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

Hirokazu Yanagihara\textsuperscript{1}, Ken-ichi Kamo\textsuperscript{2} and Tetsuji Tonda\textsuperscript{3}

\textsuperscript{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

\textsuperscript{2}Division of Mathematics
School of Medicine Liberal Arts and Sciences
Sapporo Medical University
South 1, West 17, Chuoku, Sapporo 060-8543, Japan

\textsuperscript{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.

AMS 2000 subject classifications. Primary 62H12; Secondary 62F07.
Key words: Akaike’s information criterion, Bias correction, Jackknife method, Kullback-Leibler information, Model misspecification, Normal assumption, Overspecified model, Selection of variables, Predicted residual sum of squares, Robustness.

\textsuperscript{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 \( y = (y_1, \ldots, y_p)' \). Since we would like to specify the factors affecting \( y \) 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 \( y \) 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 \( -2 \times \text{the sample log-likelihood} \); 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\textsubscript{J}). Although AIC\textsubscript{J} is the first-order bias-corrected AIC, as is TIC and EIC, the bias of AIC\textsubscript{J} tends to be smaller than the others. Especially, we can see that the bias of AIC\textsubscript{J} becomes very small numerically. However, the $n^{-1}$ term in the bias, which consists of the multivariate kurtosis, still remains in AIC\textsubscript{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\textsubscript{J} to $O(n^{-2})$ 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\textsubscript{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\textsubscript{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\textsubscript{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 $Y = (y_1, \ldots, y_n)'$ be an $n \times p$ observation matrix. We consider the following candidate model with $k$ explanatory variables which are the subset of $\{x_1, \ldots, x_K\}$.

$$M : Y \sim N_{n \times p}(X \Theta, \Sigma \otimes I_n), \quad (2.1)$$

where $X = (x_1, \ldots, x_n)'$ is an $n \times k$ matrix of nonstochastic explanatory variables with the full rank $k$. Then, the joint probability density function of $Y$ is shown by
\[ \Pi_{i=1}^{n} f(y_i|x_i, \Theta, \Sigma), \]

where

\[ f(y_i|x_i, \Theta, \Sigma) = \left( \frac{1}{\sqrt{2\pi}} \right)^{p/2} |\Sigma|^{-1/2} \exp \left\{ -\frac{1}{2} (y_i - \Theta' x_i)' \Sigma^{-1} (y_i - \Theta' x_i) \right\}. \]  

(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 \( Y \) may be not distributed according to the normal distribution. That is, the true model is denoted by

\[ M^* : Y = \eta^* + \mathcal{E} \Sigma^{*1/2}, \]  

(2.3)

where \( \eta^* \) and \( \Sigma^* \) are \( n \times p \) and \( p \times p \) unknown parameter matrices, respectively, and \( p \times 1 \) row vectors \( \varepsilon_1, \ldots, \varepsilon_n \) of error matrix \( \mathcal{E} \) are independent, and copies of a random variable vector \( \varepsilon \) which is distributed following an unknown distribution with the mean \( 0 \) and covariance matrix \( I_p \). In particular, we will assume the candidate model \( M \) is the overspecified model, satisfying \( P_X \eta^* = \eta^* \), where \( P_X \) is the projection matrix of \( X \), i.e., \( P_X = X(X'X)^{-1}X' \). 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 \( X \) in \( M \) and \( \mathcal{E} \) in \( M^* \) (see Appendix A.1) hold to guarantee valid expansions on the biases of information criteria.

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

\[ \hat{\Theta} = (X'X)^{-1}X'Y, \quad \hat{\Sigma} = \frac{1}{n} Y' (I_n - P_X) Y. \]

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

\[ -2\ell(\hat{\Theta}, \hat{\Sigma}|Y, X) = -2 \sum_{i=1}^{n} \log f(y_i|x_i, \hat{\Theta}, \hat{\Sigma}) \]

\[ = np \log 2\pi + n \log |\hat{\Sigma}| + np. \]  

(2.4)

Let \( U = (u_1, \ldots, u_n)' \) be an \( n \times p \) future observation matrix, which is independent of \( Y \), and let \( p \times 1 \) row vectors \( u_1, \ldots, u_n \) of \( U \) be independently distributed according to the same distributions of \( y_1, \ldots, y_n \), respectively. Then, the risk based on the predictive K-L information measuring the discrepancy between \( M^* \) and \( M \) under the normal assumption

