Bootstrap Estimate of Kullback-Leibler Information for Model Selection

By

Ritei Shibata

Technical Report No.424
January 1995

Department of Statistics
University of California
Berkeley, California

Bootstrap Estimate of Kullback-Leibler Information for Model Selection

Ritei SHIBATA *
Department of Mathematics, Keio University
3-14-1 Hiyoshi, Kohoku, Yokohama, 223, Japan

Abstract

Estimation of Kullback-Leibler amount of information is a crucial part of deriving a statistical model selection procedure which is based on likelihood principle like AIC. To discriminate nested models, we have to estimate it up to the order of constant while the Kullback-Leibler information itself is of the order of the number of observations. A correction term employed in AIC is an example to fulfill this requirement but it is a simple minded bias correction to the log maximum likelihood. Therefore there is no assurance that such a bias correction yields a good estimate of Kullback-Leibler information. In this paper as an alternative, bootstrap type estimation is considered. We will first show that both bootstrap estimates proposed by Efron (1983,1986,1993) and Cavanaugh and Shumway(1994) 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's Information Criterion) which is a generalization of AIC.

Key Words and Phrases: Kullback-Leibler Information, Information Criterion, Bootstrap, Bias Estimation

*Now staying at Department of Statistics, U.C.Berkeley.

1 Introduction

Estimation of Kullback-Leibler information is a key to deriving so called *information criterion* which is now widely used for selecting a statistical model. In particular, Kullback-Leibler information defined as in the following (1.1) is considered a measure of goodness of fit of a statistical model. Therefore, one of strategies 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,

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

where $\mathbf{x} = (x_1, x_2, ..., x_n)^T$ and $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_p)^T$. We assume, on the other hand, that the joint distribution of independent observations $\mathbf{y} = (y_1, y_2, ..., y_n)^T$ is G which has a density $g(\mathbf{y}) = \prod_i g_i(y_i)$ with respect to μ. Denoting $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\mathbf{y})$ the maximum likelihood estimate of $\boldsymbol{\theta}$ under a model M, we define Kullback-Leibler information for model M as

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

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,

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

By a simple Taylor expansion, we have an approximation of T,

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

where $\bar{\boldsymbol{\theta}}$ is a pseudo true parameter, that is, the $\boldsymbol{\theta}$ which minimizes $I(g(\cdot), f(\cdot, \boldsymbol{\theta}))$ or maximizes

$$\int g(\mathbf{x}) \log f(\mathbf{x}, \boldsymbol{\theta}) d\mu(\mathbf{x}).$$

Here we have 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

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

However 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 it can be a good platform. It is approximated as

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

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

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 as far as competitive models are not nested each other. However, if models $M_1 \subset M_2$ are nested and $g(\cdot)$ is a member of M_1, then the pseudo true parameter $\bar{\boldsymbol{\theta}}$ becomes the same for both models, so that only the last term $\frac{1}{2}Q$ in (1.4) remains significant. In fact, denoting the maximum likelihood estimate of $\boldsymbol{\theta}$ under each model by $\hat{\boldsymbol{\theta}}_1$ and $\hat{\boldsymbol{\theta}}_2$ we can write the difference of the corresponding log maximum likelihoods as

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

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

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

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),

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

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

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

follows.

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

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

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 non-asymptotic bias correction is important in selecting a discrete model like binomial or multinomial models, where the distribution is often skewed and normal approximation works well only for quite large number of observations. But in this paper we don't go further into this problem.

The author showed an optimality of the selection so as to minimize AIC under the assumption that the number 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 model. Then more and more parameters are needed to get closer approximation to $g(\cdot)$. Under such an assumption, the approximate standard deviation $\sqrt{2p}$ of Q becomes small relative to its mean p, and an asymptotic optimality of the selection follows. Otherwise, the random fluctuation of Q remains significant even if the bias is corrected. This is also one of reasons why AIC is apt to select an overfitting model. In this respect, it is worth to note *paradox of AIC* pointed out by Shimizu(1978). If the right hand sides of the equations in (1.5) and (1.6) are compared, the correlation of those two variables is -1. Therefore, the log maximum likelihood behaves to an opposite direction to that the target variable T behaves to. This can be thought of a paradox, because our aim is to select a model so as to maximize the target variable T.

