Statistica Sinica **7**(1997), 375-394

BOOTSTRAP ESTIMATE OF KULLBACK-LEIBLER INFORMATION FOR MODEL SELECTION

Ritei Shibata

Keio University

Abstract: Estimation of Kullback-Leibler information is a crucial part of deriving a statistical model selection procedure which, like AIC, is based on the likelihood principle. To discriminate between nested models, we have to estimate Kullback-Leibler information up to the order of a constant, while Kullback-Leibler information itself is of the order of the number of observations. A correction term employed in AIC is an example of how to fulfill this requirement; however the correction is a simple minded bias correction to the log maximum likelihood and there is no assurance that such a bias correction yields a good estimate of Kullback-Leibler information. In this paper we investigate a bootstrap type estimate of Kullback-Leibler information as an alternative. We first show that both bootstrap estimates proposed by Efron (1983, 1986) and by Cavanaugh and Shumway (1997) are at least asymptotically equivalent and there exist many other equivalent bootstrap estimates. We also show that all such methods are asymptotically equivalent to a non-bootstrap method known as TIC (Takeuchi (1976)), which is a generalization of AIC when the re-sampling method is non-parametric. Otherwise, for example, if the re-sampling method is parametric they are asymptotically equivalent to AIC. Therefore, the use of a bootstrap type estimate is not advantageous if enough observations are available and simple calculations of a non-bootstrap estimate AIC or TIC is not a burden. At the same time, it is also true that the use of a bootstrap estimate in place of a non-bootstrap estimate is reasonable and advantageous if the non-bootstrap estimate is too complicated to evaluate analytically.

Key words and phrases: Bias estimation, bootstrap, information criterion, Kullback-Leibler information.

1. Introduction

Estimation of Kullback-Leibler information is a key to deriving the so called *information criterion* which is now widely used for selecting a statistical model. In particular, Kullback-Leibler information defined in the following formula (1.1) is considered a measure of goodness of fit of a statistical model. Therefore, one strategy is to select a model so as to minimize (1.1). Throughout this paper, we mean by a statistical model a parametric family of densities with respect to a σ-finite measure μ on n dimensional Euclidean space,

$$M = \Big\{ f(\mathbf{x}, \boldsymbol{\theta}) = \prod_i f_i(x_i, \boldsymbol{\theta});\ \boldsymbol{\theta} \in \Theta \Big\},$$

where $\mathbf{x} = (x_1, \ldots, x_n)^T$ is a running variable and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^T$ is a vector of parameters. We assume, on the other hand, that the joint distribution of independent observations $\mathbf{y} = (y_1, \ldots, y_n)^T$ is an unknown G which has a density $g(\mathbf{y}) = \prod_i g_i(y_i)$ with respect to the μ. The density is not necessarily in M. Therefore, a model M here is considered to be a way of approximation to the $g(\cdot)$ rather than to know it exactly. Denoting $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\mathbf{y})$ as the maximum likelihood estimate of $\boldsymbol{\theta}$ under model M, we define Kullback-Leibler information for the model M as

$$\begin{aligned} I_n(g(\cdot), f(\cdot, \hat{\boldsymbol{\theta}}(\mathbf{y}))) &= \int g(\mathbf{x}) \log \frac{g(\mathbf{x})}{f(\mathbf{x}, \hat{\boldsymbol{\theta}}(\mathbf{y}))} d\mu(\mathbf{x}) \\ &= \int g(\mathbf{x}) \log g(\mathbf{x}) d\mu(\mathbf{x}) - \int g(\mathbf{x}) \log f(\mathbf{x}, \hat{\boldsymbol{\theta}}(\mathbf{y})) d\mu(\mathbf{x}). \quad (1.1) \end{aligned}$$

We may compare different models $M_1, M_2, \ldots$ based on the values of Kullback-Leibler information defined in (1.1) through corresponding estimates $\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2, \ldots$ Since the first term on the right hand side of the last equation in (1.1) is independent of any particular model, minimizing the Kullback-Leibler information (1.1) is equivalent to maximizing a target variable,

$$T = T(\mathbf{y}) = \int g(\mathbf{x}) \log f(\mathbf{x}, \hat{\boldsymbol{\theta}}(\mathbf{y})) d\mu(\mathbf{x}). \quad (1.2)$$

By Taylor expansion around $\bar{\boldsymbol{\theta}}$ which is a pseudo true parameter or a projection of $g(\cdot)$ on M, we have an approximation of T,

$$T = \int g(\mathbf{x}) \log f(\mathbf{x}, \bar{\boldsymbol{\theta}}) d\mu(\mathbf{x}) - \frac{1}{2} Q + o_p(1). \quad (1.3)$$

An explicit definition of $\bar{\boldsymbol{\theta}}$ is the $\boldsymbol{\theta}$ which minimizes $I(g(\cdot), f(\cdot, \boldsymbol{\theta}))$ or maximizes $\int g(\mathbf{x}) \log f(\mathbf{x}, \boldsymbol{\theta}) d\mu(\mathbf{x})$. We implicitly assumed that any necessary regularity conditions for $f(\cdot, \boldsymbol{\theta})$, including differentiability and existence of $\bar{\boldsymbol{\theta}}$ in M, hold true. We also used the notations,

$$Q = (\hat{\boldsymbol{\theta}}(\mathbf{y}) - \bar{\boldsymbol{\theta}})^T \hat{J}(\mathbf{y}, \bar{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}(\mathbf{y}) - \bar{\boldsymbol{\theta}})$$

and

$$\hat{J}(\mathbf{y}, \boldsymbol{\theta}) = -\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \log f(\mathbf{y}, \boldsymbol{\theta}).$$

In practice we have to estimate T because the T depends on an unknown $g(\cdot)$. The log maximum likelihood is a naive estimate of T and can be a good candidate. It is approximated as

$$\begin{aligned}\log f(\mathbf{y},\hat{\boldsymbol{\theta}}(\mathbf{y})) &= \log f(\mathbf{y},\bar{\boldsymbol{\theta}}) + \frac{1}{2}Q + o_p(1)\\ &= \int g(\mathbf{x})\log f(\mathbf{x},\bar{\boldsymbol{\theta}})d\mu(\mathbf{x})\\ &\quad+\Big\{\log f(\mathbf{y},\bar{\boldsymbol{\theta}}) - \int g(\mathbf{x})\log f(\mathbf{x},\bar{\boldsymbol{\theta}})d\mu(\mathbf{x})\Big\} + \frac{1}{2}Q + o_p(1). \qquad (1.4)\end{aligned}$$

The order of magnitude of the first three terms on the right hand side of the last equation in (1.4) are $O(n)$, $O_p(\sqrt{n})$ and $O_p(1)$, respectively. Therefore, only the first term is significant if competitive models M_1 and M_2 yield different pseudo true parameters $\bar{\boldsymbol{\theta}}_1 \neq \bar{\boldsymbol{\theta}}_2$ respectively. Otherwise, for example, if models $M_1 \subset M_2$ are nested and $g(\cdot)$ is a member of M_1, then the pseudo true parameters $\bar{\boldsymbol{\theta}}_1$ and $\bar{\boldsymbol{\theta}}_2$ are the same. Then only the last term $\frac{1}{2}Q$ in (1.4) remains significant. In this case, denoting the maximum likelihood estimate of $\boldsymbol{\theta}$ by $\hat{\boldsymbol{\theta}}_1$ and $\hat{\boldsymbol{\theta}}_2$ for models M_1 and M_2 respectively, we can write the difference of the corresponding log maximum likelihoods as

$$\log f(\mathbf{y},\hat{\boldsymbol{\theta}}_1) - \log f(\mathbf{y},\hat{\boldsymbol{\theta}}_2) = \frac{1}{2}(Q_1 - Q_2) + o_p(1). \qquad (1.5)$$

On the other hand, the difference of values of the target variable T is written as

$$T_1 - T_2 = -\frac{1}{2}(Q_1 - Q_2) + o_p(1). \qquad (1.6)$$

Therefore, a simple minded correction to the log maximum likelihood is correcting only a significant part of the bias of (1.5) to (1.6) for the case when $\bar{\boldsymbol{\theta}}_1 = \bar{\boldsymbol{\theta}}_2$,

$$-\mathrm{E}\,(Q_1 - Q_2), \qquad (1.7)$$

which is asymptotically equal to $-(p_1 - p_2)$. Here p_1 and p_2 are the number of parameters of the models M_1 and M_2 respectively. This yields a bias correction $-p$ to the maximum log likelihood $\log f(\mathbf{y},\hat{\boldsymbol{\theta}}(\mathbf{y}))$. It is known that if the corrected log maximum likelihood is multiplied by -2 for convenience, Akaike's information criterion (Akaike (1973)),

$$AIC = -2\log f(\mathbf{y},\hat{\boldsymbol{\theta}}(\mathbf{y})) + 2p$$

follows.