4
is defined by

\[
R_{\text{KL}} = E_Y^*E_U^* \left[ -2\ell(\hat{\Theta}, \hat{\Sigma}|U, X) \right] = np \log 2\pi + nE_Y^* \left[ \log |\hat{\Sigma}| \right] + R_{\text{FP}},
\]

(2.5)

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

\[
R_{\text{FP}} = E_U^*E_Y^* \left[ \sum_{i=1}^{n} \left( u_i - \hat{\Theta}'x_i \right)'\hat{\Sigma}^{-1} \left( u_i - \hat{\Theta}'x_i \right) \right].
\]

(2.6)

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

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

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

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 \(-2\ell(\hat{\Theta}, \hat{\Sigma}|Y, X)\), and thus, the AIC-type criterion is specified by the individual \( \hat{B} \).

### 2.2. Akaike’s Information Criterion

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

\[
\text{AIC} = -2\ell(\hat{\Theta}, \hat{\Sigma}|Y, X) + \hat{B}_{\text{AIC}} = np(\log 2\pi + 1) + n\log |\hat{\Sigma}| + 2pk + p(p+1).
\]

(2.7)

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

\[
B_{\text{AIC}} = R_{\text{KL}} - E_Y^*[\text{AIC}] = \kappa_{4}^{(1)} + O(n^{-1}),
\]

where \( \kappa_{4}^{(1)} \) is the multivariate kurtosis (see, e.g., Mardia, 1970; Isogai, 1983), which is defined by

\[
\kappa_{4}^{(1)} = E_\varepsilon^* \left[ (\varepsilon'\varepsilon)^2 \right] - p(p+2).
\]

Since Akaike evaluated \( \hat{B}_{\text{AIC}} \) under the assumption that the candidate model is not misspecified, \( \kappa_{4}^{(1)} \) appears in the top term of \( B_{\text{AIC}} \). Notice that \( \kappa_{4}^{(1)} \) measures the variation from the normal distribution. Thus, the breakdown of normality makes the constant bias
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 $Y_{(-i)}$ and $X_{(-i)}$ be $(n - 1) \times p$ and $(n - 1) \times k$ matrices obtained from $Y$ and $X$ by deleting $y_i$ and $x_i$, respectively. Further, let the MLEs of $\Theta$ and $\Sigma$ evaluated from $Y_{(-i)}$ and $X_{(-i)}$ be denoted as $\hat{\Theta}_{[-i]}$ and $\hat{\Sigma}_{[-i]}$, respectively, which are defined by

$$
\hat{\Theta}_{[-i]} = \left\{X'_{(-i)}X_{(-i)}\right\}^{-1}X'_{(-i)}Y_{(-i)}, \quad \hat{\Sigma}_{[-i]} = \frac{1}{n}Y'_{(-i)}\left\{I_{n-1} - PX_{(-i)}\right\}Y_{(-i)}.
$$

Then, PRESS is expressed as

$$
PRESS = \sum_{i=1}^{n}(y_i - \hat{\Theta}'_{[-i]}x_i)\hat{\Sigma}^{-1}_{[-i]}(y_i - \hat{\Theta}'_{[-i]}x_i).
$$

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

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

(2.8)

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

$$
\hat{r}_i^2 = \frac{1}{1 - (P_{X})_{ii}}(y_i - \hat{\Theta}'x_i)'\hat{\Sigma}^{-1}(y_i - \hat{\Theta}'x_i),
$$

(2.9)

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

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

(2.10)

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

$$
PRESS = a(0)\sum_{i=1}^{n}\{(1 - (P_{X})_{ii})\}^{-1}Q(\hat{r}_i^2; 1).
$$

(2.11)
Yanagihara (2006b) evaluated $R_{FP}$ by using the jackknife method, i.e., using PRESS, and approximated $B$ by $\hat{B}_{AIC_J} = \{c/a(0)\} \text{PRESS} - np$, where the coefficient $c$ is defined by

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

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

$$\text{AIC}_J = -2\ell(\Theta, \Sigma|Y, X) + \hat{B}_{AIC_J}$$

$$= np \log 2\pi + n \log |\hat{\Sigma}| + c \sum_{i=1}^{n} \{1 - (P_X)_{ii}\}^{-1} Q(r_i; 1). \quad (2.13)$$

From Yanagihara (2006b), the bias of AIC$_J$ becomes as follows:

$$B_{AIC_J} = R_{KL} - E_Y[AIC_J] = \left\{ \begin{array}{ll} 0 & (\varepsilon_1, \ldots, \varepsilon_n \sim \text{i.i.d.} \ N_p(0, I_p)) \\ -\frac{1}{n^{k_4(1)}} + O(n^{-2}) & (\text{otherwise}) \end{array} \right.$$.

