J. KSIAM Vol.2, No.1, 41-60, 1998

A CONSISTENT AND BIAS CORRECTED EXTENSION OF AKAIKE'S INFORMATION CRITERION(AIC) : $AIC_{bc}(k)$

Soon H. Kwon*, M. Ueno, and M. Sugeno*****

* : School of Electrical and Electronic Eng., Yeungnam Univ.
214-1 Dae-dong, Kyongsan, Kyongbuk 712-749, Korea
** : Dept of Cognitive & Information Science, Chiba Univ.
1-33 Yayoi-cho, Inage-ku, Chiba, 263, Japan
*** : Dept of Computational Intelligence & Systems Science,
Tokyo Institute of Technology
4259 Nagatsuta, Midori-ku, Yokohama 226, Japan.

Abstract

This paper derives a consistent and bias corrected extension of Akaike's Information Criterion (AIC), AIC_{bc}, based on Kullback-Leibler information. This criterion has terms that penalize the overparametrization more strongly for small and large samples than that of AIC. The overfitting problem of the asymptotically efficient model selection criteria for small and large samples will be overcome. The AIC_{bc} also provides a consistent model order selection. Thus, it is widely applicable to data with small and/or large sample sizes, and to cases where the number of free parameters is a relatively large fraction of the sample size. Relationships with other model selection criteria such as AIC_c of Hurvich, CAICF of Bozdogan and etc. are discussed. Empirical performances of the AIC_{bc} are studied and discussed in better model order choices of a linear regression model using a Monte Carlo experiment.

Key words : Model selection, Kullback-Leibler information, AIC, bias correction, consistency.

1. Introduction

As is well known, modeling may be considered as a process approximating a system, where the approximation may be governed by some objectives, in order to understand

observations of or predictions of a system's future behavior. Though there are some differences in their approaches, the complexity of a system to be modeled is an important issue to study in the fields of conventional modeling. In statistical modeling, the choice of an appropriate model, which intrinsically includes the choice of a class of potential models, determination of the order of a model, and estimation of parameters of the model, is a fundamental difficulty[4,26,31]. A general principle for addressing this problem, Occam's razor, states that an adequate but parsimonious model is preferable to others.

In this sense, the introduction of Akaike's Information Criterion, AIC[1], has called our attention to the statistical model selection problem and triggered the development of many statistical modeling techniques in various fields during the last two decades. This criterion may be viewed as an asymptotically unbiased estimator of the expected Kullback-Leibler information which is a measure of discrepancy between statistical models[19]. Thus, selection of a model minimizing AIC means that the selected model may be the best approximating model to the true model.

Since Akaike's influential paper, several approaches to model selection were developed and are still being refined. As examples, (i) the Bayesian approach (e.g., Schwarz[23], Akaike[3], and Kashyap[18]), (ii) the cross-validation approach by Stone[30], (iii) the prequential approach by Dawid[14], (iv) the criterion autoregressive transfer function by Parzen[20], (v) Hannan and Quinn's approach[15], (vi) the coding theoretic approach by Rissanen[21], and (vii) informational complexity (ICOMP) criterion of Bozdogan[8,9,10,11] are representative approaches. Some of these are briefly summarized by Sclove[24,25] and Tong[33].

The asymptotic behaviors of these model selection criteria have been extensively analyzed by many researchers (e.g., Schwarz[23], Shibata[27,28] and Bozdogan and Haughton[12]). For the data for which the true model has infinite order, AIC provides an asymptotically efficient selection of a finite order model. However, for the data for which the true model has finite order, minimizing AIC does not produce consistent model order selection, which pursues the selection of the most parsimonious model. This defect is more evident when the sample size is very large. In other words, the existing asymptotically efficient criteria (e.g., AIC) which do not provide consistent order selection tend to overfit unless the maximum allowable order of the model is specified. This overfitting problem leads to more unsatisfactory model order selection when the sample size is small, or when the number of free parameters is a relatively larger than the sample size. In this case, the overfitting stems from the fact that AIC is strongly negatively biased. This bias is attributed

to the deterioration in the accuracy of Taylor series expansions of the Kullback-Leibler information used in the derivation of AIC by Hurvich[17].

Shibata[27] has pointed out that AIC, unlike the consistent model selection criteria, will not necessarily select the most parsimonious true model asymptotically (i.e., it has tendency to overestimate the order). However, because any real problem generally has a finite samples and any consistent criterion assumes that there exists true order of a model, consistency may not be so attractive in some cases. On the contrary, if there exist certain circumstances requiring avoidance of overfitting, consistent model selection needs to be advocated.

This background increases the necessity of introducing a consistent and bias corrected model selection criterion. In this paper, we will obtain a consistent and bias corrected model selection criterion by extending the methods proposed by Bozdogan[7] and Hurvich[17], which are both extensions of Akaike's Information Criterion, AIC. Since our work is primarily based on Akaike's AIC[1], Bozdogan's CAICF[7] which is an extension of AIC, and Hurvich's AIC_c[17], we will label its consistent and bias corrected extension as AIC_{bc} not to create any confusion on the part of the readers. These criteria are widely applicable to data with small and/or large samples and to the cases where the number of free parameters is a relatively larger than the sample size, and it still provides a consistent model order selection. Furthermore, we will show the empirical performance of the proposed model selection criterion to some other criteria by Monte Carlo experiments.

2. A consistent and bias corrected model selection criterion

In the situation of statistical modeling based on a set of given observations, we proceed under the assumption that these observations are the values of random variables whose probability distributions are generally unknown. In this case, we assume a model in the form of a probability distribution and estimate the true probability distribution using a set of given observations. This estimation is to determine the number and values of unknown parameters of the assumed model so that the model has good fitting ability to the given data and good predictive ability. In this section, we derive a consistent and bias corrected model selection criterion based on a measure of the distance between the model and the true distribution, which is Boltzmann's generalized entropy[5] or the Kullback-Leibler information quantity[19].

