Second-Order Bias-Corrected AIC in Multivariate Normal Linear Models under Nonnormality

Hirokazu YANAGIHARA[1], Ken-ichi KAMO[2] AND Tetsuji TONDA[3]

[1]*Department of Social Systems and Management*
Graduate School of Systems and Information Engineering
University of Tsukuba
1-1-1 Tennodai, Tsukuba, Ibaraki 305-8573, Japan

[2]*Division of Mathematics*
School of Medicine Liberal Arts and Sciences
Sapporo Medical University
South 1, West 17, Chuoku, Sapporo 060-8543, Japan

[3]*Department of Environmetrics and Biometrics*
Research Institute for Radiation Biology and Medicine
Hiroshima University
1-2-3 Kasumi, Minami-ku, Hiroshima 734-8553, Japan

Abstract

This paper deals with correcting a bias of Akaike's information criterion (AIC) for selecting variables in multivariate normal linear regression models when the true distribution of observation is an unknown nonnormal distribution. It is well known that the bias of AIC is $O(1)$, and there are several information criteria which improve the bias to $O(n^{-1})$, where n is the sample size. By slightly adjusting the first-order bias-corrected AIC, we propose a new information criterion. Although the adjustment merely uses constant coefficients, the bias of the new criterion is reduced to $O(n^{-2})$. Through numerical experiments, we verify that our criterion is superior to others.

---

[1]Corresponding author, E-mail address: *yanagi@sk.tsukuba.ac.jp*. (Last Modified: February 2, 2006)

1. Introduction

A multivariate linear regression model has been used universally to study the effects of K explanatory variables $x_1, \ldots, x_K$ on a $p \times 1$ response variable vector $\boldsymbol{y} = (y_1, \ldots, y_p)'$. Since we would like to specify the factors affecting $\boldsymbol{y}$ in the regression analysis, searching for the optimal subset of $\{x_1, \ldots, x_K\}$ is essential. In the first phase of the analysis, it is common to assume that observations are distributed according to the normal distribution. This is because we are not able to see the distribution of observations in most cases. On the other hand, the number of all subsets, i.e., the number of candidate models considered, grows explosively by increasing K. If repetition of the estimating parameters is required simultaneously, more computational complexity becomes necessary. Since the estimators in the normal linear model can be obtained without the repetition, the normal assumption is appropriate for searching the optimal model thoroughly because the computational complexity is decreased dramatically. However, if the gap between the normal distribution and the true distribution results undesirably affects the selecting of variables, we should consider a way to reduce it. Therefore, in this paper, we deal with the model selection in multivariate normal linear models under the condition that the true distribution of $\boldsymbol{y}$ may not be the normal distribution.

Choosing the best model by minimizing the information criterion is well known and widely used. Most of the information criteria estimate the risk based on the predictive Kullback-Leibler (K-L) information (Kullback & Leibler, 1951), which measures the discrepancy between the true model and the candidate model. The simplest estimator for the risk is the "$-2\times$ the sample log-likelihood"; however, this has a constant bias. Since it is necessary to reduce this kind of bias, an usual information criterion is defined by adding a bias-correction term to the simplest estimator. The most universal information criterion is Akaike's information criterion (AIC, Akaike, 1973). Since the bias-correction term of AIC is evaluated under the assumption that the candidate model is not misspecified, the bias of AIC becomes O(1) when the normal assumption is not satisfied. Several information criteria correcting the AIC's bias to $O(n^{-1})$, where n is the sample size, were also proposed under the assumption that the candidate model may be misspecified, e.g., Takeuchi's information criterion (TIC, Takeuchi, 1976), extended information criterion (EIC, Ishiguro et al., 1997) and the cross-validation (CV) criterion (Stone, 1974, 1977). Recently, Yanagihara (2006b) proposed a bias-corrected AIC which consists of a jackknife

estimate of the bias. He evaluated the bias from the predicted residual sum of the squares (PRESS) and made a bias-correction term that was an exact unbiased term by multiplying the constant coefficient when the candidate model includes the true model. In this paper, we call this bias-corrected AIC the jackknifed AIC (AIC$_\mathrm{J}$). Although AIC$_\mathrm{J}$ is the first-order bias-corrected AIC, as is TIC and EIC, the bias of AIC$_\mathrm{J}$ tends to be smaller than the others. Especially, we can see that the bias of AIC$_\mathrm{J}$ becomes very small numerically. However, the n^{-1} term in the bias, which consists of the multivariate kurtosis, still remains in AIC$_\mathrm{J}$. Therefore, the bias may become large when the sample size is small or the kurtosis is large. Currently, there is no information criterion with second-order accuracy under nonnormality that has a numerically small bias.

In this paper, our purpose is to propose a new information criterion that has a bias which is not only small numerically but also theoretically. We correct the bias of AIC$_\mathrm{J}$ to $\mathrm{O}(n^{-2})$ without adding any correction term, while maintaining the desirable character of being numerically very small. First, in order to correct the bias theoretically, we slightly adjust the PRESS by constant coefficients. Next, as in the AIC$_\mathrm{J}$, the adjusted PRESS is multiplied by another constant coefficient which makes the exact unbiased estimator for the risk when the candidate model includes the true one.

This paper is organized in the following way. In Section 2, we describe the models, the risk based on the predictive K-L information, and AIC$_\mathrm{J}$. In Section 3, we propose a new information criterion which is the second-order accuracy of the risk. In Section 4, we verify that our criterion is better than AIC and AIC$_\mathrm{J}$ by conducting numerical experiments. In Section 5, we conclude our discussion. Technical details are provided in the Appendix.

2. Preliminaries

2.1. Models and Risk

Let $\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)'$ be an $n \times p$ observation matrix. We consider the following candidate model with k explanatory variables which are the subset of $\{x_1, \ldots, x_K\}$.

$$M : \boldsymbol{Y} \sim \mathrm{N}_{n \times p}(\boldsymbol{X\Theta}, \boldsymbol{\Sigma} \otimes \boldsymbol{I}_n), \tag{2.1}$$

where $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)'$ is an $n \times k$ matrix of nonstochastic explanatory variables with the full rank k. Then, the joint probability density function of $\boldsymbol{Y}$ is shown by

3

$\prod_{i=1}^{n} f(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{\Theta}, \boldsymbol{\Sigma})$, where

$$f(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{\Theta}, \boldsymbol{\Sigma}) = \left(\frac{1}{2\pi}\right)^{p/2} |\boldsymbol{\Sigma}|^{-1/2} \exp\left\{-\frac{1}{2}(\boldsymbol{y}_i - \boldsymbol{\Theta}'\boldsymbol{x}_i)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_i - \boldsymbol{\Theta}'\boldsymbol{x}_i)\right\}. \qquad (2.2)$$

Although the normal distribution is assumed, we are not able to see whether the assumption is actually correct. Therefore, it is more appropriate to set the situation in which the true distribution of $\boldsymbol{Y}$ may be not distributed according to the normal distribution. That is, the true model is denoted by

$$M^*: \ \boldsymbol{Y} = \boldsymbol{\eta}^* + \boldsymbol{\mathcal{E}}\boldsymbol{\Sigma}^{*1/2}, \qquad (2.3)$$

where $\boldsymbol{\eta}^*$ and $\boldsymbol{\Sigma}^*$ are $n \times p$ and $p \times p$ unknown parameter matrices, respectively, and $p \times 1$ row vectors $\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n$ of error matrix $\boldsymbol{\mathcal{E}}$ are independent, and copies of a random variable vector $\boldsymbol{\varepsilon}$ which is distributed following an unknown distribution with the mean $\boldsymbol{0}$ and covariance matrix $\boldsymbol{I}_p$. In particular, we will assume the candidate model M is the overspecified model, satisfying $\boldsymbol{P_X}\boldsymbol{\eta}^* = \boldsymbol{\eta}^*$, where $\boldsymbol{P_X}$ is the projection matrix of $\boldsymbol{X}$, i.e., $\boldsymbol{P_X} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$. In our general nonnormal setting, the terminology of "overspecified model" is not necessary to imply the candidate model includes the true model. In a normal setting, as in Fujikoshi and Satoh (1997), the terminology corresponds to the inclusion of the true model. Also, we assume that four assumptions for $\boldsymbol{X}$ in M and $\boldsymbol{\mathcal{E}}$ in M^* (see Appendix A.1) hold to guarantee valid expansions on the biases of information criteria.

Let the maximum likelihood estimators (MLEs) of $\boldsymbol{\Theta}$ and $\boldsymbol{\Sigma}$ under the normal assumption be denoted by $\hat{\boldsymbol{\Theta}}$ and $\hat{\boldsymbol{\Sigma}}$, respectively, which are defined by

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

Then, the $-2 \times$ sample log-likelihood function is given by

$$-2\ell(\hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\Sigma}}|\boldsymbol{Y}, \boldsymbol{X}) = -2\sum_{i=1}^{n}\log f(\boldsymbol{y}_i|\boldsymbol{x}_i, \hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\Sigma}})$$
$$= np\log 2\pi + n\log|\hat{\boldsymbol{\Sigma}}| + np. \qquad (2.4)$$

Let $\boldsymbol{U} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n)'$ be an $n \times p$ future observation matrix, which is independent of $\boldsymbol{Y}$, and let $p \times 1$ row vectors $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$ of $\boldsymbol{U}$ be independently distributed according to the same distributions of $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$, respectively. Then, the risk based on the predictive K-L information measuring the discrepancy between M^* and M under the normal assumption

is defined by

$$R_{\mathrm{KL}} = \mathrm{E}_{\boldsymbol{Y}}^* \mathrm{E}_{\boldsymbol{U}}^* \left[-2\ell(\hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\Sigma}} | \boldsymbol{U}, \boldsymbol{X}) \right]$$
$$= np \log 2\pi + n\mathrm{E}_{\boldsymbol{Y}}^* \left[\log |\hat{\boldsymbol{\Sigma}}| \right] + R_{\mathrm{FP}}, \qquad (2.5)$$

where R_{FP} is an expectation of the future predicted residual sum of the squares, denoted as

$$R_{\mathrm{FP}} = \mathrm{E}_{\boldsymbol{U}}^* \mathrm{E}_{\boldsymbol{Y}}^* \left[\sum_{i=1}^{n} \left(\boldsymbol{u}_i - \hat{\boldsymbol{\Theta}}' \boldsymbol{x}_i \right)' \hat{\boldsymbol{\Sigma}}^{-1} \left(\boldsymbol{u}_i - \hat{\boldsymbol{\Theta}}' \boldsymbol{x}_i \right) \right]. \qquad (2.6)$$

Here, E^* means the expectation under the true model M^*.

It is easy to obtain the simplest estimator of R_{KL} by $-2\ell(\hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\Sigma}} | \boldsymbol{Y}, \boldsymbol{X})$. However, when we estimate R_{KL} by $-2\ell(\hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\Sigma}} | \boldsymbol{Y}, \boldsymbol{X})$, the constant bias appears as follows:

$$B = R_{\mathrm{KL}} - \mathrm{E}_{\boldsymbol{Y}}^* \left[-2\ell(\hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\Sigma}} | \boldsymbol{Y}, \boldsymbol{X}) \right] = R_{\mathrm{FP}} - np.$$

In order to reduce this kind of bias, many authors have obtained an estimator of B, called $\hat{B}$, under several conditions. The AIC-type information criterion is defined by adding $\hat{B}$ to $-2\ell(\hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\Sigma}} | \boldsymbol{Y}, \boldsymbol{X})$, and thus, the AIC-type criterion is specified by the individual $\hat{B}$.

