega, \boldsymbol{\Sigma}_* | \boldsymbol{Y}, \boldsymbol{X}_\omega) - np + \mathrm{tr}\big\{(n\boldsymbol{I}_p + \boldsymbol{\mathcal{E}}'\boldsymbol{P}_{\boldsymbol{X}_\omega}\boldsymbol{\mathcal{E}})\,\boldsymbol{S}_\omega^{-1}\big\} \\
&= np(\log 2\pi - 1) + n\log|\boldsymbol{\Sigma}_*| \\
&\quad + \mathrm{tr}(\boldsymbol{\mathcal{E}}'\boldsymbol{\mathcal{E}}) + \mathrm{tr}\big\{(n\boldsymbol{I}_p + \boldsymbol{\mathcal{E}}'\boldsymbol{P}_{\boldsymbol{X}_\omega}\boldsymbol{\mathcal{E}})\,\boldsymbol{S}_\omega^{-1}\big\}\,,
\end{aligned}
\tag{A.13}
$$

where $\boldsymbol{\Pi}_\omega = (\boldsymbol{X}_\omega'\boldsymbol{X}_\omega)^{-1}\boldsymbol{X}_\omega'\boldsymbol{\Gamma}_*$. Using the equations in (A.12) and (A.13) yields

$$
\begin{aligned}
E_*\big[\mathcal{L}(\hat{\boldsymbol{\Xi}}, \hat{\boldsymbol{\Sigma}} | \boldsymbol{X})\big] - E_*\big[\mathcal{L}(\hat{\boldsymbol{\Xi}}_\omega, \hat{\boldsymbol{\Sigma}}_\omega | \boldsymbol{X}_\omega)\big] &\ge \\
&\ge n\log|\boldsymbol{I}_p + \boldsymbol{\Omega}| + 2np \\
&\quad - E_*\big[\mathrm{tr}(\boldsymbol{\mathcal{E}}'\boldsymbol{\mathcal{E}})\big] - E_*\big[\mathrm{tr}\big\{(n\boldsymbol{I}_p + \boldsymbol{\mathcal{E}}'\boldsymbol{P}_{\boldsymbol{X}_\omega}\boldsymbol{\mathcal{E}})\,\boldsymbol{S}_\omega^{-1}\big\}\big] \\
&= n\log|\boldsymbol{I}_p + \boldsymbol{\Omega}| + np - E_*\big[\mathrm{tr}\big\{(n\boldsymbol{I}_p + \boldsymbol{\mathcal{E}}'\boldsymbol{P}_{\boldsymbol{X}_\omega}\boldsymbol{\mathcal{E}})\,\boldsymbol{S}_\omega^{-1}\big\}\big]\,.
\end{aligned}
\tag{A.14}
$$

Hence, in order to evaluate the right side of equation (A.14), we have to evaluate $nE_*[\mathrm{tr}(\boldsymbol{S}_\omega^{-1})]$ and $E_*[\boldsymbol{\mathcal{E}}'\boldsymbol{P}_{\boldsymbol{X}_\omega}\boldsymbol{\mathcal{E}}\boldsymbol{S}_\omega^{-1}]$. In the same way as in the proof of Lemma 1 in [11], $\boldsymbol{S}_\omega^{-1}$ can be expressed as

$$
\boldsymbol{S}_\omega^{-1} = \boldsymbol{I}_p - \frac{1}{\sqrt{n}}\,\boldsymbol{S}_\omega^{-1}\boldsymbol{V}_\omega\,,
$$

where $\boldsymbol{V}_\omega = n^{1/2}(\boldsymbol{S}_\omega - \boldsymbol{I}_p)$. By using the Hölder's inequality, we have

