Ann. Inst. Statist. Math.
Vol. 49, No. 3, 411–434 (1997)

BOOTSTRAPPING LOG LIKELIHOOD AND EIC, AN EXTENSION OF AIC

MAKIO ISHIGURO, YOSIYUKI SAKAMOTO AND GENSHIRO KITAGAWA

The Institute of Statistical Mathematics, 4-6-7 Minami-Azabu, Minato-ku, Tokyo 106, Japan

(Received November 18, 1994; revised February 18, 1997)

Abstract. Akaike (1973, *2nd International Symposium on Information Theory*, 267–281, Akademiai Kiado, Budapest) proposed AIC as an estimate of the expected log likelihood to evaluate the goodness of models fitted to a given set of data. The introduction of AIC has greatly widened the range of application of statistical methods. However, its limit lies in the point that it can be applied only to the cases where the parameter estimation are performed by the maximum likelihood method. The derivation of AIC is based on the assessment of the effect of data fluctuation through the asymptotic normality of MLE. In this paper we propose a new information criterion EIC which is constructed by employing the bootstrap method to simulate the data fluctuation. The new information criterion, EIC, is regarded as an extension of AIC. The performance of EIC is demonstrated by some numerical examples.

Key words and phrases: Log likelihood, AIC, bootstrap, MLE, information criterion, model selection, estimator selection, bias correction, expected log likelihood, penalized log likelihood, predictive distribution.

1. Introduction

This paper is addressed to an extension of the Akaike Information Criterion, AIC. AIC was proposed as an estimate of (minus twice) the expected log likelihood. When the parameters of a model are estimated by the maximum likelihood method, the maximum value of the log likelihood has a positive bias as an estimator of the expected log likelihood. AIC is then obtained by approximately correcting the bias (Akaike (1973), Sakamoto *et al.* (1986)).

Hence, the application of the AIC is restricted to the models with maximum likelihood estimates. In principle, however, the basic idea is applicable to the evaluation of the models fitted by much wider class of estimation procedures, if the bias can be evaluated. In the derivation of AIC, based on the Taylor expansion of both the log likelihood and the expected log likelihood and the asymptotic normality of the maximum likelihood estimators, the estimate of the bias is obtained analytically. Obviously, this type of analytic approach does not necessarily apply to all the class of models and the estimation procedures.

In this paper, we exploit the bootstrap method (Efron (1979)) for the evaluation of the bias. Our method has several advantages such as:

1. It can be regarded as an extension of AIC.
2. It can be applied to the models estimated by non-MLE type procedures including Bayesian procedures.
3. Analytic approximations or asymptotic theorems are not directly used in the bootstrapping. Therefore, it can be expected that even for the MLE case, it may provide a better bias estimate depending on a particular case.

There are attempts to improve AIC (Takeuchi (1976), Sugiura (1978), Hurvich and Tsai (1989)). Their works are to provide better estimates of the bias of the maximum log likelihood as an estimator of the expected log likelihood. Their criteria give better answers than AIC, in some situations for those problems where parameters are estimated by the maximum likelihood method. In this sense, their purposes differ from ours to expand the range of applicability. Shibata (1989) proposes RIC as an extension of AIC for the case where parameter estimation is performed by maximum penalized likelihood method. His purpose is close to ours. The difference is that his method still employs analytic method to assess the fluctuation of the parameter estimate, which is done by bootstrap method in our case. AIC_I by Hurvich *et al.* (1990) extended the AIC criterion to non-MLE estimators for AR model order selection. The "modified likelihood" discussed by Wong (1983) is, in a sense, most close to our approach. His criterion for the choice of kernel width can be seen as a special case of our approach applied to a specific problem of kernel estimation of density.

The construction of this paper is as follows. In Section 2, giving a brief review of the derivation of AIC, a new information criterion EIC is defined. Numerical examples are given in Section 3. Concluding remarks are given in Section 4.

2. AIC and EIC

2.1 *Predictive distributions*

In this paper, to handle a broader class of models and estimators, we use a notion of a predictive distribution, $h(y \mid x)$, which is the distribution of a future observation y given the present observation x.

The predictive distribution can be constructed by various ways. Two typical examples are shown below.

Example 1. If we have a Bayesian model consists of a parametric model $f(y \mid \theta)$ and a prior distribution $\pi(\theta)$ of the parameter, then $h(y \mid x)$ can be defined by

$$h(y \mid x) = \int g(y \mid \theta)\pi(\theta \mid x)d\theta, \tag{2.1}$$

where $g(y \mid \theta)$ is the model of the future observation y, and $\pi(\theta \mid x)$ is the posterior distribution of θ given by

$$\pi(\theta \mid x) = f(x \mid \theta)\pi(\theta) / \int f(x \mid \theta)\pi(\theta)d\theta.$$

We temporarily use the term of Bayesian predictive distribution to refer to this special type of predictive distribution. In many practical situations, g and f are one and the same.

Example 2. Given a parametric model $f(y \mid \theta)$ and a point estimator of θ, a predictive distribution can be obtained by

$$h(y \mid x) = g(y \mid \hat{\theta}(x)). \tag{2.2}$$

Here $\hat{\theta}(x)$ denotes an estimator of θ based on the data x. $\hat{\theta}(x)$ can be MLE or the posterior mode, Arg. $\max \pi(\theta \mid x)$, or anything else. Note that $h(x \mid x)$ can be defined when g and f are the same, and it is the maximum likelihood when $\hat{\theta}(x)$ is MLE.

2.2 *A brief review of the derivation of AIC*

From the point of view of the entropy maximization principle proposed by Akaike (1973), the goodness of the predictive distribution can be evaluated by the expected log likelihood

$$E_Y \log h(Y \mid x) = \int \log h(y \mid x) dG(y), \tag{2.3}$$

where $G(y)$ denotes the true distribution of y. A natural estimate of the expected log likelihood is provided by $\log h(x \mid x)$. Its bias is defined by

$$C = E_X\{\log h(X \mid X) - E_Y \log h(Y \mid X)\}. \tag{2.4}$$

This bias appears since the same data set X is used for both the estimation of the parameter and the estimation of the expected log likelihood. In actual statistical problems, the true distribution is seldom known and only a sample X drawn from $G(x)$ is given. If an estimate $\hat{C}$ of C is available, a bias corrected log likelihood of the predictive distribution is obtained by

$$\log h(X \mid X) - \hat{C}. \tag{2.5}$$

In particular, in parametric modeling where a model has a parameter (vector) θ, the entropy maximization principle naturally leads to the maximization of the log likelihood function

$$\ell_X(\theta) = \log f(X \mid \theta) \tag{2.6}$$

and to the maximum likelihood estimate, $\hat{\theta} = \hat{\theta}(X)$, and then the predictive distribution $f(Y \mid \hat{\theta}(X))$. The unbiased estimate of the expected log likelihood is given by

$$\log f(X \mid \hat{\theta}(X)) - C. \tag{2.7}$$

Akaike (1973) showed that C can be asymptotically approximated by the dimension of the parameter vector θ, and defined AIC as

$$\text{AIC} = -2\ell_X(\hat{\theta}(X)) + 2(\text{dimension of } \theta). \tag{2.8}$$

Here is a brief review of the derivation of AIC. Note that $\hat{\theta}(X)$ in the following context does not necessarily represent MLE. If we are dealing with the special case of MLE, we will explicitly say so.