2.2. Akaike's Information Criterion

Akaike (1973) approximated B as $\hat{B}_{\mathrm{AIC}} = 2pk + p(p+1)$. By using $\hat{B}_{\mathrm{AIC}}$, AIC is defined by

$$\mathrm{AIC} = -2\ell(\hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\Sigma}} | \boldsymbol{Y}, \boldsymbol{X}) + \hat{B}_{\mathrm{AIC}}$$
$$= np(\log 2\pi + 1) + n \log |\hat{\boldsymbol{\Sigma}}| + 2pk + p(p+1). \qquad (2.7)$$

However, Fujikoshi et al. (2005) reported that the bias of AIC is expressed as

$$B_{\mathrm{AIC}} = R_{\mathrm{KL}} - \mathrm{E}_{\boldsymbol{Y}}^*[\mathrm{AIC}] = \kappa_4^{(1)} + \mathrm{O}(n^{-1}),$$

where $\kappa_4^{(1)}$ is the multivariate kurtosis (see, e.g., Mardia, 1970; Isogai, 1983), which is defined by

$$\kappa_4^{(1)} = \mathrm{E}_{\boldsymbol{\varepsilon}}^* \left[(\boldsymbol{\varepsilon}' \boldsymbol{\varepsilon})^2 \right] - p(p+2).$$

Since Akaike evaluated $\hat{B}_{\mathrm{AIC}}$ under the assumption that the candidate model is not misspecified, $\kappa_4^{(1)}$ appears in the top term of B_{AIC}. Notice that $\kappa_4^{(1)}$ measures the variation from the normal distribution. Thus, the breakdown of normality makes the constant bias

5

of AIC. From the expansion of B_{AIC}, we can see that the bias of AIC is the first-order accuracy of R_{KL} if $\kappa_4^{(1)}$ is 0. Therefore, AIC is the first-order accuracy of R_{KL} when the candidate model M includes the true model M^*. However, that is not the necessary and sufficient condition because the distribution with $\kappa_4^{(1)} = 0$ exists, except for the normal distribution (see Kale & Sebastian, 1996). On the other hand, if $\kappa_4^{(1)}$ is not 0, AIC is the constant-order accuracy of R_{KL}. In other words, the more severe the breakdown of normality, the more the bias grows.

2.3. Jackknifed AIC

Let $\boldsymbol{Y}_{(-i)}$ and $\boldsymbol{X}_{(-i)}$ be $(n-1) \times p$ and $(n-1) \times k$ matrices obtained from $\boldsymbol{Y}$ and $\boldsymbol{X}$ by deleting $\boldsymbol{y}_i$ and $\boldsymbol{x}_i$, respectively. Further, let the MLEs of $\boldsymbol{\Theta}$ and $\boldsymbol{\Sigma}$ evaluated from $\boldsymbol{Y}_{(-i)}$ and $\boldsymbol{X}_{(-i)}$ be denoted as $\hat{\boldsymbol{\Theta}}_{[-i]}$ and $\hat{\boldsymbol{\Sigma}}_{[-i]}$, respectively, which are defined by

$$\hat{\boldsymbol{\Theta}}_{[-i]} = \left\{ \boldsymbol{X}'_{(-i)} \boldsymbol{X}_{(-i)} \right\}^{-1} \boldsymbol{X}'_{(-i)} \boldsymbol{Y}_{(-i)}, \quad \hat{\boldsymbol{\Sigma}}_{[-i]} = \frac{1}{n} \boldsymbol{Y}'_{(-i)} \left\{ \boldsymbol{I}_{n-1} - \boldsymbol{P}_{\boldsymbol{X}_{(-i)}} \right\} \boldsymbol{Y}_{(-i)}.$$

Then, PRESS is expressed as

$$\text{PRESS} = \sum_{i=1}^{n} (\boldsymbol{y}_i - \hat{\boldsymbol{\Theta}}'_{[-i]} \boldsymbol{x}_i)' \hat{\boldsymbol{\Sigma}}^{-1}_{[-i]} (\boldsymbol{y}_i - \hat{\boldsymbol{\Theta}}'_{[-i]} \boldsymbol{x}_i).$$

Let the coefficient $a(j)$ be defined by

$$a(j) = \frac{n+j-1}{n+j}, \tag{2.8}$$

and let $\tilde{r}_i^2$ denote a squared norm of the ith estimated residual defined by

$$\tilde{r}_i^2 = \frac{1}{1 - (\boldsymbol{P}_{\boldsymbol{X}})_{ii}} (\boldsymbol{y}_i - \hat{\boldsymbol{\Theta}}' \boldsymbol{x}_i)' \hat{\boldsymbol{\Sigma}}^{-1} (\boldsymbol{y}_i - \hat{\boldsymbol{\Theta}}' \boldsymbol{x}_i), \tag{2.9}$$

where $(\boldsymbol{P}_{\boldsymbol{X}})_{ij}$ is the (i,j)th element of the matrix $\boldsymbol{P}_{\boldsymbol{X}}$, i.e., $(\boldsymbol{P}_{\boldsymbol{X}})_{ij} = \boldsymbol{x}'_i (\boldsymbol{X}'\boldsymbol{X}) \boldsymbol{x}_j$. Furthermore, we define a function $Q(x; \lambda)$ as

$$Q(x; \lambda) = x \left(1 - \frac{x}{n} \right)^{-\lambda}. \tag{2.10}$$

From the formulas in Fujikoshi et al. (2003), PRESS can be expressed without $\hat{\boldsymbol{\Theta}}_{[-i]}$ and $\hat{\boldsymbol{\Sigma}}_{[-i]}$ as follows:

$$\text{PRESS} = a(0) \sum_{i=1}^{n} \{(1 - (\boldsymbol{P}_{\boldsymbol{X}})_{ii})\}^{-1} Q(\tilde{r}_i^2; 1). \tag{2.11}$$

Yanagihara (2006b) evaluated R_{FP} by using the jackknife method, i.e., using PRESS, and approximated B by $\hat{B}_{\text{AIC}_\text{J}} = \{c/a(0)\}\text{PRESS} - np$, where the coefficient c is defined by

$$c = \frac{(n+k)(n-k-p-2)}{(n-k-p-1)\sum_{i=1}^{n}\{1-(\boldsymbol{P_X})_{ii}\}^{-1}}. \tag{2.12}$$

The coefficient $c/a(0)$ adjusts PRESS to an unbiased estimator of R_{FP} when the candidate model M includes the true model M^*, i.e., when the true distribution is the normal distribution. Then, AIC_J is defined by

$$\begin{aligned}
\text{AIC}_\text{J} &= -2\ell(\hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\Sigma}}|\boldsymbol{Y}, \boldsymbol{X}) + \hat{B}_{\text{AIC}_\text{J}} \\
&= np\log 2\pi + n\log|\hat{\boldsymbol{\Sigma}}| + c\sum_{i=1}^{n}\{1-(\boldsymbol{P_X})_{ii}\}^{-1}Q(\tilde{r}_i^2; 1).
\end{aligned} \tag{2.13}$$

From Yanagihara (2006b), the bias of AIC_J becomes as follows:

$$B_{\text{AIC}_\text{J}} = R_{\text{KL}} - \text{E}_{\boldsymbol{Y}}^*[\text{AIC}_\text{J}] = \begin{cases} 0 & (\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n \sim i.i.d.\ \text{N}_p(\boldsymbol{0}, \boldsymbol{I}_p)) \\ -\dfrac{1}{n}\kappa_4^{(1)} + \text{O}(n^{-2}) & (\text{otherwise}) \end{cases}.$$

AIC_J is the first-order accuracy of R_{KL} in a general nonnormal case. Therefore, we can see that AIC_J corrects the AIC's bias by replacing $\hat{B}_{\text{AIC}}$ with a renewal term, as in TIC and EIC. However, compared with the first terms in asymptotic expansions of biases, it seems that the bias of AIC_J tends to be smaller than the ones of TIC and EIC (see Yanagihara, 2006b). Moreover, from the numerical study in Yanagihara (2006b), we can see that the bias of AIC_J becomes the smallest among the biases of AIC, TIC and EIC. Generally, if the dimension p or the number of variables k becomes large, the bias of the information criterion will also become large. This increase may be shrunk by the coefficient c, which is one of the reasons why AIC_J works well.

3. New Criterion

3.1. Bias-Corrected AIC_J

In this section, we propose a bias-corrected AIC_J (Corrected AIC_J or CAIC_J). From the previous section, the bias of AIC_J becomes numerically very small; however, AIC_J is the first-order accuracy of R_{KL}. The n^{-1} term in the asymptotically expanded bias consists of the multivariate kurtosis of the true model. Therefore, if the kurtosis becomes large, the bias of AIC_J also has the danger of becoming large. Our goal is to improve the

bias of AIC_J to $O(n^{-2})$ while maintaining the desirable behavior that the bias becomes numerically very small. Since the high quality of AIC_J depends on the coefficient c in (2.12), we must correct PRESS in (2.11) to the second-order accuracy of R_{FP} in (2.6), so that the corrected PRESS may be adjusted to be unbiased by multiplying some constant coefficient when $\varepsilon_1, \ldots, \varepsilon_n$ are $i.i.d.$ $N_p(\mathbf{0}, \boldsymbol{I}_p)$. Needless to say, it makes no sense to correct the bias of AIC_J by $AIC_J - \hat{\kappa}_4^{(1)}/n$, where $\hat{\kappa}_4^{(1)}$ is an estimator of $\kappa_4^{(1)}$ as

$$\hat{\kappa}_4^{(1)} = \frac{1}{n}\sum_{i=1}^{n}\left\{(\boldsymbol{y}_i - \hat{\boldsymbol{\Theta}}'\boldsymbol{x}_i)'\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{y}_i - \hat{\boldsymbol{\Theta}}'\boldsymbol{x}_i)\right\}^2 - p(p+2),$$

(see, e.g., Mardia, 1970; Isogai, 1983). The reason is the following: The bias of $\hat{\kappa}_4^{(1)}$ becomes large unless the sample size n is huge (see Yanagihara, 2006a). Besides this, when $\hat{\kappa}_4^{(1)}$ has a small bias, i.e., the sample size n is large, the bias-correction term $-\hat{\kappa}_4^{(1)}/n$ becomes very small. Consequently, the correction method of using $\hat{\kappa}_4^{(1)}/n$ will become meaningless. Therefore, it is important to correct the bias without estimating $\kappa_4^{(1)}$ for a small sample. From these two significant points for bias correction, we adjust PRESS by the constant coefficients $b_1, \ldots, b_n$ and d as follows:

$$\text{PRESS}^+ = \sum_{i=1}^{n} b_i Q(\tilde{r}_i^2; 1 - d/n), \tag{3.1}$$

where $\tilde{r}_i^2$ is given by (2.9) and the function $Q(x; \lambda)$ is given by (2.10). From Yanagihara (2006a), we can see that the expectation of $Q(\tilde{r}_i^2; 1-d/n)$ can be calculated exactly when $\varepsilon_1, \ldots, \varepsilon_n$ are $i.i.d.$ $N_p(\mathbf{0}, \boldsymbol{I}_p)$. Moreover, it is known that

$$\sum_{i=1}^{n} Q(\tilde{r}_i^2; 1 - d/n) = \sum_{i=1}^{n} Q(\tilde{r}_i^2; 1) - \frac{d}{n}\left\{\hat{\kappa}_4^{(1)} + p(p+2)\right\} + O_p(n^{-2}).$$

Then, we can make $\hat{\kappa}_4^{(1)}$ appear without estimating $\kappa_4^{(1)}$ directly. We only need to determine the coefficients $b_1, \ldots, b_n$ and d so that the n^{-1} term disappears from the asymptotic expansion of the bias of the corrected criterion. Notice that

$$Q(\tilde{r}_i^2, 0) = \frac{1}{1 - (\boldsymbol{P_X})_{ii}}(\boldsymbol{y}_i - \hat{\boldsymbol{\Theta}}'\boldsymbol{x}_i)'\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{y}_i - \hat{\boldsymbol{\Theta}}'\boldsymbol{x}_i),$$

$$Q(\tilde{r}_i^2, 1) = \frac{1}{a(0)}\{1 - (\boldsymbol{P_X})_{ii}\}(\boldsymbol{y}_i - \hat{\boldsymbol{\Theta}}'_{[-i]}\boldsymbol{x}_i)'\hat{\boldsymbol{\Sigma}}^{-1}_{[-i]}(\boldsymbol{y}_i - \hat{\boldsymbol{\Theta}}'_{[-i]}\boldsymbol{x}_i).$$