In this paper, we will investigate bootstrap type correction to the log maximum likelihood with a hope that it can be a cure to such a paradoxical behavior of the bias correction. Of course, advantage of bootstrap correction is not limited to such a point. From the definition, it is free from any expansion, while AIC or other related criteria are based on an expansion with respect to parameters. This means that it is also applicable for any discrete parameter model like CART. Furthermore, in principle, it is free from type of the target variable and the way of estimation of parameters. There are various possibilities of extending it over the framework of likelihood principle or of the maximum likelihood estimate. Also, we should note that most important advantage of the use of bootstrapping is the easiness of calculation. It can be calculated by Monte Carlo simulations even when asymptotic approximation is too much complicated to evaluate analytically.

2 Correction by Bootstrapping

A naive bootstrap estimate of T in (1.2) can be obtained by replacing the $\mathbf{x}$ in $\log f(\mathbf{x}, \hat{\boldsymbol{\theta}}(\mathbf{y}))$ by a bootstrap sample $\mathbf{y}^* = (y_1^*, y_2^*, ..., y_n^*)^T$ and taking bootstrap expectation E_*. However the estimate $\mathrm{E}_* \log f(\mathbf{y}^*, \hat{\boldsymbol{\theta}}(\mathbf{y}))$ turns out to be equal to the log maximum likelihood $\log f(\mathbf{y}, \hat{\boldsymbol{\theta}}(\mathbf{y}))$, because the resampling size m here is equal to the number of observations n. 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, 1993). To explain his idea in our context, let $\mathbf{x}$ be a random variable which is independent of $\mathbf{y}$ but distributed as same as $\mathbf{y}$. By denoting expectation with respect to $\mathbf{x}$ or $\mathbf{y}$ by $\mathrm{E}^{\mathbf{x}}$ or $\mathrm{E}^{\mathbf{y}}$ respectively, we can rewrite the bias of the log maximum likelihood $\log f(\mathbf{y}, \hat{\boldsymbol{\theta}}(\mathbf{y}))$ to the target variable T as the following.

$$\begin{aligned} \mathrm{E}^{\mathbf{y}}\{T(\mathbf{y}) - \log f(\mathbf{y}, \hat{\boldsymbol{\theta}}(\mathbf{y}))\} &= \mathrm{E}^{\mathbf{y}}\mathrm{E}^{\mathbf{x}} \log \frac{f(\mathbf{x}, \hat{\boldsymbol{\theta}}(\mathbf{y}))}{f(\mathbf{y}, \hat{\boldsymbol{\theta}}(\mathbf{y}))} \qquad (2.1) \\ &= \mathrm{E}^{\mathbf{y}}\mathrm{E}^{\mathbf{x}} \log \frac{f(\mathbf{y}, \hat{\boldsymbol{\theta}}(\mathbf{x}))}{f(\mathbf{x}, \hat{\boldsymbol{\theta}}(\mathbf{x}))} \end{aligned}$$

Therefore, the expectation with respect to $\mathbf{y}$ of a bootstrap estimate

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

is expected quite close to the bias (2.1) and an estimate of T,

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

follows. In practice, the bootstrap expectation E_* is replaced by an average of the results of a number of Monte Carlo simulations. This is the basic idea of Efron. Here we should note that the expectation of B_1 is not necessarily equal to the bias since the bootstrap expectation E_* depends on $\mathbf{y}$. The same bootstrap estimate is proposed by Ishiguro and Sakamoto(1991), and they call it WIC. A successful application to practical problems is reported in Ishiguro, Morita and Ishiguro(1991).

Recently Cavanaugh and Shumway(1994) proposed a different method of bias correction in a context of Gaussian state space model selection. Their idea is to estimate Q appeared in (1.3) or (1.4) by a bootstrapping. In the paper, they proved that the expectation with respect to $\mathbf{y}$ of

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

is asymptotically equal to that of $-Q$, and the use of $-B_2$ in place of the p in AIC is proposed.

In this paper, we first prove that the bootstrap bias estimates B_1 and B_2 are asymptotically equivalent and there exist many other equivalent methods. These are also equivalent to a non-bootstrap criterion TIC proposed by Takeuchi(Shibata(1989)) in most cases.