Assume, for simplicity, that the log likelihood function of the parameter vector θ is given (or at least locally approximated within the range of variation of $\hat{\theta}(X)$) by

$$\ell_X(\theta) = \ell_X - \frac{1}{2}(\theta - \theta_X)^T H(\theta - \theta_X), \tag{2.9}$$

where ℓ_X and θ_X fluctuate depending on the realization of data X, but the non-negative definite Hessian matrix H does not (see Appendix A). H does depend on the distribution of X, but not on its realization. Then the expected log likelihood function of the parameter vector θ is given by

$$\ell_0(\theta) \equiv \ell_0 - \frac{1}{2}(\theta - \theta_0)^T H(\theta - \theta_0) \equiv E_X\{\ell_X(\theta)\}, \tag{2.10}$$

where θ_0 and ℓ_0 are given by

$$\theta_0 = E_X\{\theta_X\}$$

and

$$\ell_0 = E_X\{\ell_X(\theta_0)\},$$

respectively. In the following, $\Delta\theta_X$ denotes the difference $\theta_X - \theta_0$.

The bias term C can be decomposed into three terms:

$$\begin{aligned} C &= E_X\{\ell_X(\hat{\theta}(X)) - \ell_X(\theta_0)\} \\ &\quad + E_X\{\ell_X(\theta_0) - \ell_0(\theta_0)\} \\ &\quad + E_X\{\ell_0(\theta_0) - \ell_0(\hat{\theta}(X))\} \end{aligned} \tag{2.11}$$

$$\begin{aligned} &\equiv C_1 + C_2 + C_3 \\ &= E_X\{C_1(X)\} + E_X\{C_2(X)\} + E_X\{C_3(X)\}. \end{aligned} \tag{2.12}$$

By definition, $E_X\{C_2(X)\} = 0$. For the calculation of C_1, if the approximations (2.9) and (2.10) are good enough for $\theta = \hat{\theta}(X)$, we have

$$\begin{aligned} C_1(X) &\equiv \ell_X(\hat{\theta}(X)) - \ell_X(\theta_0) \\ &= -\frac{1}{2}(\hat{\theta}(X) - \theta_X)^T H(\hat{\theta}(X) - \theta_X) \\ &\quad + \frac{1}{2}(\theta_0 - \theta_X)^T H(\theta_0 - \theta_X) \end{aligned} \tag{2.13}$$

$$
\begin{aligned}
&= -\frac{1}{2}\hat{\theta}(X)^T H\hat{\theta}(X) + \hat{\theta}(X)^T H\theta_X + \frac{1}{2}\theta_0^T H\theta_0 - \theta_0^T H\theta_X \\
&= -\frac{1}{2}\hat{\theta}(X)^T H\hat{\theta}(X) + \hat{\theta}(X)^T H(\Delta\theta_X + \theta_0) \\
&\quad + \frac{1}{2}\theta_0^T H\theta_0 - \theta_0^T H(\Delta\theta_X + \theta_0) \\
&= -\frac{1}{2}\hat{\theta}(X)^T H\hat{\theta}(X) + \hat{\theta}(X)^T H\Delta\theta_X + \hat{\theta}(X)^T H\theta_0 \\
&\quad \frac{1}{2}\theta_0^T H\theta_0 - \theta_0^T H\Delta\theta_X \\
&= -\frac{1}{2}(\hat{\theta}(X) - \theta_0)^T H(\hat{\theta}(X) - \theta_0) \\
&\quad + \hat{\theta}(X)^T H\Delta\theta_X - \theta_0^T H\Delta\theta_X.
\end{aligned}
$$

Since

$$
\begin{aligned}
C_3(X) &\equiv \ell_0(\theta_0) - \ell_0(\hat{\theta}(X)) \\
&= \frac{1}{2}(\hat{\theta}(X) - \theta_0)^T H(\hat{\theta}(X) - \theta_0),
\end{aligned} \tag{2.14}
$$

we have

$$
C_1(X) + C_3(X) = \hat{\theta}(X)^T H\Delta\theta_X - \theta_0^T H\Delta\theta_X. \tag{2.15}
$$

Hence, we have

$$
C = E_X\{C_1(X) + C_3(X)\} = E_X\{\hat{\theta}(X)^T H\Delta\theta_X\}. \tag{2.16}
$$

Define

$$
U = E_X\{\Delta\theta_X \ddot{\theta}(X)^T\}
$$

then C is given by

$$
C = \operatorname{tr} HU.
$$

If $\hat{\theta}(X)$ is MLE, assuming asymptotic normality, we approximately have $\Delta\theta_X \sim N(0, H^{-1})$, which means $U = H^{-1}$ and C equals the dimension of θ. This gives AIC.

2.3 *Bootstrapping*

A sample of size n, $X = (X_1, \ldots, X_n)$ is drawn from the true distribution $G(x)$. The empirical distribution $G_*(x)$ is then defined by

$$
G_*(x) = \frac{1}{n}\sum_{i=1}^{n} I(x, X_i),
$$

where $I(x, a)$ is the function defined by $I(x, a) = 0$ if $x < a$ and $I(x, a) = 1$ otherwise. A random sample of size m (usually we put $m = n$) from the empirical

distribution G_* is called a bootstrap sample (Efron (1979)) and is denoted by $X^* = (X_1^*, \ldots, X_m^*)$.

In the bootstrapping, the true distribution $G(x)$ is replaced by the empirical distribution $G_*(x)$. Therefore, we will apply the following replacement;

$$\begin{aligned} G &\to G_* \\ X, Y \sim G &\to X^*, Y^* \sim G_* \\ E_Y \log h(Y \mid \cdot) &\to E_{Y^*} \log h(Y^* \mid \cdot). \end{aligned} \tag{2.17}$$

Here E_{Y^*} denotes the expectation under the empirical distribution $G_*(x)$.

2.4 *EIC*

The bootstrap estimate of the bias C is given by

$$C^*(X) = E_{X^*}\{\log h(X^* \mid X^*) \quad E_{Y^*} \log h(Y^* \mid X^*)\}. \tag{2.18}$$

Then the bootstrap bias correction for the log likelihood is

$$\log h(X \mid X) - C^*(X), \tag{2.19}$$

and following the definition of AIC, we define the bootstrap version of the information criterion as follows:

$$\mathrm{EIC} = -2 \log h(X \mid X) + 2C^*(X). \tag{2.20}$$

If $E_X\{C^*(X)\} = C$ holds, EIC defines an unbiased estimator of the (minus twice) expected log likelihood. In the case of $h(Y \mid X) = f(Y \mid \hat{\theta}(X))$, the unbiasedness of EIC is proved under a mild condition on the estimator $\hat{\theta}(X)$ of θ.

Let an estimate of parameter θ based on the bootstrap sample X^* be denoted by $\hat{\theta}(X^*)$ and assume that the log likelihood function based on the same sample is approximated (at least locally) by

$$\ell_{X^*}(\theta) = \ell_{X^*} - \frac{1}{2}(\theta - \theta_{X^*})^T H (\theta - \theta_{X^*}).$$

Define

$$\begin{aligned} \theta_0^* &= E_{X^*}\{\theta_{X^*}\}, \\ \Delta\theta_{X^*} &= \theta_{X^*} - \theta_0^* \end{aligned}$$

and

$$U^*(X) = E_{X^*}\{\Delta\theta_{X^*} \hat{\theta}(X^*)^T\}. \tag{2.21}$$

If

$$E_X\{U^*(X)\} = U \tag{2.22}$$

holds, $C^*(X)$ provides an unbiased estimate of C.

A wide range of estimators including maximum penalized likelihood estimators satisfy the condition (2.22).

Remark A. In the simple i.i.d. situation,

$$E_{Y^*} \log h(Y^* \mid \cdot) = \int \log h(y^* \mid \cdot) dG_*(y^*) \tag{2.23}$$
$$= \frac{n}{n} \sum_{i=1}^{n} \log h(x_i \mid \cdot) \equiv \log h(X \mid \cdot)$$

Therefore the bootstrap estimate of the bias becomes simply

$$C^* = E_{X^*}\{\log h(X^* \mid X^*) - \log h(X \mid X^*)\}. \tag{2.24}$$

EIC in this case is equivalent to WIC proposed by Ishiguro and Sakamoto (1991).