Suppose that independent random variables $X_1, \ldots, X_n$ form random samples $x_1, \ldots,$

x_n from a discrete or continuous distribution for which the probability density function is $f(\mathbf{x}|\boldsymbol{\theta})$, where the parameter vector $\boldsymbol{\theta} = \boldsymbol{\theta}_K=(\theta_1, \theta_2, \ldots, \theta_K)$ belongs to some K-dimensional parameter space Ω. In the following, we will assume that $X_1, \ldots, X_n$ are continuous random variables. The probability density function $f(\mathbf{x}|\boldsymbol{\theta})$ is a model with K parameters, i.e.,

$$\text{MODEL(K)} : f(\mathbf{x}|\boldsymbol{\theta}),\ \boldsymbol{\theta}=\boldsymbol{\theta}_K=(\theta_1, \theta_2, \ldots, \theta_K). \tag{1}$$

Assume that a true parameter vector $\boldsymbol{\theta}^*$ of $\boldsymbol{\theta}$ with its probability density function $f(\mathbf{x}|\boldsymbol{\theta}^*)$ is included in the K-dimensional parameter space Ω. A model defined by restricting the parameter space with $\theta_h=0$ for all $h > k$ is given by

$$\text{MODEL(k)} : f(\mathbf{x}|\boldsymbol{\theta}_k),\ \boldsymbol{\theta}_k =\{(\theta_1, \theta_2, \ldots, \theta_K) \mid \theta_h=0 \text{ for all } h > k\}. \tag{2}$$

In the case, the statistical model identification may be carried out by selecting a restricted model $f(\mathbf{x}|\boldsymbol{\theta}_k)$, where the $\boldsymbol{\theta}_k$ is the closest to the true parameter vector $\boldsymbol{\theta}^*$, based on the given n observations. Thus, the derivation of a criterion which gives an optimal value of k so as to minimize the average estimation error is needed in the problem of statistical model identification. The estimation error comes from both the deterministic bias due to the selection of a restricted model $f(\mathbf{x}|\boldsymbol{\theta}_k)$ and the random error due to the use of the maximum likelihood estimator $\hat{\boldsymbol{\theta}}_k$ of $\boldsymbol{\theta}_k$. Thus, minimization of the average estimation error can be done by the appropriate compromise between the bias and the random error. Following Akaike, we will use the entropy for the derivation of a criterion to minimize the average estimation error[2].

To develop this further, we will introduce the generalized entropy B of Boltzmann, or the Kullback-Leibler information quantity I as an objective measure of the distance between the model $f(\mathbf{x}|\boldsymbol{\theta})$ and the true distribution $f(\mathbf{x}|\boldsymbol{\theta}^*)$ as:

$$\begin{aligned} I(\boldsymbol{\theta}^*;\boldsymbol{\theta}) &= -B(\boldsymbol{\theta}^*;\boldsymbol{\theta}) \equiv E\{\log f(\mathbf{X}|\boldsymbol{\theta}^*) - \log f(\mathbf{X}|\boldsymbol{\theta})\} \\ &= \int f(\mathbf{x}|\boldsymbol{\theta}^*)\log f(\mathbf{x}|\boldsymbol{\theta}^*)d\mathbf{x} - \int f(\mathbf{x}|\boldsymbol{\theta}^*)\log f(\mathbf{x}|\boldsymbol{\theta})d\mathbf{x}, \end{aligned} \tag{3}$$

where E denotes the expectation with respect to the true distribution $f(\mathbf{x}|\boldsymbol{\theta}^*)$ and log means natural logarithm. Following Bozdogan, we will minimize the Kullback-Leibler information quantity I instead of maximizing the entropy B for the derivation of a criterion to minimize the average estimation error. Since the first term in (3) is a constant, we only have to estimate the second term (i.e., the expected log likelihood) in (3) with respect to $\boldsymbol{\theta}$. Unfortunately it is not directly observable but can be consistently estimated from the observed data.

Proposition 1. The consistent and bias corrected extension of Akaike's information criterion(AIC), $AIC_{bc}(k)$, is

$$\begin{aligned} AIC_{bc}(k) &= -2\ell(\hat{\theta}_k) + k\log\frac{n}{2\pi} + \log\left|J(\hat{\theta}_k)\right| + 2tr\left[J(\hat{\theta}_k)^{-1}R(\hat{\theta}_k)\right] \\ &= -2\ell(\hat{\theta}_k) + k\log\frac{n}{2\pi} + \log\left|J(\hat{\theta}_k)\right| + 2\frac{nk}{n-k-2}. \end{aligned} \tag{4}$$

***Proof*.** Hereafter, we follow the notation of Sakamoto[22] and Bozdogan[7]. Under the assumption that a given data set is a realization of a random variable vector **X** of which random variables are independent and identically distributed (i.i.d.), the likelihood function for the set of data is given by

$$\mathrm{L}(\boldsymbol{\theta}) = \mathrm{f}(\mathrm{x}_1, \mathrm{x}_2, \cdots, \mathrm{x}_\mathrm{n}|\boldsymbol{\theta}) = \prod_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{f}(\mathrm{x}_\mathrm{i}|\boldsymbol{\theta}). \tag{5}$$

By taking the natural logarithm of the likelihood function $\mathrm{L}(\boldsymbol{\theta})$, we have

$$\ell(\boldsymbol{\theta}) \equiv \log \mathrm{L}(\boldsymbol{\theta}) = \sum_{\mathrm{i}=1}^{\mathrm{n}} \log \mathrm{f}(\mathrm{x}_\mathrm{i}|\boldsymbol{\theta}), \tag{6}$$

the log likelihood function, $\ell(\boldsymbol{\theta})$. It can also be regarded as a random variable. By dividing the log likelihood function $\ell(\boldsymbol{\theta})$ by the sample size n, we get

$$\ell_\mathrm{n}(\boldsymbol{\theta}) \equiv \frac{1}{\mathrm{n}}\ell(\boldsymbol{\theta}) = \frac{1}{\mathrm{n}}\log \mathrm{L}(\boldsymbol{\theta}) = \frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}} \log \mathrm{f}(\mathrm{x}_\mathrm{i}|\boldsymbol{\theta}), \tag{7}$$

the mean log likelihood of the sample. From the efficiency of the maximum likelihood estimator, we observe that the mean log likelihood in (7) is a natural estimator of the expected log likelihood in (3). The expected mean log likelihood is given by

$$\mathrm{nE}\{\ell_\mathrm{n}(\boldsymbol{\theta})\} = \mathrm{E}\{\log \mathrm{L}(\boldsymbol{\theta})\} = \int \mathrm{f}(\mathbf{x}|\boldsymbol{\theta}^*) \log \mathrm{f}(\mathbf{x}|\boldsymbol{\theta}) \mathrm{d}\mathbf{x}. \tag{8}$$

Therefore, we have to minimize the expected mean log likelihood of the true model given by

$$\ell^*_\mathrm{n}(\mathrm{k}) \equiv \mathrm{E}\{\ell^*(\hat{\boldsymbol{\theta}}_\mathrm{k})\}, \tag{9}$$

where $\ell^*(\boldsymbol{\theta}^*) \equiv \mathrm{nE}\{\ell_\mathrm{n}(\boldsymbol{\theta}^*)\} = \mathrm{E}\{\ell(\boldsymbol{\theta}^*)\} = \mathrm{E}\{\sum_{\mathrm{i}=1}^{\mathrm{n}} \log \mathrm{f}(\mathrm{x}_\mathrm{i}|\boldsymbol{\theta}^*)\}$.