Therefore, $Q(\tilde{r}_i^2, 1-d/n)$ means the ith squared predictive residual in which the influence of $\boldsymbol{y}_i$ remains for a while. Actually, we decide $b_i = a(1)\{1 + a(1)(\boldsymbol{P_X})_{ii}\}$ and $d = 1$, where

8

the coefficient $a(j)$ is defined by (2.8). Finally, by using equation (3.1), we approximate B by

$$\hat{B}_{\text{CAIC}_{\text{J}}} = c^+ \sum_{i=1}^{n} \{1 + a(1)(\boldsymbol{P_X})_{ii}\} Q(\tilde{r}_i^2; a(0)) - np,$$

where the coefficient c^+ is defined by

$$c^+ = \frac{(n+k)\{n-k-p-2a(0)\}\Gamma\left(\frac{n-k}{2} + \frac{1}{n}\right)\Gamma\left(\frac{n-k-p}{2}\right)}{\{n+a(1)k\}(n-k-p-1)\Gamma\left(\frac{n-k}{2}\right)\Gamma\left(\frac{n-k-p}{2} + \frac{1}{n}\right)}, \tag{3.2}$$

and $\Gamma(x)$ is the gamma function. Thus, if the dimension of observation p is even, i.e., $p = 2m$, the coefficient c^+ can be obtained without calculating the gamma function, as follows:

$$c^+ = \frac{(n+k)\{n-k-p-2a(0)\}}{\{n+a(1)k\}(n-k-p-1)} \prod_{j=1}^{m} \left\{1 + \frac{2}{n(n-k-2j)}\right\},$$

(We omit writing a proof of the above equation, because that it is easy to derive from the property of the gamma function). We can see that $c^+/a(1)$ makes PRESS^+ an unbiased estimator of R_{FP} when the candidate model M includes the true model M^*. Consequently, CAIC_{J} is defined as follows.

DEFINITION. *The second-order bias-corrected AIC is defined by*

$$\text{CAIC}_{\text{J}} = -2\ell(\hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\Sigma}}|\boldsymbol{Y}, \boldsymbol{X}) + \hat{B}_{\text{CAIC}_{\text{J}}}$$

$$= np\log 2\pi + n\log|\hat{\boldsymbol{\Sigma}}| + c^+ \sum_{i=1}^{n} \{1 + a(1)(\boldsymbol{P_X})_{ii}\} Q(\tilde{r}_i^2; a(0)). \tag{3.3}$$

The order of the bias of CAIC_{J} is given by the following theorem.

THEOREM. *Suppose that the assumptions A1 $\sim$ A4 in Appendix A.1 hold. Then, the bias of $CAIC_J$ becomes*

$$B_{\text{CAIC}_{\text{J}}} = R_{\text{KL}} - \text{E}_{\boldsymbol{Y}}^*[\text{CAIC}_{\text{J}}] = \begin{cases} 0 & (\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n \sim i.i.d.\ \text{N}_p(\boldsymbol{0}, \boldsymbol{I}_p)) \\ \text{O}(n^{-2}) & (\text{otherwise}) \end{cases}.$$

We describe the outline of the proof of the above theorem in the next subsection.

3.2. Proof of Theorem

First, we show $B_{\text{CAIC}_{\text{J}}} = 0$ when $\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n$ are *i.i.d.* $\text{N}_p(\boldsymbol{0}, \boldsymbol{I}_p)$. Let T be a random variable distributed according to Hotelling's T^2 distribution with $N = n - k - 1$ and p

degrees of freedoms, whose probability density is

$$g(t; N, p) = \frac{\Gamma\left(\frac{N+1}{2}\right)}{N\Gamma\left(\frac{p}{2}\right)\Gamma\left(\frac{N-p+1}{2}\right)} \left(\frac{t}{N}\right)^{p/2-1} \left(1 + \frac{t}{N}\right)^{-(N+1)/2},$$

(see, e.g., Siotani et al., 1985, p. 196). From Yanagihara (2006a), when $\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n$ are i.i.d. $N_p(\boldsymbol{0}, \boldsymbol{I}_p)$, then $\tilde{r}_i^2$ is rewritten by T as

$$\tilde{r}_i^2 = \frac{n}{N}T\left(1 + \frac{T}{N}\right)^{-1}. \tag{3.4}$$

Substituting (3.4) into (2.10) yields

$$Q(\tilde{r}_i^2; a(0)) = \frac{n}{N}T\left(1 + \frac{T}{N}\right)^{-1/n}.$$

Therefore, we obtain the following expectation:

$$E_{\boldsymbol{Y}}^*[Q(\tilde{r}_i^2; a(0))] = \int_0^\infty \frac{n\Gamma\left(\frac{N+1}{2}\right)}{N\Gamma\left(\frac{p}{2}\right)\Gamma\left(\frac{N-p+1}{2}\right)} \left(\frac{t}{N}\right)^{(p+2)/2-1} \left(1 + \frac{t}{N}\right)^{-(N+2/n+1)/2} dt.$$

Let $N_0 = N + 2/n$ and $p_0 = p + 2$. Then the following equations hold:

$$\Gamma\left(\frac{p_0}{2}\right) = \frac{p}{2}\Gamma\left(\frac{p}{2}\right), \quad \Gamma\left(\frac{N_0 - p + 1}{2}\right) = \frac{N_0 - p - 1}{2}\Gamma\left(\frac{N_0 - p_0 + 1}{2}\right).$$

By using the above equations and replacing t/N with t/N_0, we derive

$$\begin{aligned}
E_{\boldsymbol{Y}}^*[Q(\tilde{r}_i^2; a(0))] &= \frac{np\Gamma(\frac{N+1}{2})\Gamma(\frac{N_0-p+1}{2})}{(N_0 - p - 1)\Gamma(\frac{N_0+1}{2})\Gamma(\frac{N-p+1}{2})} \int_0^\infty g(t; N_0, p_0) dt \\
&= \frac{n(n+k)p}{c^+\{n + a(1)k\}(n - k - p - 1)}. \tag{3.5}
\end{aligned}$$

Notice that

$$E_{\boldsymbol{Y}}^*[\text{CAIC}_\text{J}] = np\log 2\pi + nE_{\boldsymbol{Y}}^*[\log|\hat{\boldsymbol{\Sigma}}|] + c^+ \sum_{i=1}^n \{1 + a(1)(\boldsymbol{P_X})_{ii}\} E_{\boldsymbol{Y}}^*[Q(\tilde{r}_i^2; a(0))]. \tag{3.6}$$

Substituting (3.5) into (3.6) yields

$$E_{\boldsymbol{Y}}^*[\text{CAIC}_\text{J}] = np\log 2\pi + nE_{\boldsymbol{Y}}^*[\log|\hat{\boldsymbol{\Sigma}}|] + \frac{n(n+k)p}{n - k - p - 1}. \tag{3.7}$$

From the result in Fujikoshi and Satoh (1997), R_{KL} can be precisely written as

$$R_{\text{KL}} = np\log 2\pi + nE_{\boldsymbol{Y}}^*[\log|\hat{\boldsymbol{\Sigma}}|] + \frac{n(n+k)p}{n - k - p - 1}. \tag{3.8}$$

10

We can see that equation (3.8) corresponds to (3.7). Therefore, the bias of CAIC$_J$ is exactly 0 when $\varepsilon_1, \ldots, \varepsilon_n$ are $i.i.d.$ N$_p(\mathbf{0}, \boldsymbol{I}_p)$.

Next, we show $B_{\mathrm{CAIC_J}} = \mathrm{O}(n^{-2})$ in a general nonnormal case. By using the commutative formula, as in Yanagihara et al. (2006), we can decompose $\mathrm{E}^*_{\boldsymbol{Y}}[\mathrm{PRESS}]$ into

$$\mathrm{E}^*_{\boldsymbol{Y}}[\mathrm{PRESS}] = a(0)\left\{R_{\mathrm{FP}} + \mathrm{E}^*_{\boldsymbol{Y}}\mathrm{E}^*_{\boldsymbol{U}}[R_1] + 2\mathrm{E}^*_{\boldsymbol{Y}}\mathrm{E}^*_{\boldsymbol{U}}[R_2] + \mathrm{E}^*_{\boldsymbol{Y}}[R_3]\right\}, \tag{3.9}$$

where R_1, R_2 and R_3 are defined by

$$
\begin{aligned}
R_1 &= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{1-(\boldsymbol{P_X})_{ii}}\left\{(\boldsymbol{u}_i - \hat{\boldsymbol{\Theta}}'\boldsymbol{x}_i)'\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{y}_i - \hat{\boldsymbol{\Theta}}'\boldsymbol{x}_i)\right\}^2\left(1 - \frac{\tilde{r}_i^2}{n}\right)^{-1}, \\
R_2 &= \sum_{i=1}^{n}\frac{(\boldsymbol{P_X})_{ii}}{1-(\boldsymbol{P_X})_{ii}}(\boldsymbol{u}_i - \hat{\boldsymbol{\Theta}}'\boldsymbol{x}_i)'\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{y}_i - \hat{\boldsymbol{\Theta}}'\boldsymbol{x}_i)\left(1 - \frac{\tilde{r}_i^2}{n}\right)^{-1}, \\
R_3 &= \sum_{i=1}^{n}\frac{\{(\boldsymbol{P_X})_{ii}\}^2}{1-(\boldsymbol{P_X})_{ii}}Q(\tilde{r}_i^2;1),
\end{aligned}
\tag{3.10}
$$

(The derivation of this decomposition is described in Appendix A.2,) Let's define PRESS† as

$$\mathrm{PRESS}^\dagger = \frac{1}{a(0)}\mathrm{PRESS} - R_3 = \sum_{i=1}^{n}\{1 + (\boldsymbol{P_X})_{ii}\}Q(\tilde{r}_i^2;1).$$

Then $\mathrm{E}^*_{\boldsymbol{Y}}[\mathrm{PRESS}^\dagger]$ can be decomposed into

$$\mathrm{E}^*_{\boldsymbol{Y}}[\mathrm{PRESS}^\dagger] = R_{\mathrm{FP}} + \mathrm{E}^*_{\boldsymbol{Y}}\mathrm{E}^*_{\boldsymbol{U}}[R_1] + 2\mathrm{E}^*_{\boldsymbol{Y}}\mathrm{E}^*_{\boldsymbol{U}}[R_2]. \tag{3.11}$$

From Appendix A.3, we can see that

$$\mathrm{E}^*_{\boldsymbol{Y}}\mathrm{E}^*_{\boldsymbol{U}}[R_1] = p + \frac{1}{n}\left\{2\kappa_4^{(1)} + p(2p + 3k + 3)\right\} + \mathrm{O}(n^{-2}), \quad \mathrm{E}^*_{\boldsymbol{Y}}\mathrm{E}^*_{\boldsymbol{U}}[R_2] = \mathrm{O}(n^{-2}). \tag{3.12}$$

Substituting (3.12) into (3.11) yields

$$\mathrm{E}^*_{\boldsymbol{Y}}[\mathrm{PRESS}^\dagger] = R_{\mathrm{FP}} + p + \frac{1}{n}\left\{2\kappa_4^{(1)} + p(2p + 3k + 3)\right\} + \mathrm{O}(n^{-2}).$$

