Journal of the Korean
Statistical Society
Vol. 25, No. 2, 1996

On the Bias of Bootstrap Model Selection Criteria †

Kee-Won Lee [1] **and Songyong Sim** [2]

Abstract

A bootstrap method is used to correct the apparent downward bias of a naive plug-in bootstrap model selection criterion, which is shown to enjoy a high degree of accuracy. Comparison of bootstrap method with the asymptotic method is made through an illustrative example.

Key Words : AIC; Bootstrap; Kullback-Leibler Discrepancy.

†This work was supported by a research grant from Korea Science and Engineering Foundation 1995-1996.

[1]Department of Statistics, Hallym University, Chunchon, 200-702, Korea
[2]Department of Statistics, Hallym University, Chunchon, 200-702, Korea

1. INTRODUCTION

Akaike's Information Criterion(AIC) was first introduced by the pioneering works of Akaike(1973) in an effort to correct the apparent downward bias of $-2 \times$ (maximum log-likelihood) as an estimator of $E[2nE\{-\log g(Z, \hat{\theta})\}]$, where $\hat{\theta}$ is the maximum likelihood estimator based on n independent and identically distributed random sample from a distribution with a density $g(\cdot, \theta)$ and Z is a future observation independent of random sample. Linhart and Zucchini(1986) refined AIC by taking misspecified operating models into consideration. An algorithm is also suggested in Linhart and Zucchini(1986) to construct a version of bootstrap model selection criterion, which was employed later in Schall and Zucchini(1990) for the analysis of odds ratio from cross-classified frequencies in the presence of extraneous factors without any theoretical justification.

Chung, Lee, and Koo(1996) shows that such a naive bootstrap model selection criterion can be very misleading in the sense that it still has a downward bias of amount roughly equal to the number of parameters in the approximating model, hence it will eventually lead to an overfitted model. A major reason is that the bias of the obvious estimate is still large relative to its standard error. A correction term should be introduced to reduce bias.

In section 2, a summary of model selection terminologies is given, and the main result of Chung, Lee, and Koo(1996) is reviewed. Then, it is shown that the bootstrap estimated bias is asymptotically correct, and therefore it can be used to construct an alternative model selection criterion comparable to those traditional model selection criteria based on asymptotic approach such as AIC, which uses $2 \times$ (# of parameters) as an estimator of the bias.

We restrict our attention to the Kullback-Leibler measure of discrepancy and take a closer look at the relationship with traditional AIC. But the argument can be readily extended to other types of discrepancies. In section 3, a simple example is provided to illustrate the points discussed.

2. BOOTSTRAP MODEL SELECTION CRITERIA

2.1 Asymptotic Approach

Suppose that a set of random vectors $X_1, \ldots, X_n$ are independent and identically distributed with distribution function F. A family of such models

is called an operating family of models. See Linhart and Zucchini (1986) for detailed explanation of model selection terminologies. A family of approximating models $G(\cdot, \theta)$, which is indexed by a p-dimensional vector of parameters $\theta \in \Theta$, is used to describe the observations in a parsimonious way. Suppose also that the probability density function of $G(\cdot, \theta)$ exists, and is given by $g(\cdot, \theta)$. Kullback-Leibler measure of discrepancy is used to express some aspects of the dissimilarity between two models F and $G(\cdot, \theta)$. The discrepancy due to approximation, which arises from the fact that the approximating models are employed to simplify the operating models, is given by

$$\inf_{\theta \in \Theta} \{E_F [-\log g(Z, \theta)]\} = E_F [-\log g(Z, \theta_0)], \tag{2.1}$$

where Z is also distributed as F but independent of $X_1, \ldots, X_n$. Z is usually termed as a future observation. The employment of a future observation reflects the idea that the model selection should be based on the performance of the model over the new observation rather than what we currently observe. The minimizing argument θ_0 depends on unknown F only, say $\theta(F)$. We shall write θ_0 instead of $\theta(F)$ to simplify the notation. As can be seen, (2.1) reflects the minimum discrepancy between the two models before we take any observation. This discrepancy will typically decrease as the number of parameters in the approximating model increases.