However, such a simple minded correction does not necessarily yield a good estimate. A lot of work has been done to find a better correction. One such approach is to evaluate the bias as precisely as possible. Inspired by the pioneering work of Sugiura (1978), Hurvich and Tsai (1989, 1991, 1993) derived a bias correction,

$$p + \frac{(p+1)(p+2)}{n-p-2}$$

which is more precise than the p in AIC for normal linear models. In practice, such a correction is quite effective, particularly when the p is close to n. Also we note that non-asymptotic bias correction is important in selecting a discrete model like binomial or multinomial, where the distribution is often skewed and normal approximation does not work well unless a quite large number of observations is available. Another approach is to take account of the possibility that $g(\cdot)$ is outside of any models given. The expectation (1.7) is then different from $-(p_1 - p_2)$ even asymptotically. Based on such an observation, Takeuchi (1976) proposed the use of a criterion,

$$TIC = -2 \log f(\mathbf{y}, \hat{\boldsymbol{\theta}}(\mathbf{y})) + 2\hat{Q}, \tag{1.8}$$

where $\hat{Q}$ is an estimate of $\mathrm{tr}(\bar{I}(\bar{\boldsymbol{\theta}})\bar{J}(\bar{\boldsymbol{\theta}})^{-1})$. Explicit definition of $\bar{I}(\bar{\boldsymbol{\theta}})$ and $\bar{J}(\bar{\boldsymbol{\theta}})$ will be given later. The same criterion was proposed later by Linhart and Zucchini (1986).

The author showed optimality of selecting the model minimizing AIC under the assumption that the number p of parameters of $\bar{\boldsymbol{\theta}}$ increases as the number of observations n increases (Shibata (1980, 1981)). This is, for example, the case when $g(\cdot)$ is outside of any models provided. Then more and more parameters become necessary to get a closer approximation to $g(\cdot)$. Under such a framework, the approximate standard deviation $\sqrt{2p}$ of Q becomes small relative to its mean p, and asymptotic optimality of the selection follows. Otherwise, the random fluctuation of Q remains significant even if bias is corrected. This is also one of the reasons why the minimum AIC procedure is apt to pick up an over fitted model. In this respect, it is worth remembering *the paradox of AIC* pointed out by Shimizu (1978). Comparing the right hand sides of the equations in (1.5) and (1.6) he found that the correlation of both sides is almost -1. Therefore, the log maximum likelihood behaves in a direction opposite to that of the behavior of the target variable T. This can be thought of as a paradox, because our aim is to select a model so as to maximize the target variable T. Although it is not widely known, Shimizu also suggested a resolution of the paradox at the end of his paper. If the observed samples $\mathbf{y}$ is split into several subsamples, $\mathbf{y}^{(1)} = (y_1, \ldots, y_r), \mathbf{y}^{(2)} = (y_{r+1}, \ldots, y_{2r}), \ldots, \mathbf{y}^{(k)} = (y_{(k-1)r+1}, \ldots, y_n)$, an averaged AIC,

$$\overline{AIC} = \frac{1}{k} \sum_{i=1}^{k} AIC_i$$

is available from AIC_i's each of which is the AIC for subsample $\mathbf{y}^{(i)}, i = 1, \ldots, k$. We can then avoid the paradox by using $\overline{AIC}$ in place of AIC with an increasing k to infinity of the order of $o(n)$, for example, $k = \log n$, because the Q_i's in each AIC_i are averaged to p for large numbers of observations, by the law of large numbers. It is clear that selecting the model minimizing $\overline{AIC}$ yields a consistent model selection as long as $g(\cdot)$ is inside of one of competitive models. The reader may wonder if $\overline{AIC}$ is similar to modifications proposed by various authors for AIC to be consistent, since $\overline{AIC}$ can be rewritten as

$$\overline{AIC} = \frac{1}{k}\left\{-2\sum_{i=1}^{k}\log f(\mathbf{y}^{(i)}, \hat{\boldsymbol{\theta}}(\mathbf{y}^{(i)})) + 2kp\right\}.$$

For example, Shibata (1989) dealt with such modifications in a unified formula,

$$AIC_\alpha = -2\log f(\mathbf{y}, \hat{\boldsymbol{\theta}}(\mathbf{y})) + 2\alpha p,$$

where $\alpha = \alpha(n)$ is a divergent sequence with n. However, the main difference between $\overline{AIC}$ and AIC_α is that $\overline{AIC}$ retains the meaning as an estimate of Kullback-Leibler information although it is not for full observation $\mathbf{y}$ but for subsamples with size $r = n/k$. However AIC_α has no such a meaning.

In this paper, we investigate several bootstrap type estimates with the hope that it can be a resolution for such a paradoxical behavior of bias correction. However it turns out not to be so. Bootstrap estimates considered here are all asymptotically equivalent to a non-bootstrap estimate AIC or TIC. Nevertheless there are various advantages of using a bootstrap estimate. By definition, it is free from any expansion, while AIC or other related criteria are based on an expansion with respect to parameters. Therefore it has wider applicability than the conventional bias correction. Also it can be extended via the framework of the likelihood principle or of the maximum likelihood estimate. Probably the most important advantage of the use of bootstrapping is the ease of calculation. Only Monte Carlo simulations on computers is needed even when asymptotic approximation is too complicated to evaluate analytically. An implication of our result is that bootstrap and non-bootstrap methods are compatible asymptotically. Therefore, we are free to use a bootstrap estimate in place of a non-bootstrap criterion when the use of bootstrapping is really advantageous as regards calculation. However we should note that the asymptotic behavior depends on the method of bootstrapping, parametric, semi-parametric, or non-parametric. Finite sample behavior can be seen from the results of simulations in Section 3.

2. Bootstrap Correction

A naive bootstrap estimate of T in (1.2) can be obtained by replacing the running variable $\mathbf{x}$ in $\log f(\mathbf{x}, \hat{\boldsymbol{\theta}}(\mathbf{y}))$ by a bootstrap sample $\mathbf{y}^* = (y_1^*, \ldots, y_n^*)^T$ and taking bootstrap expectation E_*. However, the estimate $\mathrm{E}_* \log f(\mathbf{y}^*, \hat{\boldsymbol{\theta}}(\mathbf{y}))$ obtained by such a replacement turns out to be no more than the log maximum likelihood $\log f(\mathbf{y}, \hat{\boldsymbol{\theta}}(\mathbf{y}))$ as far as $\mathbf{y}^*$ is generated according to the empirical distribution of $\mathbf{y}$, that is, non-parametric re-sampling is used. As is seen in the previous section, such an estimate can not be a good estimate of T. One of other well known bootstrap estimates is that proposed by Efron (1983, 1986), which can be found also in Efron and Tibshirani (1993). Throughout this paper, the re-sampling size m is taken to be the same as the number of observations n. To explain his idea in our context, let us assume $\mathbf{x}$ is a random variable which is independent of $\mathbf{y}$ but distributed the same as $\mathbf{y}$. By denoting expectation with respect to $\mathbf{x}$ and $\mathbf{y}$ by $\mathrm{E}^{\mathbf{x}}$ and $\mathrm{E}^{\mathbf{y}}$ respectively, we can rewrite the bias of the log maximum likelihood $\log f(\mathbf{y}, \hat{\boldsymbol{\theta}}(\mathbf{y}))$ with respect to the target variable T as the following:

$$\mathrm{E}^{\mathbf{y}}\{T(\mathbf{y}) - \log f(\mathbf{y}, \hat{\boldsymbol{\theta}}(\mathbf{y}))\} = \mathrm{E}^{\mathbf{y}}\mathrm{E}^{\mathbf{x}} \log \frac{f(\mathbf{x}, \hat{\boldsymbol{\theta}}(\mathbf{y}))}{f(\mathbf{y}, \hat{\boldsymbol{\theta}}(\mathbf{y}))} = \mathrm{E}^{\mathbf{y}}\mathrm{E}^{\mathbf{x}} \log \frac{f(\mathbf{y}, \hat{\boldsymbol{\theta}}(\mathbf{x}))}{f(\mathbf{x}, \hat{\boldsymbol{\theta}}(\mathbf{x}))}. \tag{2.1}$$

Therefore, the expectation of a bootstrap estimate

$$B_1 = \mathrm{E}_* \log \frac{f(\mathbf{y}, \hat{\boldsymbol{\theta}}(\mathbf{y}^*))}{f(\mathbf{y}^*, \hat{\boldsymbol{\theta}}(\mathbf{y}^*))}$$

with respect to $\mathbf{y}$ is expected to be quite close to the bias (2.1) and the use of

$$T_1 = \log f(\mathbf{y}, \hat{\boldsymbol{\theta}}(\mathbf{y})) + B_1$$

is justified as an estimate of T. In practice, the bootstrap expectation E_* is replaced by an average of the results of a number of Monte Carlo simulations. However we should note that the expectation of B_1 is not exactly the same as the bias in (2.1) because the bootstrap expectation E_* depends on $\mathbf{y}$. The same bootstrap estimate as above is proposed by Ishiguro and Sakamoto (1991), and it is called WIC. A successful application to practical problems is reported in Ishiguro, Morita and Ishiguro (1991).

Recently Cavanaugh and Shumway (1997) proposed a different method of bias correction in the context of Gaussian state space model selection, which is based on the result by Stoffer and Wall (1991). Their idea is to estimate Q in (1.3) or (1.4) by bootstrapping. They proved that the expectation of

$$B_2 = 2\mathrm{E}_* \log \frac{f(\mathbf{y}, \hat{\boldsymbol{\theta}}(\mathbf{y}^*))}{f(\mathbf{y}, \hat{\boldsymbol{\theta}}(\mathbf{y}))}$$

with respect to $\mathbf{y}$ is asymptotically equal to that of $-Q$, and justified the use of $-B_2$ in place of the p in AIC. In this paper, we first prove that both bootstrap bias estimates B_1 and B_2 are asymptotically equivalent and there exist many other equivalent methods under several assumptions. These are also equivalent to a non-bootstrap criterion AIC or TIC under suitable assumptions.