From Yanagihara (2006b), R_{FP} is expanded as

$$R_{\mathrm{FP}} = np + \kappa_4^{(1)} + p(p + 2k + 1) + \mathrm{O}(n^{-1}).$$

Therefore, we can see that

$$a(1)\mathrm{E}^*_{\boldsymbol{Y}}[\mathrm{PRESS}^\dagger] = R_{\mathrm{FP}} + \frac{1}{n}\left\{\kappa_4^{(1)} + p(p + k + 2)\right\} + \mathrm{O}(n^{-2}).$$

11

Notice that $\text{PRESS}^+ = a(1) \sum_{i=1}^{n}\{1 + a(1)(\boldsymbol{P_X})_{ii}\}Q(\tilde{r}_i^2; a(0))$. $\text{E}_{\boldsymbol{Y}}^*[\text{PRESS}^+]$ can be expanded as

$$\text{E}_{\boldsymbol{Y}}^*[\text{PRESS}^+] = a(1)\text{E}_{\boldsymbol{Y}}^*[\text{PRESS}^\dagger] - \frac{1}{n}\text{E}_{\boldsymbol{Y}}^*\left[\sum_{i=1}^{n}\left\{(\boldsymbol{P_X})_{ii}\tilde{r}_i^2 + \frac{1}{n}\tilde{r}_i^4\right\}\right] + \text{O}(n^{-2}). \quad (3.13)$$

It is easy to obtain the following expansions:

$$\text{E}_{\boldsymbol{Y}}^*\left[\sum_{i=1}^{n}(\boldsymbol{P_X})_{ii}\tilde{r}_i^2\right] = pk + \text{O}(n^{-1}), \quad \text{E}_{\boldsymbol{Y}}^*\left[\frac{1}{n}\sum_{i=1}^{n}\tilde{r}_i^4\right] = \kappa_4^{(1)} + p(p+2) + \text{O}(n^{-1}). \quad (3.14)$$

Substituting (3.14) into (3.13) yields

$$\text{E}_{\boldsymbol{Y}}^*[\text{PRESS}^+] = R_{\text{FP}} + \text{O}(n^{-2}). \quad (3.15)$$

Notice that $\text{CAIC}_{\text{J}} = np\log 2\pi + n\log|\hat{\boldsymbol{\Sigma}}| + \{c^+/a(1)\}\text{PRESS}^+$. From Appendix A.4, we can see that $c^+/a(1) = 1 + \text{O}(n^{-3})$. This expansion and equation (3.15) make $\{c^+/a(1)\}\text{E}_{\boldsymbol{Y}}^*[\text{PRESS}^+] = R_{\text{FP}} + \text{O}(n^{-2})$. Therefore, we obtain

$$\text{E}_{\boldsymbol{Y}}^*[\text{CAIC}_{\text{J}}] = np\log 2\pi + n\text{E}_{\boldsymbol{Y}}^*\left[\log|\hat{\boldsymbol{\Sigma}}|\right] + R_{\text{FP}} + \text{O}(n^{-2}) = R_{\text{KL}} + \text{O}(n^{-2}). \quad (3.16)$$

Taking (3.16) from R_{KL} yields $B_{\text{CAIC}_{\text{J}}} = \text{O}(n^{-2})$ in a general nonnormal case.

4. Numerical Examination

In this section, we examine the numerical studies for average biases, frequencies of the model selected by the criteria, and square root mean square errors (RMSE). The information criteria used were AIC, AIC_{J} and CAIC_{J}. For numerical results of TIC and EIC, see Yanagihara (2006b).

We prepared the models with $n = 30$ as the candidate model M in (2.1). First, we constructed the $n \times 7$ explanatory variable matrix $\boldsymbol{X}_0$. The first column of this matrix was $\boldsymbol{1}_n$ and next columns were generated by $\text{U}(-1, 1)$, where $\boldsymbol{1}_n$ is an $n \times 1$ vector, and all of the vector's elements are 1. The true model was determined as

$$\boldsymbol{\eta}^* = \boldsymbol{X}_0\boldsymbol{\mu}^*\boldsymbol{1}_p', \quad \boldsymbol{\Sigma}^* = \boldsymbol{I}_p,$$

where $\boldsymbol{\mu}^* = (0, 1, 2, 3, 0, 0, 0)'$. Since the number of explanatory variables is 7, the number of all the candidate models is $2^7 - 1 = 127$. In order to save space, we classified the candidates according to the following score:

1. If the candidate model is not the overspecified model, i.e., $\boldsymbol{P_X \eta^*} \neq \boldsymbol{\eta^*}$, the score is calculated by

$$\text{score} = \frac{\text{the number of true explanatory variables contained in the candidate model}}{\text{the number of true explanatory variables } (= 3)}.$$

2. If the candidate model is the overspecified model, i.e., $\boldsymbol{P_X \eta^*} = \boldsymbol{\eta^*}$, the score is calculated by

$$\text{score} = \frac{\text{the number of explanatory variables in the candidate model } (= k)}{\text{the number of true explanatory variables } (= 3)}.$$

Table 1 shows the groups classified by the above scores and the number of candidate models included in each group. From this table, we can see that the group 4 has only one model, and that is the true model. Moreover, the models included in groups 4–8 are the overspecified models.

$$\boxed{\text{Insert Table 1 around here}}$$

Next, we set the distribution of the true model M^* in (2.3). Since our model is a multivariate nonnormal model, we prepare the data model proposed by Yuan and Bentler (1997) for generating multivariate nonnormal data.

DATA MODEL: *Let $w_1, \ldots, w_g$ $(g \geq p)$ be independent random variables with $\mathrm{E}[w_j] = 0$, $\mathrm{E}[w_j^2] = 1$ and the 4th cumulant ψ, and $\boldsymbol{w} = (w_1, \ldots, w_g)'$. Further, let r be a random variable which is independent of $\boldsymbol{w}$, $\mathrm{E}[r^2] = 1$ and the 4th moment β. Then, we generate an error vector by*

$$\boldsymbol{\varepsilon} = r\boldsymbol{H'w}, \tag{4.1}$$

where $\boldsymbol{H}$ is a $g \times p$ matrix defined by $\boldsymbol{H} = (\boldsymbol{h}_1, \ldots, \boldsymbol{h}_g)'$ with the full rank p and $\boldsymbol{H'H} = \boldsymbol{I}_p$. The multivariate kurtosis of this model becomes

$$\kappa_4^{(1)} = \beta\psi \sum_{j=1}^{g} (\boldsymbol{h}_j'\boldsymbol{h}_j)^2 + (\beta - 1)p(p+2).$$

Let χ_f be a random variable from the chi-square distribution with f degrees of freedom, and let $\boldsymbol{H}_0$ be a $(p+1) \times p$ matrix defined by

$$\boldsymbol{H}_0 = \begin{pmatrix} \boldsymbol{I}_p \\ \boldsymbol{1}_p' \end{pmatrix} (\boldsymbol{I}_p + \boldsymbol{1}_p\boldsymbol{1}_p')^{-1/2}.$$

13

Then, the error vector $\boldsymbol{\varepsilon} = r\boldsymbol{H}'_0\boldsymbol{w}$ is written as

$$\boldsymbol{\varepsilon} = r\left(\boldsymbol{I}_p - \rho\boldsymbol{1}_p\boldsymbol{1}'_p\right)\begin{pmatrix} w_1 + w_{p+1} \\ \vdots \\ w_p + w_{p+1} \end{pmatrix},$$

where

$$\rho = \frac{1}{p}\left(1 - \frac{1}{\sqrt{p+1}}\right).$$

The above equations also lead to $\sum_{j=1}^g (\boldsymbol{h}'_j\boldsymbol{h}_j)^2 = p^2/(p+1)$. By using the data model in (4.1), we generate error vectors with the following six models:

- *Model 1 (Normal Distribution)*: $w_j \sim \mathrm{N}(0,1)$, $r = 1$ and $\boldsymbol{H} = \boldsymbol{I}_p$ ($\kappa_4^{(1)} = 0$).

- *Model 2 (Laplace Distribution)*: w_j is generated from a Laplace distribution with mean 0 and standard deviation 1, $r = \sqrt{6/\chi_8^2}$ and $\boldsymbol{H} = \boldsymbol{H}_0$ ($\kappa_4^{(1)} = 4.5 \times p^2(p+1)^{-1} + p(p+2)/2$).

- *Model 3 (Uniform Distribution)*: w_j is generated from the uniform (-5,5) distribution divided by the standard deviation $5/\sqrt{3}$, $r = 1$ and $\boldsymbol{H} = \boldsymbol{H}_0$ ($\kappa_4^{(1)} = -1.2 \times p^2(p+1)^{-1}$).

- *Model 4 (Skew-Laplace Distribution)*: w_j is generated from a skew Laplace distribution with location parameter 0, dispersion parameter 1 and skew parameter 1 standardized by mean 3/4 and standard deviation $\sqrt{23}/4$, $r = \sqrt{6/\chi_8^2}$ and $\boldsymbol{H} = \boldsymbol{H}_0$ ($\kappa_4^{(1)} = 2583/529 \times p^2(p+1)^{-1} + p(p+2)/2 \approx 4.88 \times p^2(p+1)^{-1} + p(p+2)/2$).

- *Model 5 (Chi-Square Distribution)*: w_j is generated from a chi-squared distribution with 4 degrees of freedom standardized by mean 4 and standard deviation $2\sqrt{2}$, $r = \sqrt{6/\chi_8^2}$ and $\boldsymbol{H} = \boldsymbol{H}_0$ ($\kappa_4^{(1)} = 4.5 \times p^2(p+1)^{-1} + p(p+2)/2$).

- *Model 6 (Log-Normal Distribution)*: w_j is generated from a lognormal distribution such that $\log w_j \sim \mathrm{N}(0,1/4)$ standardized by mean $e^{1/8}$ and standard deviation $e^{1/8}\sqrt{e^{1/4} - 1}$, $r = \sqrt{6/\chi_8^2}$ and $\boldsymbol{H} = \boldsymbol{H}_0$ ($\kappa_4^{(1)} = 1.5 \times p^2(p+1)^{-1}(e + 2e^{3/4} + 3e^{1/2} - 6) + p(p+2)/2 \approx 8.85 \times p^2(p+1)^{-1} + p(p+2)/2$).

The skew-Laplace distribution was proposed by Balakrishnan and Ambagaspitiya (1994) (for the probability density function, see, e.g., Yanagihara & Yuan, 2005). It is easy to see that data models 1, 2 and 3 are symmetric distributions, and data models 4, 5 and 6

are skewed distributions. Moreover, the size of the kurtosis $\kappa_4^{(1)}$ in each model satisfies as follows: model 3 < model 1 < model 2 = model 5 < model 4 < model 6.

Figures 1 and 2 show R_{KL} and $\mathrm{E}_{\boldsymbol{Y}}^*[\mathrm{AIC}]$, $\mathrm{E}_{\boldsymbol{Y}}^*[\mathrm{AIC_J}]$ and $\mathrm{E}_{\boldsymbol{Y}}^*[\mathrm{CAIC_J}]$ in the cases of $p = 2$ and 6, respectively. These values were obtained for 10,000 times iterations. The transverse axis of the figures expresses the score. We drew the averages of R_{KL}, $\mathrm{E}_{\boldsymbol{Y}}^*[\mathrm{AIC}]$, $\mathrm{E}_{\boldsymbol{Y}}^*[\mathrm{AIC_J}]$ and $\mathrm{E}_{\boldsymbol{Y}}^*[\mathrm{CAIC_J}]$ of the candidate models with the same score. From these figures, we can see that the bias of $\mathrm{CAIC_J}$ is very small in the overspecified models; similarly, the bias of $\mathrm{AIC_J}$ is very small in these models. On the other hand, the bias of AIC grows when the dimension p, the score and the kurtosis $\kappa_4^{(1)}$ increase.