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 $\hat{B}_{AIC}$ 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. $N_p(0, I_p)$. 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

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

(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:

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

(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. $N_p(0, I_p)$. 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

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

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

Therefore, $Q(\tilde{r}_i^2, 1 - d/n)$ means the $i$th squared predictive residual in which the influence of $y_i$ remains for a while. Actually, we decide $b_i = a(1)a(1)(P_X)_{ii}$ and $d = 1$, where
the coefficient $a(j)$ is defined by (2.8). Finally, by using equation (3.1), we approximate $B$ by

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

where the coefficient $c^+$ is defined by

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

(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:

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

(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

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

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

(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

$$
B_{\text{CAIC}_J} = R_{\text{KL}} - E_Y[\text{CAIC}_J] = \begin{cases} 
0 & (\varepsilon_1, \ldots, \varepsilon_n \sim \text{i.i.d. } N_p(0, I_p)) \\
O(n^{-2}) & \text{(otherwise)}
\end{cases}
$$

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

### 3.2. Proof of Theorem

First, we show $B_{\text{CAIC}_J} = 0$ when $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. $N_p(0, I_p)$. 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 \( \varepsilon_1, \ldots, \varepsilon_n \) are \( i.i.d. N_p(0, I_p) \), then \( \tilde{r}_i^2 \) is rewritten by \( T \) as
\[
\tilde{r}_i^2 = \frac{n}{N} T \left( 1 + \frac{T}{N} \right)^{-1}.
\]
(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:
\[
E_Y^*[Q(\tilde{r}_i^2; a(0))] = \int_0^{\infty} \frac{n \Gamma \left( \frac{N+1}{2} \right)}{N \Gamma \left( \frac{p}{2} \right) \Gamma \left( \frac{N-p+1}{2} \right)} \left( \frac{t}{N} \right)^{(p+2)/2-1} \left( 1 + \frac{t}{N} \right)^{-(N+2+n+1)/2} dt.
\]
Let \( N_0 = N + 2/n \) and \( p_0 = p + 2 \). Then the following equations hold:
\[
\Gamma \left( \frac{p_0}{2} \right) = \frac{p_0}{2} \Gamma \left( \frac{p}{2} \right), \quad \Gamma \left( \frac{N_0 - p + 1}{2} \right) = \frac{N_0 - p}{2} \Gamma \left( \frac{N_0 - p + 1}{2} \right).
\]
By using the above equations and replacing \( t/N \) with \( t/N_0 \), we derive
\[
E_Y^*[Q(\tilde{r}_i^2; a(0))] = \frac{np \Gamma \left( \frac{N+1}{2} \right) \Gamma \left( \frac{N_0 - p + 1}{2} \right)}{(N_0 - p - 1) \Gamma \left( \frac{N_0 + 1}{2} \right) \Gamma \left( \frac{N-p+1}{2} \right)} \int_0^{\infty} g(t; N_0, p_0) dt
\]
\[
= \frac{n(n+k)p}{c^+ \{ n + a(1)k \} (n - k - p - 1)}.
\]
(3.5)
Notice that
\[
E_Y^*[\text{CAIC}_j] = np \log 2\pi + nE_Y^*[\log |\hat{\Sigma}|] + c^+ \sum_{i=1}^{n} \{ 1 + a(1)(P_{X})_{ii} \} E_Y^*[Q(\tilde{r}_i^2; a(0))].
\]
(3.6)
Substituting (3.5) into (3.6) yields
\[
E_Y^*[\text{CAIC}_j] = np \log 2\pi + nE_Y^*[\log |\hat{\Sigma}|] + \frac{n(n+k)p}{n - k - p - 1}.
\]
(3.7)
From the result in Fujikoshi and Satoh (1997), \( R_{KL} \) can be precisely written as
\[
R_{KL} = np \log 2\pi + nE_Y^*[\log |\hat{\Sigma}|] + \frac{n(n+k)p}{n - k - p - 1}.
\]
(3.8)
We can see that equation (3.8) corresponds to (3.7). Therefore, the bias of CAIC\textsubscript{J} is exactly 0 when \(\varepsilon_1, \ldots, \varepsilon_n\) are i.i.d. \(N_p(0, I_p)\).