First of all, we have to establish the consistency of both estimates $\hat{\boldsymbol{\theta}}(\mathbf{y})$ and $\hat{\boldsymbol{\theta}}(\mathbf{y}^*)$ for a model M. The consistency here means that the estimate converges to the pseudo true value $\bar{\boldsymbol{\theta}}$ as the sample size approaches infinity. Let Θ be a subset of p dimensional Euclidean space, and define the log likelihood ratio statistic,

$$Z_i(y_i, \boldsymbol{\theta}, U) = \inf_{\boldsymbol{\theta}' \in U} \log \frac{f_i(y_i, \boldsymbol{\theta})}{f_i(y_i, \boldsymbol{\theta}')}$$

for a neighborhood U in Θ. We assume that the limit

$$\bar{I}(\bar{\boldsymbol{\theta}}, U) = \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \mathrm{E}\ Z_i(y_i, \bar{\boldsymbol{\theta}}, U) \tag{2.2}$$

exists and is finite in a neighborhood $U = U_{\boldsymbol{\theta}}$ for any $\boldsymbol{\theta}$ in Θ. It is clear from the Lebesgue monotone convergence theorem that

$$\lim_{k \to \infty} \bar{I}(\bar{\boldsymbol{\theta}}, U_{\boldsymbol{\theta}}^{(k)}) = \bar{I}(\bar{\boldsymbol{\theta}}, \boldsymbol{\theta}) = \lim_{n \to \infty} \frac{1}{n} \{ I_n(g(\cdot), f(\cdot, \boldsymbol{\theta})) - I_n(g(\cdot), f(\cdot, \bar{\boldsymbol{\theta}})) \} \tag{2.3}$$

holds true for a monotone decreasing sequence of neighborhoods $U_{\boldsymbol{\theta}}^{(k)}, k = 1, 2, \ldots$ to a parameter $\boldsymbol{\theta}$, provided that $f(\cdot, \boldsymbol{\theta})$ is continuous with respect to $\boldsymbol{\theta}$, that is, $\lim_{\boldsymbol{\theta}' \to \boldsymbol{\theta}} f(\mathbf{x}, \boldsymbol{\theta}') = f(\mathbf{x}, \boldsymbol{\theta})$ for any $\boldsymbol{\theta} \in \Theta$. Here we note that the right hand side of (2.3) is nonnegative from the definition of $\bar{\boldsymbol{\theta}}$. We use similar notations for the bootstrap sample $\mathbf{y}^*$. However,

$$\bar{I}_B(\bar{\boldsymbol{\theta}}, U) = \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \mathrm{E}\ Z_i(y_i^*, \bar{\boldsymbol{\theta}}, U)$$

does not necessarily coincide with $\bar{I}(\bar{\boldsymbol{\theta}}, U)$ unless the re-sampling method is non-parametric. The same thing happens also for

$$\lim_{k \to \infty} \bar{I}(\bar{\boldsymbol{\theta}}, U_{\boldsymbol{\theta}}^{(k)}) = \bar{I}_B(\bar{\boldsymbol{\theta}}, \boldsymbol{\theta}) = \lim_{n \to \infty} \frac{1}{n} \mathrm{E} \log \frac{f(\mathbf{y}^*, \bar{\boldsymbol{\theta}})}{f(\mathbf{y}^*, \boldsymbol{\theta})} .$$

For example, if the re-sampling method is parametric, that is, $\mathbf{y}^*$ is generated according to $f(\cdot, \hat{\boldsymbol{\theta}}(\mathbf{y}))$, then $\bar{I}_B(\bar{\boldsymbol{\theta}}, \boldsymbol{\theta})$ does not coincides with $I(\bar{\boldsymbol{\theta}}, \boldsymbol{\theta})$ as in (2.3) but with the limit $\lim_{n \to \infty} \frac{1}{n} I_n(f(\cdot, \bar{\boldsymbol{\theta}}), f(\cdot, \boldsymbol{\theta}))$.

We need further assumptions to prove the consistency of $\hat{\boldsymbol{\theta}}(\mathbf{y})$ and $\hat{\boldsymbol{\theta}}(\mathbf{y}^*)$.

Assumption 1.

(i) The closure $\bar{M}$ of the model M is compact with respect to a weak topology.

(ii) Both $\frac{1}{n} \sum_{i=1}^{n} Z_i(y_i, \bar{\boldsymbol{\theta}}, U_{\boldsymbol{\theta}})$ and $\frac{1}{n} \sum_{i=1}^{n} Z_i(y_i^*, \bar{\boldsymbol{\theta}}, U_{\boldsymbol{\theta}})$ almost surely converge to $\bar{I}(\bar{\boldsymbol{\theta}}, U_{\boldsymbol{\theta}})$ and $\bar{I}_B(\bar{\boldsymbol{\theta}}, U_{\boldsymbol{\theta}})$ respectively in a neighborhood $U_{\boldsymbol{\theta}}$ for any $\boldsymbol{\theta} \in \Theta$.

(iii) $\bar{I}(\bar{\boldsymbol{\theta}}, \boldsymbol{\theta}) > 0$ and $\bar{I}_B(\bar{\boldsymbol{\theta}}, \boldsymbol{\theta}) > 0$ for any $\boldsymbol{\theta} \neq \bar{\boldsymbol{\theta}} \in \Theta$.

In the assumption (i) we identify a neighborhood $U_{\boldsymbol{\theta}}$ in Θ with the corresponding neighborhood in M. Various conditions are known for the assumption (i) holding true (for example, see Pfanzagl (1994)) . One example is the tightness condition of the family of probability measures specified by M. The assumption (ii) clearly holds true when the observations are independent and identically distributed and also for any models of independent and identically distributed observations. In other words, this is the case when densities $g_i(\cdot)$ are the same and $f_i(\cdot)$ are the same for all i. We hereafter refer to such a case as an *i.i.d. case*. One of the other important cases when the assumption (ii) holds true is a *regression case*. The regression case will be discussed into detail later in this section. The assumption (iii) is an identifiability condition. The proof of the following lemma is similar to that in Zacks (1971).

Lemma 1. *Under Assumption* 1*, both* $\hat{\boldsymbol{\theta}}(\mathbf{y})$ *and* $\hat{\boldsymbol{\theta}}(\mathbf{y}^*)$ *almost surely converge to* $\bar{\boldsymbol{\theta}}$ *as* n *tends to infinity.*

Proof. Let $U_{\bar{\boldsymbol{\theta}}}$ be a neighborhood of $\bar{\boldsymbol{\theta}}$. Then from (2.3) together with the assumption (iii) of Assumption 1, we see that there exists a neighborhood $U_{\boldsymbol{\theta}}$ such that $\bar{I}(\bar{\boldsymbol{\theta}}, U_{\boldsymbol{\theta}}) > 0$ for any $\boldsymbol{\theta} \notin U_{\bar{\boldsymbol{\theta}}}$. Therefore $V = \bar{M} - U_{\bar{\boldsymbol{\theta}}}$ is covered by such neighborhoods, and from the Heine-Borel theorem we can find a finite cover of V by $U_{\boldsymbol{\theta}_1}, \ldots, U_{\boldsymbol{\theta}_k}$ with the condition that $\bar{I}(\bar{\boldsymbol{\theta}}, U_{\boldsymbol{\theta}_\nu}) > 0$ for $\nu = 1, \ldots, k$. Then

$$\left\{ \frac{\prod_i f_i(y_i, \bar{\boldsymbol{\theta}})}{\sup_{\boldsymbol{\theta} \in V} \prod_i f_i(y_i, \boldsymbol{\theta})} \leq 1 \right\} \subset \left\{ \sum_{i=1}^n Z_i(y_i, \bar{\boldsymbol{\theta}}, V) \leq 0 \right\}$$
$$\subset \left\{ \max_{1 \leq \nu \leq k} \frac{1}{n} \sum_{i=1}^n Z_i(y_i, \bar{\boldsymbol{\theta}}, U_{\boldsymbol{\theta}_\nu}) \leq 0 \right\}.$$

From (2.2) together with the condition (ii) of Assumption 1, we see that such an event does not happen almost surely for large enough n, so that

$$\frac{\prod_i f_i(y_i, \bar{\boldsymbol{\theta}})}{\sup_{\boldsymbol{\theta} \in V} \prod_i f_i(y_i, \boldsymbol{\theta})} > 1$$

holds true almost surely for large enough n. This means that the estimate $\hat{\boldsymbol{\theta}}(\mathbf{y})$ which maximizes $\prod_i f_i(y_i, \boldsymbol{\theta})$ falls into the neighborhood $U_{\bar{\boldsymbol{\theta}}}$. This proves the convergence of $\hat{\boldsymbol{\theta}}(\mathbf{y})$ to $\bar{\boldsymbol{\theta}}$. The proof for $\hat{\boldsymbol{\theta}}(\mathbf{y}^*)$ is similar.

We further need the following two assumptions to prove our theorems.

Assumption 2.

(i) Both $\hat{J}(\mathbf{y}, \boldsymbol{\theta})/n$ and $\hat{J}(\mathbf{y}^*, \boldsymbol{\theta})/n$ almost surely converge to a positive definite matrix $\bar{J}(\boldsymbol{\theta})$ and $\bar{J}_B(\boldsymbol{\theta})$ respectively. The convergence is uniform in a neighborhood of $\bar{\boldsymbol{\theta}}$.

(ii) The log likelihood $\log f(\mathbf{y}, \boldsymbol{\theta})$ has up to third order derivatives with respect to $\boldsymbol{\theta}$, which are bounded by an integrable function.