Insert Figures 1 and 2 around here

Table 2 shows the probabilities for selecting the model with the smallest risk. From this table, we can see that the selection-probability of $\mathrm{AIC_J}$ is improved in every distribution in comparison with AIC. All selection-probabilities of $\mathrm{CAIC_J}$ are higher than those of $\mathrm{AIC_J}$. Moreover, although the selection-probabilities of AIC are almost the same in every distribution, the probabilities of $\mathrm{AIC_J}$ and $\mathrm{CAIC_J}$ in the uniform and normal distributions, i.e., the distribution with a small $\kappa_4^{(1)}$, are remarkably high compared with those in the other distributions. In $\mathrm{AIC_J}$ and $\mathrm{CAIC_J}$, the probabilities in a skewed distribution tend to be higher than those in a non-skewed distribution.

Insert Table 2 around here

Figures 3 and 4 shows the RMSEs of AIC, $\mathrm{AIC_J}$ and $\mathrm{CAIC_J}$. From these figures, we can see that the RMSEs of $\mathrm{CAIC_J}$ get smaller than those of $\mathrm{AIC_J}$ in most cases. The difference becomes clear whenever either the score, p or $\kappa_4^{(1)}$ increases. This is mainly because $\mathrm{CAIC_J}$ is smaller than $\mathrm{AIC_J}$ when the size of these criteria is large. Notice that $Q(\tilde{r}_i^2; a(0)) < Q(\tilde{r}_i^2; 1)$ $(i = 1, \ldots, n)$ holds, where $Q(\tilde{r}_i^2; a(0))$ and $Q(\tilde{r}_i^2; 1)$ are random parts of $\hat{B}_{\mathrm{CAIC_J}}$ and $\hat{B}_{\mathrm{AIC_J}}$, respectively. This inequality causes $\mathrm{CAIC_J} < \mathrm{AIC_J}$ for a large value. Figure 5 shows scatter plots of $\mathrm{CAIC_J}$ and $\mathrm{AIC_J}$ in the normal and log-normal distributions when $p = 6$. From this figure, we can confirm that $\mathrm{CAIC_J}$ gets smaller than $\mathrm{AIC_J}$, and the difference becomes clear whenever $\mathrm{CAIC_J}$ increases. On the other hand, the RMSE of AIC tends to be smaller than those of $\mathrm{AIC_J}$ and $\mathrm{CAIC_J}$ in the case of $p = 2$, because AIC is dispersed more narrowly than $\mathrm{AIC_J}$ and $\mathrm{CAIC_J}$. The relation of RMSE

is reversed in the case $p = 6$, because then AIC has a huge bias. For a common property among all criteria, the more $\kappa_4^{(1)}$ increases, the larger the RMSE becomes. Moreover, the RMSE in a skewed distribution is likely to have a larger bias than in a non-skewed distribution.

Insert Figures 3, 4 and 5 around here

5. Conclusion

We proposed the second-order bias-corrected AIC, called CAIC$_J$, which is a bias-corrected version of AIC$_J$. Our correction merely uses the predicted residual sum of the squares (PRESS) adjusted by the constant coefficients. Also, the adjusted PRESS is obtained by estimating unknown parameters only once. Although it is not necessary to estimate the kurtosis, the bias of our criterion is improved to $\mathrm{O}(n^{-2})$ under nonnormality. CAIC$_J$ reduces AIC$_J$'s bias theoretically while maintaining the desirable behavior of AIC$_J$, in which the bias becomes numerically very small. Simultaneously, CAIC$_J$ improves the RMSE and the probability for selecting the model with the smallest risk. With regard to the nonnormal distribution, the well known C_p criterion (Mallows, 1973, 1995) should be discussed. Since there is no assumption of a concrete distribution in the definition of C_p, using C_p may be more suitable for selecting variables in nonnormal data. However, Fujikoshi and Satoh (1997) and Fujikoshi et al. (2003, 2005) reported that the selection-probability of a bias-corrected C_p is lower than that of a bias-corrected AIC. In summary, we encourage the use of our CAIC$_J$ for thoroughly searching the best combination of explanatory variables when the true distribution is not known.

In this paper, we dealt with the bias correction in overspecified models. However, there may be a bias that is not negligible when the candidate model is not an overspecified model, i.e., when the candidate model is an underspecified model. Actually, the order of the biases, $\mathrm{O}(1)$, of AIC, AIC$_J$ and CAIC$_J$ is the same in the underspecified model. Sawa (1978), Noda et al. (1996, 2003), Fujikoshi and Satoh (1997) and Reschenhofer (1999) proposed new AICs correcting the bias not only in the overspecified model but also in the underspecified model when the true distribution is the normal distribution. It is necessary to consider such a bias-corrected AIC also under nonnormality.

Appendix

A.1. On Assumptions

In order to guarantee the validity of asymptotic expansions of the bias of information criteria, we give in this paper the following assumptions which are the same as in Wakaki et al. (2002):

ASSUMPTION. *Let $\| \cdot \|$ denote the Euclidean norm. Then, the following assumptions A1, A2, A3 and A4 are:*

A1. *For some integer $s \geq 3$, $\mathrm{E}^*_{\boldsymbol{\varepsilon}}[\|\boldsymbol{\varepsilon}\|^{2s}] < \infty$,*

A2. *For some integer $s \geq 3$, $\displaystyle\limsup_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \|\boldsymbol{x}_i\|^s < \infty$,*

A3. *$\displaystyle\liminf_{n \to \infty} \frac{\lambda_n}{n} > 0$, where λ_n is the smallest eigenvalue of $\boldsymbol{X}'\boldsymbol{X}$,*

A4. *For some constant $0 < \delta \leq 1/2$, $\displaystyle\max_{i=1,\dots,n} \|\boldsymbol{x}_i\| = \mathrm{O}(n^{1/2-\delta})$.*

A.2. Decomposition of $\mathrm{E}^*_{\boldsymbol{Y}}[\mathrm{PRESS}]$

Since the distribution of $\boldsymbol{u}_i$ is the same as that of $\boldsymbol{y}_i$ and since $\boldsymbol{Y}$ and $\boldsymbol{U}$ are independent of each other, the following commutative equation, as in Yanagihara et al. (2006), holds:

$$\mathrm{E}^*_{\boldsymbol{Y}}[\mathrm{PRESS}] = \sum_{i=1}^{n} \mathrm{E}^*_{\boldsymbol{Y}} \mathrm{E}^*_{\boldsymbol{U}} \left[(\boldsymbol{u}_i - \hat{\boldsymbol{\Theta}}'_{[-i]}\boldsymbol{x}_i)' \hat{\boldsymbol{\Sigma}}^{-1}_{[-i]} (\boldsymbol{u}_i - \hat{\boldsymbol{\Theta}}'_{[-i]}\boldsymbol{x}_i) \right]. \tag{A.1}$$

From Fujikoshi et al. (2003), we can see that

$$\boldsymbol{u}_i - \hat{\boldsymbol{\Theta}}'_{[-i]}\boldsymbol{x}_i = \boldsymbol{u}_i - \hat{\boldsymbol{\Theta}}'\boldsymbol{x}_i + \frac{(\boldsymbol{P_X})_{ii}}{1 - (\boldsymbol{P_X})_{ii}}(\boldsymbol{y}_i - \hat{\boldsymbol{\Theta}}'\boldsymbol{x}_i), \tag{A.2}$$

$$\hat{\boldsymbol{\Sigma}}^{-1}_{[-i]} = a(0) \left[\hat{\boldsymbol{\Sigma}}^{-1} + \frac{1}{n\{1 - (\boldsymbol{P_X})_{ii}\}} \left(1 - \frac{\tilde{r}_i^2}{n} \right)^{-1} \hat{\boldsymbol{\Sigma}}^{-1/2} \hat{\boldsymbol{\varepsilon}}_i \hat{\boldsymbol{\varepsilon}}_i' \hat{\boldsymbol{\Sigma}}^{-1/2} \right], \tag{A.3}$$

where the coefficient $a(j)$ is given by (2.8) and $\hat{\boldsymbol{\varepsilon}}_i = \hat{\boldsymbol{\Sigma}}^{-1/2}(\boldsymbol{y}_i - \hat{\boldsymbol{\Theta}}'\boldsymbol{x}_i)$. Substituting equations (A.2) and (A.3) into (A.1) yields the equation (3.9).

A.3. Asymptotic Expansions of $\mathrm{E}^*_{\boldsymbol{Y}}\mathrm{E}^*_{\boldsymbol{U}}[R_1]$ and $\mathrm{E}^*_{\boldsymbol{Y}}\mathrm{E}^*_{\boldsymbol{U}}[R_2]$

Let $\boldsymbol{Z}$ and $\boldsymbol{V}$ be random variable matrices having asymptotic normality, where

$$\boldsymbol{Z} = (\boldsymbol{X}'\boldsymbol{X})^{-1/2}\boldsymbol{X}'\boldsymbol{\mathcal{E}}, \quad \boldsymbol{V} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i' - \boldsymbol{I}_p),$$

and let $\boldsymbol{q}_i$ be a $k \times 1$ constant vector defined by

$$\boldsymbol{q}_i = \sqrt{n}(\boldsymbol{X}'\boldsymbol{X})^{-1/2}\boldsymbol{x}_i.$$

Notice that $\boldsymbol{P_X}\boldsymbol{\eta}^* = \boldsymbol{\eta}^*$. Therefore, the following conditional expectations are obtained:

$$\mathrm{E}_{\boldsymbol{U}}^*\left[\left.(\boldsymbol{u}_i - \hat{\boldsymbol{\Theta}}'\boldsymbol{x}_i)\right|\boldsymbol{Y}\right] = -\frac{1}{\sqrt{n}}\boldsymbol{\Sigma}^{*1/2}\boldsymbol{Z}'\boldsymbol{q}_i,$$

$$\mathrm{E}_{\boldsymbol{U}}^*\left[\left.(\boldsymbol{u}_i - \hat{\boldsymbol{\Theta}}'\boldsymbol{x}_i)(\boldsymbol{u}_i - \hat{\boldsymbol{\Theta}}'\boldsymbol{x}_i)'\right|\boldsymbol{Y}\right] = \boldsymbol{\Sigma}^{*1/2}\left(\boldsymbol{I}_p + \frac{1}{n}\boldsymbol{Z}'\boldsymbol{q}_i\boldsymbol{q}_i'\boldsymbol{Z}\right)\boldsymbol{\Sigma}^{*1/2}.$$

From the above equations, the conditional expectations of R_1 and R_2 for $\boldsymbol{Y}$ are derived as

$$\mathrm{E}_{\boldsymbol{U}}^*\left[R_1|\boldsymbol{Y}\right] = W_{10} + \frac{1}{\sqrt{n}}W_{11} + \frac{1}{n}W_{12} + \mathrm{O}_p(n^{-3/2}),$$

$$\mathrm{E}_{\boldsymbol{U}}^*\left[R_2|\boldsymbol{Y}\right] = \frac{1}{\sqrt{n}}W_{21} + \frac{1}{n}W_{22} + \mathrm{O}_p(n^{-3/2}),$$

where

$$W_{10} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\varepsilon}_i'\boldsymbol{\varepsilon}_i,$$

$$W_{11} = -\frac{2}{n}\sum_{i=1}^{n}\left(\boldsymbol{q}_i'\boldsymbol{Z}\boldsymbol{\varepsilon}_i + \boldsymbol{\varepsilon}_i'\boldsymbol{V}\boldsymbol{\varepsilon}_i\right),$$

$$W_{12} = \frac{1}{n}\sum_{i=1}^{n}\left\{\boldsymbol{q}_i'\boldsymbol{q}_i\boldsymbol{\varepsilon}_i'\boldsymbol{\varepsilon}_i + \boldsymbol{q}_i'\boldsymbol{Z}\boldsymbol{Z}'\boldsymbol{q}_i + (\boldsymbol{\varepsilon}_i'\boldsymbol{\varepsilon}_i)^2 + 2\boldsymbol{\varepsilon}_i'\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{\varepsilon}_i\right.$$

$$\left.+(\boldsymbol{q}_i'\boldsymbol{Z}\boldsymbol{\varepsilon}_i)^2 + 4\boldsymbol{q}_i'\boldsymbol{Z}\boldsymbol{V}\boldsymbol{\varepsilon}_i + 3\boldsymbol{\varepsilon}_i'\boldsymbol{V}^2\boldsymbol{\varepsilon}_i\right\},$$

$$W_{21} = -\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{q}_i'\boldsymbol{q}_i\boldsymbol{q}_i'\boldsymbol{Z}\boldsymbol{\varepsilon}_i,$$

$$W_{22} = -\frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{q}_i'\boldsymbol{q}_i\boldsymbol{q}_i'\boldsymbol{Z}\boldsymbol{Z}'\boldsymbol{q}_i - \boldsymbol{q}_i'\boldsymbol{q}_i\boldsymbol{q}_i'\boldsymbol{Z}\boldsymbol{V}\boldsymbol{\varepsilon}_i\right).$$