Expanding $\ell(\boldsymbol{\theta})$ in (6) in a Taylor series about a maximum likelihood estimator $\hat{\boldsymbol{\theta}}_\mathrm{K}$ and ignoring higher order terms, we obtain

$$\ell(\boldsymbol{\theta}) \cong \ell(\hat{\boldsymbol{\theta}}_K) - \frac{n}{2}\left\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_K\right\|_J^2, \tag{10}$$

where $\left\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_K\right\|_J^2 = (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_K)^T J(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_K)$, T denotes the transpose of a vector $(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_K)$ and J is the positive definite Fisher's information matrix which is defined by

$$J \equiv -E\left[\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\log f(\mathbf{X}|\boldsymbol{\theta})\right]_{\boldsymbol{\theta}^*}. \tag{11}$$

By the Pythagorean theorem, we have

$$\left\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_K\right\|_J^2 = \left\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_k\right\|_J^2 + \left\|\hat{\boldsymbol{\theta}}_k - \hat{\boldsymbol{\theta}}_K\right\|_J^2. \tag{12}$$

From (10) and (12), we further get

$$\begin{aligned}
\ell(\hat{\boldsymbol{\theta}}_k) &\cong \ell(\hat{\boldsymbol{\theta}}_K) - \frac{n}{2}\left\|\boldsymbol{\theta}_k^* - \hat{\boldsymbol{\theta}}_K\right\|_J^2 + \frac{n}{2}\left\|\boldsymbol{\theta}_k^* - \hat{\boldsymbol{\theta}}_k\right\|_J^2 \\
&= \ell(\hat{\boldsymbol{\theta}}_K) - \frac{n}{2}\left\|\boldsymbol{\theta}_k^* - \boldsymbol{\theta}^* + \boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}_K\right\|_J^2 + \frac{n}{2}\left\|\boldsymbol{\theta}_k^* - \hat{\boldsymbol{\theta}}_k\right\|_J^2 \\
&= \ell(\hat{\boldsymbol{\theta}}_K) - \frac{n}{2}\left\|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}_K\right\|_J^2 - (\boldsymbol{\theta}_k^* - \boldsymbol{\theta}^*)nJ(\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}_K)^T - \frac{n}{2}\left\|\boldsymbol{\theta}_k^* - \boldsymbol{\theta}^*\right\|_J^2 + \frac{n}{2}\left\|\boldsymbol{\theta}_k^* - \hat{\boldsymbol{\theta}}_k\right\|_J^2
\end{aligned} \tag{13}$$

Using (10) and (13), we obtain

$$\begin{aligned}
\ell(\hat{\boldsymbol{\theta}}_k) &\cong \ell(\boldsymbol{\theta}^*) - (\boldsymbol{\theta}_k^* - \boldsymbol{\theta}^*)nJ(\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}_K)^T - \frac{n}{2}\left\|\boldsymbol{\theta}_k^* - \boldsymbol{\theta}^*\right\|_J^2 + \frac{n}{2}\left\|\boldsymbol{\theta}_k^* - \hat{\boldsymbol{\theta}}_k\right\|_J^2 \\
&= \ell^*(\boldsymbol{\theta}^*) - \frac{n}{2}\left\|\boldsymbol{\theta}_k^* - \boldsymbol{\theta}^*\right\|_J^2 + \ell(\boldsymbol{\theta}^*) - \ell^*(\boldsymbol{\theta}^*) - (\boldsymbol{\theta}_k^* - \boldsymbol{\theta}^*)nJ(\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}_K)^T + \frac{n}{2}\left\|\boldsymbol{\theta}_k^* - \hat{\boldsymbol{\theta}}_k\right\|_J^2
\end{aligned} \tag{14}$$

Similarly the Pythagorean theorem, we have

$$\left\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\right\|_J^2 = \left\|\boldsymbol{\theta} - \boldsymbol{\theta}_k^*\right\|_J^2 + \left\|\boldsymbol{\theta}_k^* - \boldsymbol{\theta}^*\right\|_J^2. \tag{15}$$

Expanding $\ell^*(\boldsymbol{\theta})$ in a Taylor series about $\boldsymbol{\theta}_k^*$ and ignoring higher order terms, we have

$$\ell^*(\boldsymbol{\theta}) \cong \ell^*(\boldsymbol{\theta}_k^*) - \frac{n}{2}\left\|\boldsymbol{\theta} - \boldsymbol{\theta}_k^*\right\|_J^2, \tag{16}$$

and

$$\ell^*(\boldsymbol{\theta}_k^*) \cong \ell^*(\boldsymbol{\theta}^*) - \frac{n}{2}\left\|\boldsymbol{\theta}_k^* - \boldsymbol{\theta}^*\right\|_J^2. \tag{17}$$

Thus, from (14) and (17), we obtain

$$\ell(\hat{\boldsymbol{\theta}}_k) \cong \ell^*(\boldsymbol{\theta}_k^*) + \ell(\boldsymbol{\theta}^*) - \ell^*(\boldsymbol{\theta}^*) - (\boldsymbol{\theta}_k^* - \boldsymbol{\theta}^*)nJ(\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}_K)^T + \frac{n}{2}\left\|\boldsymbol{\theta}_k^* - \hat{\boldsymbol{\theta}}_k\right\|_J^2. \tag{18}$$

Solving (18) for $\ell^*(\boldsymbol{\theta}_k^*)$, we get

$$\ell^*(\boldsymbol{\theta}_k^*) \cong \ell(\hat{\boldsymbol{\theta}}_k) - \ell(\boldsymbol{\theta}^*) + \ell^*(\boldsymbol{\theta}^*) + (\boldsymbol{\theta}_k^* - \boldsymbol{\theta}^*)nJ(\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}_K)^T - \frac{n}{2}\left\|\boldsymbol{\theta}_k^* - \hat{\boldsymbol{\theta}}_k\right\|_J^2. \tag{19}$$