First of all, we have to prove the consistency of both estimates, $\hat{\boldsymbol{\theta}}(\mathbf{y})$ and $\hat{\boldsymbol{\theta}}(\mathbf{y}^*)$. Assuming Θ be a subset of p dimensional Euclidean space, we define the log likelihood ratio statistic,

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

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

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

exists and is finite for a neighborhood $U = U_{\boldsymbol{\theta}}$ of any $\boldsymbol{\theta}$ in Θ which is compactified by adding a point ∞ if necessary. It is clear that

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

holds true for a monotone decreasing sequence of neighborhoods $U_{\boldsymbol{\theta}}^{(k)}, k = 1, 2, ...$ to a parameter $\boldsymbol{\theta}$. We need the following assumption to prove the consistency.

Assumption 1

(i). 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 $\bar{I}(\bar{\boldsymbol{\theta}}, U_{\boldsymbol{\theta}})$ for a neighborhood $U_{\boldsymbol{\theta}}$ of any $\boldsymbol{\theta} \in \Theta$.

(ii). $\mu(f(\mathbf{x}, \boldsymbol{\theta}) \neq f(\mathbf{x}, \boldsymbol{\theta}')) > 0$ for any $\boldsymbol{\theta}' \neq \boldsymbol{\theta} \in \Theta$.

(iii). $\lim_{\theta' \to \theta} f(\mathbf{x}, \boldsymbol{\theta}') = f(\mathbf{x}, \boldsymbol{\theta})$ for any $\boldsymbol{\theta} \in \Theta$.

(iv). $\lim_{||\theta||\to\infty} f(\mathbf{x}, \boldsymbol{\theta}) = 0$ for any $\boldsymbol{\theta} \in \Theta$.

The assumption (i) clearly holds true when observations are independent and identically distributed and models are of such independent and identically distributed observations. In other words, this is the case when densities $g_i(\cdot)$ or $f_i(\cdot)$ are the same for any i. We hereafter refer such a case as an *i.i.d. case.* One of other important cases when the assumption (i) holds true is a *regression case.* We discuss this regression case into detail later in this section. The assumption (ii) is an identifiability condition, the assumption (iii) is a continuity assumption and the assumption (iv) is a regularity condition. The proof of the following lemma is similar to that in Zacks(1971).

Lemma 1

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

Proof

From the definition of $\bar{\boldsymbol{\theta}}$ and the assumption (ii), we see that $\bar{I}(\bar{\boldsymbol{\theta}}, \boldsymbol{\theta}) > 0$ for any $\boldsymbol{\theta} \neq \bar{\boldsymbol{\theta}} \in \Theta$. Let $U_{\bar{\boldsymbol{\theta}}} = \{\boldsymbol{\theta}; ||\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}|| < \epsilon\}$ be a neighborhood of $\bar{\boldsymbol{\theta}}$. From Heine-Borel theorem, $V = \Theta - U_{\bar{\boldsymbol{\theta}}}$ can be covered by a finite number of neighborhoods, $U_{\boldsymbol{\theta}_1}, U_{\boldsymbol{\theta}_2}, ..., U_{\boldsymbol{\theta}_k}$ with the condition that $\bar{I}(\bar{\boldsymbol{\theta}}, U_{\boldsymbol{\theta}_\nu}) > 0$ for $i = 1, 2, .., k$. On the other hand,

$$\left\{ \frac{\prod_i f_i(y_i^*, \bar{\boldsymbol{\theta}})}{\sup_{\theta \in \mathrm{V}} \prod_i f_i(y_i^*, \boldsymbol{\theta})} \leq 1 \right\} \subset \left\{ \sum_{i=1}^{n} Z_i(y_i^*, \bar{\boldsymbol{\theta}}, V) \leq 0 \right\}$$

$$\subset \left\{ \max_{1 \le \nu \le k} \frac{1}{n} \sum_{i=1}^{n} Z_i(y_i^*, \bar{\boldsymbol{\theta}}, U_{\boldsymbol{\theta}_\nu}) \le 0 \right\}.$$

From the assumption (i), we see that such an event does not happen almost surely for large enough n, so that

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

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

We need further the following assumption to prove 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 $\bar{J}(\boldsymbol{\theta})$, uniformly in a neighborhood of $\bar{\boldsymbol{\theta}}$.

(ii). $\log f(\mathbf{y}, \boldsymbol{\theta})$ has up to the 3rd order derivatives with respect to $\boldsymbol{\theta}$, which are bounded by an integrable function.

The assumption (i) holds true not only for the case of i.i.d. but also holds true for regression case. The latter case is discussed later in this section. The assumption (ii) is a commonly used regularity condition to allow an expansion with respect to $\boldsymbol{\theta}$.