2.4.1 *Example 1: MPLE*

Let a penalized log likelihood is approximately given by

$$\ell_P(\theta) = \ell_X(\theta) - \frac{1}{2}(\theta - \theta_P)^T W(\theta - \theta_P) \tag{2.25}$$
$$= \ell_X - \frac{1}{2}(\theta - \theta_X)^T H(\theta - \theta_X) - \frac{1}{2}(\theta - \theta_P)^T W(\theta - \theta_P),$$

with an arbitrarily fixed vector θ_P and a non-negative definite matrix W. The maximum penalized log likelihood estimate, denoted by MPLE, is given by

$$\hat{\theta}(X) = (H + W)^{-1} H\theta_X + (H + W)^{-1} W\theta_P, \tag{2.26}$$

if the inversion is possible. Note that we can choose W so that $(H + W)$ is nonsingular. Using the well-known relation

$$E_X\{\Delta\theta_X \Delta\theta_X^T\} = H^{-1}$$

it is shown that U in this case is given by

$$U = (H + W)^{-1}.$$

Assume that the log likelihood function based on the bootstrap sample is given by

$$\ell_{X^*}(\theta) = \ell_{X^*} - \frac{1}{2}(\theta - \theta_{X^*})^T H^*(\theta - \theta_{X^*}) \tag{2.27}$$

and the approximation $H^* = H$ holds, the condition (2.22) is satisfied. Note that the ordinary MLE is a special case of MPLE, and then the above argument proves, when the conditions are met, that EIC is an extension of AIC.

2.4.2 *Example 2: MAICE*

There are estimators which fail to meet the condition (2.22).

Let x be an n-vector $(x_1, x_2, \ldots, x_n)$ of independent samples from normal distribution with mean μ_0 and variance 1 which is denoted by $N(\mu_0, 1)$. Let us consider two models $N(0,1)$ and $N(\mu,1)$. The log likelihood of the model $N(\mu,1)$ is given by

$$\begin{aligned}(2.28)\qquad \log f(x \mid \mu) &= -\frac{n}{2}\log 2\pi - \frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2 \\ &= -\frac{n}{2}\log 2\pi - \frac{1}{2}\sum_{i=1}^{n}x_i^2 + \frac{1}{2n}\left(\sum_{i=1}^{n}x_i\right)^2 \\ &\quad -\frac{1}{2}n\left(\mu - \frac{1}{n}\sum_{i=1}^{n}x_i\right)^2.\end{aligned}$$

The maximum likelihood estimator of μ is defined by

$$(2.29)\qquad \hat{\mu}_X = \frac{1}{n}\sum_{i=1}^{n}x_i.$$

AIC's for two models are given by

$$\mathrm{AIC}_0 = n\log 2\pi + \sum_{i=1}^{n}x_i^2$$

and

$$\mathrm{AIC}_1 = n\log 2\pi + \sum_{i=1}^{n}x_i^2 - n\hat{\mu}_X^2 + 2,$$

respectively. Then the minimum AIC estimator is defined by

$$(2.30)\qquad \hat{\mu}(X) = \begin{cases}\mu_X & (\text{if } n\mu_X^2 > 2)\\ 0 & (\text{otherwise}).\end{cases}$$

Let $t = \sqrt{n}\hat{\mu}_X$ and

$$\Delta\mu_X = \hat{\mu}_X - \mu_0 = \frac{t}{\sqrt{n}} - \mu_0.$$

Since $t \sim N(\sqrt{n}\mu_0, 1)$,

$$\begin{aligned}(2.31)\qquad U &= E_X\{\Delta\mu_X\hat{\mu}(X)^T\} \\ &= \frac{1}{n}\int_{t^2>2} t(t-\sqrt{n}\mu_0)\frac{1}{\sqrt{2\pi}}e^{-(t-\sqrt{n}\mu_0)^2/2}dt.\end{aligned}$$

When $\mu_0 = 0$,

$$(2.32)\qquad U = \frac{1}{n}\int_{t^2>2} t^2\phi(t)dt,$$

where ϕ is the probability density function of the standard normal distribution. Comparing (2.9) and (2.28), we have $H = n$ and

$$C = \int_{t^2>2} t^2 \phi(t) dt.$$

Let μ_{X^*} denote the MLE based on the bootstrap sample X^*. If n is large, $s = \sqrt{n}\mu_{X^*}$ is approximately normally distributed with mean t and variance 1 (see Appendix B). Then,

$$U^*(X) = E_{X^*}\{\Delta\mu_{X^*}\hat{\mu}(x^*)^T\} = \frac{1}{n}\int_{s^2>2} s(s-t)\frac{1}{\sqrt{2\pi}}e^{-(s-t)^2/2}ds. \tag{2.33}$$

When $\mu_0 = 0$,

$$E_X\{U^*(X)\} = \frac{1}{n}\int_{s^2>2}\int s(s-t)\frac{1}{\sqrt{2\pi}}e^{-(s-t)^2/2}\frac{1}{\sqrt{2\pi}}e^{-t^2/2}dtds = \frac{1}{n}\int_{s^2>1} s^2\phi(s)ds.$$

This means that the condition (2.22) is violated in this case. The expectation of $C^*(x)$ is calculated to be

$$E_X\{C^*(X)\} = \int_{s^2>1} s^2\phi(s)ds.$$

It is clear that $C^*(X)$ is biased as an estimator of C for this example. And it would be the case for all the minimum AIC estimates.

2.5 *Reduction of the computational cost*

As seen from the derivation of AIC, $C^*(X)$ can be estimated by

$$C^{**}(X) = E_{X^*}\{C_1(X^*)\} + E_{X^*}\{C_3(X^*)\} = E_{X^*}\{C_1(X^*) + C_3(X^*)\}.$$

Since it can be seen that, especially for large n, the omitted term $C_2(X^*)$ has the largest variance, the variance of $C^{**}(X)$ is significantly less than that of the original bias estimate $C^*(X)$ given in (2.18). Therefore, the number of necessary bootstrap replication to attain a certain accuracy can be reduced with this modification.

3. Numerical examples

3.1 *CATDAP model selection*

To show that EIC can be applicable to those problems that are effectively handled with AIC, we apply it to CATDAP model selection (Sakamoto *et al.* (1986), Sakamoto (1991)). CATDAP (A Categorical Data Analysis Program) is a FORTRAN program that searches for the best explanatory variable of a categorical response variable. The basic model of this program is as follows:

We denote the response variable by I_0, and any subset of explanatory variables $I = \{I_1, \ldots, I_k\}$ by J, and denote respective realizations by i_0, i and j. Let $\theta(i_0, i)$ be the probability that the variable I_0 and the set of variables I take a set of values (i_0, i) and let $n(i_0, i)$ be the corresponding cell frequency. Then, the probability $\Pr(\{n(i_0,i)\} \mid \{\theta(i_0,i)\})$ of getting the cell frequencies $\{n(i_0,i)\}$ under a set of probabilities $\{\theta(i_0,i)\}$ is obtained from the multinomial distribution as

$$\Pr(\{n(i_0,i)\} \mid \{\theta(i_0,i)\}) = \prod_{i_0,i} \theta(i_0,i)^{n(i_0,i)}, \tag{3.1}$$

where the constant term independent of the parameter $\theta(i_0, i)$ is ignored. The log likelihood with respect to the parameter $\theta(i_0, i)$ is given by

$$\sum_{i_0,i} n(i_0,i) \log \theta(i_0,i). \tag{3.2}$$

If we denote by $\theta(i_0 \mid i)$ the conditional probability of i_0 given a value i of I, then

$$\theta(i_0, i) = \theta(i_0 \mid i)\theta(i) \tag{3.3}$$

and (3.2) can be written as

$$\sum_{i_0,i} n(i_0,i) \log \theta(i_0 \mid i) + \sum_{i} n(i) \log \theta(i), \tag{3.4}$$

where $\theta(i)$ and $n(i)$ denote the marginal probability and cell frequency with respect to I, respectively. Since the term of interest is not $\theta(i)$ but $\theta(i_0 \mid i)$, we consider the conditional log likelihood defined by

$$\sum_{i_0,i} n(i_0,i) \log \theta(i_0 \mid i). \tag{3.5}$$

The evaluation of any subset of explanatory variables J can be performed by evaluating the goodness of the model

$$\text{MODEL}(I_0 \mid J): \quad \theta(i_0 \mid i) = \theta(i_0 \mid j). \tag{3.6}$$

The AIC for this model can be written as

$$\text{AIC}(I_0 \mid J) = (-2) \sum_{i_0,j} n(i_0,j) \log \frac{n(i_0,j)}{n(j)} + 2C_J(C_0 - 1), \tag{3.7}$$

where the notation is as follows:

$n(i_0, j)$: the cell frequency for the cell (i_0, j) that is a realization of the variables I_0 and J,

$n(j)$: the marginal frequency with respect to J,

C_0 and C_J : the number of categories for I_0 and J, respectively.

We assume here that $n(\Phi) = n$ (n is the sample size) and $C_\Phi = 1$.

The EIC for this case is given by

$$\text{EIC}(I_0 \mid J) = (-2)\sum_{i_0,j} n(i_0,j)\log\frac{n(i_0,j)}{n(j)} + 2\mathrm{E}_{n^*(i_0,j)}\left[\sum_{i_0,j} n^*(i_0,j)\log\frac{n^*(i_0,j)}{n^*(j)} - \sum_{i_0,j} n(i_0,j)\log\frac{n^*(i_0,j)}{n^*(j)}\right], \tag{3.8}$$

where $n^*(i_0, j)$ and $n^*(j)$ are the cell frequencies with respect to the relevant variables for the bootstrap sample.

To evaluate the performance of EIC, we assumed that

$$\text{MODEL}(I_0 \mid I_1) : \theta(i_0 \mid i_1, i_2, i_3) = \theta(i_0 \mid i_1)$$

is the true model, and conducted the experiment of detecting this true model from data that follow this model. We generated random samples of size 100, 200 and 1000, and repeated 100 times for each case. Table 1.1 summarizes the results obtained from the experiment for the case of $n = 1000$. In this table "$\mathrm{E}\{-2n$ E(LL)$\}$" stands for $-2n$ times the mean of the expected log likelihood of the relevant model. The entry in the parentheses is the standard deviations of the difference in $-2n$ E(LL) between the relevant model and the true one. "Freq. of min $-2n$ E(LL)" is the frequency with which the quantity $-2n$ E(LL) of the relevant model took the minimum among eight models. Therefore, it is seen that MODEL($I_0 \mid I_1$) was judged to be the best model in terms of both measures.

The entry in the fourth column is the mean of AIC value of each model. The fifth column under the header "Freq. of MAICE" shows the frequency with which the corresponding models were chosen as the best model by the minimum AIC procedure. The table shows that MODEL($I_0 \mid I_1$), the true structure model, was chosen 73 times out of a hundred tries by the minimum AIC estimate (MAICE) procedure. The next frequently chosen model is MODEL($I_0 \mid I_1, I_3$), and MODEL($I_0 \mid I_1, I_2$) follows. The performance of the minimum EIC procedure, which is shown in the last two columns, is almost the same with that of MAICE procedure. Tables 1.2 and 1.3 summarize experiments for smaller data set of $n = 200$ and $n = 100$, respectively. From these tables it is observed that EIC behaves slightly better than AIC, especially in the case of small sample.

Table 1.1. CATDAP model selection ($n = 1000$).

MODEL	E{$-2n$ E(LL)}	Freq. of min $-2n$ E(LL)	AIC	Freq. of MAICE	EIC	Freq. of min EIC
(I_0)	1387.74	0	1386.85	0	1385.47	0
	(1.44)		(12.19)		(12.19)	
$(I_0 \mid I_1)$	1348.70	100	1348.64	73	1347.32	74
	(0.00)		(0.00)		(0.00)	
$(I_0 \mid I_2)$	1388.87	0	1387.72	0	1386.37	0
	(1.98)		(12.30)		(12.30)	
$(I_0 \mid I_3)$	1388.81	0	1387.79	0	1386.42	0
	(1.88)		(12.27)		(12.27)	
$(I_0 \mid I_1, I_2)$	1350.88	0	1350.45	11	1349.21	11
	(2.02)		(2.00)		(2.00)	
$(I_0 \mid I_1, I_3)$	1351.10	0	1350.27	12	1348.98	11
	(1.92)		(1.90)		(1.90)	
$(I_0 \mid I_2, I_3)$	1389.59	0	1388.07	0	1386.73	0
	(3.67)		(12.49)		(12.49)	
$(I_0 \mid I_1, I_2, I_3)$	1355.94	0	1353.57	4	1352.48	4
	(3.68)		(3.50)		(3.50)	

Table 1.2. CATDAP model selection ($n = 200$).

MODEL	E{$-2n$ E(LL)}	Freq. of min $-2n$ E(LL)	AIC	Freq. of MAICE	EIC	Freq. of min EIC
(I_0)	278.30	0	278.22	4	277.96	4
	(1.37)		(5.55)		(5.55)	
$(I_0 \mid I_1)$	271.32	99	271.59	64	271.37	68
	(0.00)		(0.00)		(0.00)	
$(I_0 \mid I_2)$	279.23	0	279.30	1	279.08	0
	(1.86)		(5.75)		(5.75)	
$(I_0 \mid I_3)$	279.51	0	279.05	3	278.84	2
	(1.89)		(5.79)		(5.79)	
$(I_0 \mid I_1, I_2)$	273.41	0	273.62	11	273.60	10
	(2.00)		(1.90)		(1.90)	
$(I_0 \mid I_1, I_3)$	273.55	1	273.50	14	273.46	13
	(2.05)		(1.97)		(1.97)	
$(I_0 \mid I_2, I_3)$	281.20	0	280.77	2	280.73	3
	(2.96)		(6.13)		(6.13)	
$(I_0 \mid I_1, I_2, I_3)$	277.88	0	277.62	1	278.66	0
	(4.02)		(3.45)		(3.41)	

Table 1.3. CATDAP model selection ($n = 100$).