With observations at hand, we should take random variability into consideration. The discrepancy due to estimation, which reflects the variability within the approximating model, typically increases as the number of parameters increases. In order to proceed, we should set up a consistent estimator of the discrepancy based on our observations. By using the empirical cumulative distribution function as a consistent estimator of the unknown distribution function F, we obtain a consistent estimator of (2.1), called empirical discrepancy. In our case, the empirical discrepancy is given by

$$E_{\hat{F}} [-\log g(Z, \theta)] = -n^{-1} \sum_{i=1}^{n} \log g(X_i, \theta), \tag{2.2}$$

where $\hat{F}$ is the empirical cumulative distribution function based on the observations. A minimum discrepancy estimator of θ_0 can be obtained by substituting $\hat{F}$ in place of F in the defining equation (2.1), which can be formally expressed as

$$\theta(\hat{F}) = \arg\min_{\theta \in \Theta} \{ -n^{-1} \sum_{i=1}^{n} \log g(X_i, \theta)\}. \tag{2.3}$$

We shall write $\hat{\theta}$ instead of $\theta(\hat{F})$ to simplify the notation. Under some regularity conditions, usual asymptotic properties can be established. See the appendix of Linhart and Zucchini (1986) for the list of such regularity conditions and the asymptotic properties.

We would select the model which minimizes the combined effects of the discrepancy due to approximation and the discrepancy due to estimation, called the overall discrepancy. In our case with n independent and identically distributed observations, the expected overall discrepancy is given by

$$E\left[2nE_F\{-\log g(Z,\hat{\theta})\}\right], \tag{2.4}$$

where the multiplier 2 is included for some historical reason. The inner expectation reflects an ideal goodness of fit measure as to how a future observation may deviate from the fitted approximating model on the average. Then, the number of observations and possible sampling variation are taken into account through outer expectation.

In the derivation of the model selection criterion, the following two matrices $\Omega(F)$ and $\Sigma(F)$ occur frequently;

$$\begin{aligned}\Omega(F) &= -E_F[\nabla^2 \log g(Z,\theta_0)],\\ \Sigma(F) &= E_F\left[\{\nabla \log g(Z,\theta_0)\}\{\nabla \log g(Z,\theta_0)\}^t\right],\end{aligned}$$

where ∇ and ∇^2 denote the first derivative operator and the second derivative operator respectively. Unless there is a possibility of misunderstanding, we shall use the notations Ω and Σ for simplicity. The two matrices coincide when the unknown distribution function F is in fact a member of the approximating family of models, that is, when F itself becomes $G(\cdot,\theta_0)$.

Except for some special cases, a closed form expression for (2.4) is not readily available. It is tempting to estimate (2.4) by $2nE_{\hat{F}}\{-\log g(Z,\hat{\theta})\}$, which can be rewritten as $-2\sum_{i=1}^{n}\log g(X_i,\hat{\theta})$. Obviously, this apparent estimator of the expected overall discrepancy must be biased downwards, since the same set of observations are used for both fitting and evaluating the model. In order to assess the amount of bias, the following two expansions are frequently used;

$$-2\sum_{i=1}^{n}\log g(X_i,\hat{\theta}) = -2\sum_{i=1}^{n}\log g(X_i,\theta_0)$$

$$+(\hat{\theta}-\theta_0)^t\Big\{\sum_{i=1}^{n}\nabla^2\log g(X_i,\theta_0)\Big\}(\hat{\theta}-\theta_0) \quad (2.5)$$
$$+o_p(1),$$

and

$$\begin{aligned} 2nE\{-\log g(Z,\hat{\theta})\} &= 2nE\{-\log g(Z,\theta_0)\} \\ &\quad +nE\left[(\hat{\theta}-\theta_0)^t\{-\nabla^2\log g(Z,\theta_0)\}(\hat{\theta}-\theta_0)\right] \quad (2.6) \\ &\quad +o_p(1). \end{aligned}$$

In establishing (2.5), note that $\sum_{i=1}^{n}\nabla\log g(X_i,\theta_0)$ can be expanded around $\hat{\theta}$ as

$$\sum_{i=1}^{n}\nabla\log g(X_i,\hat{\theta})+\Big\{-\nabla^2\log g(X_i,\theta_0)\Big\}(\hat{\theta}-\theta_0)+o_p(1),$$

where the first term vanishes. This type of expansion will be used frequently in investigating the theoretical behavior of the bootstrap in model selection.