Next, we show \(B_{\text{CAIC}\textsubscript{J}} = O(n^{-2})\) in a general nonnormal case. By using the commutative formula, as in Yanagihara et al. (2006), we can decompose \(E_Y^*\text{[PRESS]}\) into

\[
E_Y^*\text{[PRESS]} = a(0) \{ R_{FP} + E_Y^*E_U^*[R_1] + 2E_Y^*E_U^*[R_2] + E_Y^*[R_3] \},
\]

where \(R_1, R_2\) and \(R_3\) are defined by

\[
R_1 = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{1 - (P_x)_{ii}} \left\{ (u_i - \hat{\theta}'x_i)'\Sigma^{-1}(y_i - \hat{\theta}'x_i) \right\}^2 \left( 1 - \frac{\bar{r}_i^2}{n} \right)^{-1},
\]

\[
R_2 = \sum_{i=1}^{n} \frac{(P_x)_{ii}}{1 - (P_x)_{ii}} (u_i - \hat{\theta}'x_i)'\Sigma^{-1}(y_i - \hat{\theta}'x_i) \left( 1 - \frac{\bar{r}_i^2}{n} \right)^{-1},
\]

\[
R_3 = \sum_{i=1}^{n} \frac{(P_x)_{ii}^2}{1 - (P_x)_{ii}} Q(\bar{r}_i^2; 1),
\]

(The derivation of this decomposition is described in Appendix A.2.) Let’s define PRESS\textsuperscript{†} as

\[
\text{PRESS}^\dagger = \frac{1}{a(0)} \text{PRESS} - R_3 = \sum_{i=1}^{n} \{ 1 + (P_x)_{ii} \} Q(\bar{r}_i^2; 1).
\]

Then \(E_Y^*\text{[PRESS}^\dagger]\) can be decomposed into

\[
E_Y^*\text{[PRESS}^\dagger]\] = \(R_{FP} + E_Y^*E_U^*[R_1] + 2E_Y^*E_U^*[R_2] \).
\]

From Appendix A.3, we can see that

\[
E_Y^*E_U^*[R_1] = p + \frac{1}{n} \left\{ 2\kappa_4^{(1)} + p(2p + 3k + 3) \right\} + O(n^{-2}), \quad E_Y^*E_U^*[R_2] = O(n^{-2}).
\]

Substituting (3.12) into (3.11) yields

\[
E_Y^*\text{[PRESS}^\dagger]\] = \(R_{FP} + p + \frac{1}{n} \left\{ 2\kappa_4^{(1)} + p(2p + 3k + 3) \right\} + O(n^{-2}).
\]

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

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

Therefore, we can see that

\[
a(1)E_Y^*\text{[PRESS}^\dagger]\] = \(R_{FP} + \frac{1}{n} \left\{ \kappa_4^{(1)} + p(p + k + 2) \right\} + O(n^{-2}).
\]
Notice that \( \text{PRESS}^+ = a(1) \sum_{i=1}^{n} \{1 + a(1)(P_{X})_{ii}\} Q(\hat{r}_{i}^2; a(0)). \) \( E_Y^*[\text{PRESS}^+] \) can be expanded as

\[
E_Y^*[\text{PRESS}^+] = a(1)E_Y^*[\text{PRESS}^+] - \frac{1}{n} \sum_{i=1}^{n} \left\{ (P_{X})_{ii}\hat{r}_{i}^2 + \frac{1}{n} \hat{r}_{i}^4 \right\} + O(n^{-2}). \tag{3.13}
\]

It is easy to obtain the following expansions:

\[
E_Y \left[ \sum_{i=1}^{n} (P_{X})_{ii}\hat{r}_{i}^2 \right] = pk + O(n^{-1}), \quad E_Y \left[ \frac{1}{n} \sum_{i=1}^{n} \hat{r}_{i}^4 \right] = \kappa_4^{(1)} + p(p+2) + O(n^{-1}). \tag{3.14}
\]