MODEL	E{$-2n$ E(LL)}	Freq. of min $-2n$ E(LL)	AIC	Freq. of MAICE	EIC	Freq. of min EIC
(I_0)	139.73	3	139.54	17	139.42	19
	(1.78)		(4.60)		(4.58)	
$(I_0 \mid I_1)$	136.64	95	136.36	49	136.33	53
	(0.00)		(0.00)		(0.00)	
$(I_0 \mid I_2)$	140.76	0	140.57	0	140.54	0
	(2.35)		(4.87)		(4.84)	
$(I_0 \mid I_3)$	140.70	0	140.62	4	140.60	5
	(2.25)		(4.84)		(4.81)	
$(I_0 \mid I_1, I_2)$	138.97	1	138.21	16	138.71	11
	(2.66)		(2.31)		(2.27)	
$(I_0 \mid I_1, I_3)$	139.07	1	138.36	12	138.82	11
	(2.20)		(1.98)		(1.94)	
$(I_0 \mid I_2, I_3)$	142.60	0	142.54	2	143.00	1
	(3.28)		(5.38)		(5.33)	
$(I_0 \mid I_1, I_2, I_3)$	144.31	0	142.40	0	143.39	0
	(4.95)		(3.66)		(4.33)	

3.2 *AR model order selection*

Next example is the AR model order selection. This example is to show that EIC can be used in some aspect of time series analysis. AR model of time series data is defined by

$$x_i = \sum_{j=1}^{m} a_j x_{i-j} + \varepsilon_i \tag{3.9}$$

The least squares estimates of $a_1, a_2, \ldots, a_m$ are obtained by minimizing

$$\|X\beta_m(a)\|^2, \quad \text{where} \quad X = \begin{pmatrix} X_{M+1} \\ X_{M+2} \\ \vdots \\ X_n \end{pmatrix},$$

$$\beta_m(a) = (-1, a_1, \ldots, a_m, 0, \ldots, 0)^T \quad (m \leq M)$$

and

$$X_i = (x_i, x_{i-1}, \ldots, x_{i-M}).$$

Then an estimate of σ^2, the variance of the innovation ε_i is defined by

$$\hat{\sigma}_m^2 - \frac{1}{n-M}\|X\beta_m(\hat{a})\|^2.$$

The resampling in this case is applied to the rows $X_{M+1}, \ldots, X_n$ of the matrix X. We define *resampled matrix* X^* by

$$X^* = \begin{pmatrix} X_{(M+1)^*} \\ X_{(M+2)^*} \\ \vdots \\ X_{n^*} \end{pmatrix},$$

where $(M+1)^*, \ldots, n^*$ are $(n-M)$ independent realizations of uniform integer valued random numbers on the interval $[M+1, n]$.

Estimates of $a_1, a_2, \ldots, a_m$ and σ_m^2 are defined by

$$(\hat{a}_1^*, \hat{a}_2^*, \ldots, \hat{a}_m^*) = \text{Arg. min} \|X^* \beta_m(a)\|^2$$

and

$$\hat{\sigma}_m^{*2} = \frac{1}{n-M} \|X \beta_m(\hat{a}^*)\|^2,$$

respectively.

EIC for the AR model of order m is given by

$$\begin{aligned} \text{EIC}(m) = & -2 \times \log f(X \mid \hat{a}_1, \hat{a}_2, \ldots, \hat{a}_m, \hat{\sigma}_m^2) \\ & + 2 \times E_{X^*}\{\log f(X^* \mid \hat{a}_1^*, \hat{a}_2^*, \ldots, \hat{a}_m^*, \hat{\sigma}_m^{*2}) \\ & \qquad - \log f(X \mid \hat{a}_1^*, \hat{a}_2^*, \ldots, \hat{a}_m^*, \hat{\sigma}_m^{*2})\}, \end{aligned} \tag{3.10}$$

where

$$f(X \mid a_1, a_2, \ldots, a_m, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^{(n-M)} \exp\left\{-\frac{1}{2\sigma^2} \|X \beta_m(a)\|^2\right\}.$$

For a numerical example, hundred series of data of length 100 were generated assuming a second order AR model with coefficients $a_1 = 1.5$, $a_2 = -0.7$ and $\sigma^2 = 1.0$, as the true structure and analyzed. M is fixed at 7. Results of the experiment are summarized in Table 2. AIC_I in this table stands for AIC_I proposed by Hurvich *et al.* (1990), which is eventually defined by

$$\text{AIC}_I = \log \det \hat{\sum_m} + E_I \left\{ \text{tr} \hat{\sum_m}^{-1} \right\}$$

where $\hat{\sum}_m$ is the theoretical covariance matrix of time series, $\{x_{M+1}, \ldots, x_n\}$ generated by AR model (3.9) with estimated parameters. E_I denotes the expectation with respect to the distribution of white noise of unit variance. The entries in the parentheses are the standard deviations of differences of ELL's, AIC's, EIC's and AIC_I's from those of the true model. This table shows that at least for this example, EIC and AIC_I show similar performance and selected the true order slightly more than AIC.

Table 2. AR order selection. ELL, AIC, EIC and AIC_I. The entries in the parentheses are the standard deviations of differences of their value from those of the true model.

Order	ELL	AIC	Freq. of MAICE	EIC	Freq. of min EIC	AIC_I	Freq. of min AIC_I
0	465.81	466.24	0	466.04	0	559.78	0
	(18.85)	(21.68)		(21.70)		(210.68)	
1	328.07	330.45	0	330.26	0	423.76	0
	(9.46)	(17.04)		(17.07)		(17.04)	
2	266.85	267.62	80	267.71	83	361.23	84
	(0.0)	(0.0)		(0.0)		(0.0)	
3	267.60	268.93	4	269.28	3	362.64	5
	(0.86)	(0.77)		(0.85)		(0.77)	
4	269.04	269.67	10	270.30	9	363.99	7
	(2.56)	(2.11)		(2.27)		(2.11)	
5	270.10	270.72	5	271.74	5	365.69	3
	(2.98)	(2.49)		(2.73)		(2.49)	
6	271.09	271.87	1	273.42	0	366.92	1
	(3.34)	(2.71)		(2.92)		(2.71)	

Note. When $\{x_i\}$ is a stationary series, the expected log likelihood of the AR model of order k satisfies

$$\lim_{n\to\infty}\frac{1}{n}\log f_n(x_1,x_2,\ldots,x_n\mid\theta)$$
$$=\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\log f(x_i\mid x_{i-1},\ldots,x_{i-k})$$
$$=\int\log\underbrace{f(x_{k+1}\mid x_k,\ldots,x_1)}_{\text{model}}\underbrace{\pi(x_1,\ldots,x_{k+1})}_{\text{true structure}}dx_1,\ldots,dx_{k+1},$$

where $\pi(x_1,\ldots,\pi_{k+1})$ denotes the stationary joint probability density function of successive $(k+1)$ observations $x_1,\ldots,x_{k+1}$.

This equation suggests that the bootstrap to reproduce the fluctuation of the expected log likelihood should be of the form

$$l^*(\theta)=\sum_{i=M+1}^{n}\log f(x_{i^*}\mid x_{i^*-1},\ldots,x_{i^*-k}),$$

This is equivalent to the matrix resampling. There are cases where the rank of the *resampled* matrix X^* is less than $(M+1)$. We discard those cases for the evaluation of EIC. Our method might be called "controlled resampling". Both EIC and AIC_I are based on the idea that the bias term can be estimated using Monte Carlo simulation. They are different in two aspects. The first difference is in the choice of random data generator. The second is the procedure to calculate the bias. AIC_I utilizes the knowledge of specified structure of the AR model.

3.3 *MLE and non-MLE*

In this subsection, we show that EIC can be applicable to non ML procedures. In the presented example, the expected log likelihood can be evaluated analytically and compared with EIC. Assume that a sample of size n, $X = (x_1, \ldots, x_n)$ is drawn from the standard normal distribution

$$g_1(x) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\}. \tag{3.11}$$

The log likelihood and the expected log likelihood of the normal distribution model

$$f(x \mid \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \tag{3.12}$$

with $\theta = (\mu, \sigma^2)$ are given by

$$\begin{aligned} \log f(X \mid \theta) &= -\frac{n}{2}\log 2\pi\sigma^2 - \sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma^2}, \\ E_Y \log f(Y \mid \theta) &= -\frac{n}{2}\left\{\log 2\pi\sigma^2 + \frac{1+\mu^2}{\sigma^2}\right\}, \end{aligned} \tag{3.13}$$

respectively. Here, we consider the following six estimators, $\hat{\theta}_k = (\hat{\mu}, \hat{\sigma}_k^2)$, $k = 0, \ldots, 5$, where $\hat{\mu}$ is the maximum likelihood estimator and $\hat{\sigma}_k^2$ is defined by

$$\hat{\sigma}_k^2 = \frac{1}{n-k}\sum_{i=1}^{n}(x_i-\mu)^2. \tag{3.14}$$

Obviously, $\hat{\sigma}_0^2$ is the maximum likelihood estimator and $\hat{\sigma}_1^2$ is the unbiased estimator of the variance σ^2. In this case, it can be easily seen that the exact values of the biases are given by $C_k = 2(n-k)/(n-3)$. On the other hand, based on the asymptotic theory, AIC evaluates that $C_0 = 2$. Note that AIC can be applied to only the maximum likelihood estimator $\hat{\sigma}_0^2$.