Therefore, the asymptotic expansions of $\mathrm{E}_{\boldsymbol{Y}}^*\mathrm{E}_{\boldsymbol{U}}^*[R_1]$ and $\mathrm{E}_{\boldsymbol{Y}}^*\mathrm{E}_{\boldsymbol{U}}^*[R_2]$ are rewritten as

$$\mathrm{E}_{\boldsymbol{Y}}^*\mathrm{E}_{\boldsymbol{U}}^*[R_1] = \mathrm{E}_{\boldsymbol{Y}}^*[W_{10}] + \frac{1}{\sqrt{n}}\mathrm{E}_{\boldsymbol{Y}}^*[W_{11}] + \frac{1}{n}\mathrm{E}_{\boldsymbol{Y}}^*[W_{12}] + \mathrm{O}(n^{-2}), \tag{A.4}$$

$$\mathrm{E}_{\boldsymbol{Y}}^*\mathrm{E}_{\boldsymbol{U}}^*[R_2] = \frac{1}{\sqrt{n}}\mathrm{E}_{\boldsymbol{Y}}^*[W_{21}] + \frac{1}{n}\mathrm{E}_{\boldsymbol{Y}}^*[W_{22}] + \mathrm{O}(n^{-2}). \tag{A.5}$$

Notice that

$$\mathrm{E}_{\boldsymbol{Y}}^* \left[\frac{1}{n} \sum_{i=1}^{n} \varepsilon_i' \varepsilon_i \right] = p,$$

$$\mathrm{E}_{\boldsymbol{Y}}^* \left[\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{q}_i' \boldsymbol{q}_i \varepsilon_i' \varepsilon_i \right] = \frac{p}{n} \sum_{i=1}^{n} \boldsymbol{q}_i' \boldsymbol{q}_i = pk,$$

$$\mathrm{E}_{\boldsymbol{Y}}^* \left[\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{q}_i' \boldsymbol{Z} \varepsilon_i \right] = \frac{p}{n\sqrt{n}} \sum_{i=1}^{n} \boldsymbol{q}_i' \boldsymbol{q}_i = \frac{1}{\sqrt{n}} pk,$$

$$\mathrm{E}_{\boldsymbol{Y}}^* \left[\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{q}_i' \boldsymbol{q}_i \boldsymbol{q}_i' \boldsymbol{Z} \varepsilon_i \right] = \frac{p}{n\sqrt{n}} \sum_{i=1}^{n} (\boldsymbol{q}_i' \boldsymbol{q}_i)^2 = \frac{1}{\sqrt{n}} pq,$$

$$\mathrm{E}_{\boldsymbol{Y}}^* \left[\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{q}_i' \boldsymbol{Z} \boldsymbol{Z}' \boldsymbol{q}_i \right] = \frac{p}{n^2} \sum_{ij}^{n} (\boldsymbol{q}_i' \boldsymbol{q}_j)^2 = pk,$$

$$\mathrm{E}_{\boldsymbol{Y}}^* \left[\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{q}_i' \boldsymbol{q}_i \boldsymbol{q}_i' \boldsymbol{Z} \boldsymbol{Z}' \boldsymbol{q}_i \right] = \frac{p}{n^2} \sum_{ij}^{n} \boldsymbol{q}_i' \boldsymbol{q}_i (\boldsymbol{q}_i' \boldsymbol{q}_j)^2 = pq,$$

$$\mathrm{E}_{\boldsymbol{Y}}^* \left[\frac{1}{n} \sum_{i=1}^{n} (\varepsilon_i' \varepsilon_i)^2 \right] = \kappa_4^{(1)} + p(p+2),$$

$$\mathrm{E}_{\boldsymbol{Y}}^* \left[\frac{1}{n} \sum_{i=1}^{n} \varepsilon_i' \boldsymbol{V} \varepsilon_i \right] = \frac{1}{\sqrt{n}} \left\{ \kappa_4^{(1)} + p(p+1) \right\},$$

$$\mathrm{E}_{\boldsymbol{Y}}^* \left[\frac{1}{n} \sum_{i=1}^{n} \varepsilon_i' \boldsymbol{Z}' \boldsymbol{Z} \varepsilon_i \right] = \frac{p}{n} \sum_{i=1}^{n} \boldsymbol{q}_i' \boldsymbol{q}_i + \frac{1}{n^2} \{ \kappa_4^{(1)} + p(p+1) \} \sum_{i=1}^{n} \boldsymbol{q}_i' \boldsymbol{q}_i = pk + \mathrm{O}(n^{-1}),$$

$$\mathrm{E}_{\boldsymbol{Y}}^* \left[\frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{q}_i' \boldsymbol{Z}' \varepsilon_i)^2 \right] = \frac{p}{n^2} \sum_{ij}^{n} (\boldsymbol{q}_i' \boldsymbol{q}_j)^2 + \frac{1}{n^2} \{ \kappa_4^{(1)} + p(p+1) \} \sum_{i=1}^{n} (\boldsymbol{q}_i' \boldsymbol{q}_i)^2 = pk + \mathrm{O}(n^{-1}),$$

$$\mathrm{E}_{\boldsymbol{Y}}^* \left[\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{q}_i' \boldsymbol{Z}' \boldsymbol{V} \varepsilon_i \right] = \frac{1}{n^2} \{ \kappa_4^{(1)} + p(p+1) \} \sum_{i=1}^{n} \boldsymbol{q}_i' \boldsymbol{q}_i = \mathrm{O}(n^{-2}),$$

$$\mathrm{E}_{\boldsymbol{Y}}^* \left[\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{q}_i' \boldsymbol{q}_i \boldsymbol{q}_i' \boldsymbol{Z}' \boldsymbol{V} \varepsilon_i \right] = \frac{1}{n^2} \{ \kappa_4^{(1)} + p(p+1) \} \sum_{i=1}^{n} (\boldsymbol{q}_i' \boldsymbol{q}_i)^2 = \mathrm{O}(n^{-2}),$$

$$\mathrm{E}_{\boldsymbol{Y}}^* \left[\frac{1}{n} \sum_{i=1}^{n} \varepsilon_i' \boldsymbol{V}^2 \varepsilon_i \right] = \kappa_4^{(1)} + p(p+1) + \mathrm{O}(n^{-1}),$$

where the coefficient q is defined by

$$q = n \sum_{i=1}^{n} \{ (\boldsymbol{P}_{\boldsymbol{X}})_{ii} \}^2 .$$

From the above expectations, we obtain the following expectations:

$$\mathrm{E}_{\boldsymbol{Y}}^* [W_{10}] = p,$$

$$\mathrm{E}_{\boldsymbol{Y}}^* [W_{11}] = -\frac{2}{\sqrt{n}} \left\{ \kappa_4^{(1)} + p(p+k+1) \right\},$$

$$\mathrm{E}_{\boldsymbol{Y}}^* [W_{12}] = 4\kappa_4^{(1)} + p(4p + 5k + 5) + \mathrm{O}(n^{-1}), \tag{A.6}$$

19

$$E_{\boldsymbol{Y}}^*[W_{21}] = -\frac{1}{\sqrt{n}}pq,$$

$$E_{\boldsymbol{Y}}^*[W_{22}] = pq + O(n^{-1}).$$

By substituting expectations (A.6) into (A.4) and (A.5) we obtain the asymptotic expansions in (3.12).

A.4. Asymptotic Expansion of $c^+/a(1)$

Dividing the previous part of the right-hand side of (3.2) by $a(1)$, which does not include the gamma function, is easily expanded as

$$\frac{(n+k)\{n-k-p-2a(0)\}}{a(1)\{n+a(1)k\}(n-k-p-1)} = 1 - \frac{p}{n^2} + O(n^{-3}).$$

In order to expand the remaining part, which includes the gamma function, we apply the following Stirling formula (see, e.g., Barndorff-Nielsen & Cox, 1989, pp. 62–65) to this part:

$$\Gamma(x) = \sqrt{2\pi}e^{-x}x^{x-1/2}\left\{1 + \frac{1}{12x} + \frac{1}{288x^2} + O(x^{-3})\right\}.$$

Then, it can be rewritten by the product of G_1 and G_2 as

$$\frac{\Gamma\left(\frac{n-k}{2} + \frac{1}{n}\right)\Gamma\left(\frac{n-k-p}{2}\right)}{\Gamma\left(\frac{n-k}{2}\right)\Gamma\left(\frac{n-k-p}{2} + \frac{1}{n}\right)} = G_1 G_2,$$

where

$$G_1 = \frac{\left(\frac{n-k}{2} + \frac{1}{n}\right)^{\{n(n-k-1)+2\}/(2n)}\left(\frac{n-k-p}{2}\right)^{(n-k-p-1)/2}}{\left(\frac{n-k}{2}\right)^{(n-k-1)/2}\left(\frac{n-k-p}{2} + \frac{1}{n}\right)^{\{n(n-k-p-1)+2\}/(2n)}},$$

$$G_2 = \frac{\left\{1 + \frac{n}{6(n^2-kn+2)} + \frac{n^2}{72(n^2-kn+2)^2} + O(n^{-3})\right\}\left\{1 + \frac{1}{6(n-k-p)} + \frac{1}{72(n-k-p)^2} + O(n^{-3})\right\}}{\left\{1 + \frac{1}{6(n-k)} + \frac{1}{72(n-k)^2} + O(n^{-3})\right\}\left\{1 + \frac{n}{6(n^2-kn-pn+2)} + \frac{n^2}{72(n^2-kn-pn+2)^2} + O(n^{-3})\right\}}.$$

For G_1, the first part in the numerator is asymptotically expanded as

$$\left(\frac{n-k}{2} + \frac{1}{n}\right)^{\{n(n-k-1)+2\}/(2n)} = 1 + \frac{1}{n} - \frac{1}{2n^2} + O(n^{-3}).$$

In a similar way, G_1 is asymptotically expanded as

$$G_1 = 1 + \frac{p}{n^2} + O(n^{-3}).$$

Then, G_2 is asymptotically expanded as

$$G_2 = 1 + O(n^{-3}).$$

20

Hence, $c^+/a(1)$ is asymptotically expanded as

$$\frac{c^+}{a(1)} = G_1 G_2 \left\{ 1 - \frac{p}{n^2} + \mathrm{O}(n^{-3}) \right\} = 1 + \mathrm{O}(n^{-3}).$$

Acknowledgments

References

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

[2] Balakrishnan, N. & Ambagaspitiya, R. S. (1994). On skew Laplace distribution. *Technical Report, Department of Mathematics & Statistics, McMaster University*, Hamilton, Ontario, Canada.