Thus, from (9) and (16), we have

$$\ell_n^*(k) \equiv E\{\ell_n^*(\hat{\boldsymbol{\theta}}_k)\} = \ell^*(\boldsymbol{\theta}_k^*) - E\{\frac{n}{2}\left\|\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k^*\right\|_J^2\}. \tag{20}$$

Thus, from (19) and (20), we get

$$\ell_n^*(k) \cong \ell(\hat{\theta}_k) - \ell(\theta^*) + \ell^*(\theta^*) + (\theta_k^* - \theta^*)nJ(\theta^* - \hat{\theta}_K)^T - \frac{n}{2}\left\|\theta_k^* - \hat{\theta}_k\right\|_J^2 - E\{\frac{n}{2}\left\|\hat{\theta}_k - \theta_k^*\right\|_J^2\} \tag{21}$$

Ignoring the constant term $\ell^*(\boldsymbol{\theta}^*)$ in (21), we obtain

$$\ell_n^*(k) \cong \ell(\hat{\boldsymbol{\theta}}_k) - \ell(\boldsymbol{\theta}^*) + (\boldsymbol{\theta}_k^* - \boldsymbol{\theta}^*)nJ(\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}_K)^T - \frac{n}{2}\left\|\boldsymbol{\theta}_k^* - \hat{\boldsymbol{\theta}}_k\right\|_J^2 - E\{\frac{n}{2}\left\|\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k^*\right\|_J^2\} \tag{22}$$

Expanding $\ell(\boldsymbol{\theta}^*)$ in a Taylor series about $\boldsymbol{\theta}_k^*$ and ignoring higher order terms, we have

$$\ell(\boldsymbol{\theta}^*) \cong \ell(\boldsymbol{\theta}_k^*) - \frac{n}{2}\left\|\boldsymbol{\theta}^* - \boldsymbol{\theta}_k^*\right\|_J^2. \tag{23}$$

Thus, from (22) and (23), we obtain

$$\ell_n^*(k) \cong \ell(\hat{\theta}_k) - \ell(\theta_k^*) + \frac{n}{2}\left\|\theta^* - \theta_k^*\right\|_J^2 + (\theta_k^* - \theta^*)nJ(\theta^* - \hat{\theta}_K)^T - \frac{n}{2}\left\|\theta_k^* - \hat{\theta}_k\right\|_J^2 - E\{\frac{n}{2}\left\|\hat{\theta}_k - \theta_k^*\right\|_J^2\} \tag{24}$$

For sufficiently large n, since the maximum likelihood estimator $\hat{\boldsymbol{\theta}}_k$ is asymptotically distributed as multivariate normal with mean vector $\boldsymbol{\theta}_k^*$ and covariance matrix $(nJ(\boldsymbol{\theta}_k^*))^{-1}$, we have

$$\hat{\boldsymbol{\theta}}_k \rightarrow N(\boldsymbol{\theta}_k^*, (nJ(\boldsymbol{\theta}_k^*))^{-1}). \tag{25}$$

In this case, $\ell(\boldsymbol{\theta}_k^*)$ can be approximated as follows[7].

$$\ell(\boldsymbol{\theta}_k^*) \cong \log h(\mathbf{x}) + \frac{k}{2}\log\frac{n}{2\pi} + \frac{1}{2}\log\left|J(\hat{\boldsymbol{\theta}}_k)\right| + O(n^{-1/2}), \tag{26}$$

where h($\mathbf{x}$) is independent of a parameter vector $\boldsymbol{\theta}$.

Here, we assume that the true parameter $\boldsymbol{\theta}^*$ is situated near $\boldsymbol{\theta}_k^*$ and ignoring the constant term log h(x) in (26), from (24) and (26), we obtain

$$\ell_n^*(k) \cong \ell(\hat{\boldsymbol{\theta}}_k) - \frac{k}{2}\log\frac{n}{2\pi} - \frac{1}{2}\log\left|J(\hat{\boldsymbol{\theta}}_k)\right| - \frac{n}{2}\left\|\boldsymbol{\theta}_k^* - \hat{\boldsymbol{\theta}}_k\right\|_J^2 - E\{\frac{n}{2}\left\|\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k^*\right\|_J^2\}. \tag{27}$$

Now, multiplying both sides of (27) by -2, we obtain

$$-2\ell_n^*(k) \cong -2\ell(\hat{\boldsymbol{\theta}}_k) + k\log\frac{n}{2\pi} + \log\left|J(\hat{\boldsymbol{\theta}}_k)\right| + n\left\|\boldsymbol{\theta}_k^* - \hat{\boldsymbol{\theta}}_k\right\|_J^2 + E\{n\left\|\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k^*\right\|_J^2\}. \tag{28}$$

Since

$$n\left\|\boldsymbol{\theta}_k^* - \hat{\boldsymbol{\theta}}_k\right\|_J^2 \cong E\{n\left\|\boldsymbol{\theta}_k^* - \hat{\boldsymbol{\theta}}_k\right\|_J^2\}, \tag{29}$$

equation (28) becomes

$$-2\ell_n^*(k) \cong -2\ell(\hat{\boldsymbol{\theta}}_k) + k\log\frac{n}{2\pi} + \log\left|J(\hat{\boldsymbol{\theta}}_k)\right| + 2E\{n\left\|\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k^*\right\|_J^2\}. \tag{30}$$

To conclude the derivation, we consider the last term in (30), that is, $n\left\|\theta_k^* - \hat{\theta}_k\right\|_J^2$ under the expectation. For sufficiently large n, Akaike approximated $n\left\|\theta_k^* - \hat{\theta}_k\right\|_J^2$ by a chi-square distribution with k degrees of freedom under certain regularity conditions[1]. In this paper, we derive a general form of its expectation, which includes Akaike's approach as a special case.

$$\begin{aligned} E\{n\left\|\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k^*\right\|_J^2\} &= E\{\sqrt{n}(\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k^*)^T J\sqrt{n}(\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k^*)\} \\ &= \mathrm{tr}\{J\,\mathrm{cov}\left[\sqrt{n}(\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k^*)\right]\} + E\{\sqrt{n}(\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k^*)^T\}\,J\,E\{\sqrt{n}(\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k^*)\}. \end{aligned} \tag{31}$$

For sufficiently large n, $\sqrt{n}\left(\hat{\theta}_k - \theta_k^*\right)$ is asymptotically multivariate normal. That is,