Based on the empirical distribution function, the bootstrap sample of size n, $X^* = (X_1^*, \ldots, X_n^*)$ was drawn for $NB = 1000$ times and the bootstrap estimates of the biases C_k^* are computed. To reduce the effect of the sample, X was generated for $NS = 10000$ times and the averages of these estimates were computed. In Tables 3, these values as well as the averages of -2 times the expected log likelihood (ELL),

$$\mathrm{ELL}_k = -\frac{2}{NS}\sum_{i=1}^{NS} E_Y \log f(Y \mid \hat{\theta}_k(X_{(i)}))$$

-2 times the log-likelihood (LL)

$$\mathrm{LL}_k = -\frac{2}{NS}\sum_{i=1}^{NS} \log f(X_{(i)} \mid \hat{\theta}_k(X_{(i)}))$$

Table 3. Normal Distribution Model. Estimate of C, C^*, -2 times Expected Log Likelihood(ELL), -2 times Log Likelihood(LL) and EIC. $N = 20$, $NB = 1000$, $NS = 100000$. The second row with parenthesis shows the standard deviations of the differences of ELL's, LL's and EIC's from those of ELL-best estimates, respectively. Note that the log likelihood ratio in this case is constant and the variance is zero.

	$\hat{C}$	C^*	ELL	LL	EIC
$\hat{\sigma}_0^2$	2.39	2.32	59.38	54.60	59.25
			(1.84)	(0.00)	(0.25)
$\hat{\sigma}_1^2$	2.27	2.21	59.17	54.63	59.04
			(1.38)	(0.00)	(0.19)
$\hat{\sigma}_2^2$	2.15	2.09	59.01	54.71	58.89
			(0.92)	(0.00)	(0.13)
$\hat{\sigma}_3^2$	2.03	1.97	58.91	54.85	58.80
			(0.46)	(0.00)	(0.06)
$\hat{\sigma}_4^2$	1.91	1.86	58.89	55.06	58.78
			(0.00)	(0.00)	(0.00)
$\hat{\sigma}_5^2$	1.79	1.74	58.94	55.36	58.84
			(0.46)	(0.00)	(0.06)

and the EIC

$$\mathrm{EIC}_k = \mathrm{LL}_k + 2 \times \frac{1}{NS} \sum_{i=1}^{NS} \left[\frac{1}{B} \sum_{j=1}^{B} \{ \log f(X^*_{(i,j)}) \hat{\theta}^*_k(X^*_{(i,j)}) - \log f(X_{(i)} \mid \hat{\theta}^*_k(X^*_{(i,j)})) \} \right]$$

are shown for sample size $n = 20$. Here $X_{(i)}$ and $X^*_{(i,j)}$ denote the i-th sample and the j-th bootstrap sample drawn from $X_{(i)}$, respectively. The second row with parenthesis shows the standard deviations of the differences of ELL's, LL's and EIC's from those of ELL-best estimates, respectively.

It can be seen that LL takes the minimum at $\hat{\sigma}_0^2$ and is increasing with the increase of k. Therefore, if we use the log likelihood as the criterion, $\hat{\sigma}_0^2$ is considered as the best estimate. However, ELL takes its minimum at $k = 4$. In this case C_k^* is a decreasing function of k and by correcting this bias EIC takes its minimum at $k = 4$.

3.4 *Bayesian predictive distribution*

Preceding three subsections are concerned with the evaluation of parameter estimation. We show that EIC can be used to evaluate Bayesian prediction distribution. This example concerns with the regression analysis. We deal with a case where the regression analysis is nothing but a fitting of a model of joint distribution of an explanatory variable Z and a real random variable X. The marginal

distribution of Z is assumed to be uniform on the integers $\{1, 2, \ldots, n\}$ and the model for the conditional distribution of X for given value i of Z is $N(\mu_i, \sigma^2)$.

A data set

$$y = (y_{11}, y_{12}, y_{13}, \ldots, y_{nJ}) \tag{3.15}$$

is generated by assuming

$$f(y \mid \theta, \sigma^2) = \prod_{i=1}^{n} \prod_{j=1}^{J} \frac{1}{\sqrt{2\pi}\sigma} e^{-(y_{ij}-\mu_i)^2/2\sigma^2} \tag{3.16}$$

as the true density, where

$$\theta = (\mu_1, \mu_2, \ldots, \mu_n). \tag{3.17}$$

To generate the data, σ^2 and θ are fixed at σ_0^2 and $\theta_0 = (\mu_1^0, \ldots, \mu_n^0)$, respectively, where

$$\mu_i^0 = \sum_{m=0}^{M} a_m i^m. \tag{3.18}$$

Assuming the smoothness prior distribution

$$\pi(\theta \mid \mu_0, \omega^2) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\omega} e^{(\mu_i \quad \mu_{i-1})^2/2\omega^2}, \tag{3.19}$$

Baycsian prcdictivc distribution

$$f_{BPD}(y \mid x, v) = \int f(y \mid \theta, \hat{\sigma}^2)\hat{\pi}(\theta \mid x, v)d\theta \tag{3.20}$$

and Bayesian mode estimator

$$f_{BME}(y \mid x, v) = f(y \mid \text{Arg. max}\, \hat{\pi}(\theta \mid x, v), \hat{\sigma}^2) \tag{3.21}$$

are obtained, where

$$\hat{\pi}(\theta \mid x, v) = f(x \mid \theta, \hat{\sigma}^2)\hat{\pi}(\theta \mid v) / \int f(x \mid \theta, \hat{\sigma}^2)\hat{\pi}(\theta \mid v)d\theta. \tag{3.22}$$

The above $\hat{\pi}(\theta \mid v)$ is defined by

$$\hat{\pi}(\theta \mid v) = \pi(\theta \mid \hat{\mu}_0, \hat{\sigma}^2/v^2) \tag{3.23}$$

where $\hat{\mu}_0$ and $\hat{\sigma}^2$ are, following Akaike (1980), obtained by maximizing the likelihood of the Bayesian model BAYES(v):

$$\int f(x \mid \theta, \sigma^2)\pi(\theta \mid \mu_0, \sigma^2/v^2)d\theta \tag{3.24}$$

for fixed v. Minus twice the expected log likelihood and EIC of predictive distributions are respectively defined by

$$\text{(3.25)} \quad \mathrm{ELL}_m(x,v) = -2\int f(y \mid \theta_0, \sigma_0^2) \log f_m(y \mid x, v)dy,$$

and

$$\text{(3.26)} \quad \mathrm{EIC}_m(x,v) = -2\log f_m(x \mid x,v) + 2E_{X^*}\{\log f_m(X^* \mid X^*,v) - E_{Y^*}\log f_m(Y^* \mid X^*,v)\},$$

where m is either BPD or BME. The data shown in the upper panel of Fig. 1 is generated assuming $M = 2$, the lower panel shows the mode estimate of θ. The results are summarized in Table 4, which shows average of each quantity calculated for 100 realizations of x. It is clear that BPD is better than BME in view of the expected log likelihood and that this fact is detected by our EIC.