$$
\begin{aligned}
E_*\big[\mathrm{tr}(\boldsymbol{S}_\omega^{-1})\big] &\le p + \frac{1}{\sqrt{n}}\,E_*\big[\big|\mathrm{tr}(\boldsymbol{S}_\omega^{-1}\boldsymbol{V}_\omega)\big|\big] \\
&\le p + \sqrt{\frac{1}{n}\,E_*\big[\mathrm{tr}(\boldsymbol{S}_\omega^{-2})\big]\,E_*\big[\mathrm{tr}(\boldsymbol{V}_\omega^2)\big]}\,, \\
E_*\big[\mathrm{tr}(\boldsymbol{\mathcal{E}}'\boldsymbol{P}_{\boldsymbol{X}_\omega}\boldsymbol{\mathcal{E}}\boldsymbol{S}_\omega^{-1})\big] &\le \sqrt{E_*\big[\mathrm{tr}(\boldsymbol{S}_\omega^{-2})\big]\,E_*\big[\mathrm{tr}\big\{(\boldsymbol{\mathcal{E}}'\boldsymbol{P}_{\boldsymbol{X}_\omega}\boldsymbol{\mathcal{E}})^2\big\}\big]}\,.
\end{aligned}
$$

Let $h_{\omega,i}$ be h_i in the full model, where h_i is given by (3.3). It follows from the equation $0 \le h_{\omega,i} \le 1$ that

$$
\sum_{i=1}^n h_{\omega,i}^2 \le \sum_{i=1}^n h_{\omega,i} = n - k_\omega\,, \qquad \sum_{i=1}^n (1 - h_{\omega,i})^2 \le \sum_{i=1}^n (1 - h_{\omega,i}) = k_\omega\,.
$$

From Lemma 5 in [29], we can see that

$$
\begin{aligned}
E_*\big[\mathrm{tr}(\boldsymbol{V}_\omega^2)\big] &= \frac{1}{n}\,E_*\big[\mathrm{tr}\big(\big\{\boldsymbol{\mathcal{E}}'(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}_\omega})\boldsymbol{\mathcal{E}}\big\}^2\big)\big] - 2nE_*\big[\mathrm{tr}(\boldsymbol{S}_\omega)\big] + np \\
&= \frac{1}{n}\left\{\kappa_4^{(1)}\sum_{i=1}^n h_{\omega,i}^2 + p(p+1)\,(n - k_\omega) + p\,(n - k_\omega)^2\right\} - np + 2k_\omega p \\
&\le \left(1 - \frac{k_\omega}{n}\right)\left\{\big|\kappa_4^{(1)}\big| + p(p+1)\right\} + \frac{k_\omega^2 p}{n}\,, \\
E_*\big[\mathrm{tr}\big\{(\boldsymbol{\mathcal{E}}'\boldsymbol{P}_{\boldsymbol{X}_\omega}\boldsymbol{\mathcal{E}})^2\big\}\big] &= \kappa_4^{(1)}\sum_{i=1}^n (1 - h_{\omega,i})^2 + p(p+1)\,k_\omega + pk_\omega^2 \\
&\le k_\omega\left\{\big|\kappa_4^{(1)}\big| + p(p+1)\right\} + pk_\omega^2\,,
\end{aligned}
$$

where $\kappa_4^{(1)}$ is the multivariate kurtosis given in (2.4). The above expectations indicate that $E_*[\mathrm{tr}(\boldsymbol{V}_\omega^2)] = O(1)$ and $E_*[\mathrm{tr}\{(\boldsymbol{\mathcal{E}}'\boldsymbol{P}_{\boldsymbol{X}_\omega}\boldsymbol{\mathcal{E}})^2\}] = O(1)$. Recall that we assume $E_*[\mathrm{tr}(\boldsymbol{S}_\omega^{-2})] = O(1)$. Hence, we derive $E_*[\mathrm{tr}(\boldsymbol{S}_\omega^{-1})] = p + O(n^{-1/2})$ and $E_*[\mathrm{tr}(\boldsymbol{\mathcal{E}}'\boldsymbol{P}_{\boldsymbol{X}_\omega}\boldsymbol{\mathcal{E}}\boldsymbol{S}_\omega^{-1})] = O(1)$. Substituting the obtained orders of expectations into (A.14) yields

$$\text{(A.15)} \qquad E_*\big[\mathcal{L}(\hat{\boldsymbol{\Xi}}, \hat{\boldsymbol{\Sigma}}\,|\,\boldsymbol{X})\big] - E_*\big[\mathcal{L}(\hat{\boldsymbol{\Xi}}_\omega, \hat{\boldsymbol{\Sigma}}_\omega\,|\,\boldsymbol{X}_\omega)\big] \;\geq\; n\log|\boldsymbol{I}_p + \boldsymbol{\Omega}| + O(n^{1/2})\;.$$