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k^*) \sim N(0, J^{-1}RJ), \tag{32}$$

where $R = E\left[\left(\frac{\partial \log f(\mathbf{X}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)\left[\left(\frac{\partial \log f(\mathbf{X}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)^{T}\right]\right]_{\boldsymbol{\theta}^*}$ is the outer-product form of the Fisher information matrix. Thus, from (31) and (32), we have

$$E\{n\left\|\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k^*\right\|_J^2\} \cong \mathrm{tr}(J^{-1}R). \tag{33}$$

We note that $\mathrm{tr}(J^{-1}R)$ is the well known Lagrange-multiplier test statistic. See, for example, Takeuchi[32] and Hosking[16]. In (33), if the order of R is k and J is equal to R, which is not true in general, then $\mathrm{tr}(J^{-1}R)$ is equal to k, so that it will be the same as Akaike's result. Because $\boldsymbol{\theta}_k^*$ is not directly observable, we use its maximum likelihood estimator $\hat{\boldsymbol{\theta}}_k$. Therefore, from (30) and (33), we get the extended consistent and bias corrected model selection criterion $AIC_{bc}(k)$ as:

$$AIC_{bc}(k) \equiv -2\ell_n^*(k)$$

$$\cong -2\ell(\hat{\boldsymbol{\theta}}_k) + k\log\frac{n}{2\pi} + \log\left|J(\hat{\boldsymbol{\theta}}_k)\right| + 2\mathrm{tr}\left[J(\hat{\boldsymbol{\theta}}_k)^{-1}R(\hat{\boldsymbol{\theta}}_k)\right]. \tag{34}$$

It is worth to note that the first plus the last terms in (34) corresponds to the general definition of Takeuchi's information criterion (TIC), or sometimes also denoted by AIC_T. For this, see, e.g., Shibata[29]. Therefore, $AIC_{bc}(k)$ is a much more general criterion than TIC.

The computation of $AIC_{bc}(k)$ represented in (34) in some cases is very cumbersome. However, the new work of Bozdogan[9,10,11,12] gives us analytical way of computing the inverse of the Fisher's information matrix. It is also reasonable to approximate (34) by a simpler form for effective use, especially for small sample sizes. We achieve this by expanding $\left(\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)_{\hat{\boldsymbol{\theta}}_k}$ in a Taylor series about $\boldsymbol{\theta}_k^*$ and ignoring higher order terms, we have

$$\left(\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)_{\hat{\theta}_k} \cong \left(\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)_{\boldsymbol{\theta}_k^*} + (\hat{\theta}_k - \boldsymbol{\theta}_k^*)\left(\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}\right)_{\boldsymbol{\theta}_k^*}. \tag{35}$$

The left-hand side of (35) is zero. Thus, (35) can be rewritten as

$$(\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k^*)P(\boldsymbol{\theta}_k^*) = \frac{\left(\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)_{\boldsymbol{\theta}_k^*} \Big/ P(\boldsymbol{\theta}_k^*)}{\left(\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}\right)_{\boldsymbol{\theta}_k^*} \Big/ -P^2(\boldsymbol{\theta}_k^*)}, \tag{36}$$

where $P^2(\boldsymbol{\theta}_k^*) = -E\left[\left(\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}\right)_{\boldsymbol{\theta}_k^*}\right] = J(\boldsymbol{\theta}_k^*).$

And we have

$$\left(\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}\right)_{\boldsymbol{\theta}_k^*} = \left(\frac{\partial^2 f(\mathbf{X}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T} \Big/ f(\mathbf{X}|\boldsymbol{\theta})\right)_{\boldsymbol{\theta}_k^*} - \left(\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)^2_{\boldsymbol{\theta}_k^*}. \tag{37}$$

Taking the expectation of both sides of (37), we obtain

$$E\left[\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}\right]_{\boldsymbol{\theta}_k^*} = E\left[\frac{\partial^2 f(\mathbf{X}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T} \Big/ f(\mathbf{X}|\boldsymbol{\theta})\right]_{\boldsymbol{\theta}_k^*} - E\left[\left(\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)^2\right]_{\boldsymbol{\theta}_k^*}. \tag{38}$$

In (38), if the following is satisfied

$$E\left[\frac{\partial^2 f(\mathbf{X}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T} \Big/ f(\mathbf{X}|\boldsymbol{\theta})\right]_{\boldsymbol{\theta}_k^*} = 0, \tag{39}$$

then, we get

$$E\left[\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}\right]_{\boldsymbol{\theta}_k^*} = -E\left[\left(\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)^2\right]_{\boldsymbol{\theta}_k^*}. \tag{40}$$

From (36) and (37), we obtain

$$(\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k^*)P(\boldsymbol{\theta}_k^*) = \frac{\left(\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)_{\boldsymbol{\theta}_k^*} \Big/ P(\boldsymbol{\theta}_k^*)}{\left(\frac{\partial^2 f(\mathbf{X}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T} \Big/ f(\mathbf{X}|\boldsymbol{\theta})\right)_{\boldsymbol{\theta}_k^*} - \left(\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)^2_{\boldsymbol{\theta}_k^*} \Big/ -P^2(\boldsymbol{\theta}_k^*)}. \tag{41}$$

Thus, from (33) and (41), we get

$$\left\|\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k^*\right\|_J^2 \cong \frac{\left(\left(\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)_{\boldsymbol{\theta}_k^*} \Big/ P(\boldsymbol{\theta}_k^*)\right)^2}{\left[\left\{\left(\frac{\partial^2 f(\mathbf{X}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T} \Big/ f(\mathbf{X}|\boldsymbol{\theta})\right)_{\boldsymbol{\theta}_k^*} - \left(\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)^2_{\boldsymbol{\theta}_k^*}\right\} \Big/ -P^2(\boldsymbol{\theta}_k^*)\right]^2}. \tag{42}$$

The distribution of the numerator on the right of (42) becomes a χ^2 distribution with k degrees of freedom. The distribution of the denominator on the right of (42) also becomes a χ^2 distribution with (n-k) degrees of freedom and they are independent. Thus, $\frac{n-k}{k}\left\|\boldsymbol{\theta}_k^* - \hat{\boldsymbol{\theta}}_k\right\|_J^2$ is approximately distributed as F(k, n - k). Here, considering only the term in (42), we can see that Akaike's result is a special case of our result. This is also supported by the fact that for a large values of (n-k), if a random variable ν has an F(k, n-k) distribution, then a good approximation of ν has a $\chi^2(k)$ distribution[13]. Thus, taking the expectation of (42), then (34) becomes

$$\begin{aligned} \mathrm{AIC}_{bc}(k) &\equiv -2\ell_n^*(k) \\ &\cong -2\ell(\hat{\boldsymbol{\theta}}_k) + k\log\frac{n}{2\pi} + \log\left|J(\hat{\boldsymbol{\theta}}_k)\right| + 2\frac{nk}{n-k-2}. \end{aligned} \tag{43}$$

By (34) and (43), we get the whole proposition. Q.E.D.