[3] Barndorff-Nielsen, O. E. & Cox, D. R. (1989). *Asymptotic Techniques for Use in Statistics*. Chapman & Hall/CRC, London.

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

[5] Fujikoshi, Y., Yanagihara, H., & Wakaki, H. (2005). Bias corrections of some criteria for selection multivariate linear regression models in a general case. *Amer. J. Math. Management Sci.*, **25** (in press).

[6] Fujikoshi, Y., Noguchi, T., Ohtaki, M., & Yanagihara, H. (2003). Corrected versions of cross-validation criteria for selecting multivariate regression and growth curve models. *Ann. Inst. Statist. Math.*, **55**, 537-553.

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

[8] Isogai, T. (1983). On measures of multivariate skewness and kurtosis. *Math. Japon.*, **28**, 251–261.

[9] Kale, B. K. & Sebastian, G. (1996). On a class of symmetric nonnormal distributions with a kurtosis of three. In *Statistical Theory and Applications* (H. N. Nagaraja, P. K. Sen & D. F. Morrison eds.), 55–63, Springer, New York.

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

[11] Mallows, C. L. (1973). Some comments on C_p. *Technometrics*, **15**, 661–675.

[12] Mallows, C. L. (1995). More comments on C_p. *Technometrics*, **37**, 362–372.

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

[14] Noda, K., Miyaoka, E., & Itoh, M. (1996). On bias correction of the Akaike information criterion in linear models. *Comm. Statist. Theory Methods*, **25**, 1845–1857.

[15] Noda, K., Wang, J., Takahashi, R., & Itoh, M. (2003). Improvement to AIC as estimator of Kullback-Leibler information for linear model selection. *Comm. Statist. Theory Methods*, **32**, 2207–2225.

[16] Reschenhofer, E. (1999). Improved estimation of the expected Kullback-Leibler discrepancy in case of misspecification. *Econometric Theory*, **15**, 377–387.

[17] Sawa, T. (1978). Information criteria for discriminating among alternative regression models. *Econometrica*, **46**, 1273–1291.

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

[19] Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *J. Roy. Statist. Soc. Ser.* **B**, **36**, 111–147.

[20] Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *J. Roy. Statist. Soc. Ser.* **B**, **39**, 44–47.

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

[22] Wakaki, H., Yanagihara, H., & Fujikoshi, Y. (2002). Asymptotic expansion of the null distributions of test statistics for multivariate linear hypothesis under nonnormality. *Hiroshima Math. J.*, **32**, 17–50.

[23] Yanagihara, H. (2006a). A family of estimators for multivariate kurtosis in a nonnormal linear regression model. *J. Multivariate Anal.* (in press).

[24] Yanagihara, H. (2006b). Corrected version of *AIC* for selecting multivariate normal linear regression models in a general nonnormal case. *J. Multivariate Anal.* (in press).

[25] Yanagihara, H. & Yuan, K.-H. (2005). Four improved statistics for contrasting means by correcting skewness and kurtosis. *British J. Math. Statist. Psych.*, **58**, 209–237.

[26] Yanagihara, H., Tonda, T., & Matsumoto, C. (2006). Bias correction of cross-validation criterion based on Kullback-Leibler information under a general condition. *J. Multivariate Anal.* (in press).

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

TABLE 1. The number of candidate models included in the group

Group	1	2	3	4*	5*	6*	7*	8*
Score	0	1/3	2/3	1	4/3	5/3	2	7/3
No. of Candidates	15	48	48	1	4	6	4	1

* denotes the group which consists of only overspecified models.

TABLE 2. Probabilities for selecting the model with the smallest risk (%)