The condition (i) holds true not only for the i.i.d. case but also for the regression case. The latter case is discussed later in this section. The assumption (ii) is a commonly used regularity condition to allow Taylor expansion of $f(\mathbf{y}, \boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$. The following assumption is a key to proving the equivalence of B_1 to B_2. It follows from the assumption that $\bar{J_B}(\boldsymbol{\theta}) = \bar{J}(\boldsymbol{\theta})$ and $\bar{I_B}(\bar{\boldsymbol{\theta}}, \boldsymbol{\theta}) = \bar{I}(\bar{\boldsymbol{\theta}}, \boldsymbol{\theta})$ as far as Assumption 2 holds true.

Assumption 3.

$$\mathrm{E}_* \log f(\mathbf{y}^*, \boldsymbol{\theta}) = \log f(\mathbf{y}, \boldsymbol{\theta}) \text{ holds true for any } \boldsymbol{\theta} \in \Theta.$$

This assumption clearly holds true for the i.i.d. case when the re-sampling method is non-parametric. Otherwise it may not be trivial. As an example of the non i.i.d. case, consider a normal regression model $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{p-1})^T$ is a vector of regression parameters, and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T$ is a vector of independent and identically distributed noises as normal with mean 0 and variance σ^2. By parametric or semi-parametric re-sampling, a bootstrap sample $\mathbf{y}^*$ is generated according to the formula $\mathbf{y}^* = X\hat{\boldsymbol{\beta}} + \boldsymbol{\epsilon}^*$, where $\hat{\boldsymbol{\beta}}$ is the maximum likelihood estimate of $\boldsymbol{\beta}$ and the bootstrap sample $\boldsymbol{\epsilon}^*$ is generated following the normal distribution with mean 0 and variance $\hat{\sigma}^2$, or following the empirical distribution $\hat{G}$ of the residuals $\hat{\boldsymbol{\epsilon}} = \mathbf{y} - X\hat{\boldsymbol{\beta}}$, respectively for parametric or semi-parametric bootstrapping. Here $\hat{\sigma}^2 = \frac{1}{n}\|\hat{\boldsymbol{\epsilon}}\|^2$ is the maximum likelihood estimate of σ^2. We then have

$$\mathrm{E}_*\|\mathbf{y}^* - X\boldsymbol{\beta}\|^2 = \mathrm{E}_*\|\boldsymbol{\epsilon}^* + \hat{\boldsymbol{\epsilon}} - \boldsymbol{\epsilon}\|^2 = \|\hat{\boldsymbol{\epsilon}}\|^2 + \|\hat{\boldsymbol{\epsilon}} - \boldsymbol{\epsilon}\|^2 = \|\mathbf{y} - X\boldsymbol{\beta}\|^2.$$

This shows that Assumption 3 holds true for $f(\mathbf{y}^*, \boldsymbol{\theta})$ with $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \sigma)^T$. For the case of non-parametric bootstrapping, a bootstrap sample is a set of n pairs $(\mathbf{x}_i^{*T}, y_i^*), i = 1, \ldots, n$, randomly drawn from the pairs $(\mathbf{x}_i^T, y_i), i = 1, \ldots, n$ where $\mathbf{x}_i^T$ is the ith row vector of the design matrix X. Therefore, by defining $\boldsymbol{\epsilon}^* = \mathbf{y}^* - X^*\boldsymbol{\beta}$ we have

$$\begin{aligned} \mathrm{E}_*\|\mathbf{y}^* - X\boldsymbol{\beta}\|^2 &= \mathrm{E}_*\|X^*\boldsymbol{\beta} - X\boldsymbol{\beta} + \boldsymbol{\epsilon}^*\|^2 \\ &= \|\mathbf{y} - X\boldsymbol{\beta}\|^2 + 2\boldsymbol{\beta}^T X^T (X - \bar{X})\boldsymbol{\beta} + 2\boldsymbol{\beta}^T X^T (\boldsymbol{\epsilon} - \bar{\boldsymbol{\epsilon}}), \end{aligned}$$

where $\bar{X}$ is a design matrix whose rows are all the same vector $\bar{\mathbf{x}} = \frac{1}{n}\sum_i \mathbf{x}_i$ and $\bar{\boldsymbol{\epsilon}}$ is a vector whose elements are all the same $\bar{\epsilon} = \frac{1}{n}\sum_i \epsilon_i$. It is now clear that Assumption 3 does not hold true. However it holds true if we use the following definition of the log likelihood in place of $\log f(\mathbf{y}^*, \boldsymbol{\theta})$,

$$\log f((X^*, \mathbf{y}^*), \boldsymbol{\theta}) = -\frac{n}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}\|\mathbf{y}^* - X^*\boldsymbol{\beta}\|^2.$$

This is a well known definition of the log likelihood suitable for non-parametric bootstrapping (Efron and Tibshirani (1993)). The following theorems are not affected by such a replacement. The maximum likelihood estimate $\hat{\boldsymbol{\theta}}$ for a bootstrap sample $(X^*, \mathbf{y}^*)$ is a function of both X^* and $\mathbf{y}^*$ in this case. Hereafter, we will use the notation $\log f(\mathbf{y}^*, \boldsymbol{\theta})$ in place of $\log f((X^*, \mathbf{y}^*), \boldsymbol{\theta})$ even if non-parametric bootstrapping is used.

Before proceeding with the theorems, let us check Assumption 2 for the case of regression. Suppose that the following limit exists,

$$\lim_{n\to\infty} \frac{1}{n} X^T X = V, \tag{2.4}$$

which is a positive definite matrix, and the elements of X are uniformly $o(\sqrt{n})$, and also suppose that $y_i - \mathrm{E}\, y_i, i = 1, \ldots, n$ are independent and identically distributed. Hereafter we always assume such conditions in case of regression. Then, making use of the results in Freedman (1981) we can show that condition (i) of Assumption 2 holds true. In fact,

$$\hat{J}(\mathbf{y}, \boldsymbol{\theta}) = \begin{pmatrix} X^T X/\sigma^2 & 2(\mathbf{y} - X\boldsymbol{\beta})^T X/\sigma^3 \\ 2X^T(\mathbf{y} - X\boldsymbol{\beta})/\sigma^3 & \left(3\|\mathbf{y} - X\boldsymbol{\beta}\|^2/\sigma^2 - n\right)/\sigma^2 \end{pmatrix}$$

and both $\hat{J}(\mathbf{y}, \boldsymbol{\theta})/n$ and $\hat{J}(\mathbf{y}^*, \boldsymbol{\theta})/n$ almost surely converge to the same matrix,

$$\bar{J}(\boldsymbol{\theta}) = \bar{J}_B(\boldsymbol{\theta}) = \begin{pmatrix} V/\sigma^2 & 2(\bar{\boldsymbol{\beta}} - \boldsymbol{\beta})^T V/\sigma^3 \\ 2V(\bar{\boldsymbol{\beta}} - \boldsymbol{\beta})/\sigma^3 & (3(\bar{\boldsymbol{\beta}} - \boldsymbol{\beta})^T V(\bar{\boldsymbol{\beta}} - \boldsymbol{\beta}) + 3\bar{\sigma}^2 - \sigma^2)/\sigma^4 \end{pmatrix}$$

in a neighborhood of $\bar{\boldsymbol{\theta}}$, where $\bar{\sigma}^2 = \lim_{n\to\infty} \mathrm{E}\, \hat{\sigma}^2$. In particular,

$$\bar{J}(\bar{\boldsymbol{\theta}}) = \bar{J}_B(\bar{\boldsymbol{\theta}}) = \begin{pmatrix} V/\bar{\sigma}^2 & 0 \\ 0 & 2/\bar{\sigma}^2 \end{pmatrix}.$$

Theorem 1. *Under Assumption* 1 *and Assumption* 2*, we have*

$$\mathrm{E}_* \log \frac{f(\mathbf{y}^*, \hat{\boldsymbol{\theta}}(\mathbf{y}))}{f(\mathbf{y}^*, \hat{\boldsymbol{\theta}}(\mathbf{y}^*))} = -\frac{Q_{BB}}{2}(1 + o(1)) \;\; a.s., \tag{2.5}$$

where

$$Q_{BB} = n\, \mathrm{E}_*(\hat{\boldsymbol{\theta}}(\mathbf{y}^*) - \hat{\boldsymbol{\theta}}(\mathbf{y}))^T \bar{J}_B(\bar{\boldsymbol{\theta}})\, (\hat{\boldsymbol{\theta}}(\mathbf{y}^*) - \hat{\boldsymbol{\theta}}(\mathbf{y})),$$

and

$$\mathrm{E}_* \log \frac{f(\mathbf{y}, \hat{\boldsymbol{\theta}}(\mathbf{y}^*))}{f(\mathbf{y}, \hat{\boldsymbol{\theta}}(\mathbf{y}))} = -\frac{Q_B}{2}(1 + o(1)) \;\; a.s., \tag{2.6}$$

where

$$Q_B = n\, \mathrm{E}_*(\hat{\boldsymbol{\theta}}(\mathbf{y}^*) - \hat{\boldsymbol{\theta}}(\mathbf{y}))^T \bar{J}(\bar{\boldsymbol{\theta}})\, (\hat{\boldsymbol{\theta}}(\mathbf{y}^*) - \hat{\boldsymbol{\theta}}(\mathbf{y})).$$