(2.5) and (2.6) play key roles in proving the following Proposition 1. See the appendix of Linhart and Zucchini (1986) for the list of regularity conditions.

Proposition 1. Under proper regularity conditions, the following holds;

$$E\Big[2nE_F\{-\log g(Z,\hat{\theta})\}\Big]=E\Big[-2\sum_{i=1}^{n}\log g(X_i,\hat{\theta})\Big]+2\,\mathrm{tr}\,\Omega^{-1}\Sigma+o(1).$$

Now, Ω and Σ are the only quantities that contain the unknown parameter F, which can be consistently estimated by the empirical cumulative distribution function $\hat{F}$. Therefore, usual consistent estimators of Ω and Σ are

$$\begin{aligned} \Omega(\hat{F}) &= -n^{-1}\sum_{i=1}^{n}\nabla^2\log g(X_i,\hat{\theta}), \\ \Sigma(\hat{F}) &= n^{-1}\sum_{i=1}^{n}\{\nabla\log g(X_i,\hat{\theta})\}\{\nabla\log g(X_i,\hat{\theta})\}^t. \end{aligned}$$

We shall write $\hat{\Omega}$ and $\hat{\Sigma}$ in place of $\Omega(\hat{F})$ and $\Sigma(\hat{F})$ to simplify notation. Therefore, an asymptotic approach tells us to use the following as our model selection criterion based on expected overall discrepancy, and select the one with the smallest observed criterion value.

$$-2\sum_{i=1}^{n}\log g(X_i,\hat{\theta})+2\,\mathrm{tr}\,\hat{\Omega}^{-1}\hat{\Sigma}, \quad (2.7)$$

which reduces to the well known AIC when F coincides with $G(\cdot, \theta_0)$.

2.2 Classical Plug-in Bootstrap

In this section, we describe steps to obtain a naive bootstrap model selection criterion, which simply puts $\hat{F}$ in place of F in the expression (2.4) in a hope to obtain a reasonable approximation to the expected overall discrepancy. Note that substituting $\hat{F}$ in place of F implies $\hat{\theta}$ should be replaced by its bootstrap replicate $\hat{\theta}^*$, which is computed from the bootstrap sample.

First, we draw a bootstrap sample $X_1^*, \ldots, X_n^*$ from the empirical cumulative distribution function $\hat{F}$.

Second, the bootstrap replicate of the minimum discrepancy estimator $\hat{\theta}$ is calculated from the following bootstrap version of (2.3);

$$\hat{\theta}^* = \theta(\hat{F}^*) = \arg\min_{\theta \in \Theta} \Big\{ -2 \sum_{i=1}^{n} \log g(X_i^*, \theta) \Big\}. \tag{2.8}$$

Finally, put $\hat{F}$ in place of F in the expression (2.4). Then, the following expression of the bootstrap expected overall discrepancy is obtained;

$$E^* \Big[2n E_{\hat{F}} \{ -\log g(Z^*, \hat{\theta}^*) \} \Big] = E^* \Big[-2 \sum_{i=1}^{n} \log g(X_i, \hat{\theta}^*) \Big] \tag{2.9}$$

where Z^* is distributed as $\hat{F}$. We may use Monte-Carlo approximation to (2.9) whenever a closed form expression is not available. If we have B bootstrap repetitions, then the Monte-Carlo approximation to (2.9) can be formally written as;

$$B^{-1} \sum_{b=1}^{B} \Big\{ -2 \sum_{i=1}^{n} \log g(X_i, \hat{\theta}_b^*) \Big\},$$

where $\hat{\theta}_b^*$ is computed from the bth bootstrap sample.