Distribution	$p = 2$			$p = 6$		
	AIC	AIC_J	CAIC_J	AIC	AIC_J	CAIC_J
Normal	41.13	61.55	63.02	46.47	87.89	88.83
Laplace	40.53	49.47	50.78	47.64	63.86	65.22
Uniform	40.96	64.94	66.16	46.55	93.74	94.30
Skew-Laplace	40.46	51.28	52.83	45.76	66.87	68.12
Chi-Square	40.01	51.85	53.13	45.89	69.37	70.54
Log-Normal	40.05	49.81	51.30	46.62	66.08	67.53
Average	40.52	54.82	56.20	46.49	74.64	75.76
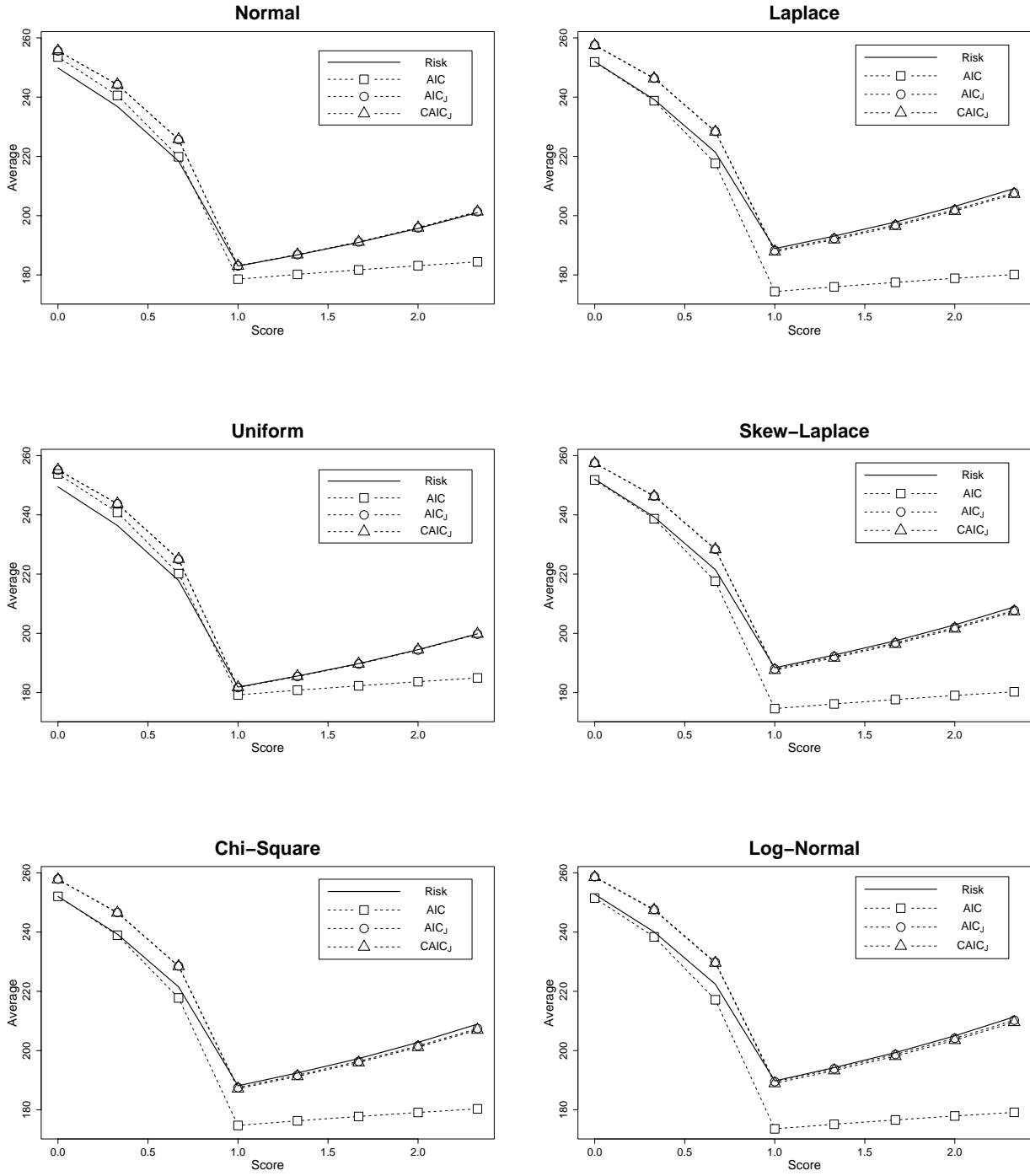
FIGURE 1. R_{KL} and $\mathrm{E}_{\boldsymbol{Y}}^*[\mathrm{AIC}]$, $\mathrm{E}_{\boldsymbol{Y}}^*[\mathrm{AIC_J}]$ and $\mathrm{E}_{\boldsymbol{Y}}^*[\mathrm{CAIC_J}]$ $(p = 2)$
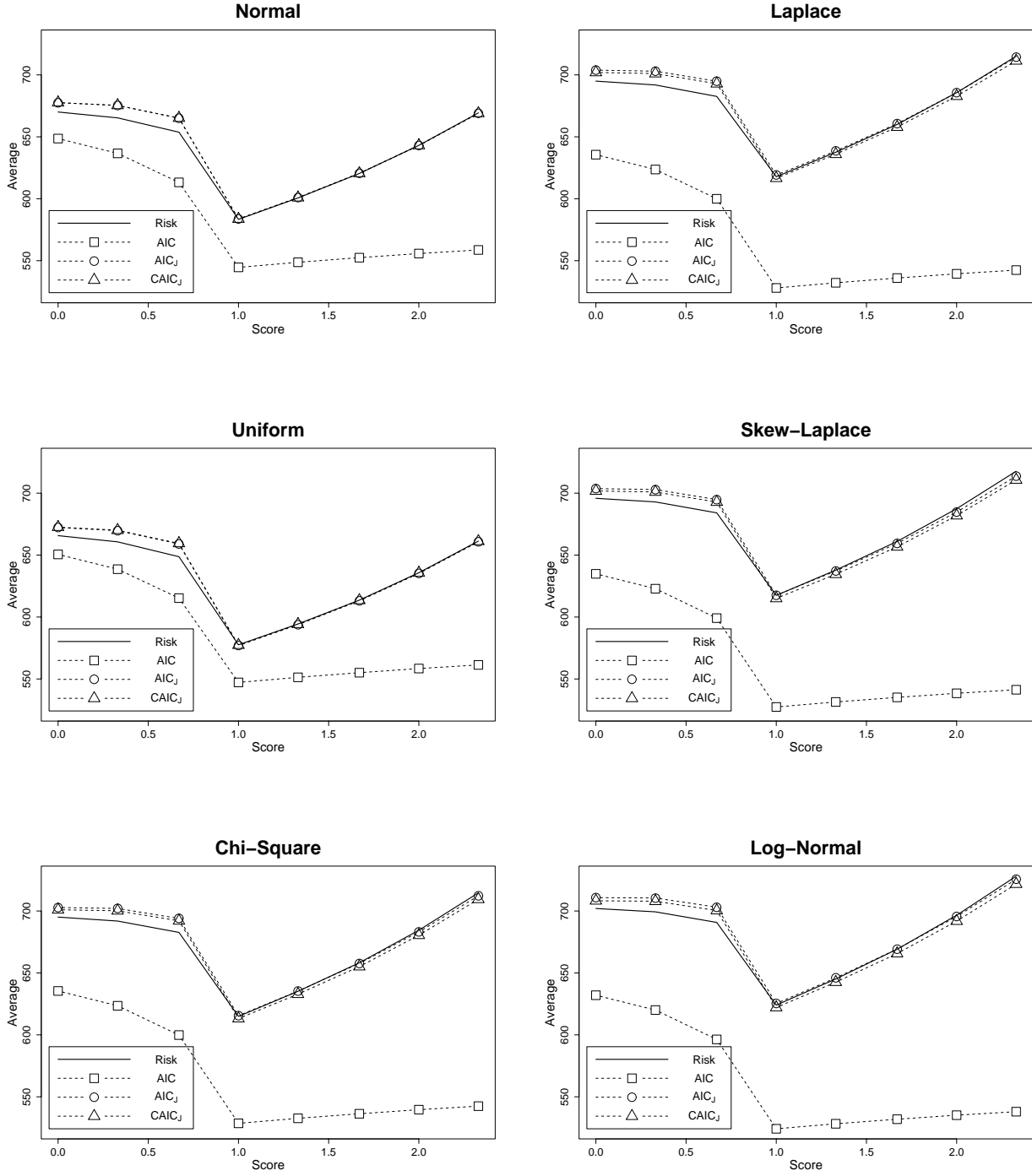
FIGURE 2. R_{KL} and $\mathrm{E}_{\boldsymbol{Y}}^{*}[\mathrm{AIC}]$, $\mathrm{E}_{\boldsymbol{Y}}^{*}[\mathrm{AIC_J}]$ and $\mathrm{E}_{\boldsymbol{Y}}^{*}[\mathrm{CAIC_J}]$ $(p = 6)$
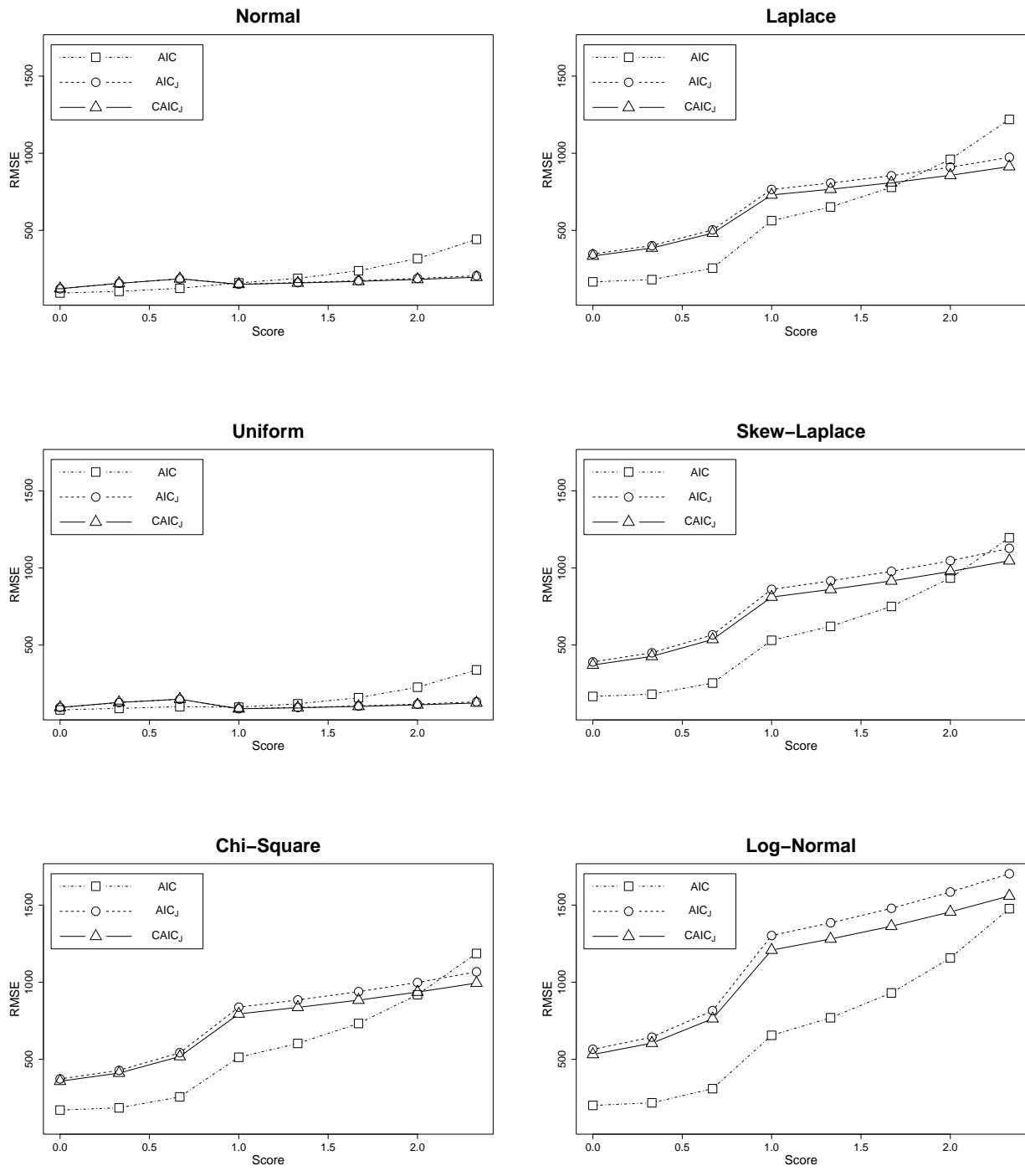
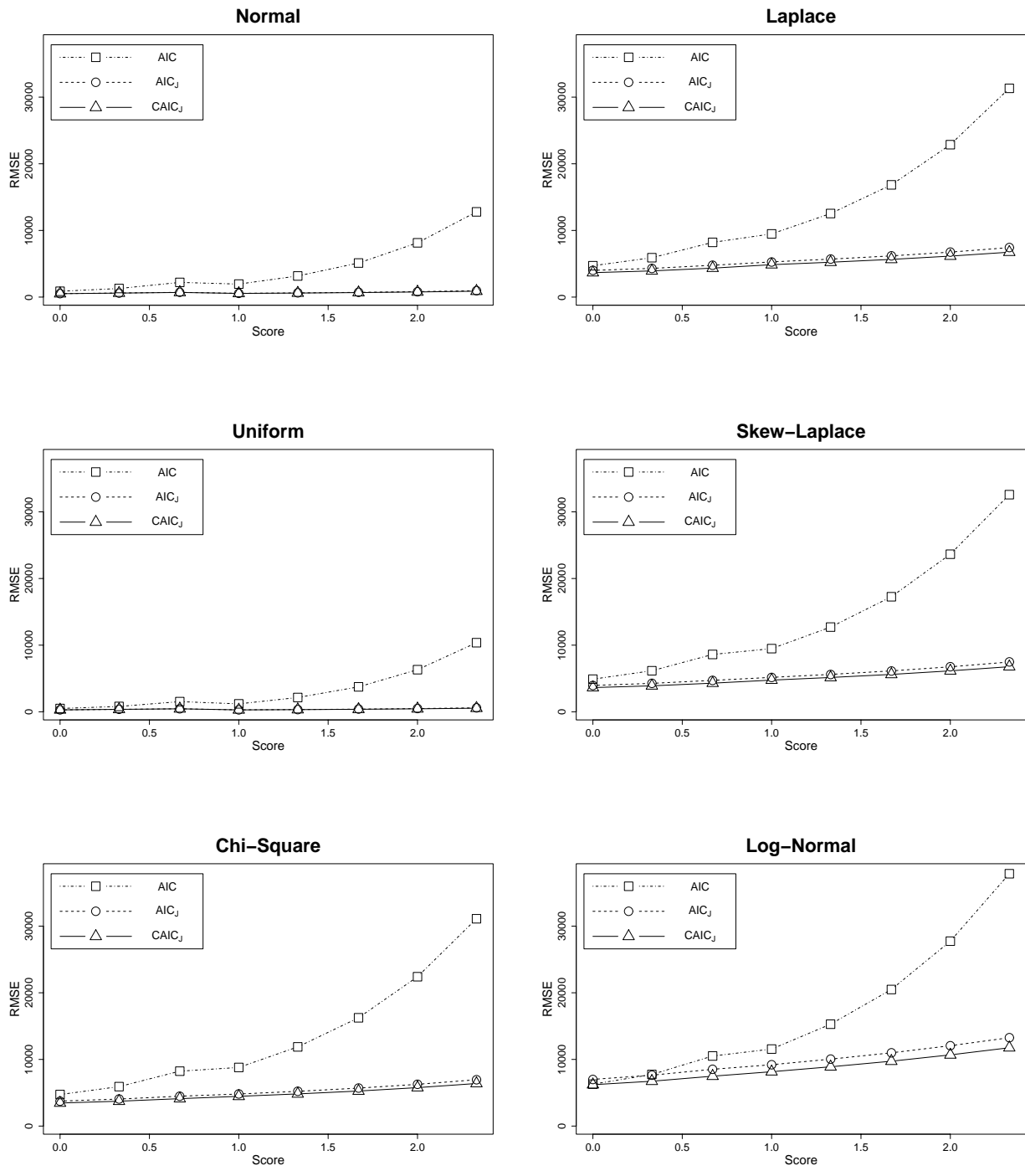FIGURE 3. RMSE of AIC, AIC$_J$ and CAIC$_J$ ($p = 2$)
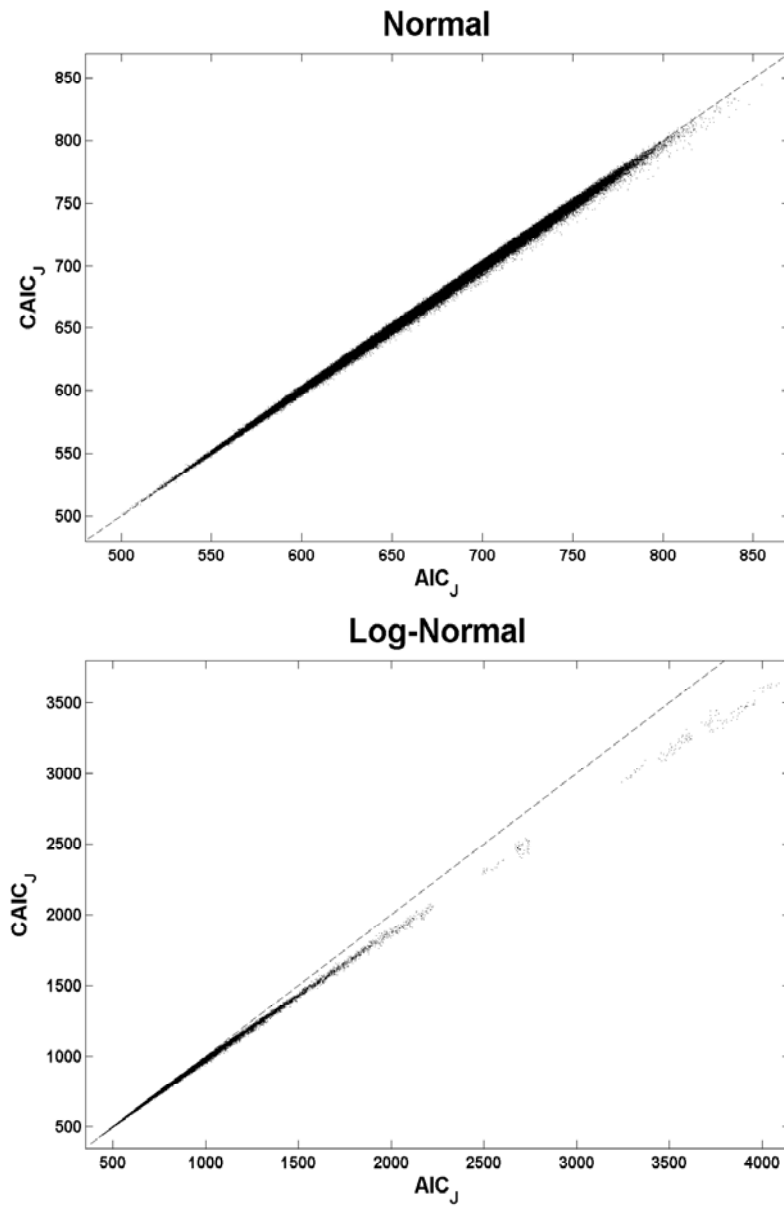
FIGURE 4. RMSE of AIC, AIC_J and $CAIC_J$ $(p = 6)$

FIGURE 5. Scatter plots of AIC$_J$ and CAIC$_J$ ($p = 6$)