Furthermore, if Assumption 3 *holds true, then* $Q_{BB} = Q_B$ *and*

$$\mathrm{E}_* \log \frac{f(\mathbf{y}, \hat{\boldsymbol{\theta}}(\mathbf{y}))}{f(\mathbf{y}^*, \hat{\boldsymbol{\theta}}(\mathbf{y}^*))} = \mathrm{E}_* \log \frac{f(\mathbf{y}^*, \hat{\boldsymbol{\theta}}(\mathbf{y}))}{f(\mathbf{y}^*, \hat{\boldsymbol{\theta}}(\mathbf{y}^*))}. \tag{2.7}$$

Proof. The equation (2.5) follows from the expansion,

$$\log f(\mathbf{y}^*, \hat{\boldsymbol{\theta}}(\mathbf{y})) = \log f(\mathbf{y}^*, \hat{\boldsymbol{\theta}}(\mathbf{y}^*)) - \frac{1}{2}(\hat{\boldsymbol{\theta}}(\mathbf{y}) - \hat{\boldsymbol{\theta}}(\mathbf{y}^*))^T \hat{J}(\mathbf{y}^*, \boldsymbol{\theta}^*)(\hat{\boldsymbol{\theta}}(\mathbf{y}) - \hat{\boldsymbol{\theta}}(\mathbf{y}^*)),$$

where $\boldsymbol{\theta}^*$ is the mid-value between $\hat{\boldsymbol{\theta}}(\mathbf{y})$ and $\hat{\boldsymbol{\theta}}(\mathbf{y}^*)$. Therefore, from Lemma 1 together with Assumption 2, we see that the equation (2.5) holds true. The proof for (2.6) is similar. In fact,

$$\log f(\mathbf{y}, \hat{\boldsymbol{\theta}}(\mathbf{y}^*)) = \log f(\mathbf{y}, \hat{\boldsymbol{\theta}}(\mathbf{y})) - \frac{1}{2}(\hat{\boldsymbol{\theta}}(\mathbf{y}) - \hat{\boldsymbol{\theta}}(\mathbf{y}^*))^T \hat{J}(\mathbf{y}, \boldsymbol{\theta}^{**})(\hat{\boldsymbol{\theta}}(\mathbf{y}) - \hat{\boldsymbol{\theta}}(\mathbf{y}^*)),$$

where $\boldsymbol{\theta}^{**}$ is the mid-value between $\hat{\boldsymbol{\theta}}(\mathbf{y})$ and $\hat{\boldsymbol{\theta}}(\mathbf{y}^*)$, and the equation (2.6) is now clear from Lemma 1 and Assumption 2. The last equation (2.7) is straightforward from Assumption 3.

From the theorem, we see that Efron's method and Cavanaugh and Shumway's method are asymptotically equivalent to each other as far as Assumption 1 to Assumption 3 hold true. In fact,

$$\begin{aligned} B_1 = \mathrm{E}_* \log \frac{f(\mathbf{y}, \hat{\boldsymbol{\theta}}(\mathbf{y}^*))}{f(\mathbf{y}^*, \hat{\boldsymbol{\theta}}(\mathbf{y}^*))} &= \mathrm{E}_* \log \frac{f(\mathbf{y}, \hat{\boldsymbol{\theta}}(\mathbf{y}^*))}{f(\mathbf{y}, \hat{\boldsymbol{\theta}}(\mathbf{y}))} + \mathrm{E}_* \log \frac{f(\mathbf{y}, \hat{\boldsymbol{\theta}}(\mathbf{y}))}{f(\mathbf{y}^*, \hat{\boldsymbol{\theta}}(\mathbf{y}^*))} \\ &= B_2\,(1 + o(1)) \ \text{ a.s.} \end{aligned}$$

It is worth noting that this equivalence holds true without taking expectation with respect to $\mathbf{y}$. Therefore the behavior of resulting model selection is the same for every observation at least asymptotically.

The theorem also suggests that there can be many other bootstrap estimates of the bias besides B_1 or B_2. The difference is only about where the bootstrap sample $\mathbf{y}^*$ is used in the definition of the log likelihood ratio. It is easily seen that only six cases remain nontrivial except for the sign difference. Let us use a notation like

$$B_1 = B\begin{pmatrix} & * \\ * & * \end{pmatrix} = \mathrm{E}_* \log \frac{f(\mathbf{y}, \hat{\boldsymbol{\theta}}(\mathbf{y}^*))}{f(\mathbf{y}^*, \hat{\boldsymbol{\theta}}(\mathbf{y}^*))}$$

to indicate by the asterisk $*$ positions where the bootstrap sample $\mathbf{y}^*$ is used. If Assumption 3 holds true, one of the six cases, $B_6 = B\begin{pmatrix} * & \\ & \end{pmatrix}$ always reduces to 0. As a result four cases remain meaningful other than the B_1;

$$B_2 = 2B\begin{pmatrix} & * \\ & \end{pmatrix}, B_3 = 2B\begin{pmatrix} * & \\ * & * \end{pmatrix}, B_4 = 2B\begin{pmatrix} & * \\ * & \end{pmatrix}, B_5 = 2B\begin{pmatrix} & \\ * & * \end{pmatrix}.$$

The positions where the bootstrap sample is used are just complementary in B_2 and B_3, and only these two estimates are always negative even before taking bootstrap expectation. This is clear from the definition. A trivial relation among those five corrections is

$$B_1 = (B_2 + B_5)/2 = (B_3 + B_4)/2. \tag{2.8}$$

If $\bar{J}_B(\bar{\boldsymbol{\theta}}) = \bar{J}(\bar{\boldsymbol{\theta}})$, B_2 and B_3 are asymptotically equivalent under Assumption 1 and Assumption 2 since $Q_{BB} = Q_B$. Consequently we see from (2.8) that B_4 and B_5 also become asymptotically equivalent. However, the equivalence between two groups $\{B_2, B_3\}$ and $\{B_4, B_5\}$ does not hold true without a stronger assumption Assumption 3. In any case we now have five different bootstrap estimates of T including T_1, $T_i = \log f(\mathbf{y}, \boldsymbol{\theta}(\mathbf{y})) + B_i \;\; i = 1, \ldots, 5$.

It is also interesting to note that a bootstrap type model selection criterion proposed by Linhart and Zucchini (1986) can be approximated as

$$\mathrm{E}_* \log f(\mathbf{y}, \hat{\boldsymbol{\theta}}(\mathbf{y}^*)) = \log f(\mathbf{y}, \hat{\boldsymbol{\theta}}(\mathbf{y})) - \frac{Q_B}{2} + o(1) \;\; \text{a.s.}$$

Therefore, this criterion in fact involves only half of the necessary correction as is discussed above. The procedure to select a model so as to maximize their criterion is then more apt to select an overfitted model than the procedure based on the log maximum likelihood criterion with one of corrections above.

Example 1. In case of a simple Gaussian model with mean μ and variance σ^2, the bias corrections above are

$$\begin{aligned}
B_1 &= \mathrm{E}_*\Big\{n - \sum_i (y_i - \bar{y^*})^2/\hat{\sigma}^{*^2}\Big\}/2,\\
B_2 &= \mathrm{E}_*\Big[n \log(\hat{\sigma}^2/\hat{\sigma}^{2^*}) + \Big\{n - \sum_i (y_i - \bar{y^*})^2/\hat{\sigma}^{*^2}\Big\}\Big],\\
B_3 &= \mathrm{E}_*\Big[n \log(\hat{\sigma}^{*^2}/\hat{\sigma}^2) + \Big\{n - \sum_i (y_i^* - \bar{y})^2/\hat{\sigma}^2\Big\}\Big],\\
B_4 &= \mathrm{E}_*\Big[n \log(\hat{\sigma}^2/\hat{\sigma}^{*^2}) + \Big\{\sum_i (y_i^* - \bar{y})^2/\hat{\sigma}^2 - \sum_i (y_i - \bar{y^*})^2/\hat{\sigma}^{*^2}\Big\}\Big]
\end{aligned}$$

and

$$B_5 = \mathrm{E}_* \; n \log(\hat{\sigma}^{*^2}/\hat{\sigma}^2),$$

where $\bar{y} = \frac{1}{n}\sum_i y_i$ and $\hat{\sigma}^2 = \frac{1}{n}\sum_i (y_i - \bar{y})^2$ are the maximum likelihood estimates of μ and σ^2, and $\bar{y^*}$ and $\hat{\sigma}^{*^2}$ are those based on the bootstrap sample $\mathbf{y}^*$.