When the assumptions in Theorem 2.1 hold, $\lim_{n\to\infty}\boldsymbol{\Omega}$ exists, because $\boldsymbol{\Gamma}_*$ can be expressed $\boldsymbol{X}_*\boldsymbol{\Xi}_*$, and $\lim_{n\to\infty}\boldsymbol{X}_*'\boldsymbol{X}_*$, $\lim_{n\to\infty}\boldsymbol{X}_*'\boldsymbol{X}$ and $\lim_{n\to\infty}\boldsymbol{X}'\boldsymbol{X}$ exist and are positive definite, where $\boldsymbol{X}_*$ is Let $\boldsymbol{\Omega}_0$ be a limiting value of $\boldsymbol{\Omega}$. Then, from (A.15), the following equation is derived:

$$\liminf_{n\to\infty}\frac{1}{n}\Big\{E_*\big[\mathcal{L}(\hat{\boldsymbol{\Xi}}, \hat{\boldsymbol{\Sigma}}\,|\,\boldsymbol{X})\big] - E_*\big[\mathcal{L}(\hat{\boldsymbol{\Xi}}_\omega, \hat{\boldsymbol{\Sigma}}_\omega\,|\,\boldsymbol{X}_\omega)\big]\Big\} \;=\; \log|\boldsymbol{I}_p + \boldsymbol{\Omega}_0| \;>\; 0\;.$$

The above result and the fact that the risk function in the true model is smaller than those in all overspecified models indicate that the risk function in the true model is the smallest among all candidate models when $n \to \infty$. Consequently, Theorem 2.1 is proved.

A.2. Relationship between the best models selected by the AIC and CAIC

Let M_j $(j = 1, ..., m_M)$ be the j-th 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 $\mathcal{J} = \{j \in \{2, ..., m_M\}\,|\; k_j \geq k_1\}$, 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 $j \in \{2, ..., m_M\}$, 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, the following inequality is derived:

$$\text{(A.16)} \qquad \mathrm{CAIC}_j - \mathrm{CAIC}_1 = \mathrm{AIC}_j - \mathrm{AIC}_1 + q(k_j) - q(k_1) > 0\,, \quad (j \in \mathcal{J})\;.$$

The result of (A.16) 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.

A.3. Asymptotic equivalence of the EIC, adjusted EIC, and TIC for an overspecified model

From [9, 27][2], when $m \to \infty$, $\hat{B}_{\mathrm{EIC}}$ and $\hat{B}_{\mathrm{EIC_A}}$ can be expanded as

$$\begin{aligned}\hat{B}_{\mathrm{EIC}} &= 2pk + p(p+1) + \hat{\kappa}_4^{(1)} + O_p(n^{-1}) \,, \\ \hat{B}_{\mathrm{EIC_A}} &= 2(k+p+1)\,\mathrm{tr}(\boldsymbol{G}) - 3\,\mathrm{tr}(\boldsymbol{G}^2) - 2\,\mathrm{tr}(\boldsymbol{G})^2 + \frac{1}{n}\sum_{i=1}^{n} \hat{r}_{\omega,i}^4 + O_p(n^{-1}) \,,\end{aligned}$$

where $\hat{\kappa}_4^{(1)}$ is given by (3.3), $\boldsymbol{G} = \hat{\boldsymbol{\Sigma}}_\omega \hat{\boldsymbol{\Sigma}}^{-1}$ and $\hat{r}_{\omega,i}^2 = (\boldsymbol{y}_i - \hat{\boldsymbol{\Xi}}_\omega' \boldsymbol{x}_{\omega,i})' \hat{\boldsymbol{\Sigma}}^{-1} (\boldsymbol{y}_i - \hat{\boldsymbol{\Xi}}_\omega' \boldsymbol{x}_{\omega,i})$. 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 $n^{-1}\sum_{i=1}^n \hat{r}_{\omega,i}^4 = p(p+2) + \kappa_4^{(1)} + O_p(n^{-1/2})$ hold, where $\kappa_4^{(1)}$ is given in (2.4). Hence, $\hat{B}_{\mathrm{EIC}}$ and $\hat{B}_{\mathrm{EIC_A}}$ can be rewritten as follows when the model M is overspecified:

$$\begin{aligned}\hat{B}_{\mathrm{EIC}} &= 2pk + p(p+1) + \kappa_4^{(1)} + O_p(n^{-1/2}) \,, \\ \hat{B}_{\mathrm{EIC_A}} &= 2pk + p(p+1) + \kappa_4^{(1)} + O_p(n^{-1/2}) \,.\end{aligned} \tag{A.17}$$