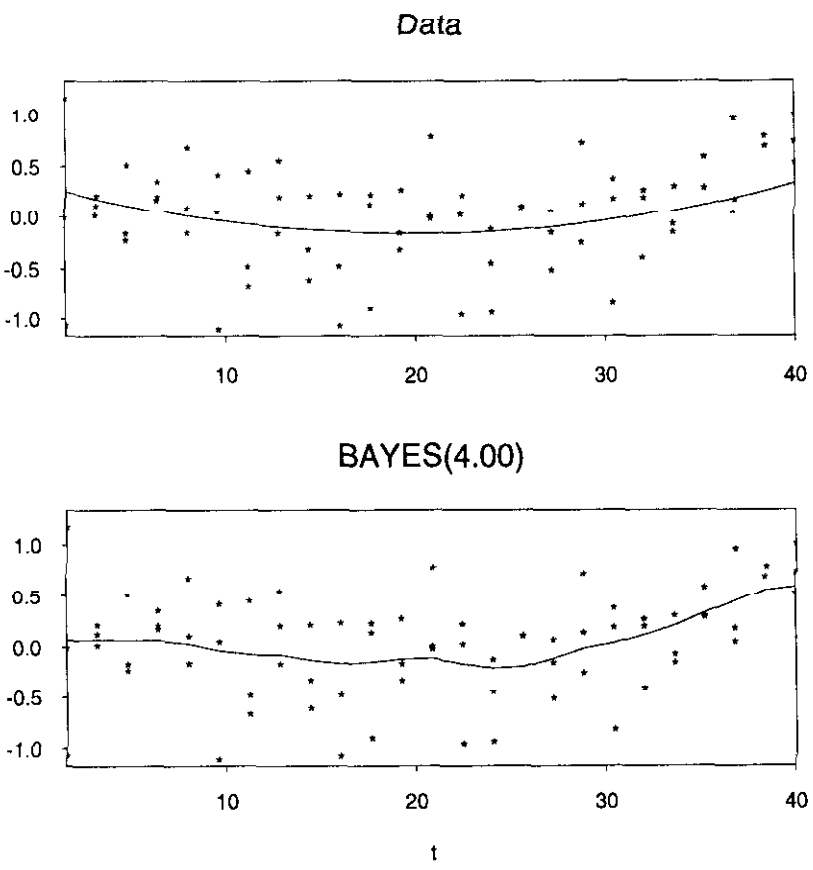

Fig. 1. Truth, data and a fitted curve.

Table 4. Comparison of BME and BPD.

v	$\mathrm{ELL}_{\mathrm{BME}}$	Frequency of min. $\mathrm{ELL}_{\mathrm{BME}}$	$\mathrm{ELL}_{\mathrm{BPD}}$	Frequency of min. $\mathrm{ELL}_{\mathrm{BPS}}$	$\mathrm{EIC}_{\mathrm{BME}}$	Frequency of min. $\mathrm{EIC}_{\mathrm{BMF}}$	$\mathrm{EIC}_{\mathrm{BPD}}$	Frequency of min. $\mathrm{EIC}_{\mathrm{BPD}}$
0.5	142.7	0	129.2	0	146.3	0	129.0	0
	(15.1)		(8.1)		(8.6)		(9.0)	
1.0	128.6	0	121.7	0	126.2	0	119.0	9
	(7.5)		(3.9)		(6.3)		(7.2)	
2.0	118.4	0	115.7	0	116.4	0	113.5	17
	(2.6)		(1.5)		(2.7)		(3.9)	
4.0	114.2	17	113.1	51	113.4	9	112.0	32
	(0.0)		(1.3)		(0.0)		(1.6)	
8.0	114.1	1	113.2	31	114.1	5	113.0	15
	(1.2)		(1.9)		(1.9)		(1.9)	
16.0	115.9	0	115.1	0	116.1	0	115.3	4
	(1.7)		(2.1)		(3.6)		(3.4)	
32.0	116.9	0	116.6	0	117.2	0	116.8	0
	(1.9)		(2.1)		(4.4)		(4.4)	
64.0	117.3	0	117.2	0	117.6	0	117.5	1
	(2.0)		(2.0)		(4.7)		(4.7)	
128.0	117.4	0	117.4	0	117.7	6	117.7	2
	(2.0)		(2.0)		(4.8)		(4.9)	

Note. Regression analysis and density estimation.

Let $f(x \mid z, \theta)$ denote a parametric model of a conditional probability density function of X given a value of an explanatory variable (vector) z, whose probability density function is supposed to be given by $g(z)$. θ denotes the parameter. When a data set

$$S = \{(x_i, z_i); i = 1, 2, \ldots, n\}$$

is given, the log likelihood is given by

$$l(\theta \mid S) = \sum_{i=1}^{n} \{\log f(x_i \mid z_i, \theta) + \log g(z_i)\}.$$

Let an estimator of the parameter θ be denoted by $\hat{\theta}(S)$, then using the following notations for two independent bootstrap samples

$$S^* = \{(x_i^*, z_i^*) \mid i = 1, 2, \ldots, n\}$$

and

$$S^\dagger = \{(x_i^\dagger, z_i^\dagger) \mid i = 1, 2, \ldots, n\},$$

the correction term of EIC of the fitted model is given by

$$
\begin{aligned}
(3.27) \qquad & E_{S^*}\{l(\hat{\theta}(S^*) \mid S^*) - E_{S^\dagger}\{l(\hat{\theta}(S^*) \mid S^\dagger)\}\} \\
& = E_{S^*}\left\{\sum_{i=1}^{n}\{\log f(x_i^* \mid z_i^*, \hat{\theta}(S^*)) + \log g(z_i^*)\}\right\} \\
& \quad - E_{S^\dagger}\left\{\sum_{i=1}^{n}\{\log f(x_i^\dagger \mid z_i^\dagger, \hat{\theta}(S^*)) + \log g(z_i^\dagger)\}\right\} \\
& = E_{S^*}\left\{\sum_{i=1}^{n}\log f(x_i^* \mid z_i^*, \hat{\theta}(S^*))\right. \\
& \qquad \left. - E_{S^\dagger}\left\{\sum_{i=1}^{n}\log f(x_i^\dagger \mid z_i^\dagger, \hat{\theta}(S^*))\right\}\right\},
\end{aligned}
$$

where

$$
x_i^* = x_{i^*}, \qquad z_i^* = z_{i^*}, \qquad x_i^\dagger = x_{i^\dagger}, \qquad z_i^\dagger = z_{i^\dagger}.
$$

i^* and $i^\dagger$ are independent integer valued random variables distributed uniformly on the interval [1,n].

The equation (3.27) shows the correction term is independent of g. In the derivation of this equation, the fact

$$
E_{S^*}\left\{\sum_{i=1}^{n}\log g(z_i^*)\right\} = E_{S^\dagger}\left\{\sum_{i=1}^{n}\log g(z_i^\dagger)\right\}
$$

was used.

Now it is clear, in this case, that EIC has the expression,

$$
\text{EIC} = -2l(\hat{\theta}(S) \mid S) + 2(\text{correction term free of } g).
$$

Note that g does appear in the first term of the above equation, but it does not matter as long as what we are interested in is the modeling of $f(x \mid z, \theta)$ and estimator of θ.

4. Concluding remarks

The range of application of EIC is very wide. It attracts our attention to the role of estimation procedures. From our present point of view, some problems which are regarded as model selection problems are rather estimator selection problems. For example, the determination of AR order is a typical one. There are many numerical examples omitted from this paper. They will be published in our forthcoming papers.