The first term in (43) is a measure of badness of fit when the parameters of the true

model with an unknown number of parameters are approximated by the maximum likelihood estimators of the k parameters of the assumed model. We can heuristically interpret functions of the penalty terms in AIC_{bc} as follows. The second term in (43) penalizes the overparametrization more strongly for small samples. Thus, the overfitting problem of the asymptotically efficient model selection criteria for small samples will be overcome. On the other hand, the third term penalizes the overparametrization more strongly for large samples. In the same way, the overfitting problem of the asymptotically efficient model selection criteria for large samples will be overcome. Here, if we retain the first two terms, AIC_{bc} is similar to Hurvich's criterion[17]. If we retain the first and the third terms in (43), AIC_{bc} is similar to Schwarz's criterion (SC).

We further note that the fourth term $\mathrm{tr}(J^{-1}R)$ in (34) is important because it provides information on the correctness of the assumed class of potential models as discussed in White[34] and Bozdogan[7]. In general, a fundamental assumption underlying classical model selection criteria is that the true model is known to lie within a specified class of potential models (i.e., the class of potential models is assumed to be correctly specified). However, in real circumstances, it is most frequently the case that we are not able to find evidence that this assumption is true. The correct specification of the class is a sufficient, but by no means a necessary condition[7].

Thus, it is very natural to introduce a term in model selection criteria, which indicates whether the assumed class of potential models is correctly specified or not. Following White[34], see also Bozdogan[7], under conditions that the class of potential models is correctly specified and that certain assumptions hold, the following information matrix equivalence theorem can be obtained. Here, we quote the theorem without proof.

Theorem 1; If $f(\mathbf{x}) = f(\mathbf{x}|\boldsymbol{\theta}^*)$ for $\boldsymbol{\theta}^*$ in K-dimensional space Ω_K, then $\boldsymbol{\theta}^* = \boldsymbol{\theta}_k^*$ and $J(\boldsymbol{\theta}_k^*) = R(\boldsymbol{\theta}_k^*)$, so that the covariance matrix $C(\boldsymbol{\theta}_k^*) = J(\boldsymbol{\theta}_k^*)^{-1} = R(\boldsymbol{\theta}_k^*)^{-1}$,

where $$J(\boldsymbol{\theta}_k^*) \equiv -E\left[\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\log f(\mathbf{X}|\boldsymbol{\theta})\right]_{\boldsymbol{\theta}_k^*},$$

$$R(\boldsymbol{\theta}_k^*) \equiv E\left[\left(\frac{\partial \log f(\mathbf{X}|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right)\left[\left(\frac{\partial \log f(\mathbf{X}|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right)^T\right]\right]_{\boldsymbol{\theta}_k^*},$$

and $C(\boldsymbol{\theta}_k^*) \equiv J(\boldsymbol{\theta}_k^*)^{-1} R(\boldsymbol{\theta}_k^*)^{-1} J(\boldsymbol{\theta}_k^*)^{-1}$ is the covariance matrix.

The information matrix equivalence theorem says that when the class of potential models is correctly specified, the information matrix can be expressed in either Hessian form, $J(\boldsymbol{\theta}_k^*)$ or outer product form $R(\boldsymbol{\theta}_k^*)$ (i.e., $J(\boldsymbol{\theta}_k^*) - R(\boldsymbol{\theta}_k^*) = 0.$). When this equality fails, it follows that the class is misspecified. Thus, from Theorem 1, we can see that we can judge whether the class of potential models are correctly specified or not by testing the relationship between the consistent estimators $J(\hat{\boldsymbol{\theta}}_k)$ and $R(\hat{\boldsymbol{\theta}}_k)$ of $J(\boldsymbol{\theta}_k^*)$ and $R(\boldsymbol{\theta}_k^*)$ respectively, which are not directly observable. If the estimated information matrix $J(\hat{\boldsymbol{\theta}}_k)$ is singular or near singular, then the Kullback-Leibler information quantity has no unique minimum at the true parameter vector. Therefore it is preferable to compute $J(\hat{\boldsymbol{\theta}}_k)$, if possible, and proceed to choose the model order.

Kashyap has obtained a similar result (i.e., $\log\left|B(\hat{\boldsymbol{\theta}}_k)\right|$) in his criterion, which took a Bayesian approach, by expanding and approximating the logarithm of the posterior probability in a Taylor series[18]. $B(\hat{\boldsymbol{\theta}}_k)$, which is the negative of the matrix of second partials of $\log L(\boldsymbol{\theta}_k)$, is finite and asymptotically positive definite and is Fisher's information matrix. Another similar result was obtained by Bozdogan[7] given by

$$\text{CAICF}(k) = -2\ell(\hat{\boldsymbol{\theta}}_k) + k(\log n + 2) + \log\left|J(\hat{\boldsymbol{\theta}}_k)\right|. \tag{44}$$

CAICF(k) which is a consistent extension of AIC penalizes the overparametrizations more strongly, in particular, for large samples. But it does not have a term correcting the bias for small samples. Thus, we can see the difference between the AIC_{bc} given in (43) with that of Bozdogan's CAICF. We note that as $n \to \infty$, the term $\frac{2n}{n-k-2} \to 2$, and AIC_{bc} reduces to CAICF in (44) since the term $-k\log(2\pi)$ can be neglected. However, it is worth to note that the term $-k\log(2\pi)$ dropped in the derivation of Bozdogan[7] cannot be dropped when the sample size is small.

In the following, we will show this by a more explicit analytical formulation. To state our results, we need the following assumption.

(A1) The parametric probability distribution of the model is correct (i.e., the estimated Fisher information matrix J is nonsingular).

(Consistency of AIC_{bc}) Suppose (A1) holds. Then, the model selection criterion AIC_{bc} is strongly consistent. That is,

$$k \to k^* \text{ as } n \to \infty, \tag{45}$$

where k^* is the true order of the model.

We will prove the consistency of AIC_{bc} in the following: Since the arguments for the underfitting and the overfitting cases are different, we will consider them separately.