Theorem 2. *Under Assumption* 1 *to Assumption* 3*, we have*

$$\lim_{n\to\infty} Q_B = \lim_{n\to\infty} Q_{BB} = \lim_{n\to\infty} \operatorname{tr}\Big(\hat{I}_n(\hat{\boldsymbol{\theta}}(\mathbf{y}))\bar{J}(\bar{\boldsymbol{\theta}})^{-1}\Big) \;\; \textit{a.s.},$$

where

$$\hat{I}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\Big\{\mathrm{E}_*\frac{\partial}{\partial\boldsymbol{\theta}}\log f_i(y_i^*,\boldsymbol{\theta})\frac{\partial}{\partial\boldsymbol{\theta}^T}\log f_i(y_i^*,\boldsymbol{\theta}) \\ -\mathrm{E}_*\frac{\partial}{\partial\boldsymbol{\theta}}\log f_i(y_i^*,\boldsymbol{\theta})\ \mathrm{E}_*\frac{\partial}{\partial\boldsymbol{\theta}^T}\log f_i(y_i^*,\boldsymbol{\theta})\Big\}.$$

Proof. From the expansion,

$$\begin{aligned}\mathbf{0} &= \frac{\partial}{\partial\boldsymbol{\theta}}\log f(\mathbf{y}^*,\hat{\boldsymbol{\theta}}(\mathbf{y}^*)) \\ &= \frac{\partial}{\partial\boldsymbol{\theta}}\log f(\mathbf{y}^*,\hat{\boldsymbol{\theta}}(\mathbf{y})) + (\hat{\boldsymbol{\theta}}(\mathbf{y}^*)-\hat{\boldsymbol{\theta}}(\mathbf{y}))^T\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\log f(\mathbf{y}^*,\boldsymbol{\theta}^{**}) \qquad (2.9)\end{aligned}$$

with the mid-value $\boldsymbol{\theta}^{**}$ between $\hat{\boldsymbol{\theta}}(\mathbf{y})$ and $\hat{\boldsymbol{\theta}}(\mathbf{y}^*)$, we see that

$$\lim_{n\to\infty} Q_B = \lim_{n\to\infty}\mathrm{E}_*\ \mathrm{tr}\Big\{\frac{1}{n}\frac{\partial}{\partial\boldsymbol{\theta}}\log f(\mathbf{y}^*,\hat{\boldsymbol{\theta}}(\mathbf{y}))\frac{\partial}{\partial\boldsymbol{\theta}^T}\log f(\mathbf{y}^*,\hat{\boldsymbol{\theta}}(\mathbf{y}))\ \bar{J}(\bar{\boldsymbol{\theta}})^{-1}\Big\}\ \text{a.s.} \qquad (2.10)$$

On the other hand, since $\hat{\boldsymbol{\theta}}(\mathbf{y})$ is the maximum likelihood estimate we have

$$\begin{aligned}&\mathrm{E}_*\frac{\partial}{\partial\boldsymbol{\theta}}\log f(\mathbf{y}^*,\hat{\boldsymbol{\theta}}(\mathbf{y}))\frac{\partial}{\partial\boldsymbol{\theta}^T}\log f(\mathbf{y}^*,\hat{\boldsymbol{\theta}}(\mathbf{y})) \\ &= \sum_{i,j}\mathrm{E}_*\frac{\partial}{\partial\boldsymbol{\theta}}\log f_i(y_i^*,\hat{\boldsymbol{\theta}}(\mathbf{y}))\ \mathrm{E}_*\frac{\partial}{\partial\boldsymbol{\theta}^T}\log f_j(y_j^*,\hat{\boldsymbol{\theta}}(\mathbf{y})) \\ &\quad -\sum_i\mathrm{E}_*\frac{\partial}{\partial\boldsymbol{\theta}}\log f_i(y_i^*,\hat{\boldsymbol{\theta}}(\mathbf{y}))\ \mathrm{E}_*\frac{\partial}{\partial\boldsymbol{\theta}^T}\log f_i(y_i^*,\hat{\boldsymbol{\theta}}(\mathbf{y})) \\ &\quad +\sum_i\mathrm{E}_*\frac{\partial}{\partial\boldsymbol{\theta}}\log f_i(y_i^*,\hat{\boldsymbol{\theta}}(\mathbf{y}))\frac{\partial}{\partial\boldsymbol{\theta}^T}\log f_i(y_i^*,\hat{\boldsymbol{\theta}}(\mathbf{y})) \\ &= \sum_i\frac{\partial}{\partial\boldsymbol{\theta}}\log f_i(y_i,\hat{\boldsymbol{\theta}}(\mathbf{y}))\sum_j\frac{\partial}{\partial\boldsymbol{\theta}^T}\log f_j(y_j,\hat{\boldsymbol{\theta}}(\mathbf{y})) + n\hat{I}_n(\hat{\boldsymbol{\theta}}(\mathbf{y})) \\ &= n\hat{I}_n(\hat{\boldsymbol{\theta}}(\mathbf{y})).\end{aligned}$$

The desired result is then obtained by combining this result with (2.10) since $Q_{BB} = Q_B$ under Assumption 3.

We can further verify the following equality for the i.i.d. case. By denoting f_i as f, we have

$$\begin{aligned}\hat{I}_n(\hat{\boldsymbol{\theta}}(\mathbf{y})) &= \frac{1}{n}\sum_i\Big\{\frac{\partial}{\partial\boldsymbol{\theta}}\log f(y_i,\hat{\boldsymbol{\theta}}(\mathbf{y}))\frac{\partial}{\partial\boldsymbol{\theta}^T}\log f(y_i,\hat{\boldsymbol{\theta}}(\mathbf{y}))\Big\} \\ &\quad -\frac{1}{n^2}\Big\{\sum_i\frac{\partial}{\partial\boldsymbol{\theta}}\log f(y_i,\hat{\boldsymbol{\theta}}(\mathbf{y}))\Big\}\Big\{\sum_j\frac{\partial}{\partial\boldsymbol{\theta}^T}\log f(y_j,\hat{\boldsymbol{\theta}}(\mathbf{y}))\Big\} \\ &= \frac{1}{n}\sum_i\Big\{\frac{\partial}{\partial\boldsymbol{\theta}}\log f(y_i,\hat{\boldsymbol{\theta}}(\mathbf{y}))\frac{\partial}{\partial\boldsymbol{\theta}^T}\log f(y_i,\hat{\boldsymbol{\theta}}(\mathbf{y}))\Big\}.\end{aligned}$$

Then, $\hat{I}_n(\hat{\boldsymbol{\theta}}(\mathbf{y}))$ almost surely converges to

$$\bar{I}(\bar{\boldsymbol{\theta}}) = \mathrm{E}\left\{\frac{\partial}{\partial\boldsymbol{\theta}}\log f(y_i,\bar{\boldsymbol{\theta}})\frac{\partial}{\partial\boldsymbol{\theta}^T}\log f(y_i,\bar{\boldsymbol{\theta}})\right\},$$

and

$$\lim_{n\to\infty} Q_{BB} = \lim_{n\to\infty} Q_B = \mathrm{tr}\left(\bar{I}(\bar{\boldsymbol{\theta}})\bar{J}(\bar{\boldsymbol{\theta}})^{-1}\right) \quad \text{a.s.}$$

Therefore, combining this result with Theorem 1, at least for the i.i.d. case we see that the bootstrap corrections B_1 through B_5 are all asymptotically equivalent to a constant under Assumption 1 to Assumption 3. In other words, such a correction is in fact correcting only the bias, and is asymptotically equivalent to the correction $-\hat{Q}$ which is employed in TIC.

Example 2. In case of Example 1, $\boldsymbol{\theta} = (\mu,\sigma)^T$, $\bar{\boldsymbol{\theta}} = (\bar{\mu},\bar{\sigma})^T$,

$$\bar{I}(\bar{\boldsymbol{\theta}}) = \begin{pmatrix} 1/\bar{\sigma}^2 & \mu(3)/\bar{\sigma}^5 \\ \mu(3)/\bar{\sigma}^5 & \mu(4)/\bar{\sigma}^6 - 1/\bar{\sigma}^2 \end{pmatrix}$$

and

$$\bar{J}(\bar{\boldsymbol{\theta}}) = \begin{pmatrix} 1/\bar{\sigma}^2 & 0 \\ 0 & 2/\bar{\sigma}^2 \end{pmatrix},$$

where $\bar{\mu} = \mathrm{E}\,y_i$ and $\bar{\sigma}^2 = \mathrm{E}\,(y_i - \bar{\mu})^2$, and $\mu(l) = \mathrm{E}\,(y_i - \bar{\mu})^l$. All corrections B_1 through B_5 based on non-parametric bootstrap almost surely converge to the same value, $-1 - \frac{1}{2}\left(\mu(4)/\bar{\sigma}^4 - 1\right)$, which is equal to -2 if $y_i, i = 1,\ldots,n$ are actually normally distributed.

The following example demonstrates the asymptotic behavior of B_1 to B_5 for the non i.i.d. case with various re-sampling methods. This example includes Example 2 as a special case.

Example 3. In case of regression, all corrections B_1 to B_5 are the same as those in Example 1 when $\hat{\sigma}^2$, $\hat{\sigma}^{2^*}$, $\sum_i(y_i - \bar{y}^*)^2$ and $\sum_i(y_i^* - \bar{y})^2$ are replaced by $\frac{1}{n}\|\mathbf{y} - X\hat{\boldsymbol{\beta}}\|^2$, $\frac{1}{n}\|\mathbf{y}^* - X\hat{\boldsymbol{\beta}}^*\|^2$, $\|\mathbf{y} - X\hat{\boldsymbol{\beta}}^*\|^2$ and $\|\mathbf{y}^* - X\hat{\boldsymbol{\beta}}\|^2$, respectively. First of all, note that the limits $\bar{J}(\bar{\boldsymbol{\theta}})$ and $\bar{J}_B(\bar{\boldsymbol{\theta}})$ remain the same matrix as

$$\begin{pmatrix} V/\bar{\sigma}^2 & 0 \\ 0 & 2/\bar{\sigma}^2 \end{pmatrix},$$

irrespective of re-sampling method. This is because $\hat{J}(\mathbf{y},\boldsymbol{\theta})$ depends on $\mathbf{y}$ only through the form of $\|\mathbf{y} - X\boldsymbol{\beta}\|^2$, and it holds true that $\bar{\sigma}^2 = \mathrm{E}\|\mathbf{y} - X\bar{\boldsymbol{\beta}}\|^2 = \mathrm{E}\|\mathbf{y}^* - X\bar{\boldsymbol{\beta}}\|^2$ at least asymptotically.