Substituting (3.14) into (3.13) yields

\[
E_Y^*[\text{PRESS}^+] = R_{FP} + O(n^{-2}). \tag{3.15}
\]

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

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

Taking (3.16) from \( R_{KL} \) yields \( B_{\text{CAIC}_{J}} = O(n^{-2}) \) 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 \( X_0 \). The first column of this matrix was \( 1_n \) and next columns were generated by \( U(-1,1) \), where \( 1_n \) is an \( n \times 1 \) vector, and all of the vector’s elements are 1. The true model was determined as

\[
\eta^* = X_0 \mu^* 1_p', \quad \Sigma^* = I_p,
\]

where \( \mu^* = (0,1,2,3,0,0,0)' \). 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., $\mathbf{P}_x \eta^* \neq \eta^*$, 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., $\mathbf{P}_x \eta^* = \eta^*$, 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.

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 $\mathbb{E}[w_j] = 0$, $\mathbb{E}[w_j^2] = 1$ and the 4th cumulant $\psi$, and $\mathbf{w} = (w_1, \ldots, w_g)'$. Further, let $r$ be a random variable which is independent of $\mathbf{w}$, $\mathbb{E}[r^2] = 1$ and the 4th moment $\beta$. Then, we generate an error vector by

$$\mathbf{\varepsilon} = r \mathbf{H}' \mathbf{w},$$

(4.1)

where $\mathbf{H}$ is a $g \times p$ matrix defined by $\mathbf{H} = (h_1, \ldots, h_g)'$ with the full rank $p$ and $\mathbf{H}' \mathbf{H} = \mathbf{I}_p$. The multivariate kurtosis of this model becomes

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

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

$$\mathbf{H}_0 = \begin{pmatrix} \mathbf{I}_p^p \\ \mathbf{1}_p^p \end{pmatrix} (\mathbf{I}_p + \mathbf{1}_p \mathbf{1}_p')^{-1/2}.$$
Then, the error vector $\varepsilon = rH'_0w$ is written as

$$
\varepsilon = r \left( I_p - \rho 1_p1_p' \right) \begin{pmatrix}
    w_1 + w_{p+1} \\
    \vdots \\
    w_p + w_{p+1}
\end{pmatrix},
$$

where

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

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

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

- **Model 2 (Laplace Distribution):** $w_j$ is generated from a Laplace distribution with mean 0 and standard deviation 1, $r = \sqrt{6/\chi^2}$ and $H = H_0$ ($\kappa_4^{(1)} = 4.5 \times p^2(p + 1)^{-1} + p(p + 2)/2$).

- **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 $H = H_0$ ($\kappa_4^{(1)} = -1.2 \times p^2(p + 1)^{-1}$).

- **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 $\sqrt{23}/4$, $r = \sqrt{6/\chi^2}$ and $H = H_0$ ($\kappa_4^{(1)} = 2583/529 \times p^2(p + 1)^{-1} + p(p + 2)/2 \approx 4.88 \times p^2(p + 1)^{-1} + p(p + 2)/2$).

- **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 $2\sqrt{2}$, $r = \sqrt{6/\chi^2}$ and $H = H_0$ ($\kappa_4^{(1)} = 4.5 \times p^2(p + 1)^{-1} + p(p + 2)/2$).

- **Model 6 (Log-Normal Distribution):** $w_j$ is generated from a lognormal distribution such that $\log w_j \sim N(0, 1/4)$ standardized by mean $e^{1/8}$ and standard deviation $e^{1/8}\sqrt{e^{1/4} - 1}$, $r = \sqrt{6/\chi^2}$ and $H = H_0$ ($\kappa_4^{(1)} = 1.5 \times p^2(p + 1)^{-1}(e + 2e^{3/4} + 3e^{1/2} - 6) + p(p + 2)/2 \approx 8.85 \times p^2(p + 1)^{-1} + p(p + 2)/2$).