(1) Overfitting case. We consider what happens to the probability of choosing an order $k^o > k^*$. The third term in (43) is of $O(1)$ and is negligible in comparison with other terms if n is very large. Thus,

$$\begin{aligned} &P\{AIC_{bc}(k^o) < AIC_{bc}(k^*)\} \\ &\quad = P\left\{2\left[\ell(\hat{\boldsymbol{\theta}}_{k^o}) - \ell(\hat{\boldsymbol{\theta}}_{k^*})\right] > \left[\frac{2n(n-2)k^o}{n-k^o-2} - \frac{2n(n-2)k^*}{n-k^*-2} + (k^o - k^*)\log\frac{n}{2\pi}\right]\right\} \\ &\qquad \to P\{\chi^2_{k^o-k^*} > \infty\} \text{ as } n \to \infty \\ &\quad = 0. \end{aligned} \tag{46}$$

(2) Underfitting case. We consider what happens to the probability of choosing an order $k^u < k^*$. In the same way, the fourth term in (43) is of $O(1)$ and is negligible in comparison with other terms if n is very large. Thus,

$$\begin{aligned} &P\{AIC_{bc}(k^u) < AIC_{bc}(k^*)\} \\ &\quad = P\left\{2\left[\ell(\hat{\boldsymbol{\theta}}_{k^*}) - \ell(\hat{\boldsymbol{\theta}}_{k^u})\right] > \left[\frac{2n(n-2)k^*}{n-k^*-2} - \frac{2n(n-2)k^u}{n-k^u-2} + (k^* - k^u)\log\frac{n}{2\pi}\right]\right\} \\ &\qquad \to P\{\chi^2_{k^*-k^u} > \infty\} \text{ as } n \to \infty \\ &\quad = 0. \end{aligned} \tag{47}$$

Then, from (46) and (47), the consistency of AIC_{bc} is proved.

3. A numerical example

In this section, to investigate the empirical performance of the AIC_{bc} and compare its performance with those of various model selection criteria, we provide a result of Monte Carlo study. For simplicity, we will consider a regression problem and assume that the true model and the approximating model are linear and given by

$$y = \sum_{j=1}^{k} \theta_j x_j + \varepsilon, \tag{48}$$

where it is assumed that ε is independent and identically distributed (i.i.d.) normal. We assume that the approximating family includes the true model, that is, the degree of the regression model in (48) (i.e., k) is less than or equal to a given K, which is an assumption universally accepted for such model selection problems. In our Monte Carlo study, 100 realizations were generated from the following model:

$$y = x_1 + 2x_2 + 3x_3 + \varepsilon \quad \varepsilon \sim N(0, \sigma^2). \tag{49}$$

Seven candidate variables stored in an $n \times 7$ matrix X of independent identically distributed normal random variables were considered. All the values of normal pseudo-random numbers of (49) were generated on a SPARC server 1000 using Box and Müller [6] method, which generates normal pseudo-random numbers from uniform pseudo-random numbers. The candidate models included the columns of X in a sequentially nested fashion; i.e. the candidate model of dimension k consisted of columns 1, . . . ,k of X. To compare the performances of the AIC_{bc} and the other criteria, we varied sample sizes (i.e., n = 10, n = 20, n = 100, and n = 500). 100 realizations for each sample size were generated. For each realization, we studied the relative performances of AIC, AIC_c, BIC, CAIC, CAICF, and AIC_{bc} as follow:

AIC[1] given by

$$AIC(k) = -2\ell(\hat{\boldsymbol{\theta}}_k) + 2k, \tag{50}$$

BIC[23] given by

$$BIC(k) = -2\ell(\hat{\boldsymbol{\theta}}_k) + k \log n, \tag{51}$$

the bias-corrected AIC, AIC_c [17] given by

$$AIC_c(k) = -2\ell(\hat{\boldsymbol{\theta}}_k) + \frac{n^2 + n(k-1)}{n-k-1}, \tag{52}$$

CAIC[7] given by

$$CAIC(k) = -2\ell(\hat{\boldsymbol{\theta}}_k) + k(\log n + 1), \tag{53}$$

and CAICF[7] given by (44).

For each sample size, the results of the Monte Carlo study are given in Table 1. In Table 1, numbers under each selected model order represent frequencies of that model order selected by each criterion among 100 realizations for each sample size.

Table 1

Frequency of model order selected by various criteria in 100 realizations of the Monte Carlo study for varying sample sizes n ($\sigma^2 = 1$)

Experiment	Criterion	Selected model order							Proportion of	
		1	2	3*	4	5	6	7	Overfitting	Underfitting
n = 10	AIC	0	0	29	5	5	15	46	0.71	0
	AIC_c	0	0	98	2	0	0	0	0.02	0
	BIC	0	0	34	5	5	15	41	0.66	0
	CAIC	0	0	53	4	7	11	25	0.47	0
	CAICF	0	0	74	4	4	7	11	0.26	0
	AIC_{bc}	0	0	99	1	0	0	0	0.01	0
n = 20	AIC	0	0	50	20	7	7	16	0.50	0
	AIC_c	0	0	84	11	4	0	1	0.16	0
	BIC	0	0	72	16	3	1	8	0.28	0
	CAIC	0	0	84	8	2	1	5	0.16	0
	CAICF	0	0	94	4	1	0	1	0.06	0
	AIC_{bc}	0	0	94	5	1	0	0	0.06	0
n = 100	AIC	0	0	47	2	17	13	21	0.53	0
	AIC_c	0	0	60	2	17	9	12	0.40	0
	BIC	0	0	90	2	5	0	3	0.02	0
	CAIC	0	0	94	1	3	0	2	0.06	0
	CAICF	0	0	98	1	1	0	0	0.02	0
	AIC_{bc}	0	0	93	1	4	0	2	0.07	0
n = 500	AIC	0	0	76	6	14	2	2	0.24	0
	AIC_c	0	0	77	5	14	2	2	0.23	0
	BIC	0	0	99	1	0	0	0	0.01	0
	CAIC	0	0	100	0	0	0	0	0	0
	CAICF	0	0	100	0	0	0	0	0	0
	AIC_{bc}	0	0	100	0	0	0	0	0	0

From the results for small samples (n= 10 and 20) in Table 1, we see that the criteria other than AIC_{bc} and AIC_c show poor performances and have a tendency to overfit the

model. On the other hand, the results show that AIC_{bc} and AIC_c have very good performances. These support our belief that the fourth term in (43) penalizes the overparametrization more strongly for small samples. For large samples (n=100 and 500), AIC_{bc}, BIC, CAIC and CAICF show consistent model order selection as we previously discussed, but AIC and AIC_c do not. Thus, we can conclude that AIC_{bc} has better performance to the other criteria studied in this paper across almost all sample sizes, and provides consistency of order selection.