The next assumption is a key to showing the equivalence of B_1 to B_2.

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 is clearly satisfied for the case of i.i.d. Otherwise it is unclear. Let us consider a normal regression model

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{2.2}$$

where $\boldsymbol{\beta} = (\beta_1, ..., \beta_{p-1})^T$, and $\boldsymbol{\epsilon} = (\epsilon_1, ..., \epsilon_n)^T$ are assumed to be a vector of independent and identically distributed noises as normal with mean 0 and variance σ^2. At least three methods of bootstrapping are currently known; parametric, semiparametric or nonparametric. By parametric or semiparametric bootstrapping, a bootstrap sample $\mathbf{y}^*$ is generated as

$$\mathbf{y}^* = X\hat{\boldsymbol{\beta}} + \boldsymbol{\epsilon}^*,$$

where $\hat{\boldsymbol{\beta}}$ is the maximum likelihood estimate of $\boldsymbol{\beta}$ and bootstrap sample $\boldsymbol{\epsilon}^*$ is generated following to normal distribution with mean 0 and variance $\hat{\sigma}^2$ or following to the empirical distribution $\hat{G}$ of residuals $\hat{\boldsymbol{\epsilon}} = \mathbf{y} - X\hat{\boldsymbol{\beta}}$, respectively for parametric or semiparametric bootstrapping. Here $\hat{\sigma}^2 = \frac{1}{n}||\hat{\boldsymbol{\epsilon}}||^2$ is the maximum likelihood estimate of σ^2. We have then

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

This shows that Assumption 1 holds true for $f(\mathbf{y}^*, \boldsymbol{\theta})$ with $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \sigma)^T$. For the case of nonparametric bootstrapping, a bootstrap sample is n pairs $(\mathbf{x}_i^{*T}, y_i^*), i = 1, ..., n$, randomly drawn from the pairs $(\mathbf{x}_i^T, y_i), i = 1, ..., n$ where $\mathbf{x}_i^T$ is the i th 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 1 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 suitable definition of the log likelihood for nonparametric bootstrapping (Efron(1993)). The following theorems are not affected by such a replacement. Of course, 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 nonparametric bootstrapping is used.

Before proceeding theorems, let us check Assumptions 1 and 2 for the case of regression. Suppose that the limit exists,

$$\lim_{n\to\infty} \frac{1}{n} X^T X = V, \tag{2.3}$$

which is a positive definite matrix, and the elements of X are uniformly $o(\sqrt{n})$, and also suppose that $y_i - \mathrm{E}\, y_i, i = 1, .., n$ 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 the assumption (i) of Assumptions 1 and 2 holds true simultaneously. In fact,

$$\hat{J}(\mathbf{y}, \boldsymbol{\theta}) = \begin{pmatrix} \frac{1}{\sigma^2} X^T X & \frac{2}{\sigma^3} (\mathbf{y} - X\boldsymbol{\beta})^T X \\ \frac{2}{\sigma^3} X^T (\mathbf{y} - X\boldsymbol{\beta}) & \frac{1}{\sigma^2} \left(\frac{3\|\mathbf{y} - X\boldsymbol{\beta}\|^2}{\sigma^2} - n \right) \end{pmatrix}$$

and $\frac{1}{n} \hat{J}(\mathbf{y}, \boldsymbol{\theta})$ almost surely converges to a matrix $\bar{J}(\boldsymbol{\theta})$ in a neighborhood of $\bar{\boldsymbol{\theta}}$. In particular,

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

where $\bar{\sigma}^2 = \lim_{n\to\infty} \mathrm{E}\, \hat{\sigma}^2$.

Theorem 1

Under Assumptions 1 through 3, we have

$$\mathrm{E}_* \log \frac{f(\mathbf{y}, \hat{\boldsymbol{\theta}}(\mathbf{y}))}{f(\mathbf{y}^*, \hat{\boldsymbol{\theta}}(\mathbf{y}^*))} = \mathrm{E}_* \log \frac{f(\mathbf{y}^*, \hat{\boldsymbol{\theta}}(\mathbf{y}))}{f(\mathbf{y}^*, \hat{\boldsymbol{\theta}}(\mathbf{y}^*))} = -\frac{Q_B}{2}(1 + o(1)) \ \text{ a.s.} \tag{2.4}$$

and

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

where

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

Proof

The first equality in (2.4) is clear from the Assumption 3. The second equality follows from the expansion,