Acknowledgements

The authors are grateful to Professor H. Akaike, the former Director General of the Institute of Statistical Mathematics, for his support to our study in this field. They are also much obliged to Professor S. Konishi of Kyushu University, by whose excellent lecture their attention was lead to the bootstrap technique. They would like to thank Dr. A. Yafune for his careful reading of the manuscript. It should be mentioned that comments from referees were very much helpful to improve the paper.

Appendix

A. *Taylor expansion of the log likelihood function*
When the log likelihood $\ell_X(\theta)$ has the form

$$\ell_X(\theta) = \sum_{i=1}^{n} \log f(X_i \mid \theta),$$

it can be expressed by the Taylor series around θ_R, a given reference point:

$$\text{(A.1)}\quad \ell_X(\theta) = \sum_{i=1}^{n} \log f(X_i \mid \theta_R) + \sum_{i=1}^{n} \left[\frac{\partial}{\partial\theta} \log f(X_i \mid \theta)\right]_{\theta_R} (\theta - \theta_R) + \sum_{i=1}^{n} \frac{1}{2}(\theta - \theta_R)^T \left[\frac{\partial^2}{\partial\theta^2} \log f(X_i \mid \theta)\right]_{\theta_R} (\theta - \theta_R) + O_p\left(\frac{1}{n}\right).$$

If θ is a random (vector) and

$$\|\theta - \theta_R\| = O_p\left(\frac{1}{\sqrt{n}}\right),$$

defining ℓ, g and H by

$$\ell = E_X\{\ell_X(\theta_R)\},$$
$$g = E_X\left\{\left[\frac{\partial}{\partial\theta}\ell_X(\theta)\right]_{\theta_R}\right\},$$
$$H = -E_X\left\{\left[\frac{\partial^2}{\partial\theta^2}\ell_X(\theta)\right]_{\theta_R}\right\}$$

and

$$\triangle\ell = \ell_X(\theta_R) - \ell$$
$$\triangle g = \left[\frac{\partial}{\partial\theta}\ell_X(\theta)\right]_{\theta_R} - g,$$

then it follows that

$$\begin{aligned}(A.2)\quad \ell_X(\theta) &= \ell + \Delta\ell + (g+\Delta g)^T(\theta-\theta_R) \\ &\quad - \frac{1}{2}(\theta-\theta_R)^T H(\theta-\theta_R) + O_p\left(\frac{1}{n}\right) \\ &= \ell + \Delta\ell \\ &\quad - \frac{1}{2}(\theta-\theta_R - H^{-1}(g+\Delta g))^T H(\theta-\theta_R - H^{-1}(g+\Delta g)) \\ &\quad + \frac{1}{2}(g+\Delta g)^T H^{-1}(g+\Delta g) + O_P\left(\frac{1}{n}\right) \\ &= \ell_X - \frac{1}{2}(\theta-\theta_X)^T H^{-1}(\theta-\theta_X) + O_p\left(\frac{1}{n}\right),\end{aligned}$$

and the approximation (2.9) is obtained by ignoring the $O_p(1/n)$ term. Definitions of ℓ_X and θ_X are

$$(A.3)\qquad \ell_X = \ell + \Delta\ell + \frac{1}{2}(g+\Delta g)^T H^{-1}(g+\Delta g)$$

$$(A.4)\qquad \theta_X = \theta_R + H^{-1}(g+\Delta g).$$

B. *Distribution of* μ_{X^*}

The sample mean and the sample variance of a set X of n elements $\{x_1, x_2, \ldots, x_n\}$ are defined by

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

and

$$S^2 - \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2,$$

respectively. S^2 is a random variable which satisfies

$$E\{S^2\} = \frac{n-1}{n}\sigma^2$$

and

$$\mathrm{Var}\{S^2\} = \frac{2(n+1)}{n^2}\sigma^4,$$

where σ^2 is the variance of $X_i (i = 1, 2, \ldots, n)$. Let μ_{X^*} be the sample mean of the bootstrap sample X^* from X then

$$\sqrt{n}\mu_{X^*} \sim N(\sqrt{n}\hat{\mu}, S^2)$$

and a realization of $\sqrt{n}\mu_{X^*}$ has an expression:

$$\sqrt{n}\mu_{X*} = \sqrt{n}\hat{\mu} + \sqrt{S^2}v = \sqrt{n}\hat{\mu} + v\sqrt{\frac{n-1}{n}\sigma^2 + \sqrt{\frac{2(n+1)}{n^2}\sigma^2 u}}$$

where

$$u \sim N(0,1) \quad \text{and} \quad v \sim N(0,1).$$

$$\begin{aligned}\sqrt{n}\mu_{X*} &= \sqrt{n}\hat{\mu} + \sqrt{\left(\frac{n-1}{n} + \frac{\sqrt{2(n+1)}u}{n}\right)\sigma^2 v} \\ &= \sqrt{n}\hat{\mu} + \sigma\left(1 + \frac{1}{2}\frac{\sqrt{2(n+1)}u}{n}\right)v + o_p\left(\frac{1}{n}\right) \\ &= \sqrt{n}\hat{\mu} + \sigma v + O_p\left(\frac{1}{\sqrt{n}}\right)\end{aligned}$$

and we have the approximation

$$\sqrt{n}\mu_{X*} \sim N(\sqrt{n}\hat{\mu}, \sigma^2).$$

In our case $\sigma^2 = 1$.

REFERENCES

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle, *2nd International Symposium on Information Theory* (eds. B. N. Petrov and F. Csaki), 267–281, Akademiai Kiado, Budapest. (Reproduced in *Breakthroughs in Statistics*, Vol. 1 (eds. S. Kotz and N. L. Johnson), Springer, New York (1992))

Akaike, H. (1980). Likelihood and Bayes procedure, *Bayesian Statistics* (eds. J. M. Bernardo, M. H. De Groot, D. U. Lindley and A. F. M. Smith), 143–166, University Press, Valencia, Spain.

Efron, B. (1979). Bootstrap methods: another look at the jackknife, *Ann. Statist.*, **7**(1), 1–26.

Huber, P. J. (1967). The behavior of maximum likelihood estimators under nonstandard conditions, *Proc. Fifth Berkely Sympos. Probab. Statist.*, **1**, 221–233, Univ. of California Press, Berkeley.

Hurvich, C. M. and Tsai, C. L. (1989). Regression and time-series model selection in small samples, *Biometrika*, **76**(2), 297–307.

Hurvich, C. M., Shumway, R. and Tsai, C.-L. (1990). Improved estimators of Kulback-Leibler information for autoregressive model selection in small samples, *Biometrika*, **77**(4), 709–719.

Ishiguro, M. and Sakamoto, Y. (1991). WIC: an estimator-free information criterion, Research Memo., No. 410, The Institute of Statistical Mathematics.

Sakamoto, Y. (1991). *Categorical Data Analysis by AIC*, Kluwer, Dordrecht.

Sakamoto, Y., Ishiguro, M. and Kitagawa, G. (1986). *Akaike Information Criterion Statistics*, Reidel, Dordrecht.

Shibata, R. (1989). Statistical aspect of model selection, *From Data to Model* (ed. J. C. Willems), 215–240, Springer, Berlin.

Sugiura, N. (1978). Further analysis of the data by Akaike's information criterion and the finite corrections, *Comm. Statist. A-Theory Methods*, **7**(1), 13–26.

Takeuchi, K. (1976). Distribution of information number statistics and criteria for adequacy of models, *Mathematical Sciences*, No. 153, 12–18 (in Japanese).

Wong, W. H. (1983). A note on the modified likelihood for density estimation, *J. Amer. Statist. Assoc.*, **78**, 461–463.