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 \( E_{Y[AIC]}^*, E_{Y[AIC_j]}^* \) and \( E_{Y[CAIC_j]}^* \) 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 \( R_{KL}, E_{Y[AIC]}^*, E_{Y[AIC_j]}^* \) and \( E_{Y[CAIC_j]}^* \) of the candidate models with the same score. From these figures, we can see that the bias of \( CAIC_j \) is very small in the overspecified models; similarly, the bias of \( AIC_j \) 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.

Table 2 shows the probabilities for selecting the model with the smallest risk. From this table, we can see that the selection-probability of \( AIC_j \) is improved in every distribution in comparison with \( AIC \). All selection-probabilities of \( CAIC_j \) are higher than those of \( AIC_j \). Moreover, although the selection-probabilities of \( AIC \) are almost the same in every distribution, the probabilities of \( AIC_j \) and \( CAIC_j \) 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 \( AIC_j \) and \( CAIC_j \), the probabilities in a skewed distribution tend to be higher than those in a non-skewed distribution.

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

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 \( O(n^{-2}) \) 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, \( O(1) \), 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 $s \geq 3$, $E^*_{\varepsilon}[\| \varepsilon \|^{2s}] < \infty$,

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

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

A4. For some constant $0 < \delta \leq 1/2$, $\max_{i=1,\ldots,n} \| x_i \| = O(n^{1/2-\delta})$.

A.2. Decomposition of $E^*_Y[\text{PRESS}]$

Since the distribution of $u_i$ is the same as that of $y_i$ and since $Y$ and $U$ are independent of each other, the following commutative equation, as in Yanagihara et al. (2006), holds:

$$E^*_Y[\text{PRESS}] = \sum_{i=1}^{n} E^*_Y E^*_U \left[ (u_i - \hat{\Theta}'_{[-i]} x_i) \hat{\Sigma}_{[-i]}^{-1} (u_i - \hat{\Theta}'_{[-i]} x_i) \right]. \quad (A.1)$$

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

$$u_i - \hat{\Theta}'_{[-i]} x_i = u_i - \hat{\Theta}' x_i + \frac{(P_X)_{ii}}{1 - (P_X)_{ii}} (y_i - \hat{\Theta}' x_i), \quad (A.2)$$

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

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

A.3. Asymptotic Expansions of $E^*_Y E^*_U[R_1]$ and $E^*_Y E^*_U[R_2]$
Let $Z$ and $V$ be random variable matrices having asymptotic normality, where

$$Z = (X'X)^{-1/2}X'\varepsilon, \quad V = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\varepsilon_i \varepsilon_i' - I_p),$$

and let $q_i$ be a $k \times 1$ constant vector defined by

$$q_i = \sqrt{n}(X'X)^{-1/2}x_i.$$ 

Notice that $P_{X\eta} = \eta$. Therefore, the following conditional expectations are obtained:

$$E_U^* \left[ (u_i - \Theta' x_i) | Y \right] = -\frac{1}{\sqrt{n}} \Sigma^{1/2} Z' q_i,$$

$$E_U^* \left[ (u_i - \Theta' x_i)(u_i - \Theta' x_i') | Y \right] = \Sigma^{1/2} \left( I_p + \frac{1}{n} Z' q_i q_i' Z \right) \Sigma^{1/2}.$$

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

$$E_U^* [R_1 | Y] = W_{10} + \frac{1}{\sqrt{n}} W_{11} + \frac{1}{n} W_{12} + O_p(n^{-3/2}),$$

$$E_U^* [R_2 | Y] = \frac{1}{\sqrt{n}} W_{21} + \frac{1}{n} W_{22} + O_p(n^{-3/2}),$$

where

$$W_{10} = \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i' \varepsilon_i,$$

$$W_{11} = -\frac{2}{n} \sum_{i=1}^{n} (q_i' Z \varepsilon_i + \varepsilon_i' V \varepsilon_i),$$

$$W_{12} = \frac{1}{n} \sum_{i=1}^{n} \left\{ q_i' q_i \varepsilon_i' \varepsilon_i + q_i' ZZ' q_i + (\varepsilon_i' \varepsilon_i)^2 + 2 \varepsilon_i' \varepsilon_i \right\},$$

$$W_{21} = -\frac{n}{n} \sum_{i=1}^{n} q_i' q_i q_i' Z \varepsilon_i,$$

$$W_{22} = -\frac{1}{n} \sum_{i=1}^{n} (q_i' q_i' ZZ' q_i - q_i' q_i' Z V \varepsilon_i).$$