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

where $\boldsymbol{\theta}^*$ is a mid value between $\hat{\boldsymbol{\theta}}(\mathbf{y})$ and $\hat{\boldsymbol{\theta}}(\mathbf{y}^*)$. We can then rewrite

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

$$= (\hat{\boldsymbol{\theta}}(\mathbf{y}) - \hat{\boldsymbol{\theta}}(\mathbf{y}^*))^T \{\ \hat{J}(\mathbf{y}, \hat{\boldsymbol{\theta}}(\mathbf{y}))\ (I +$$

$$\hat{J}(\mathbf{y}, \hat{\boldsymbol{\theta}}(\mathbf{y}))^{-1}\ (\ \hat{J}(\mathbf{y}^*, \boldsymbol{\theta}^*) - \hat{J}(\mathbf{y}, \hat{\boldsymbol{\theta}}(\mathbf{y}))\ )\ \}\,(\hat{\boldsymbol{\theta}}(\mathbf{y}) - \hat{\boldsymbol{\theta}}(\mathbf{y}^*)).$$

Therefore, from Assumption 3 and Lemma 1, we see that the last equality of (2.4) holds true. The proof of the equality in (2.5) is similar.

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

where $\boldsymbol{\theta}^{**}$ is a mid-value between $\hat{\boldsymbol{\theta}}(\mathbf{y})$ and $\hat{\boldsymbol{\theta}}(\mathbf{y}^*)$. The result is now clear from Lemma 1 and Assumption 2.

From the theorem, we see that Efron's method and Cavanaugh and Shumway's method are asymptotically equivalent. That is,

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

It is worth to note that this equivalence holds true without taking expectation with respect to $\mathbf{y}$, so that the behavior of the resulting selection is the same for every observations at least asymptotically. The theorem also suggests that there can be many other bootstrap estimates of the bias than the B_1 or B_2, although they are all equivalent to $-Q_B$ at least asymptotically. 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 are nontrivial except the sign difference. Using 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 positions by $*$ where the bootstrap sample $\mathbf{y}^*$ is used, from Theorem 1 we see that one of the six cases, $B \begin{pmatrix} * & \\ & \end{pmatrix} = o(1)$ a.s. is meaningless. The remaining four corrections other than the B_1 are as the following.

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

The positions where bootstrap sample is used are just complementary in B_2 and B_3, and only those two estimates are always negative even before taking bootstrap expectation, which is clear from the definition. It can be easily seen from Assumption 3 that the B_4 and B_2 are closely related, and the B_5 and B_3 are also so. Furthermore, any combination of those five bias corrections yields us a new correction although all of them are asymptotically equivalent.

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

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

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

Example 1

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

$$\begin{aligned} B_1 &= \frac{1}{2}\,\mathrm{E}_* \left\{ n - \frac{\sum_i (y_i - \bar{y^*})^2}{\hat{\sigma}^{*2}} \right\}, \\ B_2 &= \mathrm{E}_* \left[n \log \frac{\hat{\sigma}^2}{\hat{\sigma}^{2*}} + \left\{ n - \frac{\sum_i (y_i - \bar{y^*})^2}{\hat{\sigma}^{*2}} \right\} \right], \end{aligned}$$

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

and

$$
B_5 = \mathrm{E}_* \, n \log \frac{\hat{\sigma}^{*^2}}{\hat{\sigma}^2},
$$

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 Assumptions 1 and 2,

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

where

$$
\begin{aligned}
\hat{I}_n(\boldsymbol{\theta}) &= \frac{1}{n} \sum_{i=1}^{n} \left\{ \mathrm{E}_* \frac{\partial}{\partial \boldsymbol{\theta}} \log f(y_i^*, \boldsymbol{\theta}) \frac{\partial}{\partial \boldsymbol{\theta}^T} \log f(y_i^*, \boldsymbol{\theta}) \right. \\
&- \left. \mathrm{E}_* \frac{\partial}{\partial \boldsymbol{\theta}} \log f(y_i^*, \boldsymbol{\theta}) \, \mathrm{E}_* \frac{\partial}{\partial \boldsymbol{\theta}^T} \log f(y_i^*, \boldsymbol{\theta}) \right\}.
\end{aligned}
$$