The following proposition in Chung, Lee, and Koo (1996) tells us that simple plugging-in does not produce a desirable selection criterion.

Proposition 2. The following approximation holds for the naive plug-in bootstrap expected overall discrepancy;

$$E^* \Big[-2 \sum_{i=1}^{n} \log g(X_i, \hat{\theta}^*) \Big] = -2 \sum_{i=1}^{n} \log g(X_i, \hat{\theta}) + \operatorname{tr} \hat{\Omega}^{-1} \hat{\Sigma} + o_p(1).$$

Proof. Note the following expansion;

$$\begin{aligned} -2\sum_{i=1}^{n}\log g(X_i,\hat{\theta}^*) &= -2\sum_{i=1}^{n}\log g(X_i,\hat{\theta}) \\ &\quad +(\hat{\theta}^*-\hat{\theta})^t\{-2\sum_{i=1}^{n}\nabla\log g(X_i,\hat{\theta})\} \qquad (2.10) \\ &\quad +(\hat{\theta}^*-\hat{\theta})^t\{-\sum_{i=1}^{n}\nabla^2\log g(X_i,\tilde{\theta}^*)\}(\hat{\theta}^*-\hat{\theta}), \end{aligned}$$

where $\tilde{\theta}^*$ lies between $\hat{\theta}$ and $\hat{\theta}^*$. The second term on the right hand side of (2.10) vanishes, since $\hat{\theta}$ is chosen that way. After some analytic manipulation, the expected value of third term, with respect to the bootstrap distribution, turns out to be equivalent to $\mathrm{tr}\,\hat{\Omega}^{-1}\hat{\Sigma}$ up to order $o_p(1)$, and the proof is completed.

2.3 Refinements Based on Bias Correction

A more refined bootstrap approach estimates the bias in $-2\sum_{i=1}^{n}\log g(X_i,\hat{\theta})$ as an estimator of expected overall discrepancy (2.4), and corrects it by subtracting its estimated bias. That is, we focus on the estimation of the downward bias, which can be written as

$$\mathrm{bias}(F) = E\Big[\{-2\sum_{i=1}^{n}\log g(X_i,\hat{\theta})\} - 2nE_F\{-\log g(Z,\hat{\theta})\}\Big]. \qquad (2.11)$$

The asymptotic approach established in Proposition 1 estimates the bias by simply plugging-in $\hat{F}$ for F in the limiting form of $\mathrm{bias}(F)$, $-2\,\mathrm{tr}\,\Omega^{-1}\Sigma$. On the other hand, bootstrap estimator of the bias, by plugging-in $\hat{F}$ in place of F in (2.11), is given by

$$\mathrm{bias}(\hat{F}) = E^*\Big[\{-2\sum_{i=1}^{n}\log g(X_i^*,\hat{\theta}^*)\} - 2nE_{\hat{F}}\{-\log g(Z^*,\hat{\theta}^*)\}\Big], \qquad (2.12)$$

which can be further simplified as

$$E^*\{2\sum_{i=1}^{n}\log g(X_i,\hat{\theta}^*) - 2\sum_{i=1}^{n}\log g(X_i^*,\hat{\theta}^*)\}. \qquad (2.13)$$

Now the following Proposition 3 tells us that the bootstrap approximation is at least as good as the asymptotic approach.

Proposition 3. The bootstrap estimate of the bias is asymptotically correct, that is,

$$\text{bias}(\hat{F}) = -2\,\text{tr}\,\hat{\Omega}^{-1}\hat{\Sigma} + o_p(1).$$

Proof. The following can be easily checked using an expansion similar to (2.5);

$$E^*\{-2\sum_{i=1}^{n}\log g(X_i^*, \hat{\theta}^*)\} = -2\sum_{i=1}^{n}\log g(X_i, \hat{\theta}) - \text{tr}\,\hat{\Omega}^{-1}\hat{\Sigma} + o_p(1).$$

Then, the proof is straightforward from Proposition 2.