4. Conclusion

In this paper, we proposed a consistent and bias corrected model selection criterion (AIC_{bc}) shown in (4), which is a consistent and bias corrected extension of Akaike's information criterion, AIC. We investigated the asymptotic properties of this criterion and showed that AIC_{bc} provided a consistent model order selection. Empirical performances of AIC_{bc} over small and large sample sizes showed better order choices of a linear regression model using Monte Carlo experiments than other criteria including AIC, BIC and some modified versions of these.

It is a well-known fact that we should use consistent criteria to avoid overfitting a model and use AIC to avoid underfitting a model. However, the probability of overfitting and underfitting a model by the use of AIC_{bc} was very small over the wide range of sample sizes. From this, we conclude that it can play an important role in model selection problems.

References

[1] Akaike, H. (1974). A new look at the statistical model identification. IEEE Trans. Auto. Cont. 19, 716-723.

[2] Akaike, H. (1977). On entropy maximization principle. In P. R. Krishnaiah (Ed.), Proc. of the Symposium on Applications of Statistics, 27-47, Amsterdam : North-Holland.

[3] Akaike, H. (1979). A Bayesian extension of the minimum AIC procedure of autoregressive model fitting. Biometrika, 66, 237-242.

[4] Anderson, T. W. (1984). An introduction to multivariate statistical analysis, 2nd edition, New York: John Wiley & Sons.

[5] Boltzmann, L. (1877). Uber die Beziehung zwiscen dem zweiten Hauptsatz der mechanischen Warmetheorie und der Wahrscheinlichkeitrechung respective den Satzen

uber das Warmegleichgewicht. Winer Berichte, 76, 373-435.

[6] Box, G.E.P. and Muller M.E. (1958). A Note on the Generation of Random Normal Deviates. Ann. Math. Stat., 29, 610-611.

[7] Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. Psychometrika, 52, 345-370.

[8] Bozdogan, H. (1988). ICOMP: A New Model-Selection Criterion. In Classification and Related Methods of Data Analysis, Ed. Hans H. Bock, Amsterdam: Elsevier Science Publishers B.V. (North-Holland), 599-608.

[9] Bozdogan, H. (1990). On the information-Based Measure of Covariance Complexity and its Application to the Evaluation of Multivariate Linear Models, Communications in Statistics, Theory and Methods, A19, No. 1, 221-278.

[10] Bozdogan, H. (1994a). Choosing the number of clusters, subset selection of variables, and outlier detection in the standard mixture-model cluster analysis. In new approaches in classification and data analysis, E. Diday, Y. Lechevalier, M. Shader, P. Bertrand, and B. Burtschy (Ed.), Berlin, Springer-Verlag, 169-177.

[11] Bozdogan, H. (1994b).Mixture-model cluster analysis using a new informational complexity and model selection criteria. In Multivariate Statistical Modeling, 2, Proc. of the first US/Japan conference on the frontiers of statistical modeling : An informational approach, May 24-29, The University of Tennessee, Knoxville, TN 37996, USA. H. Bozdogan (ed.), Kluwer Academic Publishers, the Netherlands, Dordrecht, 69-113.

[12] Bozdogan, H. and Haughton, D.(1995). Informational complexity criteria for regression models. Invited paper for Statistica Sinica to appear in a special issue on Statistical Model Selection.

[13] Davis, M. H. A. and Vintor, R. B. (1985). Stochastic modeling and control, New York: Chapman and Hall.

[14] Dawid, A. P. (1984). Statistical theory: the prequential approach. J. R. Stat. Soc., A147, 278-292.

[15] Hannan, E. J. and Quinn, B. G. (1979). The determination of the order of an autoregression. J. R. Stat. Soc., B 41, 190-195.

[16] Hosking, J.R.M. (1980). Lagrange-multiplier tests of time-series models. J.R. Stat. Soc. B42, 170-181.

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

[18] Kashyap, R. L. (1982). Optimal choice of AR and MA parts in autoregressive moving average models. IEEE Tr. on Pattern Analysis and Machine Intelligence 4, 99-104.

[19] Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. Annals of Mathematical Statistics, 22, 79-86.

[20] Parzen, E. (1974). Some recent advances in time series modelling. IEEE Trans. Auto. Cont., 19, 723-729.

[21] Rissanen, J. (1983). A universal prior for integers and estimation by minimum description length. Annals of Statistics, 11, 416-431.

[22] Sakamoto, Y., Ishiguro, M., and Kitagawa, G. (1986). Akaike Information Criterion Statistics. KTK Scientific Publishers, Tokyo.

[23] Schwarz, G. (1978). Estimating the dimension of a model. Annals of Statistics, 6, 461-464.

[24] Sclove, S. L. (1987). Application of model-selection criteria to some problems in multivariate analysis. Psychometrika, 52, 333-343.

[25] Sclove, S. L. (1994). Some aspects of model-selection criteria. In Multivariate Statistical Modeling, Vol. 2, Proc. of the First US/Japan Conference on the Frontiers of Statistical Modeling: An Informational Approach, H. Bozdogan (ed.), Kluwer Academic Publishers, The Netherlands, Dordrecht, 37-67.

[26] Searle, S. R. (1982). Matrix algebra useful for statistics, New York: John Wiley & Sons.

[27] Shibata, R. (1980). Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. Annals of Statistics, 8, 147-164.

[28] Shibata, R. (1981). An optimal selection of regression variables. Biometrika, 68, 45-54.

[29] Shibata, R. (1989). Statistical Aspects of Model Selection. In From Data to Modeling, J.C. Willems (Ed.), Berlin, Springer-Verlag, 216-240.

[30] Stone, M. (1979). Cross-validatory choice and assessment of statistical predictions (with discussions). J. R. Stat. Soc., B36, 111-147.

[31] Stuart, A. and Ord, J. K. (1991). Kendall's advanced theory of statistics, vol. 2, Fifth edition, London: Edward Arnold.

[32] Takeuchi, K. (1976) Distribution of information Statistics and a Criterion of Model Fitting. Surikagaku (Mathematical Sciences), Vol. 153, 12-18. (in Japanese).

[33] Tong, H. (1989). Nonlinear Time Series Analysis. Oxford Univ. Press.

[34] White, H. (1982). Maximum likelihood estimation of misspecified models, Econometrica, 50, 1-25.