Therefore, the asymptotic expansions of $E_Y^* E_U^* [R_1]$ and $E_Y^* E_U^* [R_2]$ are rewritten as

$$E_Y^* E_U^* [R_1] = E_Y^* [W_{10}] + \frac{1}{\sqrt{n}} E_Y^* [W_{11}] + \frac{1}{n} E_Y^* [W_{12}] + O(n^{-2}), \quad (A.4)$$

$$E_Y^* E_U^* [R_2] = \frac{1}{\sqrt{n}} E_Y^* [W_{21}] + \frac{1}{n} E_Y^* [W_{22}] + O(n^{-2}). \quad (A.5)$$
Notice that

\[E_Y^* \left[ \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i' \varepsilon_i \right] = p,\]

\[E_Y^* \left[ \frac{1}{n} \sum_{i=1}^{n} q_i' q_i \varepsilon_i' \varepsilon_i \right] = \frac{p}{n} \sum_{i=1}^{n} q_i' q_i = pk,\]

\[E_Y^* \left[ \frac{1}{n} \sum_{i=1}^{n} q_i' Z \varepsilon_i \right] = \frac{p}{n \sqrt{n}} \sum_{i=1}^{n} q_i' q_i = \frac{1}{\sqrt{n}} pk,\]

\[E_Y^* \left[ \frac{1}{n} \sum_{i=1}^{n} q_i' q_i' Z \varepsilon_i \right] = \frac{p}{n \sqrt{n}} \sum_{ij} (q_i' q_j')^2 = \frac{1}{\sqrt{n}} pq,\]

\[E_Y^* \left[ \frac{1}{n} \sum_{i=1}^{n} q_i' ZZ' q_i \right] = p \sum_{ij} (q_i' q_j')^2 = pk;\]

\[E_Y^* \left[ \frac{1}{n} \sum_{i=1}^{n} (q_i' q_i' ZZ' q_i) \right] = p \sum_{ij} (q_i' q_j')^2 = pq;\]

\[E_Y^* \left[ \frac{1}{n} \sum_{i=1}^{n} (\varepsilon_i' \varepsilon_i)^2 \right] = \kappa_4^{(1)} + p(p + 2),\]

\[E_Y^* \left[ \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i' V \varepsilon_i \right] = \frac{1}{\sqrt{n}} \left\{ \kappa_4^{(1)} + p(p + 1) \right\},\]

\[E_Y^* \left[ \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i' Z' Z \varepsilon_i \right] = \frac{p}{n} \sum_{i=1}^{n} q_i' q_i + \frac{1}{n^2} \left\{ \kappa_4^{(1)} + p(p + 1) \right\} \sum_{i=1}^{n} q_i' q_i = pk + O(n^{-1}),\]

\[E_Y^* \left[ \frac{1}{n} \sum_{i=1}^{n} (q_i' Z \varepsilon_i)^2 \right] = \frac{p}{n^2} \sum_{ij} (q_i' q_j)^2 + \frac{1}{n^2} \left\{ \kappa_4^{(1)} + p(p + 1) \right\} \sum_{i=1}^{n} (q_i' q_i)^2 = pk + O(n^{-1}),\]

\[E_Y^* \left[ \frac{1}{n} \sum_{i=1}^{n} q_i' Z' V \varepsilon_i \right] = \frac{1}{n^2} \left\{ \kappa_4^{(1)} + p(p + 1) \right\} \sum_{i=1}^{n} q_i' q_i = O(n^{-2}),\]

\[E_Y^* \left[ \frac{1}{n} \sum_{i=1}^{n} q_i' q_i' Z' V \varepsilon_i \right] = \frac{1}{n^2} \left\{ \kappa_4^{(1)} + p(p + 1) \right\} \sum_{i=1}^{n} (q_i' q_i)^2 = O(n^{-2}),\]

\[E_Y^* \left[ \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i' V^2 \varepsilon_i \right] = \kappa_4^{(1)} + p(p + 1) + O(n^{-1}),\]

where the coefficient \( q \) is defined by

\[q = n \sum_{i=1}^{n} ((P x)_i)^2.\]