Therefore, a refined bootstrap method uses $-2\sum_{i=1}^{n}\log g(X_i, \hat{\theta}) - \text{bias}(\hat{F})$ as an alternative to traditional model selection criteria such as AIC.

3. EXAMPLE

Suppose that F belongs to a univariate distribution with finite fourth moment. We choose normal distribution as our approximating family of distributions. Our problem is to find out whether the mean and variance coincide with the pre specified values. Let $\theta = (\mu, \sigma^2)$ be the parameter for the normal distribution, and let μ_0, σ_0^2 be the two pre specified values. There are four available submodels to select from. The following notations will be used while exploring the example. Let

$$\begin{aligned}
\hat{\mu} &= \sum_{i=1}^{n} X_i/n,\ \hat{\mu}^* = \sum_{i=1}^{n} X_i^*/n,\\
\tilde{\sigma}^2 &= \sum_{i=1}^{n}(X_i - \mu_0)^2/n,\ \hat{\sigma}^2 = \sum_{i=1}^{n}(X_i - \hat{\mu})^2/n,\\
\tilde{\sigma}^{*2} &= \sum_{i=1}^{n}(X_i^* - \mu_0)^2/n,\ \hat{\sigma}^{*2} = \sum_{i=1}^{n}(X_i^* - \hat{\mu}^*)^2/n,\\
\tilde{\mu}_4 &= \sum_{i=1}^{n}(X_i - \mu_0)^4/n,\ \hat{\mu}_4 = \sum_{i=1}^{n}(X_i - \hat{\mu})^4/n.
\end{aligned}$$

In case of the submodel for $\theta = (\mu_0, \sigma_0^2)$, computation is straightforward and can be omitted. In case of the submodel for $\theta = (\mu, \sigma_0^2)$, we can easily check

that $\mathrm{tr}\,\hat{\Omega}^{-1}\hat{\Sigma} = \hat{\sigma}^2/\sigma_0^2$, and that the bootstrap bias (2.12) can be simplified as

$$E^*\{\sum_{i=1}^{n}(X_i^* - \hat{\mu}^*)^2 - \sum_{i=1}^{n}(X_i - \hat{\mu}^*)^2\} = -2\,\hat{\sigma}^2/\sigma_0^2,$$

justifying Proposition 3.

In case of $\theta = (\mu_0, \sigma^2)$, we can check that $\mathrm{tr}\,\hat{\Omega}^{-1}\hat{\Sigma} = (\tilde{\mu}_4 - \tilde{\sigma}^4)/(2\tilde{\sigma}^4)$, and that the bootstrap bias (2.12) can be simplified as

$$E^*\{-n(\tilde{\sigma}^2/\tilde{\sigma}^{*2} - 1)\} = -(\tilde{\mu}_4 - \tilde{\sigma}^4)/\tilde{\sigma}^4 + O_p(n^{-1/2}),$$

justifying Proposition 3. Note that the trace term is roughly the number of parameters in the approximating model.

Finally, in case of $\theta = (\mu, \sigma^2)$, we can check that $\mathrm{tr}\,\hat{\Omega}^{-1}\Sigma = (\hat{\mu}_4 + \hat{\sigma}^4)/(2\hat{\sigma}^4)$. Then, using a similar argument to the case of $\theta = (\mu_0, \sigma^2)$, we can check that the bootstrap bias (2.12) can be simplified as $-(\hat{\mu}_4 + \hat{\sigma}^4)/\hat{\sigma}^4 + O_p(n^{-1/2})$.

REFERENCES

(1) Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle, in: B.N. Petrov and F. Csáki, eds., *Proceedings of 2nd International Symposium on Information Theory* (Akadémia Kiadó, Budapest) pp. 267-281.

(2) Chung, H-Y., Lee, K-W., and Koo, J-Y. (1996). A note on bootstrap model selection criterion, *Statistics and Probability Letters*, **26**, 35–41.

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

(4) Schall, R. and Zucchini, W. (1990). Model selection and the estimation of odds ratios in the presence of extraneous factors, *Statistics in Medicine*, **9**, 1131-1141.