Proof

From the expansion,

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

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

$$
\begin{aligned}
& \lim_{n\to\infty} Q_B \qquad (2.6) \\
= & \lim_{n\to\infty} \mathrm{E}_* \operatorname{tr} \left\{ \frac{1}{n} \frac{\partial}{\partial \boldsymbol{\theta}} \log f(\mathbf{y}^*, \hat{\boldsymbol{\theta}}(\mathbf{y})) \frac{\partial}{\partial \boldsymbol{\theta}^T} \log f(\mathbf{y}^*, \hat{\boldsymbol{\theta}}(\mathbf{y}))\ \bar{J}(\bar{\boldsymbol{\theta}})^{-1} \right\} \quad \text{a.s.}
\end{aligned}
$$

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

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

The desired result is then obtained by combining this result with (2.6).

It is easy to verify the following equality for the case of i.i.d.,

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

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

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

and

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

Therefore, combining this result with Theorem 1, at least for the case of i.i.d. we see that the bootstrap bias corrections B_1 through B_5 are all asymptotically equivalent to the correction,

$$-\hat{Q} = -\mathrm{tr}\left(\hat{I}(\mathbf{y},\hat{\boldsymbol{\theta}}(\mathbf{y}))\hat{J}(\mathbf{y},\hat{\boldsymbol{\theta}}(\mathbf{y}))^{-1}\right), \tag{2.7}$$

which is turned to be equal to the correction employed in Takeuchi's Information Criterion (Shibata(1989)),

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

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 almost surely converge to the same value,

$$-1 - \frac{1}{2}\left(\frac{\mu(4)}{\bar{\sigma}^4} - 1\right),$$

which is equal to -2 when $y_i, i = 1, .., n$ are actually normally distributed.

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. However, the behavior of $\hat{I}_n(\hat{\boldsymbol{\theta}}(\mathbf{y}))$ varies with the type of bootstrapping employed. In case of parametric bootstrapping, since each y_i^* is normally distributed with mean $\mathbf{x}^T\hat{\boldsymbol{\beta}}$ and variance $\hat{\sigma}^2$ we have

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

From the condition (2.3), this matrix almost surely converges to

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

which is equal to $\bar{J}(\bar{\boldsymbol{\theta}})$. Therefore, in this case

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

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

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

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

$$\begin{pmatrix} \frac{V}{\bar{\sigma}^2} & * \\ * & \frac{1}{\bar{\sigma}^2}\left(\frac{\mu(4)}{\bar{\sigma}^4} - 1\right) \end{pmatrix}, \tag{2.8}$$

where $\mu(4)$ is the 4th moment of ϵ_i. Since $\bar{J}(\bar{\boldsymbol{\theta}})$ is a diagonal matrix,

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

This is equal to the limit of $\hat{Q}$ in (2.7). Finally, in case of nonparametric bootstrapping,

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

As is in the case of semiparametric bootstrapping, this converges to the same matrix as in (2.8) and

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

As a conclusion, bootstrap correction considered here are all asymptotically equivalent and also equivalent to a non-bootstrap correction in most cases. Therefore such a bootstrap correction can not be a cure to the paradoxical behavior of the estimate which is mentioned in Introduction. There is no positive reason why such a bootstrap correction has to be used in place of a non-bootstrap criterion like TIC, as far as the latter is too much complicated to calculate. However, our result seems natural since all bootstrap corrections here are aiming at the bias correction from the beginning. We may find a new type of bootstrap estimate which solves the paradoxical behavior of the estimate if we look for it from a different point of view. We leave this problem open for future investigation.

Acknowledgement

The author would like to thank Peter Bühlmann for his helpful suggestions on bootstrap estimation.

References

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

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

[3] B. Efron(1993), *An Introduction to the Bootstrap*, Chapman & Hall, New York.

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

[5] J. Cavanaugh and R. Shumway(1994), A bootstrap variant of AIC for state-space model selection, submitted to *Statistica Sinica.*

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

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

[8] C. Hurvich and C. Tsai(1993), A corrected Akaike information criterion for vector autoregressive model selection, *J. Time Series Analysis*, 14, 271-279.

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

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

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

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

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

[14] R. Shibata(1989), Statistical aspect of model selection, in *From Data to Model*, ed. J.C.Willems, 215-240, Springer.

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

[16] D. Stoffer and K. Wall(1991), Bootstrapping state-space models: Gaussian maximum likelihood estimation and the kalman filter, *J. Amer. Stat. Assoc.*, 86, 1024-1033.

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

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