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 $\hat{r}_i^2 = \boldsymbol{\varepsilon}_i'\boldsymbol{\varepsilon}_i + O_p(n^{-1/2})$ and $1 - h_i = O(n^{-1})$ are satisfied. Then, $\hat{B}_{\mathrm{TIC}}$ can be expanded as

$$\hat{B}_{\mathrm{TIC}} = 2pk + p(p+1) + \kappa_4^{(1)} + O_p(n^{-1/2}) \,. \tag{A.18}$$

Comparing (A.17) with (A.18) yields $\mathrm{EIC} = \mathrm{TIC} + O_p(n^{-1/2})$ and $\mathrm{EIC_A} = \mathrm{TIC} + O_p(n^{-1/2})$, when the model M is overspecified and $m \to \infty$.

A.4. Asymptotic equivalence of the CV criterion and the TIC

From [27], the last term in (3.5) can be expanded as

$$\begin{aligned}\sum_{i=1}^{n} \frac{(n-1)\,\hat{r}_i^2}{h_i(nh_i - \hat{r}_i^2)} &= np + 2pk + p(p+1) + \hat{\kappa}_4^{(1)} \\ &\quad + 2\sum_{i=1}^{n}(1-h_i)\,(\hat{r}_i^2 - p) + O_p(n^{-1}) \,,\end{aligned} \tag{A.19}$$

where $\hat{r}_i^2$, $\hat{\kappa}_4^{(1)}$, and h_i are given by (3.3). Moreover, by applying the Taylor expansion to equation (3.5), we obtain

$$\sum_{i=1}^{n} \log\left(1 - \frac{\hat{r}_i^2}{nh_i}\right) = -\frac{1}{n}\sum_{i=1}^{n}\frac{\hat{r}_i^2}{h_i} + O_p(n^{-1}) \,. \tag{A.20}$$

[2] At the bottom of p. 240 of [9], $-\,\mathrm{tr}(\hat{\boldsymbol{\Lambda}}^2)$ is missing in the equation $E[\hat{B}_A \,|\, \boldsymbol{Y}]$.

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 (A.20), we obtain

$$\sum_{i=1}^n \log\left(1 - \frac{\hat{r}_i^2}{nh_i}\right) = -p + O_p(n^{-1}) \,. \tag{A.21}$$

On the other hand, $np\log\{2\pi n/(n-1)\} = np\log 2\pi + p + O(n^{-1})$ holds. Consequently, substituting this result and equations (A.19) and (A.21) into (3.5), and comparing the obtained equation with the definition of the TIC, yields $\mathrm{CV} = \mathrm{TIC} + O_p(n^{-1})$.

A.5. Asymptotic equivalence of the best models selected by the nine information criteria

Let IC be a general notation to indicate one of the nine information criteria considered in this paper. Notice that the bias-correction terms in the information criteria expect for the CV criterion are $O_p(1)$, and $\mathrm{CV} = \mathrm{TIC} + O_p(n^{-1})$ holds. Since $\hat{\boldsymbol{\Sigma}} \xrightarrow{p} \boldsymbol{\Sigma}_* + \boldsymbol{\Sigma}_*^{1/2}\boldsymbol{\Omega}_0\boldsymbol{\Sigma}_*^{1/2}$ as $n \to \infty$, where $\boldsymbol{\Omega}_0$ is a limiting value of $\boldsymbol{\Omega}$ given by (2.5), we have

$$\begin{aligned} \frac{1}{n}\,\mathrm{IC} &\xrightarrow{p} p\log 2\pi + \log|\boldsymbol{\Sigma}_*| + \log|\boldsymbol{I}_p + \boldsymbol{\Omega}_0| + p \\ &\geq p\log 2\pi + \log|\boldsymbol{\Sigma}_*| + p\,, \qquad \text{as } n \to \infty\,, \end{aligned} \tag{A.22}$$

with equality if and only if M is an overspecified model. The equation in (A.22) indicates that underspecified models are never selected as the best model when $n \to \infty$.