From the above expectations, we obtain the following expectations:

\[E_Y^*[W_{10}] = p,\]

\[E_Y^*[W_{11}] = -\frac{2}{\sqrt{n}} \left\{ \kappa_4^{(1)} + p(p + k + 1) \right\},\]

\[E_Y^*[W_{12}] = 4\kappa_4^{(1)} + p(4p + 5k + 5) + O(n^{-1}),\quad \text{(A.6)}\]
\[
E^*_Y[W_{21}] = -\frac{1}{\sqrt{n}}pq,
\]
\[
E^*_Y[W_{22}] = pq + O(n^{-1}).
\]

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

\[
G_1 = \left(\frac{n-k}{2} + \frac{1}{n}\right)^{\frac{n(n-k-1)+2}{2n}} \left(\frac{n-k-p}{2}\right)^{\frac{n-k-p-1}{2n}}
\]

\[
\left(\frac{n-k}{n-k-1}\right)^{\frac{n-k}{2}} \left(\frac{n-k-p}{n-k-p-1}\right)^{\frac{n-k-p}{2n}}
\]

\[
G_2 = \left\{1 + \frac{1}{6(n-k+2)} + \frac{n^2}{72(n^2-\log+2)^2} + O(n^{-3})\right\} \left\{1 + \frac{1}{6(n-k-p+2)} + \frac{n^2}{72(n^2-\log-p+2)^2} + O(n^{-3})\right\}
\]

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

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

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

\[
G_2 = 1 + O(n^{-3}).
\]
Hence, \( c^+/a(1) \) is asymptotically expanded as

\[
\frac{c^+}{a(1)} = G_1G_2 \left\{ 1 - \frac{p}{n^2} + O(n^{-3}) \right\} = 1 + O(n^{-3}).
\]

Acknowledgments

The authors are grateful to N. Hanabusa, University of Notre Dame, for her helpful comments. Hirokazu Yanagihara’s research was supported by the Ministry of Education, Science, Sports and Culture Grant-in-Aid for Young Scientists (B), #17700274, 2005–2007.

References


**Table 1.** The number of candidate models included in the group

<table>
<thead>
<tr>
<th>Group</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4*</th>
<th>5*</th>
<th>6*</th>
<th>7*</th>
<th>8*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Score</td>
<td>0</td>
<td>1/3</td>
<td>2/3</td>
<td>1</td>
<td>4/3</td>
<td>5/3</td>
<td>2</td>
<td>7/3</td>
</tr>
<tr>
<td>No. of Candidates</td>
<td>15</td>
<td>48</td>
<td>48</td>
<td>1</td>
<td>4</td>
<td>6</td>
<td>4</td>
<td>1</td>
</tr>
</tbody>
</table>

* denotes the group which consists of only overspecified models.

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

<table>
<thead>
<tr>
<th>Distribution</th>
<th>$p = 2$</th>
<th>$p = 6$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AIC</td>
<td>AIC$_J$</td>
</tr>
<tr>
<td>Normal</td>
<td>41.13</td>
<td>61.55</td>
</tr>
<tr>
<td>Laplace</td>
<td>40.53</td>
<td>49.47</td>
</tr>
<tr>
<td>Uniform</td>
<td>40.96</td>
<td>64.94</td>
</tr>
<tr>
<td>Skew-Laplace</td>
<td>40.46</td>
<td>51.28</td>
</tr>
<tr>
<td>Chi-Square</td>
<td>40.01</td>
<td>51.85</td>
</tr>
<tr>
<td>Log-Normal</td>
<td>40.05</td>
<td>49.81</td>
</tr>
<tr>
<td>Average</td>
<td>40.52</td>
<td>54.82</td>
</tr>
</tbody>
</table>
Figure 1. $R_{KL}$ and $E_{Y}[AIC]$, $E_{Y}[AIC_J]$ and $E_{Y}[CAIC_J]$ ($p = 2$)
Figure 2. $R_{KL}$ and $E_Y^*[AIC]$, $E_Y^*[AIC_J]$ and $E_Y^*[CAIC_J]$ ($p = 6$)
Figure 3. RMSE of AIC, AIC\textsubscript{J} and CAIC\textsubscript{J} (p = 2)
Figure 4. RMSE of AIC, $AIC_{J}$ and $CAIC_{J}$ ($p = 6$)
Figure 5. Scatter plots of $AIC_J$ and $CAIC_J$ ($p = 6$)