In case of parametric bootstrapping, each bootstrap sample y_i^* is normally distributed with mean $\mathbf{x}^T\hat{\boldsymbol{\beta}}$ and variance $\hat{\sigma}^2$, and we have

$$\hat{I}_n(\hat{\boldsymbol{\theta}}(\mathbf{y})) = \begin{pmatrix} \frac{1}{n}X^TX/\hat{\sigma}^2 & 0 \\ 0 & 2/\hat{\sigma}^2 \end{pmatrix}.$$

From the condition (2.4), this matrix almost surely converges to the same matrix as $\bar{J}(\bar{\boldsymbol{\theta}})$ or $\bar{J}_B(\bar{\boldsymbol{\theta}})$. Therefore, in this case

$$\lim_{n\to\infty} Q_{BB} = \lim_{n\to\infty} Q_B = p \text{ a.s.}$$

In case of semi-parametric bootstrapping, $y_i^* - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}, i = 1,\ldots,n$ are distributed according to the empirical distribution of residuals $\hat{\epsilon}_i, i = 1,\ldots,n$, so that

$$\hat{I}_n(\hat{\boldsymbol{\theta}}(\mathbf{y})) = \begin{pmatrix} \frac{1}{n}X^TX/\hat{\sigma}^2 & \frac{1}{n}\sum_i \mathbf{x}_i^T\frac{1}{n}\sum_i \hat{\epsilon}_i^3/\hat{\sigma}^5 \\ \frac{1}{n}\sum_i \mathbf{x}_i\frac{1}{n}\sum_i \hat{\epsilon}_i^3/\hat{\sigma}^5 & \left(\frac{1}{n}\sum_i \hat{\epsilon}_i^4/\hat{\sigma}^4 - 1\right)/\hat{\sigma}^2 \end{pmatrix}.$$

Therefore, for example, if sufficiently higher order moments of ϵ_i exist, this matrix almost surely converges to

$$\begin{pmatrix} V/\bar{\sigma}^2 & * \\ * & \left(\mu(4)/\bar{\sigma}^4 - 1\right)/\bar{\sigma}^2 \end{pmatrix}, \tag{2.11}$$

where $\mu(4)$ is the 4th moment of ϵ_i and the $*$ indicates off-diagonal elements we are not interested in. Since $\bar{J}(\bar{\boldsymbol{\theta}})$ or $\bar{J}_B(\bar{\boldsymbol{\theta}})$ is a diagonal matrix,

$$\lim_{n\to\infty} Q_{BB} = \lim_{n\to\infty} Q_B = (p-1) + \left(\mu(4)/\bar{\sigma}^4 - 1\right)/2 \text{ a.s.}$$

This is equal to the limit of $\hat{Q}$ in (1.8). Finally, in case of non-parametric bootstrapping,

$$\hat{I}_n(\hat{\boldsymbol{\theta}}(\mathbf{y})) = \begin{pmatrix} \frac{1}{n}\sum_i \mathbf{x}_i\mathbf{x}_i^T\hat{\epsilon}_i^2/\hat{\sigma}^4 & \frac{1}{n}\sum_i \mathbf{x}_i^T\hat{\epsilon}_i^3/\hat{\sigma}^5 \\ \frac{1}{n}\sum_i \mathbf{x}_i\hat{\epsilon}_i^3/\hat{\sigma}^5 & \left(\frac{1}{n}\sum_i \hat{\epsilon}_i^4/\hat{\sigma}^4 - 1\right)/\hat{\sigma}^2 \end{pmatrix},$$

converges to the same matrix as in (2.11) and

$$\lim_{n\to\infty} Q_{BB} = \lim_{n\to\infty} Q_B = (p-1) + \left(\mu(4)/\bar{\sigma}^4 - 1\right)/2 \text{ a.s.}$$

3. Some Results of Simulations

To see small sample behavior of bootstrap estimates, several experiments were conducted by generating Gaussian random numbers with mean 0 and variance 1. However, note that the result does not lead to any definite conclusion because our experiments were limited to simple Gaussian models. The aim of this section to give the reader a rough idea about how a bootstrap estimate works in small samples. Figure 1 shows boxplots of the corrections B_1 to B_5 for a Gaussian distribution model with parameters μ and σ^2 as described in Example 1. Experiments were performed five hundred times with the sample size 50 and with bootstrapping 1000 times to approximate bootstrap expectation for each experiment. For reference, boxplot of non-bootstrap correction $-\hat{Q} = -1 - \left(\hat{\mu}(4)/\hat{\sigma}^4 - 1\right)/2$ which is used in TIC is also included in the figure, where $\hat{\mu}(4)$ is the sample 4th moment. All corrections are distributed around -2 since only two parameters are involved in the model. In terms of variance, B_3 seems superior to other bootstrap corrections and behaves similarly to $-\hat{Q}$.

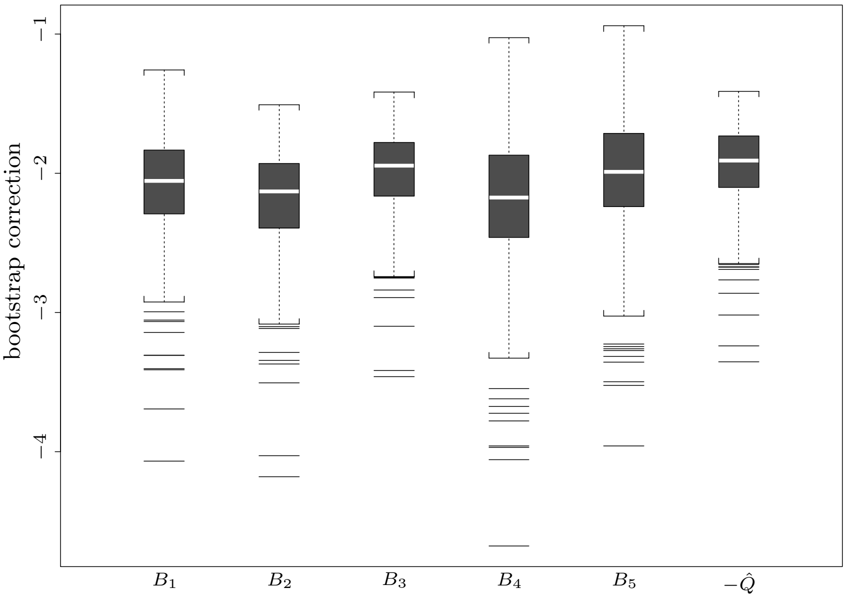

Figure 1. Distribution of bootstrap corrections; $n = 50$, $B = 1000$, $M = 500$.

However, goodness of correction can not be judged only by the distribution of the correction itself. It is significantly correlated with the log maximum likelihood. To see the real effect of bootstrapping on model selection, compare the distribution of difference of the values of each estimate for two nested models M_1 and M_2, since such differences are only relevant in selecting one of the models. The model M_2 is a Gaussian distribution model with two parameters μ and σ^2 and the M_1 is a Gaussian model with only one parameter σ^2 with $\mu = 0$. Figure 2 shows boxplots of $T_i(M_2) - T_i(M_1)$, for $i = 1, \ldots, 5$ together with boxplots for $-TIC/2$, $-AIC/2$ and the target variable T. In this experiment, bootstrap samples were generated independently for each model. Otherwise the correlation between bootstrap samples might affect the behavior of an estimate. This is also true in practice in using bootstrap type correction. Paradoxical behavior of bias correction is now clear in the figure. All estimates are distributed around -1 but in a direction opposite to that in which T is distributed. In fact, a very negative value for T corresponds to a very positive value for estimates. Such extreme values also correspond to extreme observations in which all values are almost the same. Also we see that T_3 is superior to other bootstrap corrections but not as good as non-bootstrap methods like AIC or TIC in terms of variance.

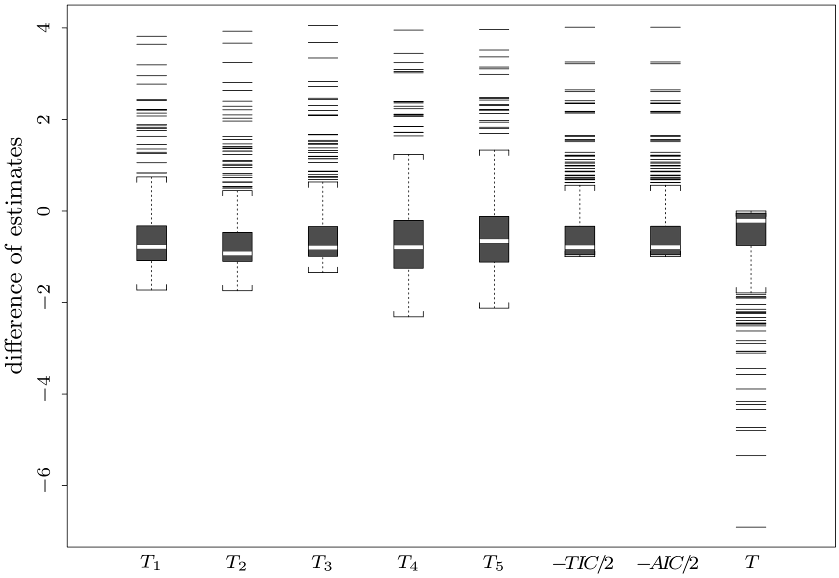

Figure 2. Distribution of difference of values of estimate for two nested models; $n = 50$, $B = 800$, $M = 500$.