Let ICA denote an information criterion proposed under normality (i.e., the AIC, CAIC, or MAIC), and let ICT denote an information criterion proposed without a normality assumption (i.e., the TIC, EIC, $\mathrm{EIC_A}$, CV criterion, $\mathrm{AIC_J}$, or $\mathrm{CAIC_J}$). Notice that when M is an overspecified model, $\mathrm{ICA} = \mathrm{AIC} + o_p(1)$, $\mathrm{ICT} = \mathrm{TIC} + o_p(1)$ and $\hat{B}_{\mathrm{TIC}} \xrightarrow{p} 2pk + p(p+1) + \kappa_4^{(1)}$ as $n \to \infty$, where $\kappa_4^{(1)}$ is the multivariate kurtosis given in (2.4). Hence, when M is an overspecified model, we derive

$$\mathrm{ICT} = \mathrm{ICA} + \kappa_4^{(1)} + o_p(1)\,. \tag{A.23}$$

It should be emphasized that $\kappa_4^{(1)}$ does not depend on the candidate model considered, i.e., $\kappa_4^{(1)}$ is common in all overspecified models. 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 equation (A.23), we obtain

$$\mathrm{ICT}_1 - \mathrm{ICT}_2 = \mathrm{ICA}_1 - \mathrm{ICA}_2 + o_p(1)\,.$$

This equation indicates 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 $n \to \infty$.

ACKNOWLEDGMENTS

The authors would like to thank the editor, the associate editor and the two reviewers for valuable comments. Especially, the authors are grateful that one of reviewers pointed out many mistakes of our paper. 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

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

[2] 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.

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

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

[5] 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.

[6] Fox, J. and Weisberg, S. (2011). *Multivariate linear models in R.* In "An Appendix to An R Companion to Applied Regression (2nd. ed.)", `http://socserv.mcmaster.ca/jfox/Books/Companion/appendix.html`, accessed April 17, 2014.

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

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

[9] 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.

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

[11] Hashiyama, Y.; Yanagihara, H. and Fujikoshi, Y. (2014). Jackknife bias correction of the AIC for selecting variables in canonical correlation analysis under model misspecification, *Linear Algebra and its Applications*, **455**, 82–106.

[12] Hurvich, C.M. and Tsai, C.-L. (1989). Regression and time series model selection in small samples, *Biometrika*, **76**, 297–307.

[13] 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.

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

[15] Konishi, S. and Kitagawa, G. (2003). Asymptotic theory for information criteria in model selection – functional approach, *Journal of Statistical Planning and Inference*, **114**, 45–61.

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

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

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

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

[20] McQuarrie, A.D.R. and Tsai, C.-L. (1998). *Regression and Time Series Model Selection*, World Scientific Publishing Co., Inc., River Edge, NJ.

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

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

[23] 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.

[24] 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.

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

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

[27] 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.

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

[29] Yanagihara, H. (2015). Conditions for consistency of a log-likelihood-based information criterion in normal multivariate linear regression models under the violation of normality assumption, *Journal of the Japan Statistical Society*, **45**, 21–56.

[30] Yanagihara, H. and Fujisawa, H. (2012). Iterative bias correction of the cross-validation criterion, *Scandinavian Journal of Statistics*, **39**, 116–130.

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

[32] 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.

[33] Yanagihara, H.; Tonda, T. and Matsumoto, C. (2006). Bias correction of cross-validation criterion based on Kullback–Leibler information under a general condition, *Journal of Multivariate Analysis*, **97**, 1965–1975.

[34] Yanagihara, H.; Wakaki, H. and Fujikoshi, Y. (2015). A consistency property of the AIC for multivariate linear models when the dimension and the sample size are large, *The Electronic Journal of Statistics*, **9**, 869–897.

[35] Yanagihara, H.; Yuan, K.-H.; Fujisawa, H. and Hayashi, K. (2013). A class of model selection criteria based on cross-validation method, *Hiroshima Mathematical Journal*, **43**, 149–177.

[36] 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).

[37] 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 –" (W. Bandilla and F. Faulbaum, Eds.), Lucius & Lucius, Stuttgart, Germany, 385–391.