In the following Table 1 and Table 2, frequencies are shown of the model M_1 selected. The model M_1 is correct and more parsimonious than the M_2 in our case. In other words, Kullback-Leibler information is less for M_1 than that for M_2. Table 1 is for the case of bootstrapping 100 times and Table 2 is for the case of bootstrapping 800 times. The estimate T_2 and T_3 are better than the others when the bootstrapping number is 100 and T_1 becomes equivalent to T_2 or T_3 if the number is increased to 800. This suggests that the rate of convergence of T_2 or T_3 is a little better than that of T_1 in terms of the number of bootstrappings. However there is no significant difference among the first three estimates and also from the non-bootstrap estimate $-TIC/2$ or $-AIC/2$. The last example is a practical one of selecting a regression model or of selecting regression variables. The data used here is "Soil Evaporation Data" by Freund (1979), which is available as S objects evap.x and evap.y on S or Splus. The number of observations is 46 and there are 10 candidates of explanatory variables: average air temperature(avat), average humidity(avh), speed of wind (wind), average soil temperature(avst), maximum soil temperature(maxst), minimum humidity(minh), maximum air temperature(maxat), maximum humidity(maxh), minimum soil temperature(minst) and minimum air temperature(minat) for soil evaporation data. Explanatory variables are rearranged to make comparison of estimates easier. The re-sampling method used for bootstrapping is semi-parametric as described in Example 3. The result in Figure 3 shows that the bootstrapping 100 times is not enough to have a good approximation to the bootstrap expectation. Figure 4 shows that the bootstrapping 1000 times seems enough. Although all estimates are downward biased compared with the non-bootstrap estimates, the result of selection so as to maximize the estimated value is the same, a model which includes explanatory variables up to the 6th variable maxst.

Table 1. Frequency of M_1 selected; $n = 50$, $B = 100$, $M = 500$.

T_1	T_2	T_3	T_4	T_5	$-TIC/2$	$-AIC/2$	T
370	426	421	324	315	421	421	500

Table 2. Frequency of M_1 selected; $n = 50$, $B = 800$, $M = 500$.

T_1	T_2	T_3	T_4	T_5	$-TIC/2$	$-AIC/2$	T
423	430	420	408	396	420	420	500

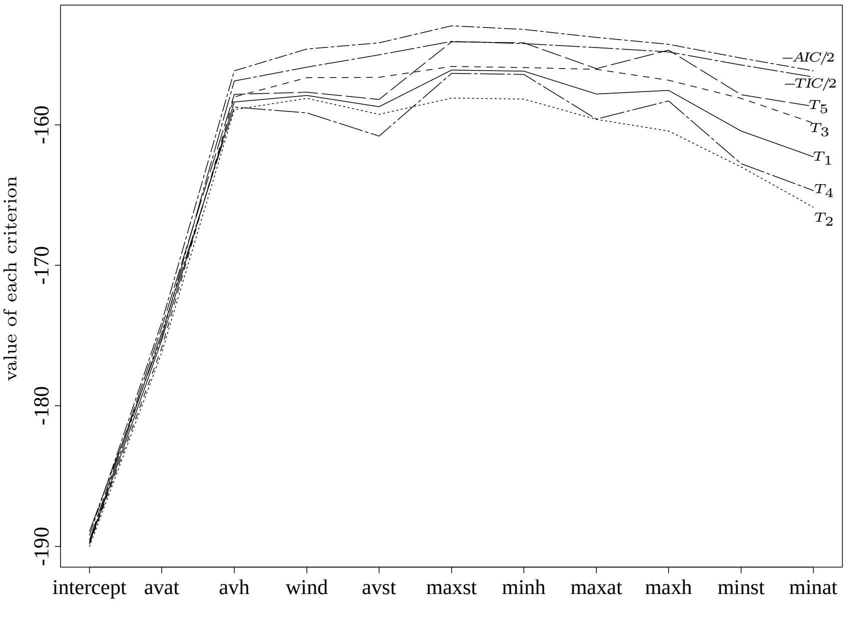

Figure 3. Soil evaporation data; $B = 100$.

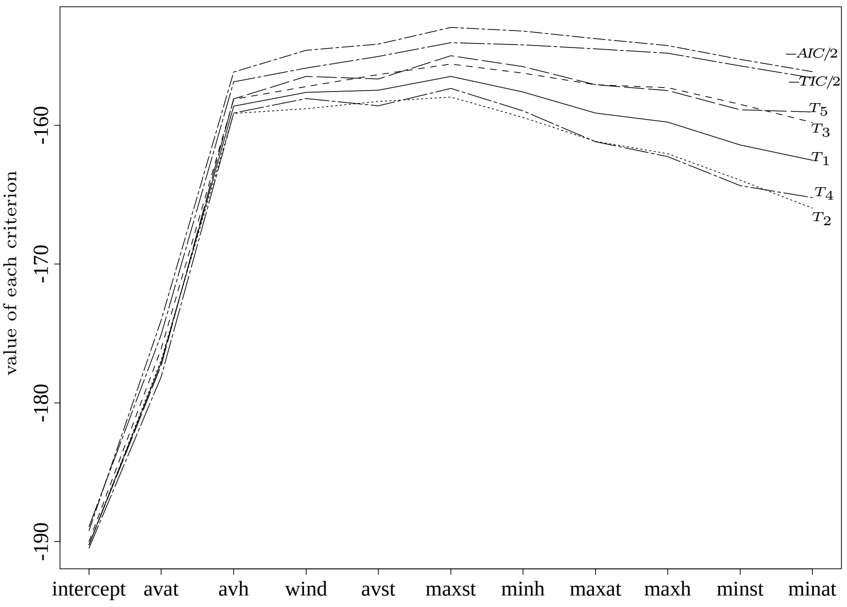

Figure 4. Soil evaporation data; $B = 1000$.

4. Conclusion

In conclusion, bootstrap type estimates considered here are all asymptotically equivalent to each other and also equivalent to a non-bootstrap criterion AIC or TIC under suitable assumptions. In this respect there is no positive reason why one of the bootstrap estimates has to be used in place of a non-bootstrap criterion when calculation of the non-bootstrap criterion is not a real problem. At the same time, our result shows that a bootstrap estimate like Efron's can be used freely in place of a non-bootstrap criterion if the calculation is much easier than that of non-bootstrap estimate. There may exist other advantages of using bootstrap estimates, for example, very small sample case or the case when the convergence to the asymptotic distribution is slow or non uniform. We leave such problems open for future investigation.

Acknowledgement

A part of this work was done during the author's stay in the Department of Statistics, UC Berkeley. The author would like to thank Peter Bühlmann and Peter Bickel for their helpful suggestion on bootstrap estimation.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In 2*nd International Symposium on Information Theory* (Edited by B. N. Petrov and F. Csáki), 267-281. Akadémia Kiadó, Budapest.

Cavanaugh, J. and Shumway, R. (1997). A bootstrap variant of AIC for state-space model selection. *Statist. Sinica* **7**, 473-496.

Efron, B. (1983). Estimating the error rate of a prediction rule: improvement on cross-validation. *J. Amer. Statist. Assoc.* **78**, 316- 331.

Efron, B. (1986). How bias is the apparent error rate of a prediction rule. *J. Amer. Statist. Assoc.* **81**, 461-470.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap.* Chapman and Hall, New York.

Freedman, D. A. (1981). Bootstrapping regression models. *Ann. Statist.* **9**, 1218-1228.

Freund, R. J. (1979). Multicollinearity etc. Some "new" Examples. *Amer. Statist. Assoc., Proceedings of Statistical Computing Section*, 111-112.

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

Hurvich, C. M. and Tsai, C. L. (1991). Bias of the corrected AIC criterion for underfitted regression and time series models. *Biometrika* **78**, 499-509.

Hurvich, C. M. and Tsai, C. L. (1993). A corrected Akaike information criterion for vector autoregressive model selection. *J. Time Ser. Anal.* **14**, 271-279.

Ishiguro, M. and Sakamoto, Y. (1991). WIC: An estimation-free information criterion. *Research Memorandum of the Institute of Statistical Mathematics, Tokyo*, 410.

Ishiguro, M., Morita, K. and Ishiguro, M. (1991). Application of an estimator-free information criterion (WIC) to aperture synthesis mixing. In *Radio Interferometry: Theory, Techniques and Application* (Edited by T. J. Cornwell and R. A. Perley), Astronomical Society of the Pacific, San Francisco.

Linhart, H. and Zucchini, W. (1986). *Model Selection.* John Wiley, New York.

Pfanzagl, J. (1994). *Parametric Statistical Theory.* Walter de Gruyter, Berlin.

Shibata, R. (1980). Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. *Ann. Statist.* **8**, 147-164.

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

Shibata, R. (1989). Statistical aspects of model selection. In *From Data to Model* (Edited by J. C. Willems), 215-240. Springer.

Shimizu, R. (1978). Entropy maximization principle and selection of the order of an autoregressive Gaussian process. *Ann. Inst. Statist. Math.* **30**, 363-270.

Stoffer, D. S. and Wall, K. D. (1991). Bootstrapping state-space models: Gaussian maximum likelihood estimation and the Kalman filter. *J. Amer. Statist. Assoc.* **86**, 1024-1033.

Sugiura, N. (1978). Further analysis of the data by Akaike's information criterion and the finite corrections. *Comm. Statist.* **A7**, 13-26.

Takeuchi, K. (1976). Distribution of information statistics and a criterion of model fitting. *Suri-Kagaku* (*Mathematical Sciences*), **153**, 12-18 (in Japanese).

Zacks, S. (1971). *The theory of statistical inference.* John Wiley, New York.

Department of Mathematics, Keio University, 3-14-1 Hiyoshi, Kohoku, Yokohama, 223, Japan.

(Received January 1995; accepted May 1996)